Available online at www.ijpe-online.com

International Journal of Performability Engineering

Vol. 13, No. 4, July 2017, pp. 511-518
DOI: 10.23940/ijpe.17.04.p18.511518

Active Learning Method for Chinese Spam Filtering

Guanglu Sun[a,b], Shaobo Li[a], Teng Chen[a], Xuhang Li[a], Suxia Zhu[a,b,*]

[a] *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*
[b] *Research Center of Information Security & Intelligent Technology, Harbin University of Science and Technology, Harbin, 150080, China*

Abstract

An active learning method is put forward to filter Chinese spam. In terms of training the filtering model, labeling all of the emails seems to be costly and time-consuming, while unlabeled emails can be easily accessed. Misclassification and a low-certainty method is proposed to reduce the number of labeled emails. The ROSVM model is also utilized as the online filtering model. The experimental results show that the proposed method not only decreases the number of training emails and the computational cost of spam filter, but also improves the accuracy of the filter.

Keywords: Spam Filtering; Active Learning; Support Vector Machine; Misclassification and Low-certainty Method

(Submitted on February 20, 2017; Revised on May 11, 2017; Accepted on June 15, 2017)

1. **Introduction**

Email has become an efficient communication tool nowadays. However, with the increase in popularity of emails, more problems have been caused by spam emails in recent years [10]. These spam emails that usually contain unpleasant content causes the decline in emails' credibility. Spam refers to a type of email that contains an unsolicited and disguised message that is sent to recipients who have no real relationship with the sender [2].

The content of emails is the most important part as it contains crucial information. Spam filtering usually works by the way of content analysis [13]. In content analysis, the text of an actual email is analyzed to determine if the given text is spam. Thus, machine learning techniques are applied to filter spam, which is the same as text classification [21]. The approach of content analysis based on a machine learning model has shown to be a bright prospect for dealing with spam [8].

It is necessary to have a host of labeled emails for training an accurate filter [6]. However, the acquisition of labeled emails is a very expensive task. Emails need to be manually labeled which consumes a lot of time. Labeling thousands of messages is wildly unrealistic in general. Chinese spam filtering is in the same situation, as labeled Chinese emails are not adequate either. The filter should make use of unlabeled data to train an accurate classifier without large amounts of labeled emails. In terms of that, active learning methods are applied to increase a produced filter's accuracy, as little labeled training messages are needed in this situation [5].

Active learning is a machine learning technique that has two functions. It can be used to build accurate classifiers by selecting the most informative examples, and it can also label unlabeled data by querying their labels from human experts [12]. In this paper, we propose a misclassification and low-certainty (MLC) based active learning method to reduce the

* Corresponding author. Tel.: +86-451-86390657.
E-mail address: zhusuxia@hrbust.edu.cn.

requirement of labeled training emails. This MLC method not only improves the accuracy of spam filter, but also decreases the computational cost of machine learning model in spam filtering.

In the remainder of this paper, Section 2 reviews the literature. Section 3 describes the spam filtering model. The MLC method based on active learning method is put forward in Section 4. Section 5 tests the proposed method on Chinese spam filtering task using the online support vector machine (SVM) filter. Finally, conclusions are drawn and future work is given.

2. **Related Work**

The traditional technologies of spam filtering are mainly based on user-defined rules [19]. With the development of machine learning theory, especially the text classification methods, machine learning methods are gradually being applied to spam filtering. Nowadays, spam filtering is mainly tackled by machine learning methods [9].

Machine learning methods can be divided into two patterns: batch learning and online learning [20]. Batch learning methods often suffer from expensive retraining cost as new training data arrives frequently. In contrast, online learning methods are much more scalable and effective in training and predicting procedures, making it especially suitable for spam filtering tasks with increasingly large amounts of training data.

Among various online learning methods, online SVM is effective in filtering spam with its high precision [4,12]. Sculley and Wachman [15] proposed a modified version of online SVM called Relaxed Online SVM, which greatly reduces the computational cost of updating the model. Blanzieri and Bryl [1] presented an SVM based filtering algorithm, which improved the accuracy by using the locality in the spam phenomenon. In this paper, we also adopt the online SVM model as the baseline method for spam filtering.

Although online SVM is satisfactory in spam filtering, it cannot adapt to the situation when labeled emails are scarce. In order to achieve fewer requirements of labeled emails for training, an active learning strategy is added to the online SVM spam filter. The active learning method can actively choose training samples according to the current knowledge rather than receive training samples passively, leading to a more optimal model [7,11,14].

3. **Spam Filtering Model**

In general, generative models (e.g. Na ïve Bayes) and discriminative models (e.g. SVM and Logistic Regression) are two types of machine learning models [18]. Goodman and Hulten indicated that discriminative models usually show better performance than generative models given the results on the PU-1 spam corpus [10]. Through the TREC Spam competition, we found that most high-performance methods used discriminative methods [2,3,17].

3.1. Framework for spam filtering

Figure 1 shows the framework for the proposed spam filter. It is grouped into four parts.

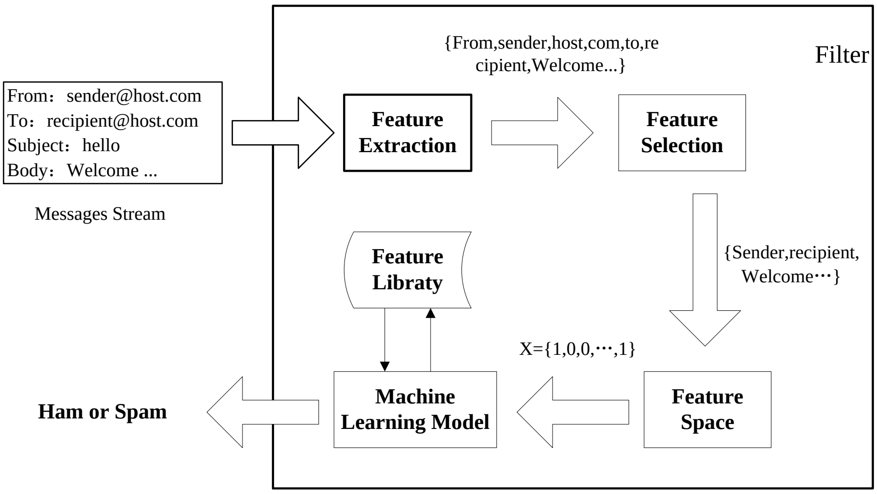


Figure 1. The typical steps of spam filtering

- Feature extraction. When an email arrives, the attachment, picture and audio are put aside. Then, the n-grams feature extraction method is applied to extract the email's features.
- Feature selection. Feature space contains a lot of irrelevant and redundant features. Therefore, the important features are chosen to decrease the feature space's dimension.
- Feature space. The set of words that present the email are transformed into a specific format required by the machine learning model.
- Machine Learning Model (Classifier). It is used as the classifier and the learner of our system.

3.2. Online support vector machine

Online SVM has given strong performance on online spam filtering. It got the best result for the competition of content-based spam filtering in TREC 2007.

Although SVM filter's performance on spam filtering task is satisfactory, the computational cost of SVM is high. Given the number of training examples, it spends quadratic training time. Naïve Bayes, the machine learning method for text classification, is often chosen by practitioners in content-based spam filtering, and only liner training time is needed. A Relaxed Online SVM (ROSVM) approach was proposed by Sculley et al. in 2007 [15], which causes training time to decline dramatically.

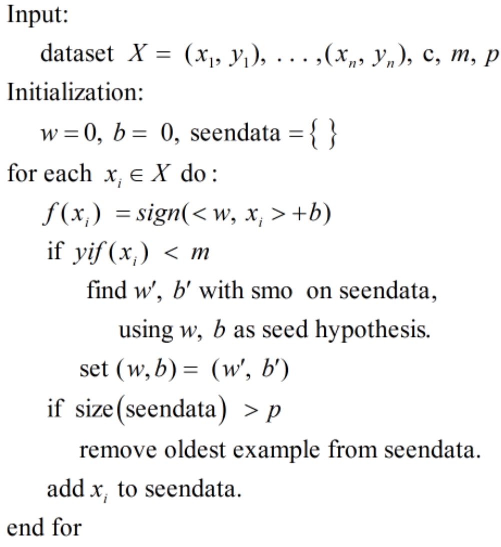
Sculley et al. considered that the full margin-maximization feature in SVM is not necessary. Computational cost can be reduced by relaxing this requirement. There are currently three ways to relax the Online SVM:

- Optimizing over the last p examples will reduce the optimization size.
- Only training on actual errors will reduce the number of training updates.
- The number of iterations will be reduced in the iterative SVM.

Let $\{(x_t, y_t)| t = 1, \ldots, T\}$ be a sequence of input patterns received over the method, where each $x_t \in R^d$ is a vector of d dimension and $y_t \in \{-1, +1\}$.

Algorithm 1 shows the steps of this method. c, m, and p are the parameters of the model.

Algorithm 1 Pseudo-code for Relaxed Online SVM



```
Input:
    dataset $X = (x_1, y_1), \ldots, (x_n, y_n)$, c, $m$, $p$.
Initialization:
    $w = 0$, $b = 0$, seendata $= \{ \}$
for each $x_i \in X$ do:
    $f(x_i) = sign(< w, x_i > + b)$
    if $yif(x_i) < m$
        find $w'$, $b'$ with smo on seendata,
            using $w$, $b$ as seed hypothesis.
        set $(w, b) = (w', b')$
    if size(seendata) $> p$
        remove oldest example from seendata.
    add $x_i$ to seendata.
end for
```

The ROSVM method is confirmed to be equal or approximate to the performance of the Online SVM on content-based spam filtering. Furthermore, the cost of computation for ROSVM filter sharply declined. But, with the comparison to logistic regression and compression models, it still has a higher cost. All of the tests on TREC06p corpus using ROSVM required 18541 seconds. However, it took just 238 seconds for the filter using the logistic regression model, which is much faster (78 times) than the ROSVM filter. In order to further reduce the cost of computation for the online SVM filter, the MLC strategy is combined with an active learning method.

4. **Misclassification and low-certainty based active learning method**

Like machine learning tasks, active learning is put forward to select the most useful examples to implement manual labeling [14]. A typical framework of active learning, called the pool-based active learning method, is shown in Figure 2. In this framework, the active learner selects the most informative unlabeled messages from the unlabeled messages pool. Then, the user labels these selected messages. The active learner adds the real label of the selected most informative messages into the labeled dataset, and removes these messages from the unlabeled messages pool. It will repeat this process until it satisfies the stopping conditions.

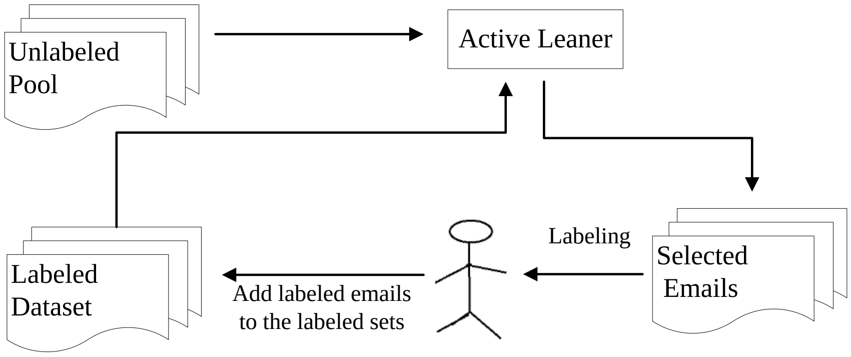


Figure 2. Pool-based active learning

Pool-based active learning has two problems involved in its application for spam filtering. Firstly, for a large email system, the cost of the method may be prohibitive. Secondly, spam filtering is typically an online learning commission. In a filter system that emails enter into, it is an email stream instead of a pool of emails. Hence, online active learning is adopted for spam filtering task [16].

Figure 3 shows the framework of online active learning. Messages get into the classifier in the way of the stream, with one message at each time. The classifier makes a predication on whether a message is spam. Requesting a label for the given message is optional so that the filter has good performance in classification and contains the least label requests at the same time. In this paper, we use online active learning settings in our filter.

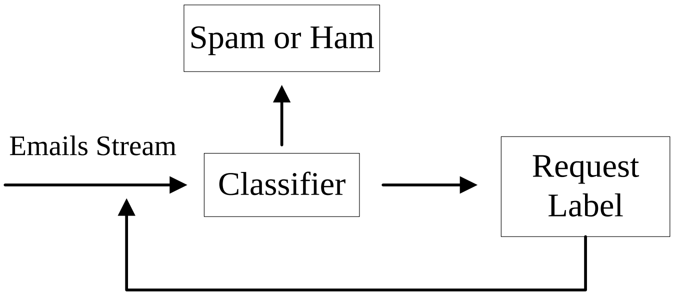


Figure 3. Online active learning

The online SVM classifier makes a predication of each message. There are two types of predication results. Firstly, the predication of messages is different from user feedback. These messages are called misclassification messages (sometimes called uncertainty messages). Secondly, the predication of messages agrees with user feedback. These messages are called certainty messages. The active learning method uses different strategies to select training messages for two types of messages.

Misclassification messages are uncertainty. The true label of these messages provides more information to optimize the hyperplane of the online SVM classifier. The traditional method selects the maximum of misclassification messages, as it lacks the suppression classification error. Hence, these messages are all selected in our paper. The other messages are certainty messages. For these messages, we select the low of certainty messages to label. That is to say, the near hyperplane messages should be labeled because they are important to the determination of the hyperplane. We use a Low-Certainty based active learning method to select messages. It only selects the near hyperplane messages without considering whether the predication is right. The method is described as follow. Given a parameter *θ,* whether a label for example x_i is requested is decided by $Q(x_i)$.

$$Q(x_i) = \begin{cases} 1, |f(x_i)| \le \theta \\ 0, |f(x_i)| > \theta \end{cases} \tag{81}$$

$Q(x_i)$ = 1 means the message requests label, 0 means not.

In this paper, we propose a Misclassification and Low-Certainty based active learning method to label messages with the above two methods. MLC method is defined by formula (2).

$$Q(x_i) = \begin{cases} 1, misclassification \\ 1, |f(x_i)| \le \theta \\ 0, |f(x_i)| > \theta \end{cases} \tag{2}$$

As the parameter *θ* approaches 0, intuitive sense is made. The more a message approaches the hyperplane, the weaker performance the classifier's predication achieves. A discussion will be held about the value of *θ*, which makes a difference in the number of selected messages, performance and computational cost in experiments.

5. **Experiments**

The proposed MLC method with the ROSVM model is tested in this section. It can be seen from the results that the MLC method shows a strong performance in Chinese spam filtering.

5.1. Experimental setting

It adopted five large benchmark Chinese data sets inform the public and developed spam filtering competitions of the TREC and SEWM, respectively: TREC06c, SEWM07/08/10/11. Table 1 shows the basic statistics of these Chinese data sets.

Table 1. Statistics of data sets

Corpus	Ham	Spam	Total
TREC06c	21766	42854	64620
SEWM07	15000	45000	60000
SEWM08	20000	50000	70000
SEWM10	15000	60000	75000
SEWM11	15000	45000	60000

In the online SVM model, the regularization parameter *C* was equal to 100, the lookback buffer was set as 10000, the maximum number of iteration was set as 1, and the margins were set to 0.8. The N-gram feature extraction method is applied to build email feature vectors based on the first 2500 characters of each email.

In terms of the parameter in the MLC method, the cost of computation and the online SVM based filter's performance were reported. It also presented the number of training emails that requested label. *θ* was set from 0 to1.

The performance is evaluated using the number of training emails (NTE), the CUP times and (1-ROCA)%. The lower the values of (1-ROCA)%, the stronger the performance of the spam filter. The CPU time of each method was computed in the same computing system.

5.2. Experimental results

The experiment compared the results of the spam filter system with and without the MLC method in order to prove correctness of the propose method (in Table 2). The best classification performance was through the Online SVM filter with MLC method.

The online SVM with Misclassification method is compared with the Low-Certainly method in email spam. In Section 4, the effects of different parameters are analyzed. The better parameters are chosen to compare the two methods. It is shown that the MLC method shows higher performance than the Misclassification method and Low-Certainty method.

The experiments also test each θ value with the MLC method put forward in Section 3 on a separate set. Figure 4, Figure 5, and Figure 6 show the test results that NTE and CPU times are decreasing when the value of θ is rising. With the value of θ ascending, the NTE will decrease in filter system. When $\theta < 0.4$, the computational cost of CPU is very little. But, when $\theta > 0.6$, it rises up sharply. However, increasing the values of θ has a great impact on the (1-ROCA) % of the filter. When $\theta < 0.3$, the performance of the filter is greatly improved. It has very little change when $\theta > 0.3$. In this paper, the value of θ is set to 0.4.

Table 2. Results compared between the MLC method and the other methods

Corpus	(1-ROCA)%		
	Misclassification	Low-Certainly	MLC
TREC06c	0.0069	0.0044	0.0004
SEWM07	0.0003	0.0000	0.0000
SEWM08	0.0037	0.0001	0.0000
SEWM10	0.0001	0.0032	0.00001
SEWM11	0.0037	0.0010	0.0000

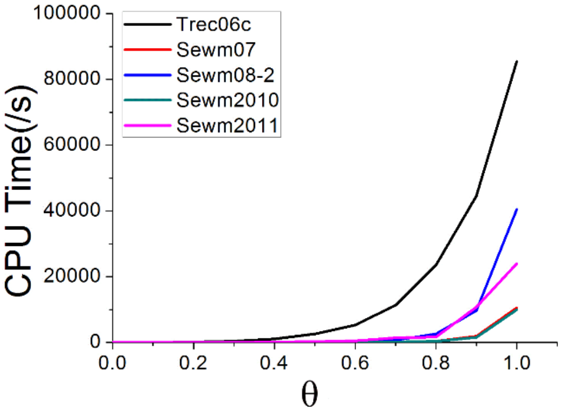


Figure 4. The change of CPU times with changed values of θ

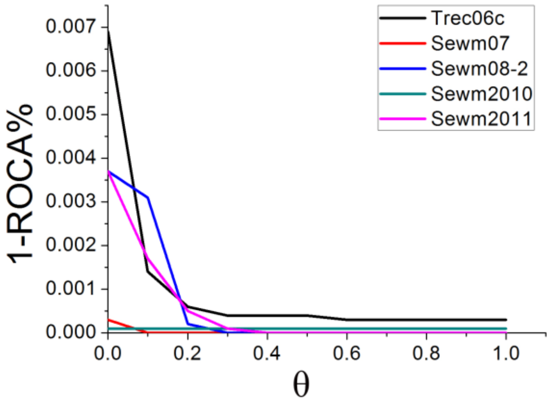


Figure 5. The change of (1-ROCA)% with changed values of θ

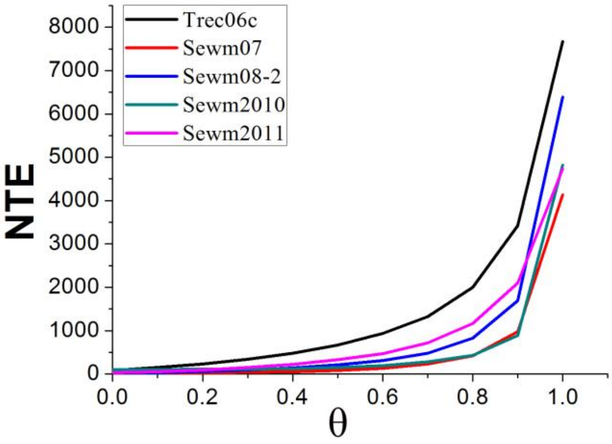


Figure 6. The change of NTE with changed values of θ

6. **Conclusions**

We have proposed a Misclassification and Low-Certainty (MLC) based active learning method to decide when to conduct label requests for new data in spam filtering. Through this method, improved results over the Misclassification method and Low-Certainty method are given. It also decreases the number of labels required to get strong performance at a small cost for computation. Moreover, the method of online active learning matches this area because spam filtering is naturally an online task. In all, the experimental results show that the method applied in this paper has achieved good performance by reducing the label requests for emails and saving computational cost on training the Model.

Acknowledgements

This work was partly financially supported through grants from the National Natural Science Foundation of China (No. 60903083 and 61502123), Scientific planning issues of education in Heilongjiang Province (No. GBC1211062), and the research fund for the program of new century excellent talents (No. 1155-ncet-008).

References

1. E. Blanzieri and A. Bryl, “Evaluation of the highest probability SVM nearest neighbor classifier with variable relative error cost,” *University of Trento*, 2007
2. G. V. Cormack and T. R. Lynam, “TREC 2005 spam track overview,” in *Proceedings of the Fourteenth Text REtrieval Conference*, pp. 500-274, 2005.
3. G. V. Cormack, “TREC 2007 spam track overview,” in *Proceedings of the The Sixteenth Text REtrieval Conference*, 2007
4. G. V. Cormack and T. R. Lynam, “Online supervised spam filter evaluation,” *ACM Transactions on Information Systems*, vol . 25, no. 3, pp. 11, 2007
5. M. Davy, “A review of active learning and co-training in text classification,” *Trinity College Dublin, Department of Computer Science*, 2005
6. S. J. Delany, M. Buckley and D. Greene, “SMS spam filtering: methods and data,” *Expert Systems with Applications*, vol. 39, no. 10, pp. 9899-9908, 2012
7. Y. Fu, X. Zhu and B. Li, “A survey on instance selection for active learning,” *Knowledge and information systems*, pp. 1-35, 2013
8. P. A. Graham, “A plan for spam,” *Available from World Wide Web: http://www. paulgraham. com/spam. html*, 2003
9. T. S. Guzella and W. M. Caminhas, “A review of machine learning approaches to spam filtering,” *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206-10222, 2009
10. G. Hulten and J. Goodman, “Tutorial on junk e-mail filtering,” in *ICML*, July 2004
11. W. Liu and T. Wang, “Active learning for online spam filtering,” *Information Retrieval Technology*, pp. 555-560, 2008
12. W. Liu and T. Wang, “Online active multi-field learning for efficient email spam filtering,” *Knowledge and Information Systems*, vol. 33, no. 1, pp. 117-136, 2012
13. N. P éRez-D íAz, D. Ruano-Ord áS, F. Fdez-Riverola and J. R. M éNdez, “SDAI: An integral evaluation methodology for content-based spam filtering models,” *Expert Systems with Applications*, vol. 39, no. 16, pp. 12487-12500, 2012
14. B. Settles, “Active learning literature survey,” *University of Wisconsin, Madison*, vol. 52, no. 55-66, pp. 11, 2010
15. D. Sculley and G. Wachman, “Relaxed online SVMs for spam filtering,” in *The Thirtieth Annual ACM SIGIR Conference Proceedings*, 2007

16. D. Sculley. “Online active learning methods for fast label efficient spam filtering,” in *Proceedings of CEAS*, 2007
17. I. Santos, C. Laorden, B. Sanz and P. G. Bringas , “Enhanced topic-based vector space model for semantics-aware spam filtering” *Expert Systems with applications*, vol. 39, no. 1, pp. 437-444, 2012
18. O. Saad, A. Darwish and R. Faraj, “A survey of machine learning techniques for Spam filtering,” International Journal of Computer Science and Network Security, vol.12, no. 2, pp. 66, 2012
19. S. Shalev-Shwartz, “Online learning and online convex optimization," *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107-194, 2012
20. J. Wang, P. Zhao, S. C. Hoi and R. Jin, “Online feature selection and its applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 698-710, 2014
21. B. Zhou, Y. Yao and J. Luo, “Cost-sensitive three-way email spam filtering,” *Journal of Intelligent Information Systems*, vol. 42, no. 1, pp. 19-45, 2014