Mail Filtering Algorithm Based on The Feedback Correction Probability Learning

Zou Xiao Yun
hubei polytechnic institute
xiaogan hubei,432000

Abstract—**With the popularity of the Internet, e-mail with its fast and convenient advantages has gradually developed into one of the important communication tools in people's lives. However, the problem of followed spam is increasingly severe, it is not only the dissemination of harmful information, but also waste of public resources. To solve this problem, the author proposed a mail filtering algorithm based on the feedback correction probability learning. The feedback correction probability training has less feedback learning data and use error-driven training in order to achieve a high classification effect. The experiment also tested the idea.**

Keywords-The feedback Probability learning; Spam; Probability

I. INTRODUCTION

Internet is an international information system and brings immeasurable economic and social benefits. At the same time, it also brings some dangerous international problems, such as e-mail spam (unsolicited e-mail)[1~3]. According to " China Internet Development Statistics Report ", China's Internet users receive 13.9 e-mail weekly in July 2004: 4.6 normal e-mail, the spam has reached 9.3. Spam spread like an epidemic in network environment, affecting normal network communications, wasting network resources, so the computer users waste a lot of time in mail identification. For users, spam has brought them a lot of harm and loss in their work, income, life, entertainment and the spiritual realm[4~6].

To solve this problem, this paper proposed a mail filtering algorithm based on the feedback correction probability learning. The feedback correction probability training has less feedback learning data and use error-driven training in order to achieve a high classification effect. The experiment also tested the idea. Comparing with the standard analytic hierarchy process, we know that the algorithm reduce the amount of calculation while ensure the correctness of the conclusion. It also makes weight to determine more reasonable and provides a scientific basis for further evaluation decisions.

II. THE MAIN IDEAS AND PRINCIPLES OF CORRECTION PROBABILITY ASSOCIATED ALGORITHM

The object of information filtering is the dynamic flow of information, thus the form and content of spam is changing over time, so mail filtering system needs to be updated according to the new changes and demands. So we introduce the feedback learning techniques to improve our spam filtering model, making the model realizing real-time updates based on the form and content of spam and the needs of users. The feedback technique is an important research methods, through the man-machine interactive way to make the output of the model return to the input, in order to improve the performance. It is a method to achieve optimized through learning misclassified part of the data , then re-training and learning. In general there are two types of feedback learning: incremental learning and re-learning. The feedback learning make full use of the results or intermediate results of the original study, and generally do not need to rescan the learned mails. Re-learning is a completely re-start learning of the new study set contains the e-mail have been learned and to be learned. E-mail sets can be divided into a training set, test set and classified set. For filtering process, the e-mail can be divided into two categories: spam and legitimate mail, which useful in spam filtering applications.

A. *The process of learning algorithm*

Feedback learning algorithm of support vector machine finds points located at the boundary of the two categories and feeds back classification error messages. Model classification processing relates to support vector of classification boundary. If system classification proved it's a correct document means that original classification model contains relevant information, no value to our feedback learning. Classification error e-mail, contain classified information not included in original model, is the focus of feedback learning. But these e-mail tend to be a small part in entire classification results set. In theory, the learning method has feedback fewer samples and improves the classification accuracy. Support vector machine classification through quadratic optimization training process to get data points set used for classification process on the boundary. These points, called support vectors, including data points correctly and incorrectly classified located on classification boundary. Support vector set K is a relatively small subset of training set E, generally speaking $K \subset E$. Support vector containing useful classified information in training document set E. Therefore, support vector machines has good incremental training learning features. Feedback learning can make full use of incremental learning advantages to re-optimize the selection of support vector set S in feedback document set F and classification model Ω, in order to gain new classification model Ω' and support set k'.

1. The analysis and optimization of support feedback vector machine

Published by Atlantis Press, Paris, France.

Construct decision function $f(x)=\operatorname{sgn}(\sum_{i=1}^{n} y_i\alpha_i^* K(x_i,x)+b^*)$, $K(.)$ is kernel function, x_i is support vector, $a=\{a_1,a_2,....a_i\}$ is optimal solution of dual problem. The KKT conditions each sample satisfied is:

$$a_i=0 \Rightarrow y_i f(x_i)\geq 1$$

$$0\leq a_i \leq C \Rightarrow y_i f(x_i)=1$$

$$a_i=C \Rightarrow y_i f(x_i)\leq 1$$

New sample satisfy KKT conditions don't change support vector set, while others do. Samples against KKT conditions can be divided into three types:

1. In classification interval, the sample and the class on the same side of classification boundary, correctly classified by original classifier, $0\leq y_i f(x_i)<1$; as S1 shown in Figure 1;

2. In classification interval, the sample and the class on the different sides of classification boundary, incorrectly classified by original classifier, $-1\leq y_i f(x_i)<0$; as S2 shown in Figure1;

3. Outside classification interval, the sample and the class on the different sides of classification boundary, incorrectly classified by original classifier, $y_i f(x_i)<-1$; as S3 shown in Figure1;

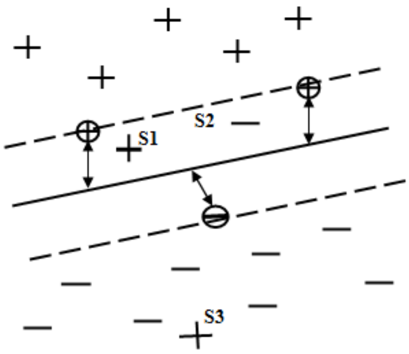

Figure 1 Samples against KKT conditions

Through analysis of the samples distribution, we found that misclassification sample is a special case against KKT conditions. Therefore, we select the wrong classified ones as the feedback learning samples, so that training programs can quickly capture new information contribute to classification and timely abandon duplicate information. Here mainly for feedback processing of training documents set E, the feedback e-mail sets need to take optimization training together with the original model of support vector set. These e-mail may have become the support vector of original classification model, and their remaining in set S will affect the model results. Therefore, the feedback learning process during e-mail training, we should firstly detect the support vector of feedback e-mail, remove the same feature vectors and add new feature vector, before the training.

B. The processes of feedback algorithm

The processes of feedback algorithm are shown as follow:

Step1: Process each e-mail $d_i \in \mathrm{R}$ in set R (set of classification results), feedback the error e-mail which also can be selected in a certain ratio, depends on our specific environment. Verify the artificial feedback classification (legitimate mail or spam) according to the e-mail characteristics, then put them into the feedback e-mail set F to prepare for the effective characteristic collection.

Step2: After document feedback processing, extract the body of the e-mail, then quantification process and feature extraction. Read support vector sets $e_j^{zx}(m_j,h_j), i=1,2,3......d$, d is the number of support vectors in the set.

Step3: Begin the feedback training learning. Get a e-mail in feedback set F and generate related document vector, then obtain support vector in set S of original classification model Ω. Feedback e-mail vector:

$$f_i^{gb}(m_i,h_i), i=1,2,3......n^g$$

Step4: Analysis the type of feedback document, if the e-mail is part of training document feedback mail, calculate the similarity of the feedback vector and the support vector. The distance between points in the space is measured by the cosine of the angle between the vectors, which represents the degree of similarity between documents. The smaller the angle the higher the similarity of the document. $sim(f_i^{gb},e_j^{zx})=\frac{f_i^{gb}e_j^{zx}}{\left\|f_i^{gb}\right\|\left\|e_j^{zx}\right\|}$, if the similarity is bigger than θ_m , two e-mail are same, so delete the feedback vector.

Step5: Repeat Step3, until all support vectors are different from each other. These vectors form set k'.

Step6: Collect recent spam characteristic phrases, then change them into the vector model denoted as M.

Step7: Take the union of F, M and k' as training set, begin Re-optimize training of support vectors.

III. Experimental Analysis

Feedback learning experiment is cumbersome and also need enough experience. We use F1 as the reference, compare the difference before and after feedback. Refer to the result of support vector before feedback, final result is average of many experiments. Preliminary experiments show that the filtering effect is improved, we will do further research experiments of this algorithm to verify the result.

In additional, when the files have strong compatible characteristics the manual classification standard will not as same as the practical training sample standard, which will cause the temple fluctuation in some specific type because the feedback learning needs the persons give related feedback types. However, this will not affect the performance of feedback learning classification. It also indicates the importance of the quality of the learning samples for classification performance.After all, we can infer that using feedback learning techniques to re-filter the E-mail is quite reasonable method based on the performance of the feedback learning in the information filter area. It utilizes the error diver to train in order to achieve good filter performance which fulfills the function of E-mail real time update and increase the filter accuracy on the base of small amount of E-mails.

IV. Conclusions

The paper proposes a scam E-mail filter algorithm based on feedback correct probability learning techniques to fix the advantage of the standard filter models. The method, which is scientific and simple, decreases the subjective factors and provides scientific basis for the scam messages filter strategy. The model can be easily applied to other area if it is combined with computers and set up appropriate evaluation systems.

Acknowledgment

Chinese book Classification Code: TP391 Document code: A.

References

[1] K.Schneider,A Comparison of Event Models for Naive Bayes Anti-Spain E-Mail Filtering[A].In: Proc.10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003) [C].Budapest,Hungary. 2003.4:307 一 314.

[2] S.S.Keerthi, K. Shevade, C. Bhattachayya et al. Improvements to Platt’ s SMO Algorithm for SVM Classifier Design[J]. Neural Computation . 13(3),2001.

[3] Lin Chunfu,Wang Shengde.Fuzzy Support Vector Machines[J]. IEEE Transactions on Neural Networks, 2002..

[4] Huang Hanpang, Liu Yihung. Fuzzy Support Vector Machines for Pattern Recognition and Data Mining [J]. International Journal of Fuzzy Systems, 4(3),2002.

[5] Li Kunlun, Huang Houkuan. An Architecture of Active Learning SVMs for Spam. Signal Processing, 2002 6th International Conference, 2002, Vol. 2, Aug. 26-30: 1247–1250.

[6] Rocchio J J. Relevance Feedback in Information Retrieval[A]. The Smart Retrieval System Experiments in Automatic Document Processing [C]. NeJersey:P rentice Hall Inc,1971. 313-323.

TABLE I. Table 1：Comparison of F1 before and after feedback

Corpus	F1 before SVM feedback	F1 after SVM feedback
Ling-spam bare	86%	87.5%
Ling-spam Lemn	86.3%	87.6%
Ling-spam Stop	86.6%	88.3%
Ling-spam Lemn_stop	87.5%	89.7%