2010 International Conference on Computational and Information Sciences

Research of a Spam Filtering Algorithm Based on Naïve Bayes and AIS

Qin Luo, Bing Liu, Junhua Yan, Zhongyue He
School of Computer Science
Southwest Petroleum University
Chengdu, 610500, China
E-mail:dorothy_lq@163.com, bing_liu001@163.com,
yanjh817@163.com, hezhongyue@yeah.net

***Abstract*—The Naïve Bayesian classifier has been suggested as an effective method to construct anti-spam filters for its strong categorization and high precision. Artificial immune system has become a new embranchment in computing intelligence for its good self-learning, self-adaptability and robustness. This paper proposes a new spam filtering means based on Naïve Bayes and AIS, and analyses the key problems of the algorithm. The accuracy rate is compared with a naïve Bayesian classifier-Bogofilter and it is shown that the proposed algorithm performs as well as Naïve Bayes and has a great potential for augmentation.**

Keywords-spam; Naïve Bayes; artificial immune; spam filtering;

I. INTRODUCTION

In recent years, the overflow of spam makes the process of using Internet greatly affected. A study in 2006[1] reported that spam messages constituted approximately 40% of the incoming messages in China and an investigated made by Anti-spam Center of ISC indicated that 17.86 spam messages were received per user per week on the 3rd quarter in 2008. The situation seems to be worsening, and without appropriate counter-measures, spam messages may undermine the usability of e-mail.

Currently, there are three main kinds of anti-spam approaches: (1) white-lists of trusted senders and black-lists of known spammers. (2) hand-crafted rules filtering approach. (3) contend-based filtering approach. White-lists reduce the risk of accidentally blocking non-spam, hereafter called ham, but they have to be combined with other techniques to discriminate between spam messages and ham messages from sender not in the white-list. Black-lists containing e-mail addresses or domain names are of little use, as spammers typically use fake sender addresses. Hand-crafted rules are also problematic: rules that are common across users can be studied by spammers, who can adjust their messages to avoid triggering the rules[2][3]. Contend-based filtering approach is based on the premise that distinctive features can be discovered and used for sake of classifying a message[4]. It can be roughly divided into the rule-based approach and the statistical-based approach. A rule-based approach expresses the domain knowledge in terms of a set of heuristic rules. However, building a rule-based system often involves acquiring and maintaining a huge set of rules with an extremely high cost. And a statistical-based approach expresses the differences among messages in terms of the likelihood of certain events. A statistical-based model performs well on one corpus may work badly on another one with quite different characteristics. Therefore, finding an efficient method to prevent spam is imperative.

Naïve Bayes is the most popular statistical-based anti-spam method for its strong categorization and high precision. But it is week in self-learning and self-adaptability. Artificial immune system has impressive performance on recognition, learning and memorizing. A better way may be combining the advantages of both approaches into a single model. In this paper, we attempts to combine the mechanism of Naïve Bayes and artificial immune system, proposes a hybrid spam filtering algorithm based on the two algorithms, and then solves the key problems of the algorithm. Our experiment results proved that this algorithm is effective to filter spam in SpamAssassin corpus.

II. KEY TECHNOLOGY

A. Naïve Bayes

The Naïve Bayes is the simplest and most-widely used algorithm that derived from Bayesian Decision Theory[Duda and Hart 1973]. From Bayes' theorem and theorem of total probability, the probability that a document d with vector x = $<x_1,\ldots,x_n>$ belongs to category c is:

$$P(c_j \mid d_x) = \frac{P(c_j)P(d_x \mid c_j)}{\sum_{j \in \{spam, ham\}} p(c_j)P(d_x \mid c_j)} \quad (1)$$

The Naïve Bayesian classifier assumes that $x_1,\ldots,x_n$ are conditionally independent given the category c. This is so-called "Naïve Bayes assumption". While this assumption is clearly false in most real-world tasks, Naïve Bayes often performs classification very well[5][6]. $p(c_j)$ is easy to estimate from the frequencies of the training corpus. Now, only a posterior probability $P(d_x \mid c_j)$ has to be estimated. Naïve Bayes classifier has two different generative models: multi-variate Bernoulli event model(MBM) and multinomial model(MM), both of which make the Naïve Bayes assumption. The $P(d_x \mid c_j)$ on the two models is different. McCalluml's test indicated that the difference in performance between the two models is not great. Multi-variate Bernoulli event model even performs better than multinomial at small vocabulary sizes. And it is simpler[7]. In the multi-variate Bernoulli event model, a document is a binary vector over the space of words. Given a vocabulary V,

This work is sponsored by Science Funds of Southwest Petroleum University

978-0-7695-4270-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICCIS.2010.43

each dimension of the space t, t{1,…, V}, corresponds to word w_t from the vocabulary. Dimension t of the vector for document d_x is written $B_{xt,}$ and is either 0 or 1, indicating whether word w_t occurs at least once in the document. Then, the probability of a document given its class from (2) is simply the product of the probability of the attribute values over all attributes:

$$P(d_x \mid c_j) = \prod_{t=1}^{|V|} (B_{xt}P(w_t \mid c_j) + (1 - B_{xt})(1 - P(w_t \mid c_j))) \quad (2)$$

Given a set of labeled training documents, D = {d1,…d|D|}, The probability of word w_t in class c_j is:

$$P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} B_{xt}P(c_j \mid d_i)}{2 + \sum_{i=1}^{|D|} P(c_j \mid d_i)} \quad (3)$$

B. AIS

Artificial immune system (AIS) is a biologically inspired paradigm for information processing. From the viewpoint of calculation, bio-immune system is a fully functional system with high parallel, distributed, self-adaption and self-organization. It has a strong ability of learning, recognizing, memorizing and feature extracting. We consider the immune system particularly suitable inspiration for spam filtering technology because of certain properties inherent in most immune inspired algorithm[8][9]. Examples of these include:

- Pattern recognition. By perceiving spam as antigens, the mails can be classified according to the recognition capability of bio-immune system.
- Self-adaption. Like living organism often facing variety of invaders, the pattern and definition of spam are always evolving continuously. So spam filters should follow up and adapt to these changes.
- Noise tolerance. The natural immune system is tolerant towards noise. AIS has the potential to filter noisy data and uncover an underlying concept.

In this paper, we used population-based immune algorithm. This immune algorithm has the following steps[10]:

a) Define the antigen.

b) Generate a set of antibodies.

c) Calculate affinity S.

d) Clonal selection.

e) Update new population of antibodies based on user feedback.

III. SPAM FILTERING ALGORITHM BASED ON NAÏVE BAYES AND AIS

A. Design of the Hybrid Algorithm

Based on above analysis, a hybrid spam filtering algorithm is proposed by combining the Naïve Bayes and AIS. The basic idea of this algorithm is: when a new mail arrives, extract the vector space model (VSM) to produce a set of antigens. These antigens are firstly recognized by memory cells detector, if match, then the mail is flagged as spam, otherwise antigens are recognized by immature cells detector. That is, affinity between antigens and antibodies is calculated. If the affinity is higher than a pre-defined threshold, the mail is non-self, that is spam, and otherwise the mail is ham. Here, calculation of affinity is very important, which decide the effect of spam filtering. Based on the structure of antibody and antigen, we design the affinity formula.

The main steps required for hybrid algorithm are summarized as follows:

a) A new mail arrives.

b) The mail produce a set of antigens after pretreatment.

c) These antigens are recognized by memory cells detector, if match, then the mail is flagged as spam.

d) Otherwise, calculate the affinity of antigens and the antibodies in immature cell detector.

e) If the affinity is higher than a certain threshold, the mail is flagged as spam, otherwise the mail is flagged as ham.

f) Based on the user feedback, the antibodies libraries of memory cells and immature cells are updated by the antigens.

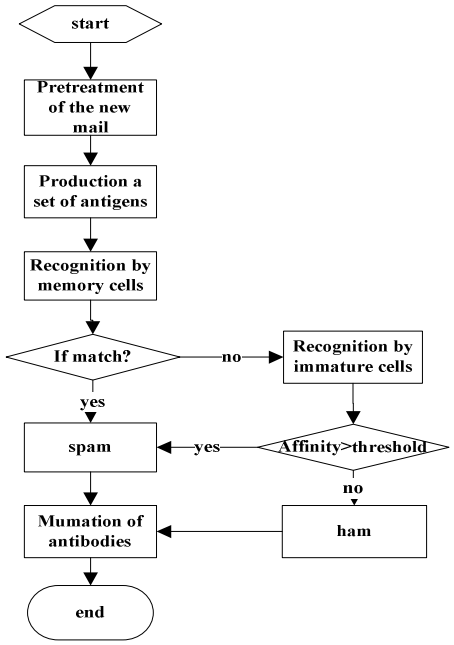

Figure 1. Flow chart of the hybrid algorithm

B. Implementation of the Hybrid Algorithm

1) Antigen/Antibody: Antigens are the feature vector of a new mail. A set of antigens are the VSM linked list of the mail. The definition of antigens is below:

```
typedef struct{
    char[ ] identifyString;
    int gfreqence;    //the number of occurrences of the antigen in ham
```

int sfrequence; //the number of occurrences of the antigen in spam

}Antigen;

We use another structure to recognize antigens. These are called antibodies. The definition of antibodies may be outlined as follows:

typedef struct{

char [ ] identifyString;

int gfrequence; //the number of occurrences of the antibody in ham

int sfrequence; //the number of occurrences of the antibody in spam

float fw; //the probability of occurrences of the antibody in spam

float pw; //the probability of occurrences of the antibody in ham

int life; //the lifecycle of the antibody

int numOfCapture; // the number of recognition of the antibodiy

int numOfSuccess; //the number of a mail recognized the antibody is spam

}Antibody;

Because the numbers of antigens and antibodies are not certain and linked list can dynamic increase, we use hash linked list to store them. Based on Naïve Bayes MBM, f_w and p_w are given by (4) and (5), where linked list l_antibody is used to store antibodies, p_antibody is the pointer pointing any cell in l_antibody, N_s and N_l are the number of spam and ham.

$$(*p\_antibody).f_w = \frac{1+sfrequence}{2+N_s} \tag{4}$$

$$(*p\_antibody).p_w = \frac{1+gfrequence}{2+N_l} \tag{5}$$

2) Affnity Measure: The affinity between two cells is a measure of the proportion of one cell's feature vector also present in the other. It is used throughout the algorithm and is guaranteed to return a value between 0 and 1. The mail is spam if the affinity is higher than threshold; otherwise it's ham, i.e. by (6), where p_{spam} refers to the affinity.

$$p = \begin{Bmatrix} 1, & p_{spam} > threshold \\ 0, & p_{spam} < threshold \end{Bmatrix} \tag{6}$$

Based on Naïve Bayes MBM, p_{spam} is given by (7), where $p(wi \mid c=1)$ is f_w of an antibody and $p(wi \mid c=0)$ is p_w of an antibody.

$$\begin{aligned} p_{spam} &= p(c=1 \mid d) = \frac{p(c=1)\times p(d \mid c=1)}{p(d)} \\ &= \frac{p(c=1)\times p(d \mid c=1)}{p(c=1)p(d \mid c=1) + p(c=0)p(d \mid c=0)} \\ &= \frac{\frac{Ns}{N}\times \prod_{i=1}^{n} p(wi \mid c=1)}{\frac{Ns}{N}\times \prod_{i=1}^{n} p(wi \mid c=1) + \frac{N_l}{N}\times \prod_{i=1}^{n} p(wi \mid c=0)} \end{aligned} \tag{7}$$

3) Memory Cell: To lower false-positive rate, the mail is flagged as spam when the affinity is higher than threshold at first response. Then, find the antibodies who meet the evolve function, defined by (8), to evolve into memory cell. If a feature word in the next mail matches a memory cell, then the mail is flagged as spam directly[8]. In our algorithm, the structure of a memory cell is the same as an antibody. But a memory cell has a longer lifecycle.

When antibodies are created, a lifecycle, an integer value representing a countdown to its death, is defined. If it equals to zero, the cell will die. The dead cells won't occupy the system resource, so it can improve the efficiency of the algorithm. The value of EVOLUTERATE must be high enough to avoid a legitimate message being mistakenly classified as spam.

$$evolute(Antibody) = \begin{Bmatrix} 1, \ if\ (Antibody.numOfSuccess \ / \ Antibody.numOfCapture) \ge EVOLUTERATE \\ 0, \ otherwise \end{Bmatrix} \tag{8}$$

Memory cell can greatly improve the efficiency of our algorithm.

IV. RESULTS

The corpus used in this algorithm is acquired from http://www.spamassassin.org, which is provided by Justin Mason of Network Associates. A total of 6047 mails are used in SpamAssassin corpus, including 4150 spam and 1897 ham.

In classification tasks, performance is often measured in terms of accuracy (Acc) or error rate (Err = 1-Acc). Let N_{ham} and N_{spam} be the total numbers of ham and spam messages, respectively, to be classified by the filter, and $n_{Y \to Z}$ be the number of messages belonging to category Y that the filter classified as belonging to category Z ($Y, Z \in \{spam, ham\}$). Then:

$$Acc = \frac{n_{ham \to ham} + n_{spam \to spam}}{N_{ham} + N_{spam}}$$

$$Err = \frac{n_{ham \to spam} + n_{spam \to ham}}{N_{ham} + N_{spam}}$$

In our test, we found that there are many factors will impact the classification, such as evolve function, the number of immature cells and memory cells, the value of threshold and lifecycle. Through repeated tests, we choose the best-performing configuration for our algorithm. 10-fold cross-validation was used in our test: the corpus was partitioned randomly into ten parts, and the experiment was repeated ten times, each time reserving a different part for testing and using the remaining nine parts for training. All the results were then averaged over the ten iterations. The parameters of our algorithm are shown in table Ⅰ. A legal range for each parameter is also indicated.

TABLE I. PARAMETERS

Parameter	*Value*	*Range*
no. of immature cells	200	>0
no. of memory cells	100	>0
threshold	0.99	0-1
EVOLUTERATE	0.97	0-1

In order to validate the results, we test our algorithm against publicly available Bogofilter, which is a fast Bayesian spam filter. We firstly used 4000 mails(2000 spam) from test corpus to train the two anti-mail filters. Then we varied the number of testing mails from 200 to 2000 by 200 for testing. There are 1500 spam in testing mails, which have many mutation words, such as "Viagra" of "Viagra". Figure 2 shows the results of the two filters.

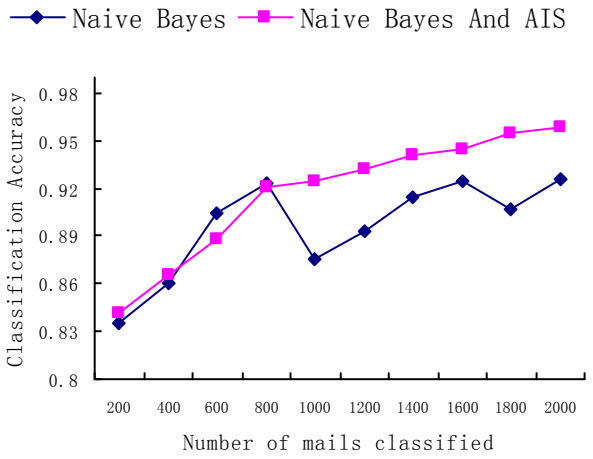

Figure 2. Comparison of Bogofilter and our algorithm

From above testing, we found that the accuracy rate of the two algorithms is closely matched in general at initial phase. This is because there is little mutation spam at initial phase and the two filters use the same corpus for training. As the number of testing mails increase, there are some areas where the changing data causes them to behave differently. Of interest are the areas 800 to 1000 and 1600 to 1800. In both situations our algorithm exhibits an increase in accuracy while there is a decrease in accuracy from Bogofilter. Our explanation of this could be that the hybrid algorithm is faster to react to sudden changes. Consider a word that has been very common among spam for example. Our algorithm will represent this detail as the presence of this word in memory cell set. The Naïve Bayesian algorithm will represent this as a high frequency of occurrence in spam class compared to other class. Consider now this word begins to be used in ham. Our algorithm will react quickly by deleting the word that would result in a misclassification. By contrast, the Bayesian algorithm will react by only incrementing the frequency count of this word in ham class by one. The method results in a negligible effect on calculation of final class probability. That is, the mail may still be misclassified.

V. CONCLUTION

In this paper, we propose a hybrid algorithm based on Naïve Bayes and AIS and compare the accuracy rate of Bogofilter and our algorithm using the same corpus. Our findings demonstrated that the hybrid algorithm not only achieved high classification accuracy at first, but also has self-learning, self-adaptability and robustness.

REFERENCES

[1] Anti-spam Center of ISC, 2007 Anti-spam Investigation Report [DB/OL]. http://www.anti-spam.cn, 2007-04-05.

[2] I.Androutsopoulos, G. Paliouras, E. Michelakis, "Learning to Filter Unsolicited Commercial E-Mail," Technical Report of National Centre for Sciential Research "Demokritos", 2004.

[3] I. Androutsopoulos, J. Koutsias, K. Chandrinos and C. Spyropoulos, "An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Personal E-mail Messages," Athens, Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), 2000, pp.160-167.

[4] Adriano Veloso, Wagner Meira Jr., "Lazy Associative Classification for Content-based Spam Detection," LA-Web '06. Fourth Latin American, Oct. 2006, pp. 154-161.

[5] P.Domingos and M.Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol.29, 1997, pp. 103-130.

[6] Nir Friedman, Dan Geiger, and Moises Goldszmidt, "Bayesian network classifiers," Machine Learning, vol.29, 1997, pp. 131-163.

[7] Andrew McCallum, Kamal Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, 1998, pp. 41-48.

[8] Kim J, Bentley P., "Towards an Artificial Immune System for Network Intrusion Detection: An Investigation of Dynamic Clonal Selection," Proc. Congress on Evolutionary Computation, Seoul, Korea, 2001, pp. 27-30.

[9] Secker A, Freitas AA, Timmis J., "AISEC:an Artificial Immune System for E-mail Classification," The 2003 Congress Of Evolutionary Computation[C], 2003(3), pp. 131-138.

[10] Dasgupta D, Forrest S., "Artificial Immune Systems in Industrial Applications," Proc. 2nd International Conference on Intelligent Processing and Manufacturing of Materials, Honolulu, 1999, pp. 257-267.

[11] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras and C.D. Spyropoulos, "An Evaluation of Naïve Bayesian Anti-Spam Filtering," Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learnin g (ECML 2000), Barcelona, Spain, 2000, pp. 9-17.

[12] Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz, "A Bayesian Approach to Filtering Junk E-mail," Proc. of AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin,1998, pp. 55-62.