# Statistical Machine Learning Used in Integrated Anti-Spam System

**Abstract:**

IASS is the integrated anti-spam system, which adopts machine learning to filter spam in a intelligent, flexible, precise, and self-adaptive way. The methods of linear classification based on optimal separating hyperplane and K-means clustering are used in action recognition layer. The method of improved naive Bayes is used in content analysis layer. The application of machine learning helps improve the performance of IASS.

SECTION 1.

## Introduction

With the rapid development of Internet, great numbers of junk emails come forth. This kind of emails is called spam. Spam disturbs the daily grind of people and anti-spam becomes a very important task. *Integrate Anti-Spam System*, supported in part by *National Project 863 of China*, is an effectual tool to stop spam-making. Just as showed by figure 1, IASS is composed of 4 layers.

In IASS, the rule-based and statistical Machine Learning is adopted broadly in combination to filter spam with the intelligent, flexible, precise, and self-adaptive method. This paper will discuss the application of statistical Machine Learning in IASS, especially in the Action Recognition Layer and Content Analysis Layer. How to understand Machine Learning? A definition is given by [1]: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. It is concluded that Machine Learning includes 3 elements: E (experience), T (tasks), and P (performance). There're many modules in which Machine Learning is used in IASS. In order to complete the design and implement of every learning sub-system in these modules, below things must be taken into account:

* The exact type of knowledge to be learned
* A representation for this target knowledge
* A learning mechanism

SECTION 2.

## Action Recognition

Machine Learning used in Action Recognition will be discussed in below two aspects.

### 2.1. Against IP-fixed spam-maker

Some spam-makers use a fixed IPaddress sending large mount of emails to a mail server in a relatively short period. This kind of attack is easily recognized by men, but very difficultly by a machine. Among this kind of attacks, Directory Harvest Attack is typical. It is reported DHA has been occupied *1/4* of all SMTP (Simple Mail Transport Protocol) traffic. The procedure of DHA is showed by figure 2.

A spam-maker sends an email to thousands of possible addresses in the same mail domain. Majority of these addresses are invalid and rejected by the mail server. Minorities of these addresses are valid and thus the spam-maker can get a legal-address list, which can be used to send more spasm. DHA can be reduced by limiting the number or frequency of invalid addresses, or the proportion between invalid and valid addresses, during a fix time window. However, how if a spam-maker has held a legal-address and sent more spam to them? Maybe a method is monitoring the number and frequency sending emails from a certain IP. But many foxy spam makers can control the sending frequency, thus determining a threshold distinguishing spam-makers and legal email subscribers becomes a hard task. Under such situation, sending frequency jitter is expected as a vane. More situations, only one or two vanes are not sufficient. Therefore, making use of multi-vane synthetically seems a better choice.

Based on the adequate log on our mail server, we observed large amount of SMTP session information about spam sending from a fix IP and below features are concluded:

* A fixed IP sends large amount of emails to a mail domain in a relatively short time
* The interval of sending emails is stable
* More than one *rcpt to* commands appear in a SMTP session in some situations
* In some situations, many receiver addresses are hypothesized, so many don't exist and correlation between them are relatively apparent

With respect to above features, 5 vanes are important to a single IP of sending emails during a time window:

* f: The frequency of sending emails, defined by number of SMTP sessions per minute
* j: The jitter of sending interval, defined by max interval subtract min interval
* e: The number of sending a email with invalid receiver addresses
* n: The average number of receivers (or the average number of *rcpt to* commands in a single SMTP session)
* s: The average receivers' similarity degree (we used a computational method, and the degree is zero when only one receiver exists)

Above every vane has approximately linear relation with the possibility that the IP is a spam-maker. At present, we have many difficulties in providing a strict mathematical testify about this relation. So the approximately linear relation is empirical. For this approximately linear relation, we can judge a SMTP session whether from a spam maker by:

*G*(*x*)=*sign* *f*(*x*)(1)

When G(x)=1, it is from a spam maker. Here,

*f*(*x*)=*xTβ*+*β*0(2)

dimensions column vector x=(f,j,e,n,s)T. So far, the task of machine learning is to find proper coefficients in equation [2](https://ieeexplore.ieee.org/document/#deqn2). We use vectorizable SMTP session log obtained from our mail server as input knowledge. These log information has been labeled as positive or negative. We use Optimal Separating Hyperplane as basic learning mechanism.[2] So, the question is equivalent to an optimization question:

max*β*,*β*0*C* *s*.*t* *yi*(*xTiβ*+*β*0)≥*C*∥*β*∥,*i*=1,⋯,*N*(3)

This optimization question can be resolved by standard mathematical software, and its resolving detail is discussed in [3]. The form of proper coefficients is:

*β*=∑*i*=1*Nαiyjxi*(4)

As showed by equation [4](https://ieeexplore.ieee.org/document/#deqn4), the coefficient is represented by the linear combination of support points. These support points form the classification hyperplane.

### 2.2. Against Dynamic IP spam-maker

More foxy spam-makers have ability to send spam with more than one IP at the same time. Generally, this includes 2 situations: 1) they own several IPs that distribute in a large scale and it is hard to find out the relationship among these IPs 2) they own a IP segment and there is some rules between the dynamic-changed IPs. Under the first situation, recognizing the attack with machine learning is a very hard task. However, under the second situation, by examining the IPs' sequence within a time window, it is possible that a program finds out the IP-changing rule and predicts whether an IP is next used IP or not.

The key of this problem is how to distill such a IP-sequence from a great many SMTP connections during the mail server is working as usual. Because spam-makers always send similar content within a period, we can detect those SMTP session which sent identical or almost-identical emails. The barrier of this method is how to decide the similarity between emails. The algorithm is always very complex and high-cost. A better method is gathering all the IPs which sent emails during a period. Although the data is not pure, for there are normal email users' connections, normal IPs distribute with a large scale and spam-makers' IPs do not so. Thus, data clustering algorithm can be used to find out these bad IPs. We use K-means Clustering and the algorithm is described below. [3]

1. examine a most-recent time window and collect all IPs which have SMTP sessions, the length of the time window is T.
2. pre-process all the IPs, produce a set which embraces many 1-dimension vectors (points). The simplest pre-process method is transferring a IP to an unsigned long integer.
3. select the number of expected clustering centers R. It means the number of IP-scales used by spam-makers we may find out is R. In simplest situation, we set R 1, which means we expect find out a spam-maker is sending spam from many IPs which distribute within a small scale and center around some IP.
4. to every expected clustering center, find out all the points whose distances from this center are nearer than those from any other center. These points compose the Cluster of this center.
5. calculating the mean of each Cluster, make it become the new expected clustering center. Then repeat step 4 until every Cluster is constringent or the number of iteration is beyond a threshold.
6. if a constringent Cluster is found out, there is a dynamic IP attack. Otherwise, there is not.
7. if a constringent Cluster is found out, when a new SMTP connection comes, check this IP belongs to this scale or not. If so, block it or send a alert email to the administrator.

SECTION 3.

## Content analysis

Content Analysis Layer in IASS makes use of Improved Naïve Bayes and SVM in combination to classify a email into system-defined and user-defined subject. In this chapter, the discussion is focused on how to create a new user-defined subject class based on user-collected email corpus with Improved Naïve Bayes [4]. Like other Bayes algorithms, we suppose the contribution for classification of one word depends on the number that the word appears in a text. But it is difficult to remove the useless common characteristic between classes.

The improved algorithm uses the CHI counting value, which will reduce the weight of characteristics that shared between classes, in order to strengthen classification performance. In order to reduce the influence of daily words, a parameter is changed as follows:

Among them: F is a characteristic vector of the text, Pr(w|D′) is the number of w in text d’ divided by the total word number in d’, Pr (Cj) is the number of articles belongs to Class Cj divided by the total number of articles used for training. CHI(w, Cj) is the *χ*2 statistical result between the word w and the class Cj.

SECTION 4.

## Conclusion

In IASS, rule-base and statistical machine learning are used in combination. In this paper, the discussion is focused on the latter. Rule-based machine learning is used typically in the User-Defined Rules model of IASS. The method of linear classification based on Optimal Separating hyperplane is used to judge whether an IP is multi-sending spam according to its 5 SMTP session features in this connection and historic many connections during a recent period. The method of K-means clustering is used to monitor recent N connections and obtain a sequence of IPs that ask for SMTP connections, and then analyze this sequence to judge whether there is a dynamic-IP spam multi-sending. Above two methods are adopted in Action Recognition Layer in IASS. The method of Improved Naïve Bayes is used to build up a new user-defined filtering subject based on user-collected corpus. This method is adopted in Content Analysis Layer in IASS. At present, these methods have shown rather good performance in IASS when protecting our mail server. Using Improved Naïve Bayes, several user-defined subject have achieved precision 90% and recall 80%. However, because there are many difficulties to evaluate the performance against dynamic real-world spam attacks, we can not provide quantitative performance specification about two methods used in Action Recognition Layer at present. We are planning a complex simulation environment so as to carry out the thorough Action Recognition Layer's performance evaluation. This is our exigent task in the near future.