IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 5 (Sep.-Oct. 2012), PP 07-10
www.iosrjournals.org

A Model for Fuzzy Logic Based Machine Learning Approach for Spam Filtering

[1]Mehdi Samiei yeganeh, [2]Li Bin, [3]G.Praveen Babu

[1, 2](*M.Tech.* (S/w. Eng.),[3] (Associate Professor, *School of Information Technology, Jawaharlal Nehru Technological University, Iran)*

Abstract: *It is definitely impossible to say, who was that first person to come up with the simple idea of sending out a public announcement to millions of people, and then at least one person will react to it no matter what is the proposal. E-mail provides a perfect way to send these millions of advertisements without any for a sender, and this fortunate fact is nowadays extensively exploited by several organizations. As a result, the e-mail boxes of millions of people get cluttered with all these so-called Unsolicited Bulk E-mail (UBE) also known as "spam" or "junk mail". E-mail spam, is a subset of electronic spam involving nearly identical messages sent to numerous recipients through e-mail. Definitions of spam usually include the aspects that e-mail is unsolicited and sent in bulk. Another subset of UBE is UCE (Unsolicited Commercial E-mail). The opposite of "spam", e-mail which one wants, is called "ham", usually when referring to a message's automated analysis (such as Bayesian filtering). Machine learning techniques now days are used to automatically filter the spam e-mail in a very successful and efficient way. In this paper we consider some of the machine learning methods such as Naïve Bayes, Artificial Neural Networks, Artificial Immune System Classifier methods, and fuzzy logic.*

Keywords*: spam, machine learning, Naïve bayes, artificial immune system, fuzzy logic.*

I. Introduction

Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. Spam is waste of time, storage space and communication bandwidth. The problem of spam e-mail has been increasing for years. In recent statistics, 40% of all e-mails are spam which is about 15.4 billion e-mails per day and that costs internet users about $355 million per year [1].

Most spam is commercial advertising, often for dubious products, get-rich-quick schemes, or quasi-legal services. Spam costs the sender very little to send -- most of the costs are paid for by the recipient or the carriers rather than by the sender. There are two main types of spam, and they have different effects on Internet users. Cancellable Usenet spam is a single message sent to 20 or more Usenet newsgroups. (Through long experience, Usenet users have found that any message posted to so many newsgroups is often not relevant to most or all of them.) Usenet spam is aimed at "lurkers", people who read newsgroups but rarely or never post and give their address away. Usenet spam robs users of the utility of the newsgroups by overwhelming them with a barrage of advertising or other irrelevant posts. Furthermore, Usenet spam subverts the ability of system administrators and owners to manage the topics they accept on their systems.

E-mail spam targets individual users with direct mail messages. E-mail spam lists are often created by scanning Usenet postings, stealing Internet mailing lists, or searching the Web for addresses. E-mail spam typically costs users money out-of-pocket to receive. Many people - anyone with measured phone service - read or receive their mail while the meter is running, so to speak, spam costs them additional money. On top of that, it costs money for ISPs and online services to transmit spam, and these costs are transmitted directly to subscribers. One particularly nasty variant of e-mail spam is sending spam to mailing lists (public or private e-mail discussion forums.) Because many mailing lists limit activity to their subscribers, spammers will use automated tools to subscribe to as many mailing lists as possible, so that they can grab the lists of addresses, or use the mailing list as a direct target for their attacks.

Machine learning techniques now-a-days are used to automatically filter the spam e-mail such as Naïve bayes[5,1], artificial immune system[1], and fuzzy logic[5]. Machine Learning is the study of methods for programming computers to learn. Computers are applied to a wide range of tasks and for most of these it is relatively easy for programmers to design and implement the necessary software [2]. Machine Learning [3] is a natural outgrowth of the intersection of Computer Science and Statistics. We might say explaining the question of Computer Science is "How can we build machines that solve problems, and which problems are inherently tractable/intractable?" The question that largely explains Statistics is "What can be inferred from data plus a set of modeling assumptions, with what reliability?" The explaining question for Machine Learning builds on both, but it is a distinct question, whereas Computer Science has focused primarily on how to manually program computers. Machine Learning focuses on the question of how to get computers to program themselves [4]. A

Machine Learning system usually starts with some knowledge and a corresponding knowledge organization, so that it can interpret, analyze, and test the knowledge acquired.

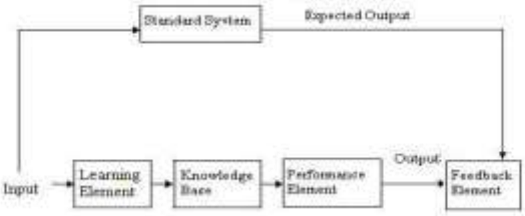

Figure 1 Typical learning system model

In general, a spam filter is an application which implements a function:

$$f(m,\theta) = \begin{cases} C_{spam}, \text{ if message } \mathbf{m} \text{ is considered spam} \\ C_{leg}, \text{ if message } \mathbf{m} \text{ is considered legitimate mail} \end{cases}$$

Where **m** is a message to be classified, θ is a vector of parameters and C_{spam}, Cleg are labels assigned to the messages. Most of the spam filters are based on machine learning classification techniques. In a learning-based technique the vector of parameters θ is the result of training the classifier on a pre-collected dataset:

$\theta = \Theta$ (M), M = $\{(m_1, y_1),(m_2, y_2),...(m_n, y_n)\}$, $y_i \in \{ C_{spam}, Cleg\}$,

Where m_1, m_2, ...m_n are previously collected messages, y_1, y_2, ...y_n are the corresponding labels, and Θ is the training function[6]. In the remain sections of this paper we describe the most important machine learning approach such as Naïve Bayes, Artificial Neural Networks, Artificial Immune System classifier methods, and also fuzzy logic method.

II. Naïve Bayes classifier method

In 1998 the Naïve Bayes classifier was proposed for spam recognition [1]. Bayesian spam filtering [12] is a statistical technique of filtering. It makes use of a Naïve Bayes classifier to identify spam e-mail. Bayesian classifiers work by correlating the use of tokens (typically words, or sometimes other things), with spam and non-spam e-mails and then using Bayesian inference to calculate a probability that whether an e-mail is or is not spam. Bayesian spam filtering is a very powerful technique for dealing with spam, that can tailor itself to the e-mail needs of individual users, and gives low false positive spam detection rates that are generally acceptable to users. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham e-mail in its database. If the total words probabilities exceed a certain limit, the filter will mark the e-mail to a specific category. Here, only two categories are necessary: spam or ham. Almost all the statistic-based spam filters use Bayesian probability calculation to combine individual token's statistics to an overall score [7], and make filtering decision based on the score [1]. Following algorithm [8] is used for filtering spam by this method:

Stage1. Training

Parse each e-mail into its constituent tokens. Generate a probability for each token W.

$S [W] = C_{spam} (W) / (C_{Ham} (W) + C_{spam} (W))$ store spam rating values to a database

Stage2. Filtering

```
For each message M
        While (M not end) do
 Scan message for the next token $T_i$
 Query the database for spam rating $S (T_i)$
 Calculate accumulated message probabilities
    S [M] and H [M]
Calculate the overall message filtering indication by: I [M] = f(S [M], H [M])
f is a filter dependent function,
Such as I [M] = 1+S [M]-H [M] □ 2
```

Where $C_{spam}(T)$ and $C_{Ham}(T)$ are the number of spam or ham messages containing token T respectively. To calculate the possibility for a message M with tokens $\{T_1,......,T_N\}$, one needs to combine the individual token's spam rating to evaluate the overall message spam rating and

(H [M] = N (1- S [TI])).

III. Artificial immune classifier method

Immune algorithms are bio-inspired algorithms which have been inspired by human immune system [9, 10]. A role of the immune system is to protect our bodies from infectious agents such as viruses, bacteria. On the surface of these agents are antigens that allow the identification of the invading agents, thus provoking an immune response Recognition in the immune system, which is performed by lymphocytes. Each lymphocyte expresses receptor molecules of one particular shape on its surface called antibody. An elaborate genetic mechanism involving combinatorial association of a number of gene segments underlies the construction of these receptors. The overall immune response involves three evolutionary methods: gene library, negative selection and clonal selection. In gene library, antibodies recognize antigens by the complementary properties that belong only to antigens. Thus, some knowledge of antigen properties is required to generate competent antibodies [1]. The following algorithm [10] is used for immune classifier:

```
 (An email message m)
For (each term t in the message) do
{
    If (there exists a detector p, based on base
    String r, matches with t) then
   {
      If (m is spam) then
      {
       Increase r’s spam score by s-rate;
       }
     Else
     {
        Increase r’s ham score by ns-rate;
      } }
Else {
If (m is spam) then {
  If (detector p recognizes t and edmf(p,t) > threshold)
    Then {
          The differing characters are added to its corresponding entry in the library of character
          Generalization rules;
     }       else {
        A new base string t is added into the library of base strings;
       }
     }}
Decrease the age of every base string by a-rate;
}
```

IV. Fuzzy logic method

The concept of fuzzy logic was introduced in 1965 by Lotfi Zadeh in his seminal paper [11]. Professor Zadeh reasoned that people do not require precise, numerical information input, yet they are capable of highly adaptive control. If feedback controllers could be programmed to accept noisy, imprecise input, they would be much more effective and perhaps easier to implement. Unfortunately, U.S. manufacturers have not been so quick to embrace this technology while the Europeans and Japanese have been aggressively building real products around it. Fuzzy logic deals with fuzzy sets that allow partial membership in a set. The membership in these vaguely defined sets is represented by the degree of relevance. This provides flexibility in dealing with uncertainty in systems such as spam filtering. Using a fuzzy similarity approach, a classification model is built from a set of pre-classified e-mail instances [5].

In this method we have three stages for filtering the spam. First one is Pre-processing, second one is Training and the last one is Classification. In first step, all HTML tags are stripped off. Then, all stop words, i.e. words that appear frequently but have low content discriminating power, are removed from each e-mail message. During the training phase, a model is built based on the characteristics of each category in a pre-classified set of e-mail messages. The training data set should be selected in such a way that it is varying in content and subject. Each sample message is labeled with a specific category. The message is then classified by comparing its fuzzy similarity measures as given in figure 2.

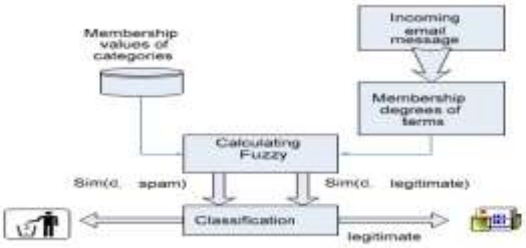

Figure 2. Classification of a new instance E-mail message.

$$\mu_d(t_i) = f_{i,d} \; / \; \max\{f_{j,d}\}$$

where the $T_j \in d$ membership degree of each token, $f_{i,d}$ is the number of occurrences of token t_i in message d. Thus, the token with the maximum number of occurrences will be assigned a value of 1 and all other tokens will be assigned proportional values. The fuzzy similarity measure (SM) is given by,

$$SM(d, c_j) = \frac{\sum_{t \in d} \mu_R(t, c_j) \otimes \mu_d(t)}{\sum_{t \in d} \mu_R(t, c_i) \oplus \mu_d(t)}$$

and if SM (d, sp M (d, legitimate) > µ that µ>1 then d is spam[5].
In the previous approach [5], the spam mails are only seen. It means a user can only view the spam mails but cannot ensure the deletion of the spam emails. Whereas in our approach, the application will enable the user to locate the spam mails and then delete give you the user an option to delete as well on its own.

V. Conclusion

The fuzzy logic considers the content of the message to predict its category rather than relying on a fixed pre-specified set of keywords. Thus, it can adapt to spammer tactics and dynamically build its knowledge base. Based on the idea in [5] we have attempted to enhance the functionality of the model proposed here. Here in this paper we have enhanced by including the feature identification of the spam mails in the e-mail account and deletion performed on its own, which was not available in previous approach.

References

[1]. MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION” W.A. Awad1 and S.M. ELseuofi, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[2]. “Machine Learning”Thomas G. Dietterich Department of Computer ScienceOregon State UniversityCorvallis, OR 97331.

[3]. “Machine-Learning Approaches for Classifying Haplogroup from Y Chromosome STR Data” Christos A. Ouzounis, King's College London, United Kingdom.

[4]. “The Discipline of Machine Learning”Tom M. Mitchell July 2006 CMU-ML-06-108.

[5]. “A Fuzzy Similarity Approach for Automated Spam Filtering” El-Sayed M. El-Alfy , Fares S. Al-Qunaieer , 978-1-4244-1968-5/08/$25.00 ©2008 IEEE.

[6]. “A Survey of Learning-Based Techniques of Email Spam Filtering” Enrico Blanzieri , Anton Bryl University of Trento, Italy.

[7]. “Binary LNS-based Naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis”, M. N. Marsono, M. W. El-Kharashi, and F. Gebali,IET Computers & Digital Techniques, 2008.

[8]. “Fast statistical spam filter by approximate classifications”, Li, K. and Zhong, Z., In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006.

[9]. "Learning and Optimization using the Clonal Selection Principle, " L.N. De Castro and F. Von Zuben, IEEE Transactions on Evolutionary Computation, vol. 6, 2002, pp. 239-251.

[10] "Theoretical Advances in Artificial Immune Systems,” j. Timmis, A. Hone, T. Stibor and E. Clark, Theoretical Computer Science, vol. 403, 2008,pp. 11-32.

[10]. ”Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks” Wu, C. Expert Syst., 2009

[11]. “Fuzzy sets,” L.A. Zadeh, Information and Control, 1965.

[12]. “Learning to filter spam e-mail: A comparison of a Naïve Bayesian and a memory-based approach,” I. Androutsopoulos, G. Paliouras, V. Karkaletsis and G. Sakkis, In Proc. of the 4th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2000.