ANFIS based Spam Filtering Model for Social Networking Websites

Dhananjay Kalbande University of Mumbai
Harsh Panchal University of Mumbai
Nisha Swaminathan University of Mumbai
Preeti Ramaraj University of Mumbai
Sardar Patel Institute of Technology, Munshi Nagar, Andheri (W), Mumbai- 400 053

ABSTRACT

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. There are various types of spam such as email spam, forum spam, online classified ads spam, attachment spam, social networking spam etc. For the purpose of this paper, we would like to concentrate more on social networking spam (SNS). SNS is when unwanted messages or posts are sent to people in bulk, or when a single click of a seemingly harmless link reposts the link on other profiles, thus spreading the spam like a virus. We plan to use an adaptive neuro fuzzy inference system (ANFIS) that incorporates the advantages of both the neural networking concepts and fuzzy logic to identify the spam messages on such websites.

General Terms

Design, algorithm, documentation, experimentation, performance

Keywords

ANFIS, adaptive neuro fuzzy systems, spam classification

1. INTRODUCTION

The Internet has almost become an indispensible part of everyday life. Internet usage will continue to grow and therefore online communication and information exchange is gaining immense popularity, impacting users' social and commercial lives. Along with the growth of e-mail and social networking websites, there has been an increased production of spam over the years. Spam is nothing but unsolicited messages. In 2010, the average global spam rate for the year was 89.1%, an increase of 1.4% compared with 2009. [1] In recent times, the types of spam have dramatically spread from simple annoying messages to the more dangerous and covert virus-attached mails.

Examples include spam sent by botnets, attached Trojans, irrelevant URLs, spam in local languages and viral marketing based spam.

Social networking spam has been the recent favorite of spammers. Utilizing the social network's search tools, the spammer can target a certain demographic segment of the users, or use common fan pages or groups to send notes to them from an account disguised as that of a real person. Such notes may include embedded links to pornographic or other product sites designed to sell their product. The attacker also posts irrelevant links on these sites which on clicking would transfer the private information of the user to the attacker. Thus spam has become a serious issue in the internet world and hence it has become important to devise ways to detect and filter them. There are various methods that have been explored to tackle the issue of spam filtering. The most popular ones are based on Naïve Bayes Theorem, Artificial Immune System, Support Vector Machines and Rule based learning algorithms. All these methods perform content based spam filtering. [2]

In this paper, we explore yet another method based on Adaptive Neuro Fuzzy Inference System (ANFIS) to detect and filter spam in social networking websites. The spam corpus contains spam that has been affecting social networking sites like Facebook, Orkut and Twitter in recent times.

2. ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM

Instead of using the concepts of only fuzzy or only neural networks, the top aspects of the two can be integrated and used to model a system. The representation of prior knowledge into a set of constraints to reduce the optimization search space is a great feature of fuzzy systems. Also, the adaptation of the neural network according to the error generated on training which is fed back to the computational layer is very useful. These two points can be incorporated to form a hybrid model called the ANFIS. Thus ANFIS provides both smoothness and adaptability using only a small training set. [4]

The ANFIS architecture for a two-rule Sugeno model is given in Figure 1. In the diagram, each layer involves a computation that takes as input the output value of the previous layer. The basic steps in computation are obtaining membership value of input sets using certain functions. These membership values give rise to input premise parameters which are then used to calculate firing strengths of different fuzzy sets based on if-then rules. After calculating firing strengths, the output consequent parameters are computed from which one can finally obtain the crisp value output for the system by performing a summation of all the signals that reach this layer. [5]

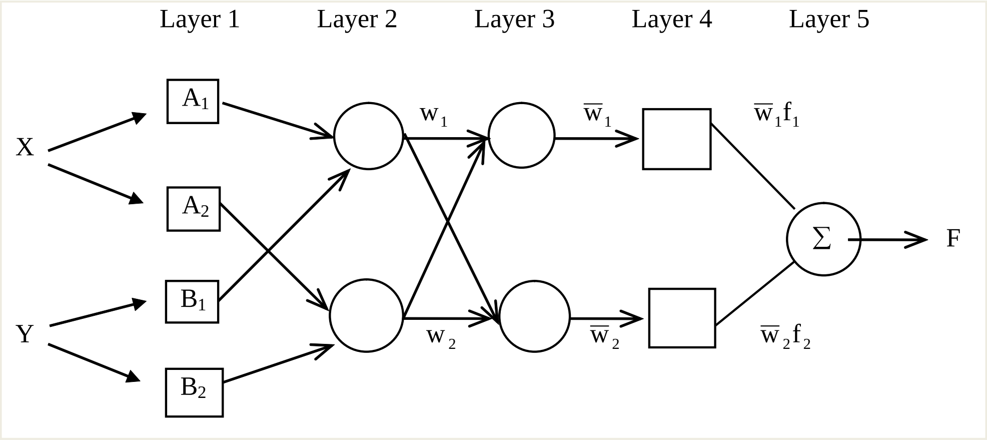

Figure 1: ANFIS architecture [3]

3. PROPOSED PROCEDURE FOR SPAM IDENTIFICATION

Here we have proposed a method to detect whether a given input is spam or not using the ANFIS classifier. The algorithm is as given in Figure 2.

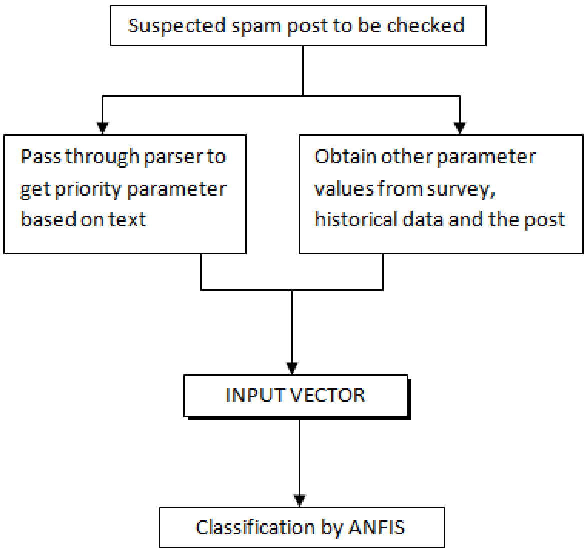
Suspected spam post to be checked
Pass through parser to get priority parameter based on text
Obtain other parameter values from survey, historical data and the post
INPUT VECTOR
Classification by ANFIS

Figure 2: Flowchart for detection of spam

The spam found on social networks is different from the traditional email spam because SNS has other parameters apart from text analysis that determine if the post is spam or not. Our algorithm contains the following modules:

3.1 Text Parser

This consists of the traditional email spam classifier that analyzes the text in the post or message and gives the probability of it being associated with a spam post.

3.2 Extraction of Other Parameters

Apart from the text, there are other parameters that are specific to social networks that are very useful in classifying spam. These parameters include the number of user pages on which a particular message or post has appeared, the number of times the users have marked the given post as spam and also the presence or absence of a hyperlink or URL. These parameters are obtained from historical data from the social networking websites and other related surveys.

3.3 Input Vector Values

For each of the input parameters, some of which have been mentioned above, a quantitative value needs to be assigned. This value is based on the amount of similarity it shows to the entries in the spam database. This input vector is then given to the ANFIS classifier which further determines whether the input is spam or not.

3.4 ANFIS Classifier

The ANFIS system is initially trained using data sets for a certain number of epochs. Then, the computed input vector is given to this system in order to determine the spam probability which will identify the input as spam or not. Once classification of the input is done and it is identified, steps can be taken to filter it from the place where it was detected.

4. INPUT VECTOR PARAMETERS

Classification of SNS requires more parameter processing than the traditional method of processing the message content. We have identified the following five parameters that decide if a given post or message can be classified as spam or not.

4.1 Number of Associated User Pages

Number of user pages on which the message has been posted is an important parameter to be considered. Since spam is known for its notorious nature of flooding user pages and user feeds, this parameter can be very useful in identifying spam.

4.2 Number of Times Marked as Spam

Our system also looks to incorporate feedback from the user. This is done by taking into account the number of times a particular post has been marked as spam. This gives the system support and confidence in classifying spam.

4.3 Text Priority

The text contained in the message plays an important role in identifying spam. As this parameter features in most of the known spam classifiers, we have extracted it using a C#

implementation of Paul Graham's Naïve Bayesian Spam Filter algorithm.[6][7] This system constitutes the text analyzer for the preprocessing of the message content. It gives the probability of the given text belonging to a spam post which forms another parameter for classification.

4.4 Presence of Hyperlink or URL

The presence or absence of a URL becomes a decisive parameter for spam classification at times. A hyperlink transfers control to an external source, thus exponentially increasing the chances of a virus or a malware entering your system.

4.5 Number of Common Timestamps

The signature characteristic of a SNS is the timestamps of the various instances of a post to be classified. When a script reposts a certain message to a number of user pages, it necessarily does this in a very short span of time. So, if a given post has a number of instances with the same timestamp, it is a clear indication that the message is script generated (usually activated by the user clicking on a particular hyperlink) and not posted by a human user. To classify, we check the timestamps of all the instances of a post and determine how many of them have common timestamps. The higher this value, the higher the chances of it being a script generated post and hence a spam. This parameter signifies the number of times the message has been spread using a script (corresponding to the number of times the user activates the script by some action like clicking on a hyperlink).

5. FUZZY INFERENCE SYSTEM STRUCTURE

We have implemented the ANFIS classifier using the MATLAB Fuzzy Logic Toolbox.[8] The ANFIS classifier needs a fuzzy inference system (FIS) that it can train using the SNS database in order to classify new instances of spam efficiently. The FIS structure consists of three parts. The input membership functions, the output membership functions and the rule set linking the two.

5.1 Input Membership Functions

There are five input parameter and each parameter has two membership functions.

5.1.1 Number of Associated User Pages

We take the range of this parameter from 0 to 400 for the purpose of demonstrating how this system works. The two fuzzy membership functions: high and low are as shown below in Figure 3.

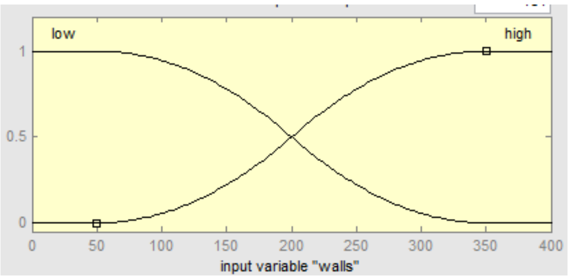

Figure 3: Membership function- number of user pages

5.1.2 Number of Times Marked as Spam by User

This parameter ranges from 0 to 200 and is divided into the membership functions high and low as shown below in Figure 4.

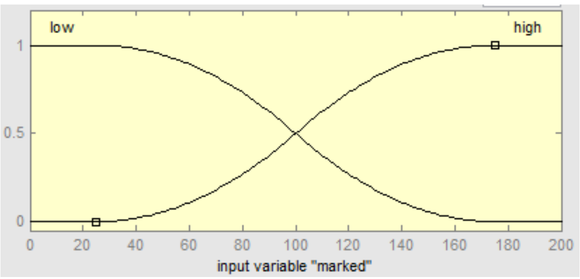

Figure 4: Membership function- times marked as spam

5.1.3 Presence of a Hyperlink or URL

If a hyperlink or URL is present in the message, this parameter is 1 else it is 0. This parameter is not exactly a fuzzy parameter but it can be represented using the fuzzy membership function as shown in Figure 5.

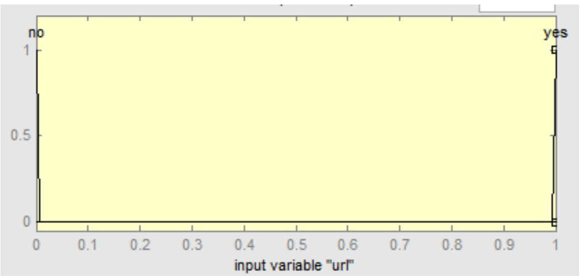

Figure 5: Membership function- URL present or not

5.1.4 Number of Instances of Common Timestamps

From the total number of instances of a given post or message, the number of common timestamps is mapped onto the membership functions showed in Figure 6. It ranges from 0 to 10 for the given system.

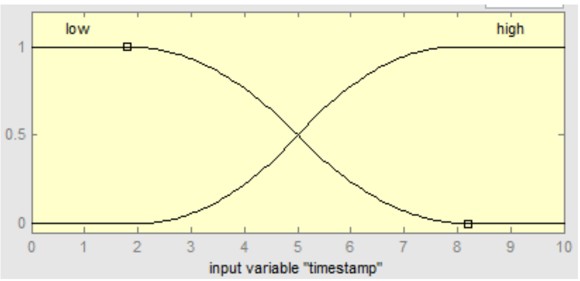

Figure 6: Membership function- common timestamps

5.1.5 Priority of Text in the Message

The Naïve Bayesian text analyzer system returns the percentage of the text belonging to the spam category. This ranges from 0 to 1 and is mapped onto fuzzy membership functions as shown below in Figure 7.

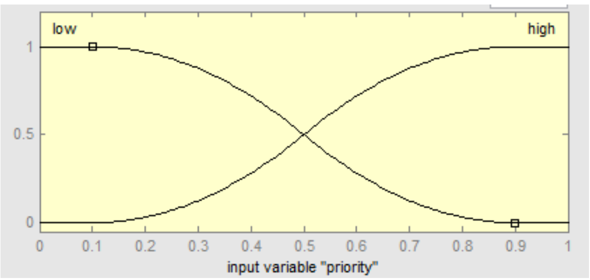

Figure 7: Membership function- Priority of text in the message

5.2 Output Membership Functions

Since the FIS structure is to be trained using the training data, we use the *Sugeno fuzzy model*. The FIS for ANFIS necessitates that the number of output membership functions equals the number of rules in the fuzzy rule set.

Each membership function is of the form:

$$\mathbf{f_i = a_ix_1+b_ix_2+c_ix_3+d_ix_4+e_ix_5+k_i}$$

where,
- f_i is the output membership function
-a_i, b_i, c_i, d_i and e_i are the coefficients of the input vector [x_1 x_2 x_3 x_4 x_5]
-k_i is the corresponding constant for each of the functions

We consider 7 rules in the rule set and hence we define the following output membership functions. The first 5 parameters of each of the membership functions are set to 0, while the value of constant k is initialized as follows:

i)	Extremely Low	*k=0*
ii)	Very Low	*k=17*
iii)	Low	*k=33*
iv)	Medium	*k=50*
v)	High:	*k=66*
vi)	Very High	*k=87*
vii)	Extremely High	*k=100*

5.3 Rule Set

The rule set used for the FIS is as shown in Table 1.

Table 1: Fuzzy Rule Set

Ins.	**Mark.**	**URL**	**T.S.**	**Pr.**	**Op**	***Spam***
Low	Low	No	-	Low	**A**	*E. low*
Low	Low	Yes	-	Low	**A**	*V. Low*
High	Low	No	Low	Low	**A**	*Low*
High	Low	Yes	Low	High	**A**	*Medium*
Low	High	Yes	Low	High	**A**	*High*
-	-	-	High	High	**O**	*V. High*
High	High	Yes	High	High	**A**	*E. High*

where,
-Ins. is the number of Instances,
-Mark. is the number of times marked as spam
-URL is the presence of URL or hyperlink
-T.S. is the number of common timestamps
-Pr. is the priority of text in the post
-Op. is the operator between the members
-Spam is the probability of post being spam
-A is AND operation
-O is OR operation

The final structure of the FIS after configuring it is as show below in Figure 8.

6. TRAINING THE FIS

In order to train the FIS, we collected different types of SNS data and a combination of spam and good posts was used to train the FIS using an adaptive-neuro method. This was implemented using the *anfisedit* functionality in the MATLAB Fuzzy Logic Toolbox that provided a GUI for training the FIS. Figure 9 shows the output of the FIS after training.

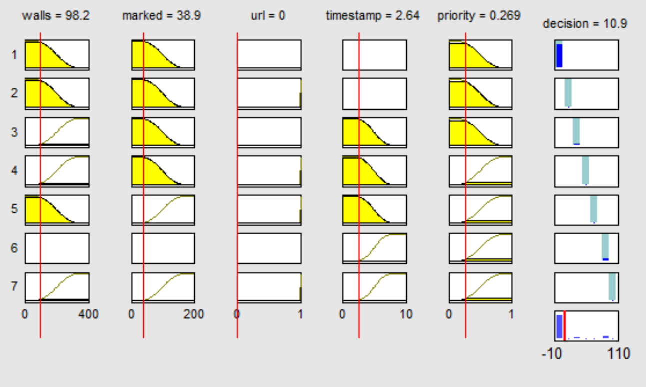

Figure 8: Fuzzy Inference System Structure in MATLAB

The average testing error is 15.0399, which is due to unavailability of actual parameter values. In spite of the error, the system successfully gives a high spam probability for spam posts and a very low probability for good posts.

The final ANFIS structure with the interconnection of the various nodes in each layer is as show in Figure 10. The connection of the nodes in the 2nd layer i.e. inputmf and the 3rd layer i.e. rules is determined by the fuzzy rule set that is decided by the user.

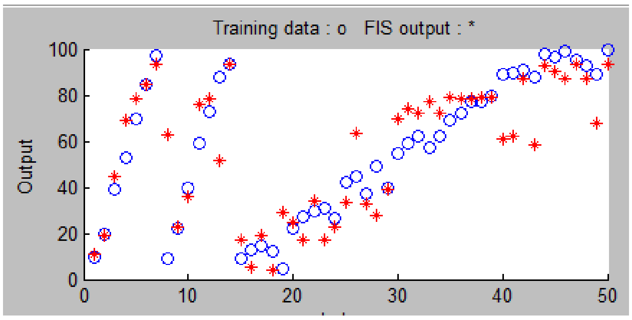

Figure 9: FIS output after training using anfisedit

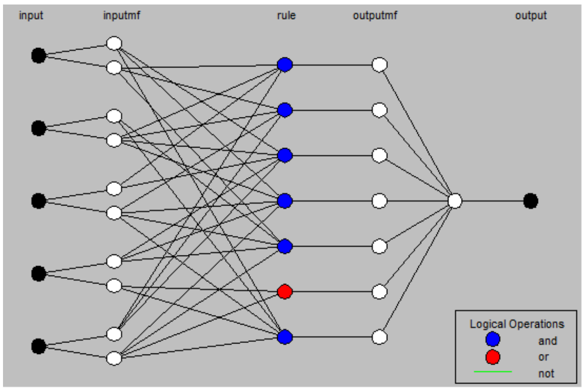

Figure 10: ANFIS structure for SNS filtering

7. LIMITATIONS

The accuracy of the method cannot be determined unless the actual values of all the parameters of the input vector are obtained. The current proposed system is not specialized for a particular social networking website. The system can be configured for a particular social networking website by including its idiosyncrasies and special cases into the FIS structure and providing proper training data set.

8. SCOPE

Social networking websites are replacing other means of communication that were previously used. Due to this reason, spammers choose to attack these websites and cause nuisance to the user. Also, these websites are easy targets for spread of viruses, Trojans and malwares. Our system aims at keeping these spammers in check. More input parameters can be identified with time and also definitions of more kinds of spam can be included in the spam database. With the increase in relevant parameters, the number of false positives would decrease and spam would be identified better.

9. CONCLUSION

Social networking spam is different from email spam since in SNS the text is not the only parameter to be used to classify the message. The parameters proposed above make it easier to classify SNS. The proposed system can classify social networking spam with efficiency when the fuzzy inference system is configured according to the website specific parameter data. The more website specific parameters are identified, the more the efficiency of the system can be improved.

10. ACKNOWLEDGEMENT

Mr. Harsh Panchal, Ms. Preeti Ramaraj and Ms. Nisha Swaminathan would like to thank Dr. Dhananjay Kalbande for his constant encouragement and help. His guidance in understanding the various aspects related to this project was vital

.

11. REFERENCES

[1] MessageLabs Intelligence, 2010, *Annual Security Report*

[2] Md. Rafiqul Islam, Morshed U. Chowdhury, 2005, *Spam Filtering using ML Algorithms*, IADIS International Conference on WWW/Internet

[3] J.S Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, *Neuro-Fuzzy and Soft Computing*, p336, Ch12, Prentice Hall

[4] José Vieira, Fernando Morgado Dias, Alexandre Mota, *Neuro-Fuzzy Systems*

[5] R. Jang, July 1992, "*Neuro-Fuzzy Modeling: Architectures, Analysis and Applications*", PhD Thesis, University of California, Berkley.

[6] Paul Graham's Naive Bayesian Spam Filter algorithm, http://www.paulgraham.com/spam.html

[7] Jason Kester, A Naïve Bayesian Spam Filter in C#, http://www.codeproject.com/Articles/23472/A-Naive-Bayesian-Spam-Filter-for-C

[8] Fuzzy Logic Toolbox, User's Guide. http://www.mathworks.com/help/pdf_doc/fuzzy/fuzzy.pdf

[9] T. M. Nazmy, H. El-Messiry, B. Al-Bokhity, *Adaptive Neuro-Fuzzy Inference System For Classification Of ECG Signals*, Journal Of Theoretical And Applied Information Technology

[10] Inan Guler, Elif Derya Ubeyli, 2005, *Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficients*, Journal of Neuroscience methods

[11] D. Puniškis, R. Laurutis, R. Dirmeikis, 2006, *"An Artificial Neural Nets for Spam e-mail Recognition*, ISSN 1392 – 1215

[12] Qin Luo, Bing Liu, Junhua Yan, Zhongyue He, 2010, *Research of a Spam Filtering Algorithm Based on Naïve Bayes and AIS* , 2010 International Conference on Computational and Information Sciences

[13] Peter Brevern, N.S.M. El-Tayeb & V.C.Vengkatesh, Sivarao, *GUI Based ANFIS Modeling: Back PropagationOptimization Method for CO2 Laser Machining*, International Journal of Intelligent Information Technology Application, 2009

[14] Ismaila Idris, *E-mail Spam Classification With Artificial Neural Network and Negative Selection Algorithm*, International Journal of Computer Science & Communication Networks, Vol. 1(3), 227-231

[15] Ashok Kusagur, Dr. S. F. Kodad, Dr. B V. Sankar Ram, September 2010, *Modeling, Design & Simulation of an Adaptive Neuro-Fuzzy Inference System (ANFIS) for Speed Control of Induction Motor*, International Journal of Computer Applications (0975 – 8887), Vol. 6– No.12