The International Journal of Engineering And Science (IJES)
||Volume|| 2 ||Issue|| 1 ||Pages|| 62-64 ||2013||
ISSN: 2319 – 1813 ISBN: 2319 – 1805

Proposed Phishing mail detection using fuzzy classification methods

[1]Ami k. Trivedi, [2]G.J.Sahani
[1](Department of Computer Engineering, Sardar Patel institute of technology, vasad, Gujarat
[2](Department of Computer Engineering, Sardar Patel institute of technology, vasad,

----------Abstract----------

Phishing is an attack that deals with social engineering methodology to illegally acquire and use someone else's data on behalf of legitimate website for own benefit If. Even a few victims fall for the pitch, the effort is profitable. Phishing mail misrepresent true sender and steal personal and financial account credentials. Most of the detection techniques use decision tree, machine learning algorithms, genetic algorithm, clustering techniques. In these techniques crisp logic is used. They classify email as spam and Not-spam email. Crisp logic is often failed because it does not provide sharp boundaries. Several Artificial Intelligence (AI) techniques including neural networks and fuzzy logic [7-9] are successfully applied to a wide variety of decision making problems in real world. Up to our knowledge, there was not developed any system to the phishing mail detection based on fuzzy classification method. In this work we would like to build a fuzzy rule generation system to detect phishing mail. It classifies email into different category like very legitimate, legitimate, suspicious, phishy, very phishy etc. The main advantage of fuzzy rule-based classification systems is that they do not require large memory storage, their inference speed is very high and the users can carefully examine each fuzzy if-then rule.

Date of Submission: 17, December, 2012 ⟺ Date of Publication: 05, January 2013

I. Introduction

Mail spammers can be categorized based on their intent. Some spammers are telemarketers, who broadcast unsolicited emails to several hundred/thousands of email users. They do not have a specific target, but blindly send the broadcast and expect a very limited rate of return. The next category of spammers comprise the opt-in spammers, who keep sending unsolicited messages though you have little or no interest in them. In some cases, they spam you with unrelated topics or marketing material. Some of the examples are conference notices, professional news or meeting announcements. The third category of spammers is called phishers. Phishing attackers misrepresent the true sender and steal the consumers' personal identity data and financial account credentials. These spammers send spoofed emails and lead consumers to counterfeit websites designed to trick recipients into divulging financial data such as credit card numbers, account usernames, passwords and social security numbers. By hijacking brand names of banks, e-retailers and credit card companies, phishers often convince the recipients to respond.

II. Literature review and related work

Research on header analysis was recently reported by Microsoft, IBM, and Cornell University in the 2004 and 2005 anti-spam conferences [4] [5]. Barry Lieba's [4] filtering technique makes use of the path travelled by an email, which is extracted from the email header. They have analysed the end units in the path but not the end-to-end path. They claim their approach complements the existing filters but does not work as a standalone mechanism. Christine E. Drake [5] in his paper, "Anatomy of Phishing Email", discusses the tricks employed by the email scammers in phishing emails. In paper "Anti-phishing techniques: A Review" [1] authors proposed content filtering methodology in which content of email is filtered out using. Machine learning methods like Bayesian Additive Regression Trees (BART), support vector machine (SVM).They also used genetic algorithm to detect phishing email but problem with this algorithm is that it is only work for one rule, if there are more than one rule algorithm becomes more complex. Character based anti-phishing approach check for the URL which mostly used by the attackers, but this approach result in false positive. In paper "Detecting Phishing in email" [2] author suggest classification on three kinds of analyses on the header: DNS-based header analysis, Social

Network analysis and Wantedness analysis. In the DNS-based header analysis, they classified the corpus into 8 buckets and used social network analysis to further reduce the false positives. They introduced a concept of wantedness and credibility, and derived equations to calculate the wantedness values of the email senders. Finally, credibility and all the three analyses to classify the phishing emails. Method resulted in far less false positives. They plan to perform two more analyses on the incoming email traffic, i) End to End path Analysis, by which they try to establish the legitimacy of the path taken by an email; ii) Relay Analysis, by which they verify the trustworthiness and reputation of the relays participating in the relaying of emails. They combine all the methods i) Path analysis, ii) Relay analysis, iii) DNS-based header analysis and iv) Social Network analysis for developing a email classifier, which classifies the incoming email traffic as i) Phishing emails ii) Socially Wanted emails (Legitimate emails) iii) Socially Unwanted emails (Spam emails).They are only concentrated on email header not in the content of the email. Paper "learning to detect phishing in email"[3] proposed machine learning approach with 10-fold cross validation training data which use random forest as classifier which only categorized spam or no-spam email. They are not going to classify mail as legitimate, spam or phishing.

III. Fuzzy Systems

Fuzzy logic was invented by Zadeh [6] in 1965 for handling uncertain and imprecise knowledge in real world applications. It has proved to be a powerful tool for decision-making, and to handle and manipulate imprecise and noisy data. The first major commercial application was in the area of cement kiln control. This requires that an operator monitor four internal states of the kiln, control four sets of operations, and dynamically manage 40 or 50 "rules of thumb" about their interrelationships, all with the goal of controlling a highly complex set of chemical interactions. One such rule is "If the oxygen percentage is rather high and the free-lime and kiln-drive torque rate is normal, decrease the flow of gas and slightly reduce the fuel rate". The notion central to fuzzy systems is that truth values (in fuzzy logic) or membership values (in fuzzy sets) are indicated by a value on the range [0.0, 1.0], with 0.0 representing absolute Falseness and 1.0 representing absolute Truth. A fuzzy system is characterized by a set of linguistic statements based on expert knowledge. The expert knowledge is usually in the form of "if-then" rules.

A fuzzy set A in X is characterized by a membership function which is easily implemented by fuzzy conditional statements. For example, if the antecedent is true to some degree of membership, then the consequent is also true to that same degree. If antecedent Then consequent.

In a fuzzy classification system, a case or an object can be classified by applying a set of fuzzy rules based on the linguistic values of its attributes. Every rule has a weight, which is a number between 0 and 1, and this is applied to the number given by the antecedent. It involves 2 distinct parts. The first part involves evaluating the antecedent, fuzzifying the input and applying any necessary fuzzy operators. The second part requires application of that result to the consequent, known as inference. To build a fuzzy classification system, the most difficult task is to find a set of fuzzy rules pertaining to the specific classification problem. A fuzzy inference system is a rule-based system that uses fuzzy logic, rather than Boolean logic, to reason about data. Its basic structure includes four main components (1) a fuzzifier, which translates crisp (real-valued) inputs into fuzzy values; (2) an inference engine that applies a fuzzy reasoning mechanism to obtain a fuzzy output; (3) a defuzzifier, which translates this latter output into a crisp value; and (4) a knowledge base, which contains both an ensemble of fuzzy rules, known as the rule base, and an ensemble of membership functions known as the database. The decision-making process is performed by the inference engine using the rules contained in the rule base. These fuzzy rules define the connection between input and output fuzzy variables.

IV. Proposed fuzzy classification based approach

To detect phishing mail, we are going to proposed four direct rule generation methods based on fuzzy classification. The first method generates fuzzy if-then rules using the mean and the standard deviation of attribute values. The second approach generates fuzzy if-then rules using the histogram of attributes values. The third procedure generates fuzzy if-then rules with certainty of each attribute into homogeneous fuzzy sets. In the fourth approach, only overlapping areas are partitioned. The first two approaches generate a single fuzzy if-then rule for each class by specifying the membership function of each antecedent fuzzy set using the information about attribute values of training patterns. The other two approaches are based on fuzzy grids with homogeneous fuzzy partitions of each attribute. To start with these methods linguistic descriptors such as high, low, medium are assigned to a range of values for each key phishing characteristic indicator. Valid ranges of the inputs are considered and divided into classes, or fuzzy sets. For example, length of URL

address can range from 'low' to 'high' with other values in between. We cannot specify clear boundaries between classes. The degree of belongingness of the values of the variables to any selected class is called the degree of membership; membership function is designed for each phishing characteristic indicator, which is a curve that defines how each point in the input space is mapped to a membership value between [0, 1]. Linguistic values are assigned for each phishing indicator as low, moderate, and high while for spam mail as Very legitimate, Legitimate, Suspicious, Phishy, and Very phishy (triangular and trapezoidal membership function). For each input, their values range from 0 to 10 while for output, range from 0 to 100.Flow of our proposed system is as follows:

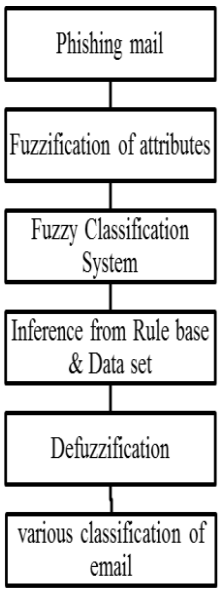

Figure 1 : Flow of Proposed Fuzzy Classification System

V. Conclusion and Future Work

A motivation behind using fuzzy rule is soft decision boundaries provide sharp transition between classes and could become a crisp one if it is needed. Machine learning techniques can be applied to spam mail detection. Many systems are built on fuzzy classification method like predicting result of basketball game using fuzzy classification method [7], fuzzy classification methods for breast cancer data [9],online customer[8].By analysing these papers we conclude that fuzzy classification provide good result in decision making problems. In future we are going to collect sample data for email and then applying different fuzzy classification methods on it and classified email as spam or Anti-spam.

References

[1] Gaurav, Madhuresh Mishra, Anurag Jain, "Anti-Phishing Techniques: A Review", *International Journal of Engineering Research and Applications, vol.2, Issue: 2, Mar-Apr 2012, pp350-355*

[2] Srikanth Palla and Ram Dantu, "Detecting Phishing in Emails"

[3] Ian fatte, Norman sadeh, Anthony Tomasic, "Learning to detect Phishing in email", *Technical Report CMU-ISRI-06-112, June-2006.*

[4] Barry Leiba, Joel Ossher, V.T. Rajan, Richard Segal, Mark Wegman, "SMTP Path Analysis" *First Conference on Email and Anti-Spam (CEAS) 2004 Proceedings.*

[5] Christine E. Drake, Jonathan J. Oliver, and Eugene J. Koontz,"Anatomy of a Phishing Email", *Proceedings of First Conference on Email and Anti-Spam (CEAS), July 2004.*

[6] Zadeh, L.A., Fuzzy Logic *IEEE Computer, pp. 83-93 (1988).*

[7] Krzysztof Trawi´nski "A Fuzzy Classification System for Prediction of the Results of the Basketball Games", *IEEE computer, 2010.*

[8] Andreas Meier and Nicolas Werro,"A Fuzzy Classification Model for Online Customers", *Informatica 31, pp. 175-182.2007.*

[9] Ravi. Jain, Ajith. Abraham,"A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data" *7th International Work Conference on Artificial and Natural Neural Networks, IWANN'03, Spain, 2003.*