Application of Rule based Fuzzy Inference System in Prediction of Internet Phishing

Shefali Paliwal
Research scholar
Hindustan Institute of Technology and Management,Agra
APJ Abdul Kalam Technical University, Lucknow

Darpan Anand
Asst. Professor
Hindustan Institute of Technology and Management,Agra
APJ Abdul Kalam Technical University, Lucknow

Sartaj Khan
Asst. Professor
Hindustan Institute of Technology and Manangement,Agra
APJ Abdul Kalam Technical University, Lucknow

ABSTRACT

Over the years many cases of internet frauds have increased and phishing is one of the techniques used by hackers to execute the frauds through internet. Many tools and techniques have been designed to detect phishing attacks and to prevent them. Phishers may have a ton of methodologies and strategies to lead an all-around composed phishing assault and thus cause access the legitimate information. The objectives of the phishing assaults, are principally on-line managing an account customers, banking customers and payment services, etc. the companies indulged in these services are confronting significant money related misfortune and absence of trust in Internet-based administrations.

Keeping in mind the end goal to beat these, there is a critical need to discover answers for battle phishing assaults. Recognizing a phishing site is a very difficult task and thus requires master learning and experience. Thus, there must be some easy ways to deal with phishing attacks. Different arrangements, design and tools have been proposed and created to address the issues of phishing attacks. The majority of these methodologies are not ready to settle on a choice progressively on whether the site is truth be told phished, and thus raising the counts of false positives. This is principally because of the limitations of the beforehand proposed approaches, which includes depending just on black and white list database, missing of human insight and specialists, poor versatility and their opportuneness.

In this work developed an intelligent phishing system by using fuzzy-based fuzzy inference system. It use UCI machine data to test inference system and found satisfactory results further it compares phishing detection system with fuzzy logic with other algorithms like J48, naïve Bayes classifier and Neuro-fuzzy based phishing detection system. Thus, the objective of this work is proposing an efficient non algorithmic anti- phishing system.

Keywords

Fuzzy Logic, Neural network, inference engine

1. INTRODUCTION

Phishing is a moderately new cyber based crime in correlation with different other cyber-crimes like hacking. Phishing attempts to gain access to sensitive information like credit card numbers, ATM card numbers, usernames, passwords etc. for carrying out frauds using these data. More number of phishing cases have risen over the years and have caused huge financial and data loss across the globe. Usually the phishing websites or attacks banking customers, payment services. Many efforts have been laid down to protect the customers from various kinds of phishing attacks. Phishing attacks make use of advertisements based micro websites, Facebook pages, twitter, mails, etc. to attract customers to click on the websites and give the customer information. Once the information is obtained these sites make use of this information to execute an attack. Many effective measures have been carried out in the past to ensure that such kinds of attacks are prevented or detected. Even social engineering based educational programs are launched to educate the users against phishing websites. This work examines the existing technologies of phishing detection and prevention system and based. It then comes up with a Mamdami based Fuzzy phishing detection system which has better efficiency of detection of phishing website than the existing system. This work also compares the fuzzy based phishing detection system using other data classification algorithms and Nero fuzzy based methods.

2. LITERATURE REVIEW

In this section, briefly survey existing anti-phishing solutions and a list of the related works.

A dynamic security skin for browser based approach was proposed Dhamija and Tygar’s (2005) [7]. In this technique a user proves his/her identity by sharing a secret image to the remote server which helps in user verification by a human. Phishers thus have the difficulty to spoof the image and thus carry out the attack. This method has the advantage that it is easy for the user to carry on the verification. But the disadvantage of this approach is that the approach needs an effort form the user and user must thus be aware of the phishing attack. The proposed approach requires changes to the entire web infrastructure, so it can succeed only if the entire industry supports it. Also this technique does not provide security for situations where the user login is from a public terminal.

In 2006 from anti-phishing work group database Dhamij et al analyzed about two hundred different phishing attacks and came up with the factors that help in phishing attacks. These were the lack of computer knowledge, deception tricks, visual deception tricks etc. using twenty two participants the researcher surveyed to identify websites from their knowledge whether the website is phishing website or not. On the basis of the research the group concluded that the users usually overlooked warnings from pop-up, invalid signatures, SSL, padlocks etc. about twenty three percent of the participants ignored the security indicator about the phishing attacks and about forty percent of the time they were prone to such kinds of attacks. Thus the researchers concluded that it is important to re-design the security system by considering the user knowledge and their issues.

In 2006 Chandrasekaran et al. (2006) [6] came up with a novel approach to classify phishing contents based on the phishing emails characteristic properties. About twenty five different characters were identified which included style markers, subject line of the email, structure of the body etc. where used based on these features 100 phishing emails and 100 non-phishing emails were analyzed. Now using simulated annealing algorithm for feature selection a feature set was formed and each feature set was given a rank based on their relevance. Now support vector machines were used to classify the emails based on the selected features. This helped to classify accurately about ninety five percent of phishing email and yielded low false positive rates.

In 2007 Abu-Nimeh et al [1] researched on classifying phishing emails by comparing six different machine learning techniques. The phishing dataset used by them comprised of 2889 emails and they selected 43 different unique variables for classification. Bag of words was used as the feature set.

The result displayed that almost all the classifier could classify the mails accurately to 92% and the bag of words methods using spam detection mechanism was the one that achieved high accuracy.

Most of these approaches concentrated on emails and their structural contents. Most of them relied on textual data. These days phishing attacks have become more complex and simply don't rely on textual data. Thus these techniques have become little irrelevant to use.

Phishing detection based tools have also been developed over the years. Some of them are the anti-phishing tools of Google Safe Browsing, McAfee Site Advisor, Microsoft IE 8 anti-phishing protection tools, etc. these tools usually rely on blacklisting of phishing URLs to help the identification of phishing sites. But the disadvantages these approach offers is that non blacklisted phishing sites get un- recognized and easily commit the attack. Thus the success of these techniques relies on the frequent update of the black list data collected from the internet monitoring.

3. METHODOLOGY

3.1 Fuzzy based Phishing website detection system methodology

In this methodology fuzzy based system is made to help in detecting the features of the phishing website. Fuzzy logic based phishing detection system makes use of certain website characteristics and factors which have been identified and extracted for the system to helps in identifying the website as phishing one or not. These factors have been identified through case studies, and survey of existing works.

The proposed system model is given in figure 3.1

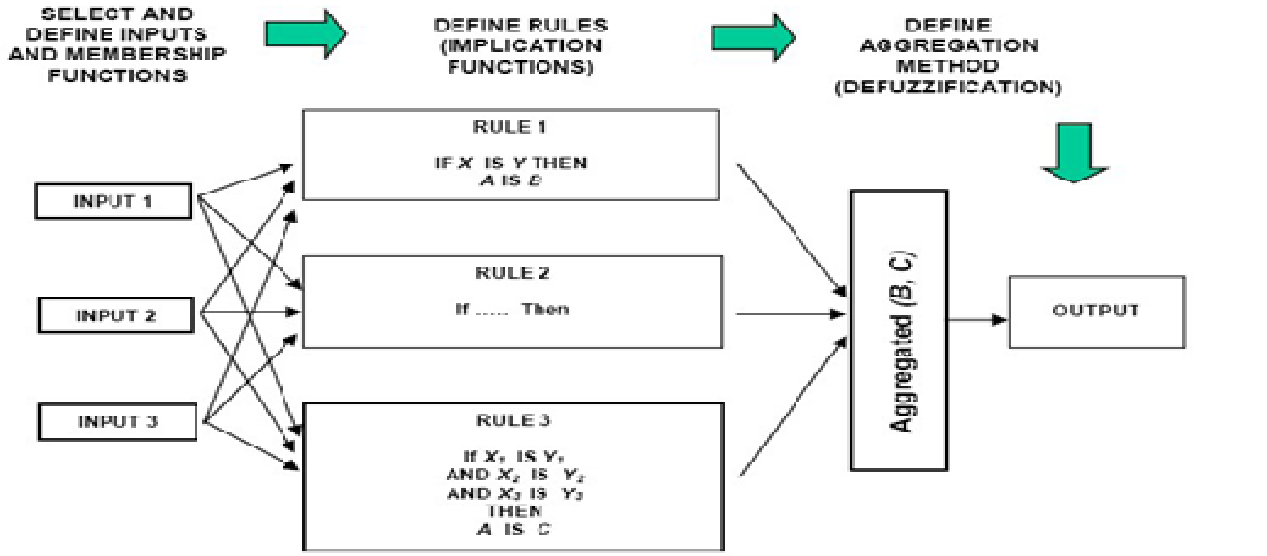

SELECT AND DEFINE INPUTS AND MEMBERSHIP FUNCTIONS
DEFINE RULES (IMPLICATION FUNCTIONS)
DEFINE AGGREGATION METHOD (DEFUZZIFICATION)
INPUT 1
INPUT 2
INPUT 3
RULE 1
IF X IS Y THEN A IS B
RULE 2
If ..... Then
RULE 3
If X_1 IS Y_1 AND X_2 IS Y_2 AND X_3 IS Y_3 THEN A IS C
Aggregated (B, C)
OUTPUT

Figure 3.1: Proposed methodology structure

Algorithm of fuzzy logic based phishing website detection system

Step 1: identify the factors for detecting the phishing website

Step 2: Fuzzification: generate membership values for fuzzy variables using membership functions.

Step 3: Make the fuzzy rule base for the given set of input variable using conjunction operators.

Step 4: perform the unification of the output of all the rules by combining the membership function of the given set of rules into a scale single output.

Step5: Defuzzification: transform the fuzzy inference system into the crisp output so that the rules are evaluated using centroid technique to give three sets of output values as phishing site, doubtful site and legitimate website.

Step 6: display the result.

Using the above algorithm the steps of the algorithms are expanded below in detail.

Fuzzification process

Each input variable is described linguistically as High, low and medium ranged values. Thus, each key identified factor for phishing detection is given these values. The values areassigned using the research over time and existing literature reviews.

The main inputs selected for the system are as follows:

- URL & DOMAIN identity: usually a website has an URL for identifying it and retrieving it. On the basis of URL analysis it is found that websites that have short URL are normally trustworthy and thus legitimate than the websites that have very long URL.
- Security & Encryption: Some phishers can delight and lure many visitors into using their forged websites by emphasizing on the security issue to gain their trust and by always asking for their prompt action to protect their personal information from being hacked.
- Page style Content & web address bar: some websites have contents like @, hexadecimal characters and address bar which makes it look phishy.

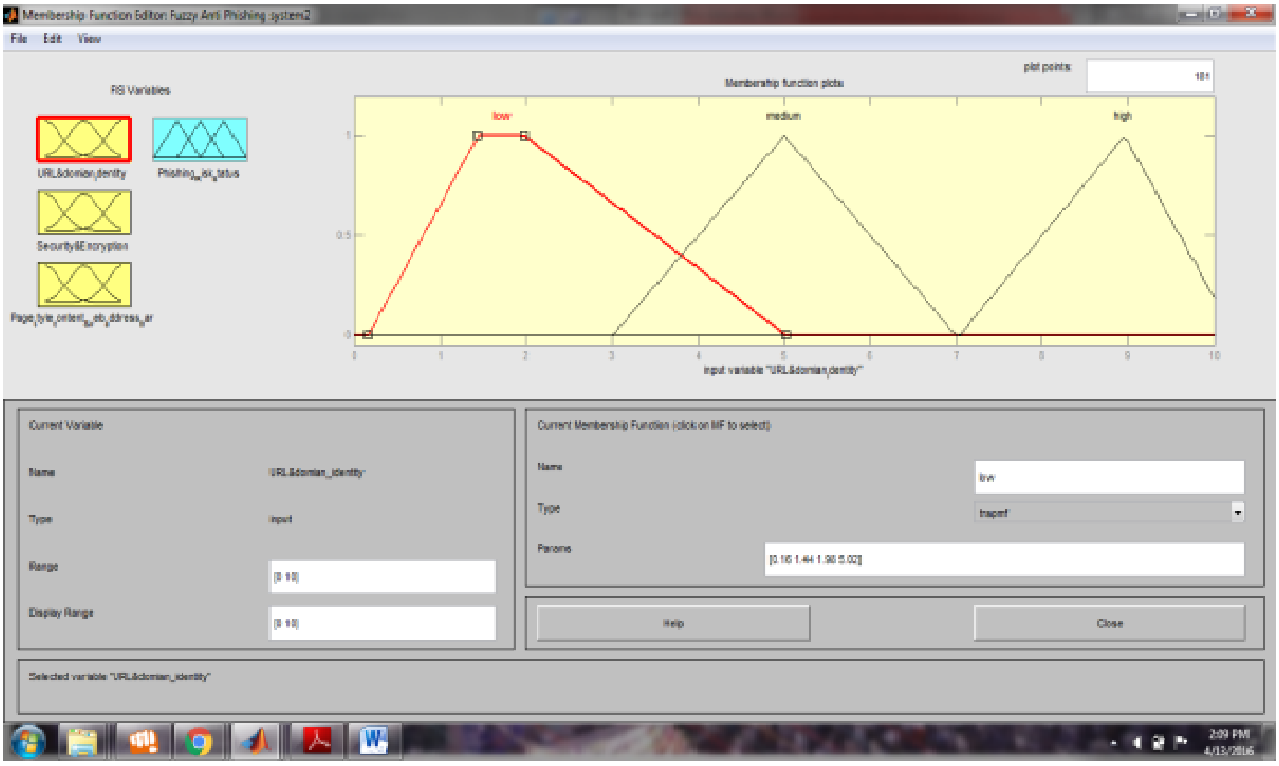
Each variable is assigned a range. The inputs are mapped in the range of 0 to 10 and outputs in the range of 0 to 100. Example of the first variable mapping is shown below

URL & Domain identity – Low, Moderate, High.

Low [0.16 1.44 1.98 5.02]

Medium [3, 5, 7]

High [7.04 8.96 10.24]



Current Variable

Name "URL&domain_identity"

Type input

Range [0 10]

Display Range [0 10]

Current Membership Function (click on MF to select)

Name low

Type trapmf

Params [0.16 1.44 1.98 5.02]

Selected variable "URL&domain_identity"

Figure 3.2: Input variables for URL and Domain identity

Similarly it is done for all other variables. This is implemented in MATLAB R 2013b.

Creating Fuzzy rule base

Each membership value with linguistic representation is used

to create the fuzzy rule base using fuzzy operators like AND and OR. Thus, in this research work I have 3 categories of inputs and one output named as phishing risk. Thus for 3 sets of inputs my total rules come down to 27.

Table 1.Rules

Rules	URL & DOMAIN identity	Security & Encryption	Page style Content & web address bar	Phishing risk
1	Low	Low	Low	Legitimate
2	Low	Low	Medium	Legitimate
3	Low	Low	High	Doubtful
4	Low	Medium	Low	Doubtful
5	Low	Medium	Medium	Doubtful
6	Low	Medium	High	Phishy
7	Low	High	Low	Doubtful
8	Low	High	Medium	Phishy
9	Low	High	High	Phishy
10	Medium	Low	Low	Legitimate
11	Medium	Low	Medium	Doubtful
12	Medium	Low	High	Phishy
13	Medium	Medium	Low	Doubtful
14	Medium	Medium	Medium	Doubtful
15	Medium	Medium	High	Phishy

16	Medium	High	Low	Phishy
17	Medium	High	Medium	Phishy
18	Medium	High	High	Phishy
19	High	Low	Low	Doubtful
20	High	Low	Medium	Doubtful
21	High	Low	High	Phishy
22	High	Medium	Low	Doubtful
23	High	Medium	Medium	Doubtful
24	High	Medium	High	Phishy
25	High	High	Low	Phishy
26	High	High	Medium	Phishy
27	High	High	High	Phishy

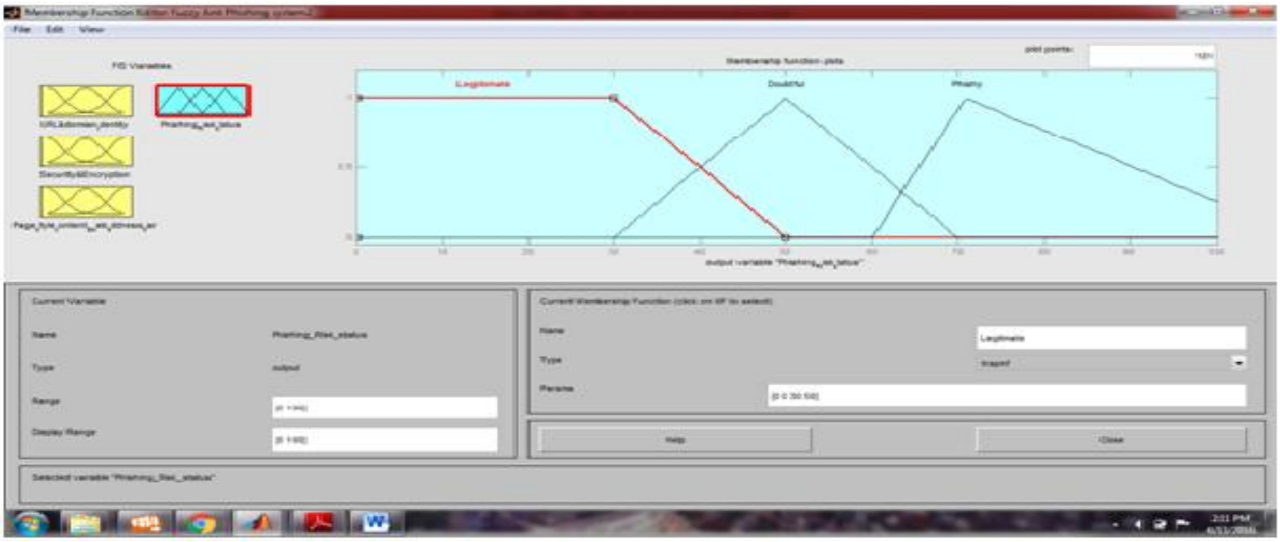
Defuzzification

The output function is named as the phishing risk and three linguistic variables are used to define the functions which are ‘phishy’, doubtful and ‘legitimate’. The fuzzy output set is then defuzzified is shown in Figure 3.3.

Legitimate [0, 0, 30, 50]

Doubtful [30, 50, 70]

Phishy [50, 70, 100]



Name Phishing_Risk_status

Type output

Range [0 100]

Display Range [0 100]

Name Legitimate

Type trapmf

Params [0 0 30 50]

Selected variable "Phishing_Risk_status"

Figure 3.3: Output function

Finally the Phishing Website Risk is calculated as = 0.3 * URL & Domain Identity crisp

+ (0.2 * Security & Encryption crisp) + (0.1 * Page Style & Contents Web Address Bar crisp).

3.2 Other classifier based phishing website detection system

The other classifier chosen are Naïve Bayes classifier and J48 classifier. These classifier are validated on Weka platform.

The basic step before using these classifier is to carry out data preprocessing where in the .arff file of the dataset is converted into the .CSV file. Each column is assigned a name and the data is arranged for processing ahead.

Next the data is transformed into nominal form using normalization so that it can be accepted. Finally the classifiers are called upon to be used for validation over the given dataset.

3.3 Neuro Fuzzy based model (EFuNN)

Neuro fuzzy based model EFuNN is selected for the given validation of the dataset of the phishing website. EFuNN is a mamdami based neuro-fuzzy system which has five neuron layers based feed forward network. Each layer of the system performs a specific task. The first layer of the system is the input layer and takes the input variables. Followed by it is the second layer which is the condition layer where each neuron is represented by triangular fuzzy membership function and this it performs fuzzification. The third layer is the evolving layer which contains the rule base created automatically. The fourth layer is the action layer which represents fuzzy membership functions of the output neuron and finally the final layer is the output layer that gives the output using center of gravity defuzzification.

The basic algorithm for EFuNN treats each layer of the neuron as a fuzzy rule and selects the largest weights among them.

4. RESULTS

4.1 Fuzzy logic based anti- Phishing model Subsections

This is implemented in MATLAB 2013 b.

For all three inputs to be 0 i.e. low low low

We get phishing risk as 24%

That means the website are legitimate

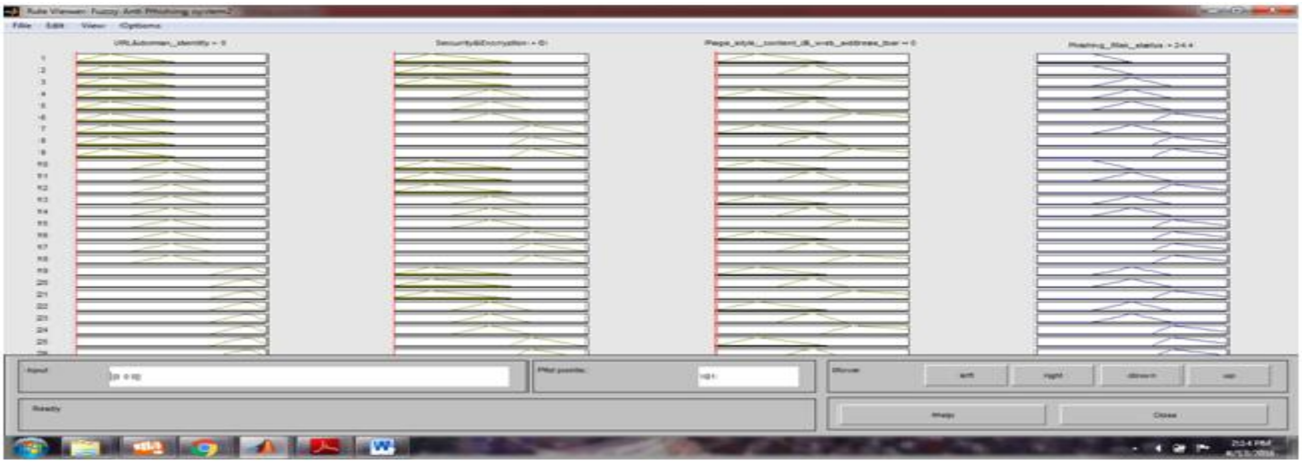


Figure 4.1: Rules bases for [0 0 0]

For inputs [5 5 5] i.e. medium

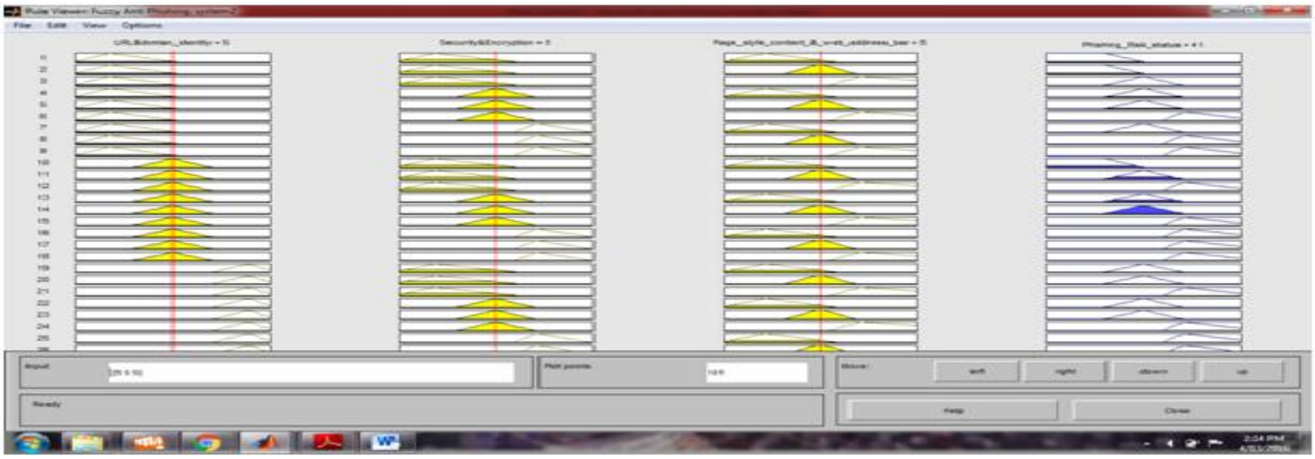


Figure 4.2: Rules bases for [555]

Phishing risk status is 41% that is it is doubtful

For all values to be high

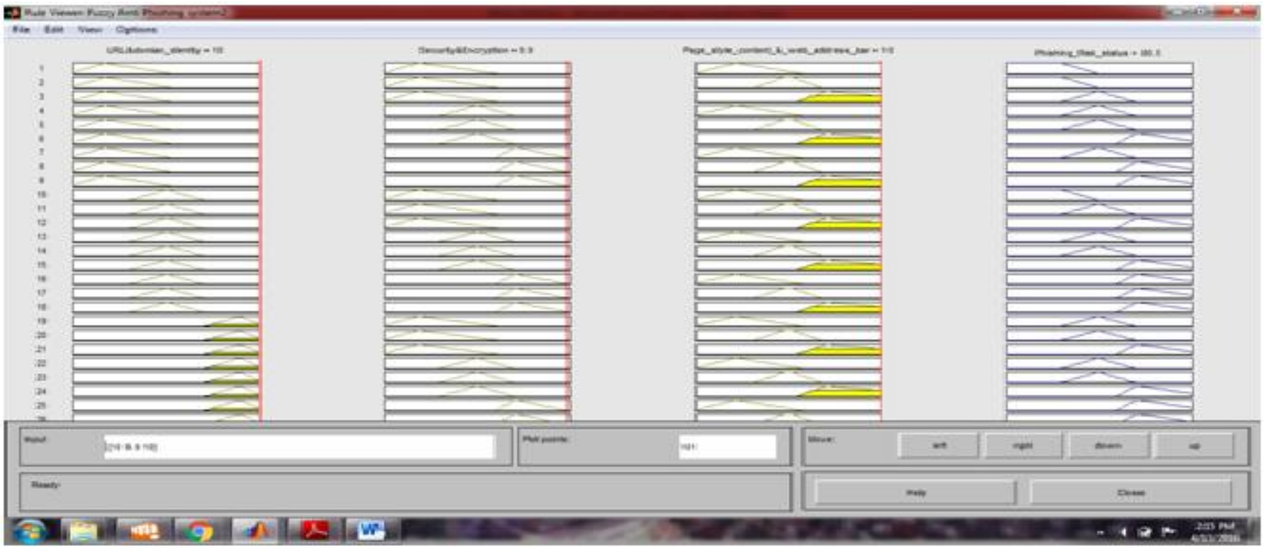
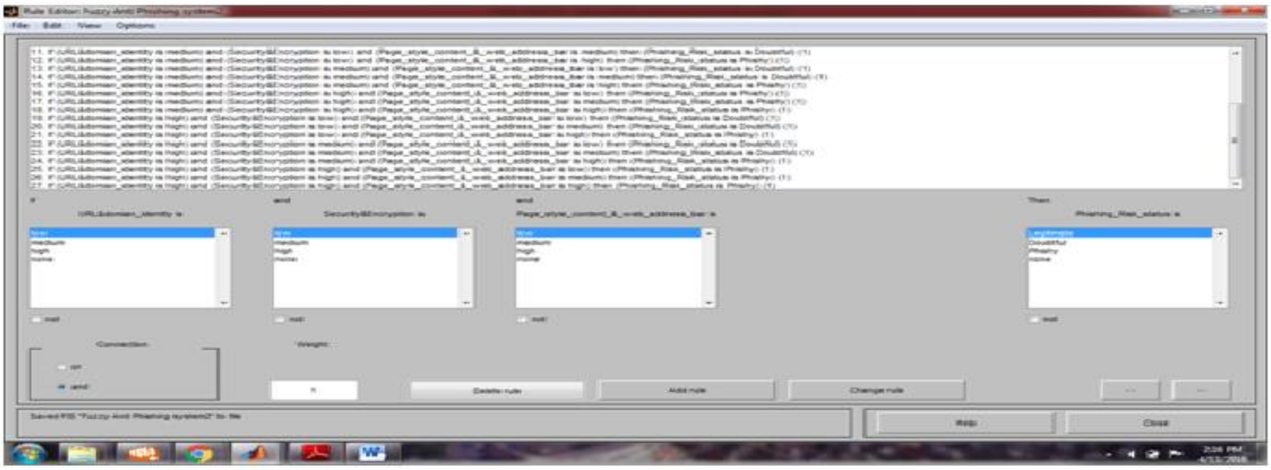


Figure 4.3: Output for [10 10 10]

The phishing risk is 80.5 % i.e. the website are phishing website.

The rules used are:



Rule Editor: Fuzzy Anti Phishing system2

11. If (URL&domain_identity is medium) and (Security&Encryption is low) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Doubtful) (1)
12. If (URL&domain_identity is medium) and (Security&Encryption is low) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)
13. If (URL&domain_identity is medium) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is low) then (Phishing_Risk_status is Doubtful) (1)
14. If (URL&domain_identity is medium) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Doubtful) (1)
15. If (URL&domain_identity is medium) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)
16. If (URL&domain_identity is medium) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is low) then (Phishing_Risk_status is Phishy) (1)
17. If (URL&domain_identity is medium) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Phishy) (1)
18. If (URL&domain_identity is medium) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)
19. If (URL&domain_identity is high) and (Security&Encryption is low) and (Page_style_content_&_web_address_bar is low) then (Phishing_Risk_status is Doubtful) (1)
20. If (URL&domain_identity is high) and (Security&Encryption is low) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Doubtful) (1)
21. If (URL&domain_identity is high) and (Security&Encryption is low) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)
22. If (URL&domain_identity is high) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is low) then (Phishing_Risk_status is Doubtful) (1)
23. If (URL&domain_identity is high) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Doubtful) (1)
24. If (URL&domain_identity is high) and (Security&Encryption is medium) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)
25. If (URL&domain_identity is high) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is low) then (Phishing_Risk_status is Phishy) (1)
26. If (URL&domain_identity is high) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is medium) then (Phishing_Risk_status is Phishy) (1)
27. If (URL&domain_identity is high) and (Security&Encryption is high) and (Page_style_content_&_web_address_bar is high) then (Phishing_Risk_status is Phishy) (1)

Saved FIS "Fuzzy Anti Phishing system2" to file

Figure 4.4: Rules

Now the model is executed over a set of 60 websites collected from phish tank archive.

The dataset is as follows:

Phishing website: 30

Legitimate website: 15

Doubtful: 15

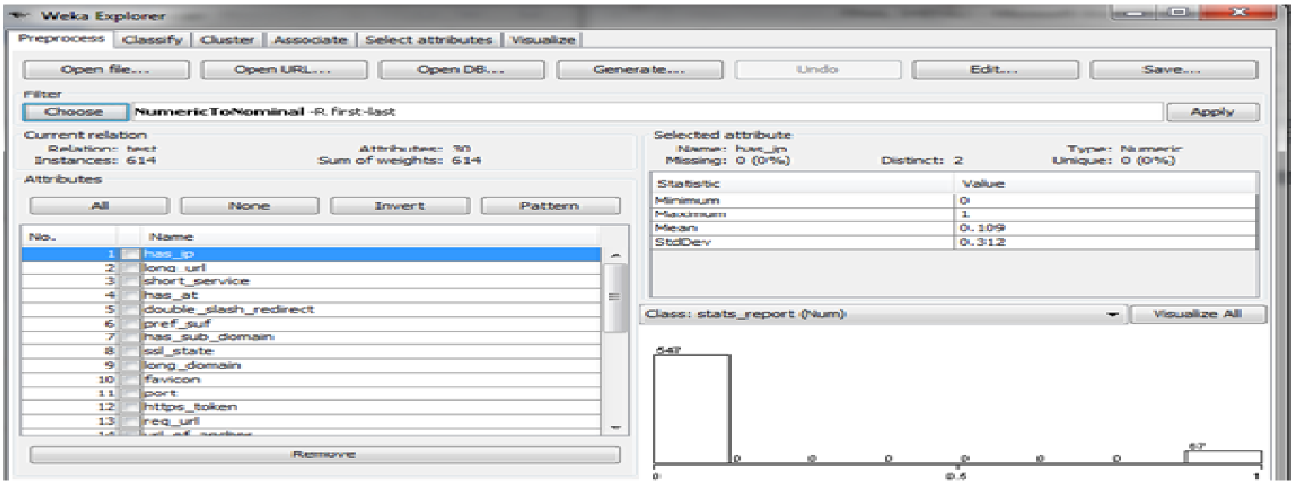
The output obtained for the system is as follows

DecisionWebsite	Legitimate	Suspicious	Phishy
Legitimate	12	2	1
Suspicious	3	11	1
Phishing	4	6	20

Thus one can see that the category of doubtful and phishy website has increased. Thus Fuzzy system seems to be more effective than the other anti- phishing system.

4.2 Naïve classifier based anti- Phishing model

This is implemented in WEKA. The first step is to load the dataset file. The dataset file is in .csv format. It is loaded successfully in the WEKA environment. Since the data is in numeric form it is needed to transform the data for classification. The transformation applied is numeric to nominal form.



Weka Explorer

Filter: NumericToNominal -R first-last

Current relation
Relation: test
Instances: 614
Attributes: 30
Sum of weights: 614

Selected attribute
Name: has_ip
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)

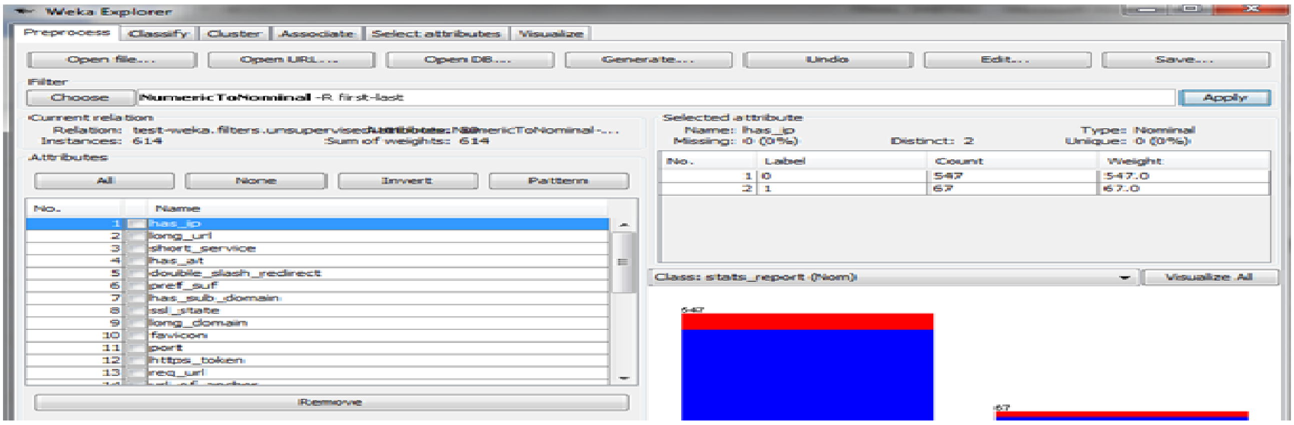
Statistic	Value
Minimum	0
Maximum	1
Mean	0.109
StdDev	0.312

No.	Name
1	has_ip
2	long_url
3	short_service
4	has_at
5	double_slash_redirect
6	pref_suf
7	has_sub_domain
8	ssl_state
9	long_domain
10	favicon
11	port
12	https_token
13	req_url

Class: stats_report (Num)

Figure 4.5: Data set

The data is now converted to nominal form and thus can be interpreted as follows



Filter: **NumericToNominal** -R first-last

Current relation
Relation: test-weka.filters.unsupervised.attribute.NumericToNominal-...
Instances: 614 Sum of weights: 614

Attributes

No.	Name
1	has_ip
2	long_url
3	short_service
4	has_at
5	double_slash_redirect
6	pref_suf
7	has_sub_domain
8	ssl_state
9	long_domain
10	favicon
11	port
12	https_token
13	req_url

Selected attribute
Name: has_ip Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	547	547.0
2	1	67	67.0

Class: stats_report (Nom)

Figure 4.6: data preprocessing

The output obtained after Naïve bayes classifier is applied over the data for 10 cross validation is as follows

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 551 89.7394 %

Incorrectly Classified Instances 63 10.2606 %

Kappa statistic 0.6758

Mean absolute error 0.1028

Root mean squared error 0.2851

Relative absolute error 35.6379 %

Root relative squared error 75.1554 %

Coverage of cases (0.95 level) 96.5798 %

Mean rel. region size (0.95 level) 56.3518 %

Total Number of Instances 614

Thus Naïve bayes claasified 462 out of phishy data as Phishing or Fraud websites and 18 of non –phishy website as

fraud website. Thus the overall accuracy of the classifier is as follows:89.7394 %

4.3 J48 classifier based anti- Phishing model

Similarly the output for J48 classifier is as follows:

Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 585 95.2769 %

Incorrectly Classified Instances 29 4.7231 %

Kappa statistic 0.8328

Mean absolute error 0.0616

Root mean squared error 0.1928

Relative absolute error 21.3332 %

Root relative squared error 50.8277 %

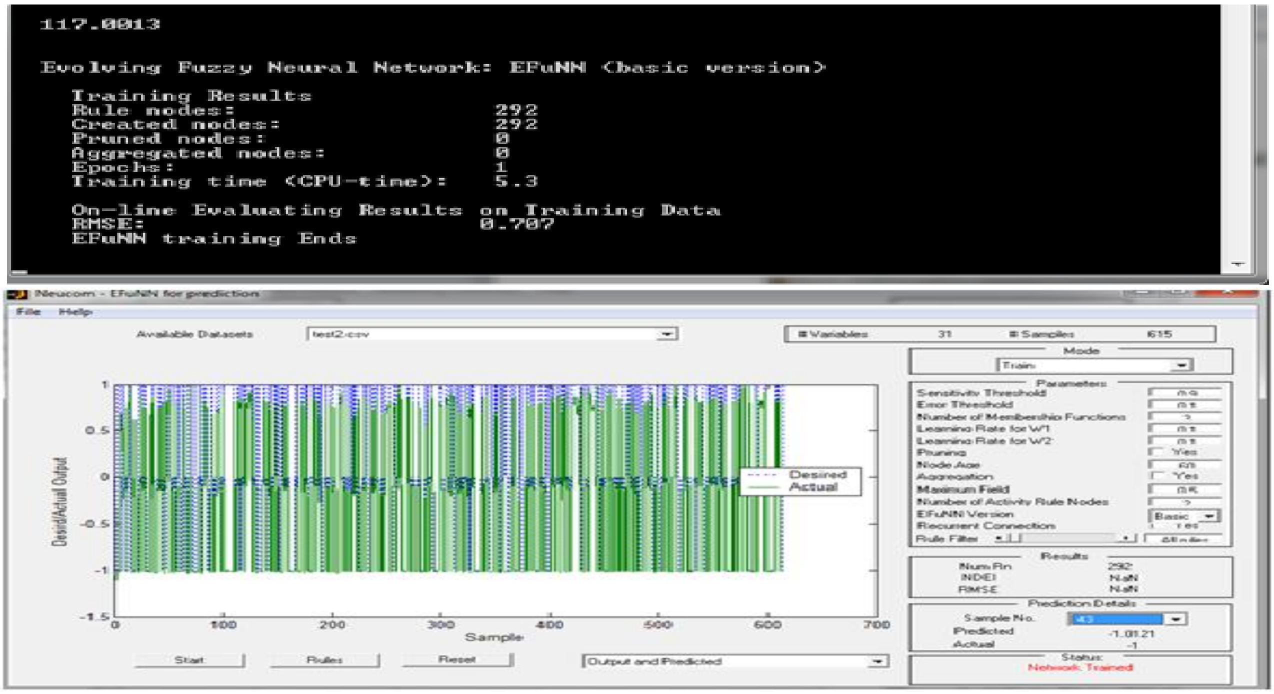
Coverage of cases (0.95 level) 99.5114 %

Mean rel. region size (0.95 level) 59.1205 %

Total Number of Instances 614

It can be seen that 495 websites out of 547 are phishy. The accuracy is 95.27%.

4.4 EFuNN based anti- Phishing model



```
117.0013

Evolving Fuzzy Neural Network: EFuNN (basic version)

   Training Results
   Rule nodes:                    292
   Created nodes:                 292
   Pruned nodes:                  0
   Aggregated nodes:              0
   Epochs:                        1
   Training time (CPU-time):      5.3

   On-line Evaluating Results on Training Data
   RMSE:                         0.707
   EFuNN training Ends
```

Neucom - EFuNN for prediction

Available Datasets: test2.csv

Variables: 31 # Samples: 615

Mode: Train

Parameters
Sensitivity Threshold: 0.9
Error Threshold: 0.1
Number of Membership Functions: 3
Learning Rate for W1: 0.1
Learning Rate for W2: 0.1
Pruning: □ Yes
Node Age: 60
Aggregation: □ Yes
Maximum Field: 0.5
Number of Activity Rule Nodes: 3
EFuNN Version: Basic
Recurrent Connection: Yes
Rule Filter: All rules

Results
Num Rn: 292
NDEI: NaN
RMSE: NaN

Prediction Details
Sample No.: 43
Predicted: -1.0121
Actual: -1

Status: Network Trained

Figure 4.7: EFuNN model output

Thus, the RMSE for EFuNN is 0.707.

5. CONCLUSION AND FUTURE WORK

The main aim of this work was to compare the fuzzy based anti-phishing system, with other snit phishing system. The work was able to successfully design a simple and efficient fuzzy based anti-phishing website detection. Further it was classified by Naïve Bayes, J48 system and EFuNN system.

Fuzzy logic has been used to successfully perform the task of phishing detection and categorization system. The model was also validated using neuro fuzyy based EFuNN model. Along with the AI model other classifier models where looked upon. These included the Naïve bayes classifier and the J48 tree based classifier system.

6. REFERENCES

[1] Abu-Nimeh, S., Nappa, D., Wang, X., and Nair, S. (2007) A comparison of machine learning techniques for phishing detection. In eCrime '07: Proceedings of the anti-phishing working group's 2nd annual eCrime researchers summit, (pp. 60–69), New York, NY, USA, ACM. .

[2] APWG, (2014) Phishing Activity Trends Report, http://antiphishing.org/reports/apwg_report_sep2014_final.pdf..

[3] Brooks, J. (2006) Anti-Phishing Best Practices: Keys to aggressively and effectively protecting your organization from Phishing Attacks, White Paper, Cyveillance.

[4] Buckley, J., and Tucker, D. (1989) Second generation fuzzy expert system. Fuzzy Sets and Systems, Vol.31, No.4, (pp. 271-284).

[5] Business Security Guidance, (2006) How to Protect Insiders from Social Engineering Threats, www.microsoft.com/technet/security/default.mspx,

[6] Chandrasekaran, M., Narayanan, K., and Upadhyaya, S. (2006) Phishing email detection based on structural properties. Proceedings of the NYS Cyber Security Conference.

[7] Dhamija, R., and Tygar, J. (2005) The battle against phishing: Dynamic security skins. In Proc. ACM Symposium on Usable Security and Privacy (SOUPS 2005), (pp. 77– 88). Dhamija, R., Tygar, J., and Marti, H. (2006) Why phishing works, In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM Press, (pp. 581-590), New York, NY, USA.

[8] Dhamija, R., Tygar, J., and Marti, H. (2006) Why phishing works, In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, ACM Press, (pp. 581-590), New York, NY, USA.

[9] Emigh, A. (2006) Online Identity Theft: PhishingTechnology, Chokepoints andCountermeasures. http://www.antiphishing.org/Phishing-dhs-report.pdf, Access date