Aslam Khan *et al*, International Journal of Computer Science and Mobile Applications,
Vol.6 Issue. 5, May- 2018, pg. 1-10 **ISSN: 2321-8363**
UGC Approved Journal
Impact Factor: 5.515

An Implementation of Phishing URL Detection Approach using Data Mining Technique

Aslam Khan[1], Rahul Sharma[2]

[1]R.K.D.F School of Engineering, Indore (M.P) India, aslamkhanashu@gmail.com
[2]R.K.D.F School of Engineering, Indore (M.P) India, sharma.rahul5656@gmail.com

Abstract: ***Now in these days the communication is frequently performed using digital channels. These channels are not much secure due to attackers and phishing. To prevent losses of the social and financial then security is required in communication. For phishing URL classification a machine learning based data model is proposed and discussed proposed work on this context. To find the features of phishing URLs, we work on the phish tank dataset with the proposed classification model, and using these features to understand new URLs, it is phishing or not. Therefore two technique of data mining is employed for train the model first the phishTank dataset is transformed into a binary dataset. In further, on the dataset the C4.5 algorithm is applied and then transformed data generates the rules using C4.5. URLs classification can be used these rules but to speed up the classification purpose required to reduce the amount of rules. Therefore the Bayesian classifier is implemented on C4.5 decision tree algorithm. For identifying the phishing URLs the Bayesian classifier prune the classification performed and the C4.5 generated rules. Java technique is used for the proposed implementation and Apriori algorithm based technique is used for comparative study. The comparative performance study demonstrates the efficient outcomes as compared to the traditional method of phishing URL classification.***

Keywords: Data Mining, Phishing, URL detection, Decision Tree, Apriori, phishTank, Website Phishing,

1. INTRODUCTION

In present time, Social networks is common and popular platforms where person interact with other person easily. For communication, share and know to each other is possible by person (users) with help of social networks. In social network platforms, there is huge amount of social and personal data available. So, privacy protection of user become more urgent research topics. A lot of privacy violation incidents that caused by phishing attacks and they still work for stealing information in traditional way. An attacker mimic electronic communications by which he get confidential information that provide by user, this social engineering form is phishing. Through emails, such type of communication that tricks users to visit those fraudulent website which is collect passwords, credit card details and confidential information of user's[1][2][3].

The main aim of phishing to steal user's identities and credentials, with malicious intention it encourages users to visit those fake webpages which is same feel like original website page and having exactly similar look. After getting user's identity many illegal activities like money laundering will be used to harm. Technique by which to trace user's sensitive information like credit card details, bank details, username, password and confidential information without user permission, is a criminal activity is done by phishing which is the main aim of phishing. From trusted website both are them design and content attributes are used in create a fake website to possible this activity. When user disclosure sensitive information and is unaware of entering into phishing zone, then it is being takes care by phishers [4][5].

In this paper work the machine learning approach is proposed for study and a new technique design. The machine learning approaches are help to understand the available patterns in available data and utilize these patterns to recognize the similar patterns in newly introduced data samples. That technique compute the essential features from the previously reported phishing

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

URLs and using these features the algorithm classify the web URLs in terms of phishing or legitimate URLs. In order to accomplish the classification task for the phishing URL classification the C4.5 decision tree algorithm and Bayesian classification algorithm is proposed for implementation. The proposed approach is promising for accurate analysis of URL data to identify the phishing URLs

2. PROPOSED WORK

This section provides the understanding about the proposed phishing detection technique. Therefore to explain the proposed work the methodology and proposed algorithm is explained in this section.

2.1. Methodology

The proposed system architecture of the phishing URL classification system is demonstrated in figure 1. Additionally their components are explained in details as:

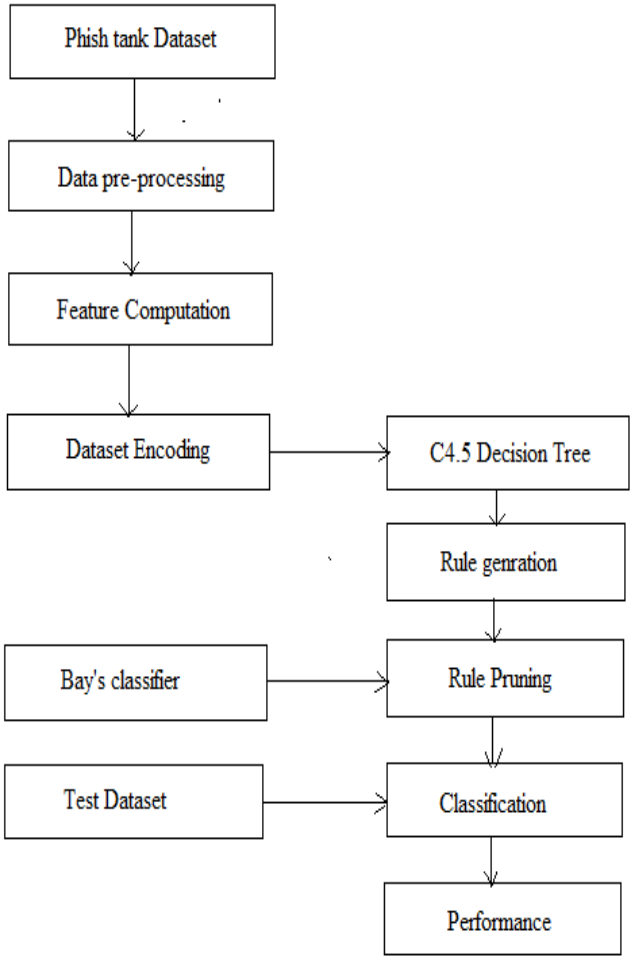

Figure 1 Proposed System Architecture

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

Phish tank dataset: phish tank dataset is initial input for the proposed system. The phish tank dataset is basically collection of different phishing URLs that are recently reported by different security institutions. The phish tank dataset includes the different information in dataset such as phish ID, URL, phish detail URL, submission date, verification time, online status, target. All these information is available either using the CSV format or by using web service API. In this system we are utilizing the CSV data format for training and testing purpose.

Data pre-processing: in this given dataset the entire information is not useful for analysis purpose. Therefore the unused attributes from the initial dataset is reduced and only we preserve the URL information for analysis.

Feature computing: after preprocessing of dataset the data contains only the URLs. For evaluation of dataset the motivation article [6] contains 14 different features for computations. These lists of features are as:

1. length of the host URL
2. number of slashes in URL
3. dots in host name of the URL
4. number of terms in the host name of the URL
5. special characters
6. IP address
7. Unicode in URL
8. transport layer security
9. Subdomain
10. certain keyword in the URL
11. top level domain
12. number of dots in the path of the URL
13. hyphen in the host name of the URL
14. URL length

Each URL is evaluated using these given feature constrains and by these evaluation a value is computed. The computed values for each URL are stored separately for further use.

Dataset encoding: after computation of features each URL return 14 values. Additionally the feature has a threshold value. Using this threshold the obtained feature values are compared. If the computed values are satisfying the threshold then the feature value is listed as 1 otherwise it is 0. In this manner all the computed features are again transformed into a binary vector. This vector transformed dataset is further used for training and testing of system.

C4.5 decision tree: the transformed dataset is consumed with the C4.5 decision tree algorithm. After applying the decision tree algorithm the entire data is mapped into a tree structure. That is further used in next process.

Rule generation: In this phase the generated tree data structure is used as input, the system process all the branches of the decision tree and produces the "if then else" rules.

Bays classifier: In this phase the "if then else" rules are consumed with the bays classification algorithm. The bays algorithm cross checks each rule according to the probability computation. Thus the less probable rules are removed from the generated classification rule list the remaining rules are further used for classification task.

Test dataset: the test dataset is derived from the initial training set and some additional legitimate URLs. Therefore the test dataset contains both the legitimate and phishing URL and by classifying these URLs the system performance is identified.

Classification: in this phase both the data the remaining classification rules and test set is produced as input. The system evaluates each URL on the basis of prepared classification rules and the system provides a class label for each URL.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

Performance: during the classification of URLs the generated class labels are observed for counting the accuracy of the classification in addition of that the misclassified URLs are counted for measuring the error rate of the system. In addition of that the required time for classification and consumed memory is also measured to demonstrate the performance of system.

2.2. Proposed Algorithm

This section described the summary steps of the proposed methodology in terms of algorithm. The table 1 contains the steps of algorithm.

Table 1 Proposed Algorithm

Input: phish tank dataset PD, test dataset TD Output: classification labels C
Process: 1. $R = readTrainingData(PD)$ 2. $P_n = preprocessData(R)$ 3. $for(i = 1; i \leq n; i++)$ a. $F = computeFeature(P_i)$ b. $E = encodeFeatures(F, Threshold_i)$ 4. $end\ for$ 5. $T = C4.5.GenarateRules(E)$ 6. $Rules_M = Bays.PruneRules(T)$ 7. $for(j = 1; j \leq m; j++)$ a. $C = ClassifyURL(Rules_j, TD)$ 8. End for 9. Return C

3. RESULT ANALYSIS

This section provides the evaluation of the performance for both kinds of algorithm namely C4.5 decision tree and Apriori. On compared different parameters to obtained performance of algorithms.

3.1. Time Complexity

The amount of time required to classify the entire test data is known as the time consumption. Time consumption of algorithm can be computed by finding the difference among the algorithm initialization time and process completion time.

This can be calculated using the following formula:

$$T_{consumed} = T_{end} - T_{start}$$

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

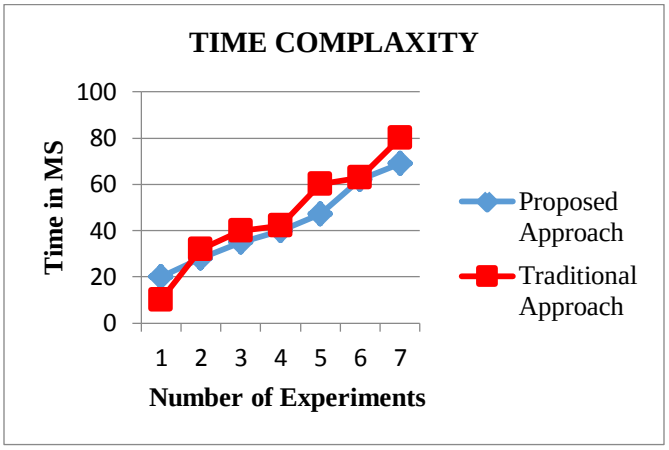

Figure 2 Time Complexity

The time complexity of both the algorithms (i.e. Apriori and C 4.5) is indicate in table 2 and figure 2. X axis contains the different number of experiments and the Y axis contains the time required to complete the process in order to represent the performance of algorithms. In terms of milliseconds time complexity is computed. On the report of obtained outputs of both the algorithms time is increases in similar ratio with increasing amount of the data. Moreover the C 4.5 requires less time to compute the classes of URLs in terms of phishing or legitimate as compared to Apriori algorithm.

Table 2 Numerical Values of Time Complexity

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	20	10
2	28	32
3	35	40
4	40	42
5	47	60
6	62	63
7	69	80

3.2 Space Complexity

Memory consumption of the system also termed as the space complexity in terms of algorithm performance. To calculated the algorithm performance we have to use following formula -

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

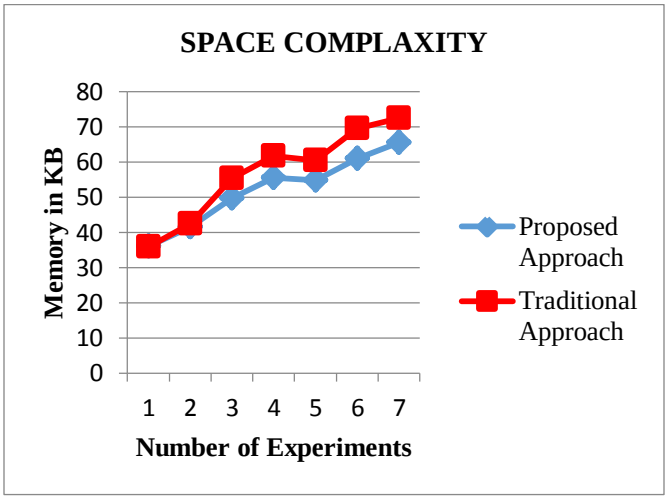

Figure 3 Space Complexity

The amount of main memory space required to compute the outcomes by any algorithm is demonstrates the space complexity of the algorithm. The space complexity of algorithms is described using figure 3 and table 3. The space complexity of algorithms is described in Y axis, which is computed in form of KB (kilobytes). Additionally different observation is listed by code execution is denoted in X axis. According to the obtained results the Apriori algorithm requires large amount of memory as compared to the C 4.5 Decision tree algorithm. The reason behind large resource consumption is that because the Apriori algorithm initially generates the candidate-sets and places them on main memory for further utilization. Therefore C4.5 is consuming less resources for tree generation.

Table 3 Tabular Values of Space Complexity

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	36.15	35.88
2	41.55	42.51
3	49.75	55.45
4	55.63	61.85
5	54.84	60.58
6	61.02	69.64
7	65.54	72.48

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

3.3. Accuracy

The accuracy is the measurement of the algorithm's correctness of recognitions. This can be calculated by finding the ratio between the correctly classified data and the total data samples are produced for classify. To compute the accuracy of algorithm the following formula can be used:

$$\text{Accuracy} = \frac{\text{Total correctly classified URLs}}{\text{total input for classify URLs}} \text{ X } 100$$

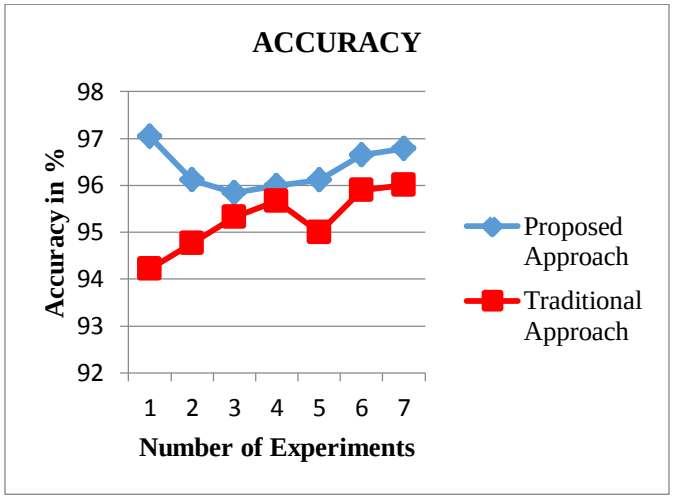

Figure 4 Accuracy

The performance of Apriori algorithm and C 4.5 Decision tree for phishing URL detection in terms of percentage accuracy is given using figure 4 and table 5. In this diagram the X axis contains different observation and the Y axis shows the amount of data correctly recognized by the algorithms. Additionally, blue line depict the proposed approach and rend line show traditional approach. According to the experimental results both the algorithms initially provides similar accuracy but as the number of data increases the difference in performance is clearly observed. The results show the accuracy of the C 4.5 increases as the amount of patterns increases for classification.

Table 4 Numerical Values for Accuracy

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	97.05	94.21
2	96.12	94.76
3	95.84	95.32
4	95.99	95.66

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

5	96.11	95
6	96.65	95.89
7	96.78	96.01

3.4. Error Rate

The error rate is the percentage amount of misclassified data over the total samples provided for classification. The error rate of the algorithm can be measured using the following formula:

$$\text{error rate} = \frac{\text{incorrectly classified data}}{\text{total data input}} \text{ X } 100$$

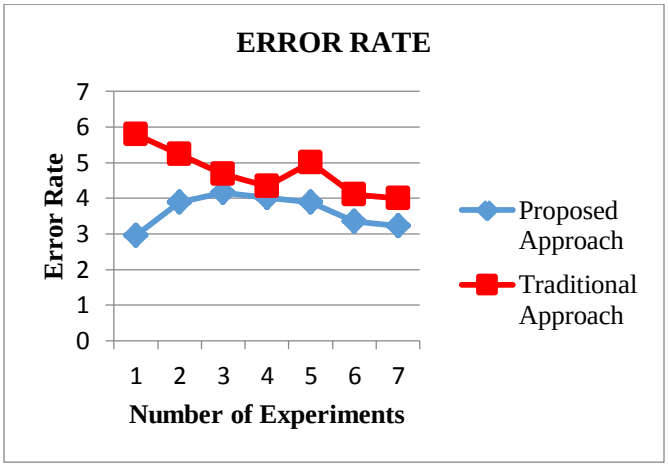

Figure 5 Error Rate

The error rate for both the algorithms namely Apriori algorithm and decision tree C4.5 algorithm for classifying the phishing URLs are given using figure 5 and table 5. The table includes the error rate values and the graph includes the lines for representing the performance values. The X axis of data contains the different number of experiments and the Y axis shows the corresponding obtained error rate produced by algorithms. According to the experimental results the C4.5 algorithm produces the less amount of error rate as compared to the Apriori algorithm.

Table 5 Error Rate

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	2.95	5.79
2	3.88	5.24
3	4.16	4.68
4	4.01	4.34

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

5	3.89	5
6	3.35	4.11
7	3.22	3.99

4. CONCLUSION AND FUTURE WORK

This section provides the conclusion of the performed study for identifying the phishing patterns of web URLs. In addition of that future extension of the proposed method is also discussed in this section

4.1. Conclusion

In this work a phishing detection tool is proposed for design and implementation. This tool is getting training from the phish tank dataset. Basically the phish tank dataset is repository where different phishing URLs are reported. In this dataset URL feature computation technique is applied and 14 different features are computed. These features are learned with the C4.5 decision tree algorithm. The decision tree algorithm generates the rules on the basis of available data. But the number of rule is in large quantity which increases the comparison and detection time. Therefore the Bayesian classifier is implemented for pruning of unused rules. This process enhances the speed of URL classification. During testing of model the random selected phish tank data objects are collected and some other known legitimate URLs are captured in a file and this test file is used testing of given data model. The proposed technique is promising for efficient and accurate classification of URL objects.

The implementation of the proposed approach is performed using JAVA technology. After implementation the performance of system is measured and compared with the similar model. The obtained performance is summarized in table 6

Table 6 Performance Summary

S. No.	Parameters	Proposed method	Traditional method
1	Accuracy	High	Low
2	Error rate	Low	High
3	Time consumption	Low	High
4	Space consumption	Low	High

According to the obtained performance the proposed technique found efficient and accurate therefore it is suitable for utilizing with the real world applications.

4.2. Future Work

The main aim of the proposed work is to design and development of efficient and accurate phishing URL classification technique is developed successfully. In near future the following extensions are feasible for the work:

✓ The proposed technique is just considers the limited 14 features for phishing URL identification need to find more phishing URL features for more accurate data analysis.

ISSN: 2321-8363
UGC Approved Journal
Impact Factor: 5.515

✓ Current approach is a rule based classification approach which is slow as compared to opaque models therefore in near future need to work on speeding up the process of classification.

REFERENCES

[1] Mao, Jian, Wenqian Tian, Pei Li, Tao Wei, and Zhenkai Liang. "Phishing-Alarm: Robust and Efficient Phishing Detection via Page Component Similarity", IEEE Access 5 (2017): pp. 17020-17030.
[2] Norouzi, Monire, Alireza Souri, and Majid Samad Zamini. "A data mining classification approach for behavioral malware detection." Journal of Computer Networks and Communications 2016 (2016): 1.
[3] Routhu Srinivasa Rao and Syed Taqi Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach", Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science 54 (2015) pp.147 – 156.
[4] Hassan Y. A. Abutair, Abdelfettah Belghith, "Using Case-Based Reasoning for Phishing Detection", 8th International Conference on Ambient Systems, Networks and Technologies, Procedia Computer Science 109C (2017) pp. 281–288
[5] R. Gowtham, Dr. Ilango Krishnamurthi, K.Sampath Sree Kumar, "An efficacious method for detecting phishing webpage through Target Domain Identification", Decision Support Systems November 30, 2013
[6] Jeeva, S. Carolin, and Elijah Blessing Rajsingh, "Intelligent phishing URL detection using association rule mining." Human-centric Computing and Information Sciences 6.1 (2016): pp. 1-19.

