Applying C4.5 Decision Tree to Classify Phishing URL

[1]Arpita Gupta, [2]Mahendra Patel
[1]M.Tech Student, [2]Assistant Professor
[1]Department of Computer Science and Engineering,
[1]Medi-Caps University, Indore, India

Abstract : ***Data mining is one of the most essential techniques for analyzing large amount of data. The analysis of data leads to provide the data patterns by which experts can predict, make decisions and understand the associations among different attributes. Therefore, in large number of applications, the data mining techniques are frequently used. In this presented work, the data mining is demonstrated for performing the classification task of web based URLs. The main aim behind this classification is to analyze the pattern of unsolicited URLs. In this context, an accurate data mining technique namely C4.5 decision tree algorithm is used. The C4.5 decision tree algorithm is a supervised learning algorithm which performs the entire data and their attributes into a kind of tree structure and using this structure can be used for predicting the data patterns. In this decision tree, the tree nodes are providing the relationship among the data attributes and the edges of this tree provides the values of the attributes to form the relationships. Finally, the leaf node of this tree demonstrates the decision of the classification. In order to provide the decision of phishing URLs, the PhishTank data set is used. This data set contains a significant amount of phishing URLs which is further evaluated on the basis of feature extraction technique. Using these feature study and obtained patterns in data, the decision tree algorithm is prepared and prediction is made. In order to justify the performance of the proposed URL classification technique. A traditional technique is also implemented which is based on Apriori algorithm. The comparative performance study demonstrates the proposed technique is efficient and accurate as compared to the Apriori based phishing URL classification technique.***

Index Terms - Apriori Algorithm, Data Mining, Decision Tree, Phishing URL Detection, PhishTank, Phishing Website.

I. INTRODUCTION

Now, in these days in most of the applications where a large amount of data is generated, the data mining techniques are employed for understanding the patterns of data. The employed techniques are either help to the administrative points or used for functional points to regulate the different events in an organization. In this presented work, the main aim of the work is to employ, the data mining technique to the cyber security technique for understanding the patterns of phishing URLs and classify them from the legitimate URLs. The phishing is a criminal act by which using false techniques and methods the malicious users are recovering the user's confidential and private information to harm the target person socially or financially. Therefore, phishing detection and prevention is an essential work in cyber security, now in these days.

In this context, the supervised learning based data mining model is proposed for design and implementation. The supervised learning techniques first use the historical data to understand the patterns and then use the learned patterns to recognize the similar patterns form the newly introduced data. In order to apply the data mining techniques for phishing URL characteristics understanding we need a collection of phishing URLs therefore phish tank database is considered for experimentation. In addition of that, some well known URLs are also considered for legitimate pattern analysis. In addition of that, for finding the patterns form the data the technique discussed in article [1] is also considered to evaluate the URLs on the basis of listed heuristics and 14 features form each URLs are extracted. Finally, the extracted 14 features are learned with the data mining algorithms to classify the patterns of URLs. In this section, the basics of the proposed phishing classification system design is provided in further the detailed study on the phishing techniques and their detection approaches are provided.

II. PROPOSED WORK

The proposed work is focused on developing a data mining technique that helps to classifying the URLs in binary classes namely phishing and legitimate. Therefore, this section includes the understanding of proposed methodology and the algorithm study.

2.1 System Overview

Phishing is a crime in which the malicious user tries to trap a web user to steal their confidential and private information. Using this information the attacker is target the person in social or financial context. In web, there are a number of approaches by which the attacker trying to trap a web user. Among them by using false URLs and web pages are much popular and effective. The attackers most of the time make a false duplicate pages of a popular web application. By using emails, SMS or other channels try to distribute the false information the normal user click on these links and passes their valuable information to the attacker. Therefore, there are some strong technique is required to find such kind of malicious URLs from web. The phishing is a classical problem in cyber security and to prevent such kind of attacks a number of efforts and contributions are placed in recent years but among most of them are not much effective or accurate to detect the phishing patterns. In this presented work, a data mining based technique is proposed for design and implement that technique first consumes the phish tank data set to recover the phishing URL features from URLs and on the basis of the recovered patterns the data mining algorithm learn and classify the URLs. The proposed technique of phishing URL classification is motivated form a research article where for classifying the phishing URL the Apriori algorithms, association rules are used. Basically, the Apriori based rules are large in quantity and development of these rules need a significant amount of computational resources. In this context, the supervised learning algorithm helps to reduce the amount of learning time and less number of rules to classify the phishing URL patterns. In this section, the proposed work and their need is explained in further sections the proposed model is explained in detail.

2.2 Algorithm Study

This section provides the study of different algorithms that are used for proposed system design. Basically, here we utilize only the C4.5 classification algorithm and their algorithm is provided in this section.

C4.5 (developed by Quinlan, 1993) an algorithm that learns the decision-tree classifiers. It is observed that C4.5 performs short in the domain where, there is pre-entrance of continuous attributes compared with the learning tasks with mostly separate attributes. For instance, a system which looks for well-defined decision tree with 2 levels and then put comments:

"The correctness of trees made with T2 is equalized or even exceed trees of C4.5 upon 8 out of all the datasets, with the entire except one that have incessant attributes only."

INPUT: An exploratory data set of data (D) portrayed with the means of discrete variables.

OUTPUT: T denoted as a decision tree say which is constructed by means of passing investigational data sets.

1. A node (X) is created;
2. Verify if the instance falls in the same class.
3. Create node (X) as the leaf node and assign a label CLASS C;
4. Check IF the attribute list is empty, THEN
5. Create node (X) a leaf node and assign a label of most customary CLASS;
6. Now, select an attribute which has highest information gain from the provided attribute List, and then marked as the test_attribute;
7. Confirming X in the role of the test_attribute;
8. In sequence to have a recognized value for every test_attribute for dividing the samples;
9. Making a fresh twig of tree that is suitable for test_attribute = atti from node X;
10. An assume that B_i is a group of test_attribute=atti in the samples;
11. Check If Bi is NULL, THEN
12. After that, add a new leaf node, with label of the most general class;
13. A leaf node, ELSE is going to be added and returned by the Generate_decision_tree.

2.3 Methodology

The proposed phishing detection model is demonstrated in Fig 1. Additionally, their intermediate steps are also explained in this section.

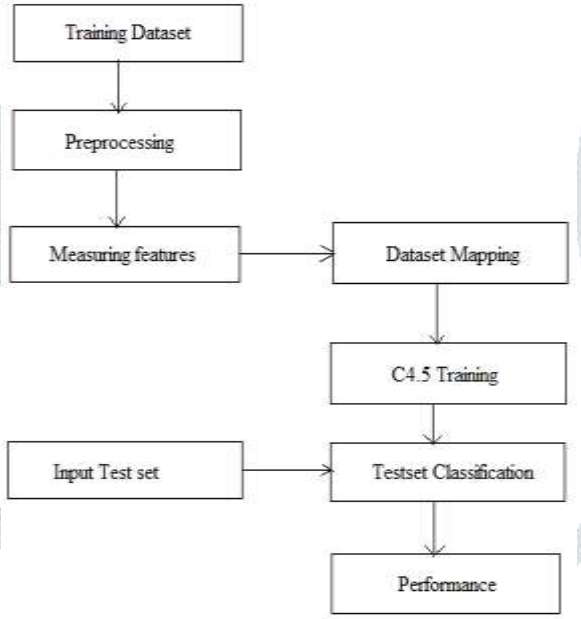

Figure 1 Proposed System Architecture

Training Dataset

In any data mining or machine learning model, the training set is key element. The implemented classifiers are performing training on the training to recognize similar patterns as given in predefined data set. In this context here the phish tank data set is considered for design and implementation. The phish tank data set is a global repository where different organizations and institutes are submitting the phishing reported URLs with their description.

Preprocessing

The phish tank data set is contains a number of different attribute which are not suitable to use for understanding the pattern of URLs. In this data set, only two attributes the class labels and the target URL is used for developing the machine learning model. Therefore, these two attributes are preserved and remaining attributes are removed from the initial input data set. After Preprocessing of input data, the significance from the input data is computed as described in next section.

Measuring features

In the motivational article [1], there are 14 different features or heuristics are defined for identifying the phishing patterns these significance are:

1. Length of the host URL
2. Number of slashes in URL
3. Dots in host name of the URL
4. No. of terms in the host name of the URL
5. Special characters
6. IP address
7. Unicode in URL
8. Transport layer security
9. Sub-domain

10. Certain keyword in the URL
11. Top level domain
12. No. of dots in the path of the URL
13. Host name of the URL is in Hyphen
14. URL length

These 14 heuristics are the basically constrains or the characteristics of the URLs which are required to recover from all the data set URLs. Therefore, in the proposed model 14 different functions are implemented which works to count the heuristics.

Dataset Mapping

Each heuristic function associated with a predefined threshold value. During evaluation of heuristic function for a URL produces values grater than the defined threshold value then the function returns the negative value says 0 and if it satisfies the given threshold then it returns the 1 value. Therefore, for each URL 14, 0 and 1 flags are computed and data set is restructured in terms of 15 attributes where 14 values are binary pattern values taken from the heuristics and remaining 1 attribute is a class label phishing or legitimate.

C4.5 Training

After transformation of data set into binary attributes, the C4.5 machine learning algorithm is employed on the data. The C4.5 algorithm generates decision tree using the input data set attributes. As described in Fig 1, the C4.5 algorithm usages the attributes and values to form a decision tree by traversing this tree the class labels of a given pattern can be identified.

Input Test Set

In order to validate trained machine learning model a set of URLs are also prepared. That set of URLs are contains some well known legitimate and phishing URLs that exist in phish tank data set. The system first transforms the data set using the described 14 heuristics and their threshold values. After transforming the test data set each instance of data set is classified using the developed decision tree.

Performance

During the classification of test set the system count how many instance of data is identified as phishing URLs and how much accurately identified as legitimate additionally that counts are verified according to the available class labels. Based on this counting, the performance of system is described in terms of accuracy and classification error rate.

III. PROPOSED ALGORITHM

This section provides the summary of the proposed methodology steps in terms of algorithm steps. Therefore, Table 1 contains the list of steps which are followed for classifying the phishing URLs.

Description

Phishing attack is the crucial problem of the today's web generation. In this scenario, most of the web information suffers from this attacks. The proposed algorithm, illustrate, the phishing URL detection using PhishTank Dataset. In this algorithm, firstly, we take training dataset and store in to some other variable. Now, pre-process dataset after passing this variable into process. After this, we get pre-processed data. This pre-processed data means number of URL. Let, assume, there is 'n' number of URL in our pre-processed data. In this dataset, there are 14 features or heuristic. Therefore, we have count these feature form 1 to 14 one by one. Heuristic function for a URL produces values greater than the defined threshold value then the function returns the negative value says 0 and if it satisfies the given threshold then it returns the 1 value. 1 values shows the value of phishing on URL. After transformation of dataset into binary attributes, we apply data mining algorithm i.e. C4.5 for training of model. We train model using pre-processed data and classify this train model using test dataset. Finally, we get the class label that demonstrate how much accurate our prediction model. Finally, we calculate accuracy of the overall process that describe how many URL accurately classified and how much are phishing URLs.

TABLE 1 PROPOSED ALGORITHM

Input: phish tank dataset D, test dataset T, list of heuristics H_{14}
Output: class labels C

Process:

1. $R = readTrainingData(D)$
2. $P_n = preprocessData(D)$
3. $for = (i = 1; i \le n; i++)$
   a. $for = (j = 1; j \le 14; j++)$
      a.i. $H' = H_j.countHeuristics(P_i)$
      a.ii. $if\ (H_j.ThresholdSatisfy(H'))$
         a.ii.1. $P.value = 1$
      a.iii. Else
         a.ii.2. $P.value = 0$
      a.iv. End if
   b. End for
4. End for
5. $Model_{train} = C4.5CreateTree(P_n)$
6. $C = Model_{train} \cdot Classify(T)$
7. Return C

IV. RESULT ANALYSIS

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

4.1 Time Complexity

The quantity 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:

Time Consumed = End Time - Start Time

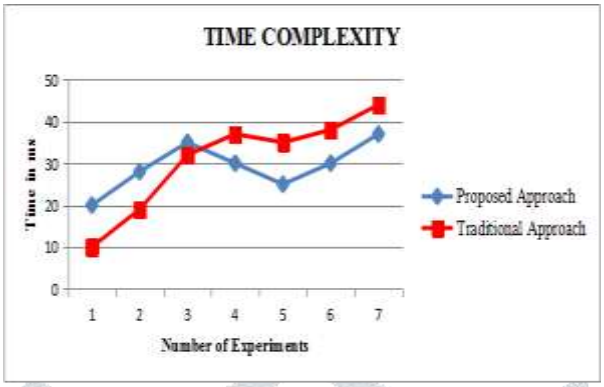

Figure 2 Time Complexity

The time complexity of both the algorithms (i.e. Apriori and C4.5) is denoted in Table 2 and Figure 2. In order to represent the performance of algorithms the X axis contains the different number of experiments and the Y axis contains the time required to complete the process. Here, the time complexity is computed in terms of milliseconds. According to the obtained results, the time of both the algorithms with increasing amount of data is increases in similar ratio. Additionally, the C4.5 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	19
3	35	32
4	30	37
5	25	35
6	30	38
7	37	44

4.2 Space Complexity

The space complexity in terms of algorithm performance is also known as memory consumption of the system. This can be calculated using the following formula:

Memory Consumption = Total Memory - Free Memory

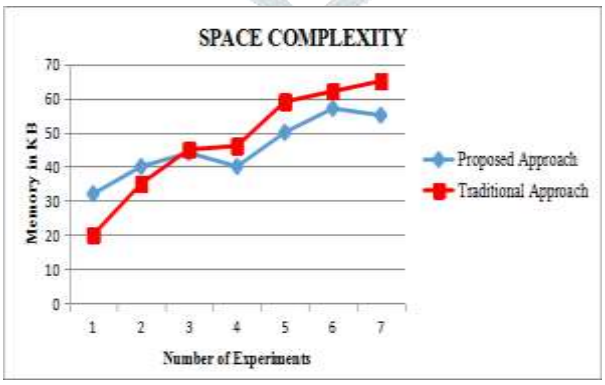

Figure 3 Space Complexity

The space complexity of the algorithm demonstrates the amount of main memory space required to compute the outcomes by any 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 C4.5 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 fewer 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	32	20
2	40	35
3	44	45
4	40	46
5	50	59
6	57	62
7	55	65

4.3 Accuracy

The accuracy is the measurement of the algorithm's correctness of recognition. 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:

$$Accuracy = \frac{\text{Total correctly classified}}{\text{Total input for classify}} \times 100$$

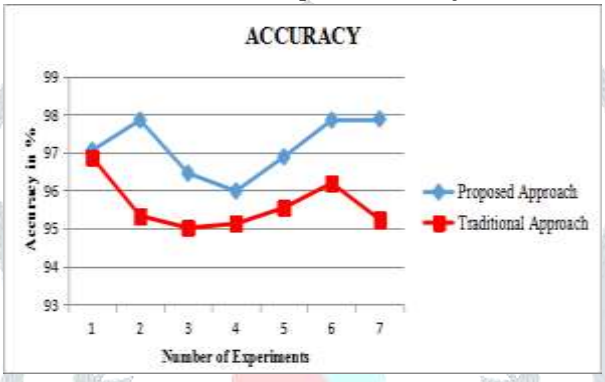

Figure 4 Accuracy

The performance of Apriori algorithm and C4.5 Decision tree for phishing URL detection in terms of percentage accuracy is given using Figure 4 and Table 4. 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 C4.5 increases as the amount of patterns increases for classification.

TABLE 4 NUMERICAL VALUES OF ACCURACY

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	97.05	96.87
2	97.85	95.33
3	96.45	95.01
4	95.98	95.12
5	96.88	95.54
6	97.85	96.19
7	97.88	95.23

4.4 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:

$$Error\ Rate = \frac{\text{Incorrectly classified data}}{\text{Total data input}} \times 100$$

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 TABULAR VALUES OF ERROR RATE

Number of Experiments	Proposed Approach (Decision Tree C4.5)	Traditional Approach (Apriori Algorithm)
1	2.95	3.13
2	2.15	4.67
3	3.55	4.99
4	4.02	4.88
5	3.12	4.46
6	2.15	3.81
7	2.12	4.77

IV. CONCLUSION

The data mining is a technique which is used for analyzing the large amount of data to reduce the cost of manual data analysis and to improve the efficiency of analysis. Therefore, in a number of new generation applications where the large amount of data generated the machine learning and data mining algorithms are employed to evaluate and recover the valuable patterns to make effective decisions, predictions and others. In this presented work, the application of data mining technique in cyber security is presented. The proposed work, provides a data mining based technique to analyze the phishing URL patterns to recognize the similar malicious URLs in real world.

The proposed technique is based on a machine learning model which is used to classify the phishing URLs using the Apriori based association rules. The Apriori algorithm is a frequent pattern analysis technique which consumes a significant amount of time and memory to generate the association rules. Therefore, in order to improve the efficiency of the Apriori algorithm in this model the supervised learning algorithm namely C4.5 decision tree algorithm is proposed for new model implementation. Additionally, the similar 14 features are used to obtain the characteristics of the URLs. After implementation of the C4.5 algorithm in this context the performance measurement and comparative study is the key area of investigation in improvement in traditional machine learning model.

The implementation of the proposed phishing URL detection technique is performed under JAVA technology. Additionally, for perform experimentation the phish tank database is used in CSV format. After performing implementation the classification results and system performance is measured in the following parameters as described in Table 6.

TABLE 6 PERFORMANCE SUMMARY

S.No.	Parameters	Proposed Technique	Traditional Technique
1	Time Complexity	Low	High
2	Space Complexity	Low	High
3	Accuracy	High	Low
4	Error Rate	Low	High

V. FUTURE WORK

The main aim of the proposed work is to design an efficient and accurate data mining model that help to recognize the URLs in terms of phishing or legitimate. In this context, a decision tree based model is implemented successfully and their performance is demonstrated. The possible future extension of the work is reported as.

1. The proposed work implements single classifier for classifying the patterns in near future the ensemble learning technique is proposed for improving the classification performance.
2. The proposed work only considers the 14 predefined features from URL pattern approximation in near future more literature is collected to identify more features for phishing URL patterns.

REFERENCES

[1] Jeeva, S. Carolin, and Elijah Blessing Rajsingh. 2016. Intelligent phishing url detection using association rule mining. Human-centric Computing and Information Sciences 6.1. pp. 1-19.

[2] K. Amarendra. May-June 2014. A Survey on Data Mining and its Applications from International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3 Issue 3.

[3] Samiddha Mukherjee, Ravi Shaw, Nilanjan Haldar and Satyasaran Changdar. 2015. A Survey of Data Mining Applications and Techniques (JCSIT) from International Journal of Computer Science and Information Technologies, Volume 6 Issue 5. pp. 4663-4666

[4] Bora, Shital P. 2011. Data mining and ware housing from In Electronics Computer Technology ICECT from 3rd International Conference on IEEE, vol. 1, pp. 1-5.

[5] Nikita Jain and Vishal Srivastava. Nov-2013. Data Mining Techniques: A Survey Paper IJRET from International Journal of Research in Engineering and Technology, Volume 02 Issue 11.

[6] Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. Data mining: concepts and techniques from Elsevier.

[7] Sumathi, Sai, and S. N. Sivanandam. 2006. Introduction to data mining and its applications from Springer, Vol. 29.

[8] Ian H. Witten; Eibe Frank; Mark A. Hall. 30 January 2011. Data Mining: Practical Machine Learning Tools and Techniques from Elsevier, 3rd Ed.

[9] Kantardzic and Mehmed. 2003. Preparing the Data. Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. 26-52.

[10] Petre, Ruxandra - Stefania. 2012. Data mining in cloud computing from Database Systems Journal. 3.3. 67-71.

[11] Sebastiani, F.. 2002 . Machine learning in automated text categorization from ACM Computing Surveys, Volume 34, Number 1. pp. 1–47