IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Phishing URL Classification Using ARM Based Association Rules

Rahul Patel[1*], Anand Rajavat[2]

[1]PG Scholar, Computer Science Engineering, Shri Vaishnav Institute of Information Technology, SVVV, Indore, India
[2]Director, Shri Vaishnav Institute of Information Technology, SVVV, Indore, India

[*]*Corresponding Author: rahulpatel@svvv.edu.in,*

DOI: https://doi.org/10.26438/ijcse/v7i5.710717 | Available online at: www.ijcseonline.org

Accepted: 18/May/2019, Published: 31/May/2019

Abstract— The usages of internet and increasing volume of internet users force us to think about the current cyber security and its infrastructure. There are a number of different kinds of attack deployed using the internet among the phishing is one of the most serious attack conditions. In this condition an innocent user can lose their financial status or social credibility. The phishing attacker still the users private, sensitive and confidential information, additionally usages these data to harm the target person. Therefore it is a serious criminal offence. In this context a number of techniques are developed for resolving the issues of phishing attacks, but most of them are not much effective due to changing strategies of the phishing attackers. The phishing attacks are mostly deployed using the malicious and forged URLs. Thus the pattern recognition of these URLs can help us to resolve the phishing attacks. In this presented work a data mining based phishing URL classification technique is proposed for design and implementation. The proposed technique usages the phish tank database for obtaining the knowledge about the phishing URL properties and then using these properties the data mining system prepare the rules for identifying the target phishing URLs. In this context the ARM algorithm is employed. The Arm algorithm first prepares the association rules using the apriori algorithm. After generation of association rules the confidence based score are used to label each rule to a score values. Finally on the basis of score threshold the unfruitful rules are pruned. The remaining rules are used for classification task. The proposed technique is implemented and their performance is measured, according to the gained performance the proposed technique is accurate and efficient as compared to the traditional apriori algorithm based classification technique.

Keywords—: *phishing attack, malicious URL classification, association rule mining, rule based classification, ARM algorithm, outlier removal.*

I. INTRODUCTION

In this age of communication and computation, a significant amount of users are now connected through the internet and it's services. Additionally every day a wide range of new users are attracted from internet and it's applications. On average a common person interacted with the internet and it's applications many times. On the other hand the cyber crime rate is also increases in the similar manner. Among most of the cases are based on online fraud or phishing. The phishing is a kind of cyber crime where the innocent user is passes their confidential and private information to the attacker. Additionally attacker usage the obtained information and harm the data owners. Therefore the phishing is a criminal offence to capture someone's personal and confidential banking or other information for intestinally harming the end user.

In this presented work the phishing attacks are studied and using the data mining techniques a fruitful technique is tried to develop. In this context a significant amount of literature is explored and it is observed, the phishing attacks are mostly deployed using the web URLs. Basically attackers prepare a clone of a target website and send the links to innocent users. The appearance of these malicious URLs can be any medium i.e. email, SMS, or any other source of information. User clicks the given URL and passes their information to the attacker. Therefore a URL classification technique is need to be designed which extract the valuable features from the raw URLs and estimate the possibility to be a URL is phishing or legitimate. In this context a data mining model is proposed that usages the historical phishing URL data to develop a data model additionally after preparation of target data model able to classify the URLs.

II. PROPOSED WORK

The proposed work is motivated to design an accurate phishing URL classification system, therefore this chapter includes the description of the proposed working model. Additionally their functional aspects are discussed in this chapter.

A. System Overview

The data mining techniques are a tool which is used to analyze the data. Here the data can be found in a specific format or it is available in an unorganized manner. The data mining techniques involve the various methodologies that enable us to evaluate and obtain the required application oriented patterns form the data. Therefore the data mining techniques offers the computational algorithms that are able to filter invaluable data and obtain the required data patterns. These algorithms can be able to learn and recognize the similar pattern of data if they found. This ability of data mining algorithms makes it suitable to use in various applications for decision making, pattern recognition, classification and prediction. In this presented work the application of data mining is provided in the domain of cyber security.

The internet attract new users and attackers to get benefits of the internet and it's applications. The normal user perform communication, banking and other task using the internet but an criminal want to still the confidential part of data and information. Therefore the attackers are smart internet user who continuously update itself but the normal user is not aware about the risk and prevention about the various attacks and frauds based on internet. Among various kinds of cyber security issues the phishing is one of the critical kinds of security issue. Additionally the developed methods need to be update day by day as the new attack strategies are appeared. Therefore the proposed work is intended to design and develop a data mining based phishing URL classification technique that evaluate the URL data and classify the data into legitimate and malicious nature of URLs. In this context the association rule mining technique is tried to implement. Additionally to reduce cost of the computational resources the modification on existing ARM based technique is employed for classifying the malicious URL data. This section provides the formal overview of the proposed work. The methodology of the URL classification is demonstrated in next section.

B. Methodology

The proposed system is demonstrated in the given figure 2.1. That diagram contains entire computational components for processing the data and obtaining the required outcomes.

Phish tank dataset: the phish tank is a database which is available online. This database is used by various cyber security companies. These companies report the identified phishing URLs and their target web URLs with the time and location information. Therefore that is a huge database which contains the historical recorded database of malicious URLs. That database can be used by online API (application programming interface) or by offline by downloading the CSV file from the phish-tank website. The dataset contains the following attributes:

1. **Phish ID:** that denotes the ID assigned by phish tank database for recognizing the phishing attack.
2. **URL:** that demonstrates the URL which is used to deploy the phishing attack.
3. **Phish detail URL:** the details about the phishing URL is available online in this link.
4. **Submission date:** the date of reporting this URL as phishing is given in this attribute
5. **Verification time:** the time when it is verified as the phishing URL
6. **Online:** that shows the detection type of URL
7. **Target:** the target company or organization for deploying the attack

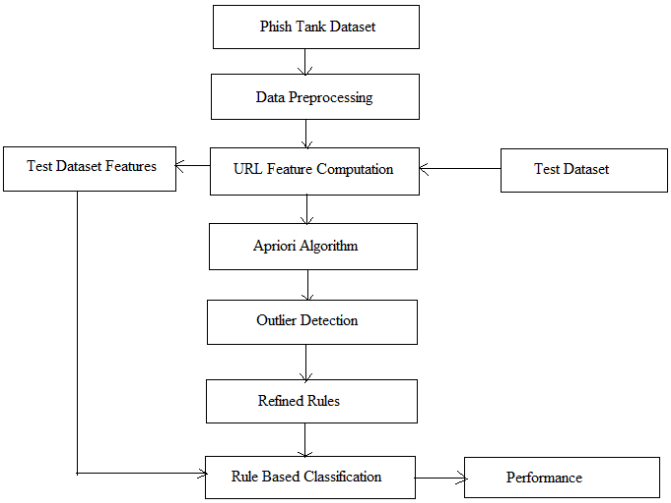

Figure 2.1 proposed system architecture

Data preprocessing: the preprocessing technique is a method by which the quality of data is improved. Therefore the noisy contents (i.e. attributes and instances) are removed from the initial dataset. In addition of that in some cases the data is transformed form one data format to another format by using the data normalization, transformation and mapping. Here the phish tank dataset contains 7 different attributes and all the attributes in data set is not much valuable for the proposed URL classification system. Therefore the required attributes URL which used for deploying the phishing attack is used for further process remaining attributes are reduced form the given dataset.

URL feature selection: the URL is an unstructured data format, which is not used for classification directly. Therefore we need to estimate some properties by which the data mining algorithms learn on the patterns. The research article [1] contains the list of features which are used here to compute the URL properties for utilizing with the classification technique. The following features are described in the target research article.

Table 2.1 features list

S. No.	Feature list
1	URL length
2	top level domain
3	number of dots in the path of the URL
4	certain keyword in the URL
5	hyphen in the host name of the URL
6	Sub-domain
7	Unicode in URL
8	transport layer security
9	length of the host URL
10	dots in host name of the URL
11	IP address
12	special characters
13	number of slashes in URL
14	number of terms in the host name of the URL

The given feature list is basically a kind of function which is used to estimate a value from a given URL. Among some of the values are in form of Boolean, or a numerical form. The individual function is used with a threshold value. Additionally if the threshold value is compared with these function based values then the Boolean value is received. Therefore if according to the properties the threshold value indicate the URL is malicious then we put for that 1 and if function recommends the URL is legitimate then we put here 0. Therefore each URL is extended to a 2 dimensional feature vector which contains the 14 attributes with 0 and 1 values, the URL and the class label phishing or legitimate.

Apriori algorithm: the above filtered and transformed data is used with the apriori algorithm to compute the association rules. The following algorithm is used for computing the association rules:

Table 2.2 apriori algorithm

Apriori (T,) { //T is the database and is the minimum support

$L1 = \{fr;$

$for\ (k = 2;\ L_{k-1}! =$

$C_k =\ candidates\ generated\ from\ L$

Ck= candidates generated from Lk-1

//that is Cartesian product and eliminating any size itemset that is not

//frequent

for each transaction t in data base c

#increment the count of all candidates in that are contained in t

= candidates in with minSupport

}//end for each

}

//end for

return

The apriori algorithm provides the association rules using the above mentioned algorithm. The generated apriori algorithm based rules are used next phase for reducing the amount of association rules.

Outlier detection: the outliers are the patterns that are not regularly fit on the target pattern as another data patterns are behaving. Therefore the removal of irregular pattern helps to improve the accuracy of the rules. In this presented work the ARM algorithm is applied for classification task therefore first the apriori algorithm is used to generate the association rules in further the association rule mining properties are used to prune the low confidence rules. Therefore first a score mapping table is prepared that categorize the confidence of the rules into a score value. The following mapping table is used for score identification.

Table 2.3 score table

Confidence in %	Score value
<=10	0
> 10 and < = 20	1
> 20 and < = 30	2
> 30 and < = 40	4
> 40 and < = 60	6

> 60 and < = 90	8
> 90	9

The calculated rules are mapped into the score values and on the basis of the threshold value here we use the threshold value = 3 to subdivide the generated rules into two groups i.e. lower association rules and higher association rules. The rules which contains the score less than or equal to 3 is categorized in lower association rules and higher then score value 3 is grouped as the higher association rules. The rules which are found in criteria higher score values are keep preserved and remaining rules are removed in this phase.

Refined rules: after removal of outlier rules the preserved rules are termed here as the refined association rules which are used here for classification task.

Test dataset: in addition of a set of URL is prepared which is used for testing of the proposed data model. That set of URLs contains some of legitimate URLs identified by us and some of the URLs are taken from the phish tank dataset.

Test dataset features: the test dataset URLs are again processed using the feature extraction phase for computing the required 14 features form each given URLs. These computed lists of features are passed into the next phase for classifying the data.

Rule based classification: the features are obtained in this phase which is obtained by processing the list of test dataset. The rules are applied on the features and by comparing the rules to the available attributes the decision is made the URL is legitimate or phishing.

Performance: based on the classification outcomes from the previous phase the performance of system is measured in terms of classification accuracy and error rate.

C. Proposed Algorithm

The above given process is summarized in this section. Therefore the algorithm steps are demonstrated using the table 2.4. The table contains all the necessary steps to process the URL data for classification.

Table 2.4 proposed algorithm

Input: phish tank dataset PD, Test Dataset T Output: classified URLs C
Process:

1. $R = readDataset(PD)$
2. $P_n = preProcessData(R)$
3. $for(i = 1; i \leq n; i++)$
    a. $F = computeFeatures(P_i)$
4. $end\ for$
5. $R_m = Apriori.GenrateRules(F)$
6. $O_p = removeOutlier(R_m)$
7. $T_D = ReadTestURLs(T)$
8. $for(j = 1; j \leq D; j++)$
    a. $TF = computeFeatures(T_j)$
9. $end\ for$
10. $for(k = 1; k \leq p; K++)$
    a. $C = O_p.classify(TF)$
11. End for
12. Return C

III. RESULT ANALYSIS

The implementation of the proposed technique is demonstrated in pervious chapter. This chapter includes the description about the measured performance with respect to the existing URL classification technique.

A. Accuracy

In data mining and machine learning applications the accuracy is an essential parameter of the performance. In this context a classifier how accurately recognize the pattern can be defined as the classification accuracy. Therefore accuracy is the ratio of the correctly recognized data samples with respect to the total samples produced for recognition. That can be measured using the following formula:

$$accuracy = \frac{total\ correctly\ recognized\ samples}{total\ samples\ to\ recognize} X100$$

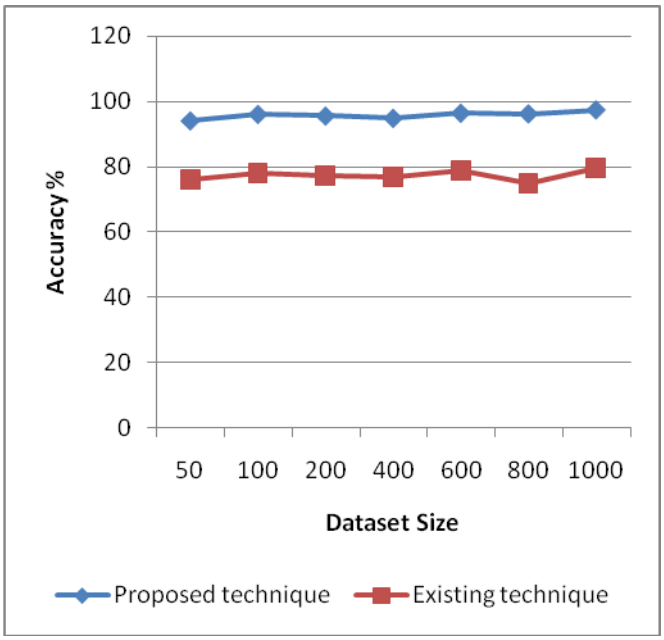

Figure 3.1 accuracy %

Table 3.1 accuracy %

Dataset Size	Proposed technique	Existing technique
50	94	76
100	96	78
200	95.5	77
400	94.8	76.8
600	96.4	78.6
800	96.2	74.8
1000	97.4	79.5

The accuracy of the proposed ARM based technique and traditionally available apriori based technique is given using figure 3.1 and table 3.1. The table 3.1 contains the experimental observation based values which is visualized using figure 3.1. The line graph contains experimental dataset size in terms of number of instances to classify in X-axis and the Y axis consist of the percentage accurately classified data. According to the obtained results both techniques are consistently securing the accuracy but the proposed ARM based technique demonstrate higher degree of accurate results as compared to traditional apriori based classification technique.

B. Error Rate

The error rate of the data mining application demonstrates the inaccurately classified instances in terms of percentage. In this context the error rate is the ratio of incorrectly recognized instances and the total instances produced for classification. The following formula can be used for finding the error rate of the proposed system.

$$error\ rate = \frac{incorrectly\ identified\ instances}{total\ instances\ to\ classify} X100$$

Or

$$error\ rate = 100 - accuracy$$

Table 3.2 error rate %

Dataset Size	Proposed technique	Existing technique
50	6	24
100	4	22
200	4.5	23
400	5.2	23.2
600	3.6	21.4
800	3.8	25.2
1000	2.6	20.5

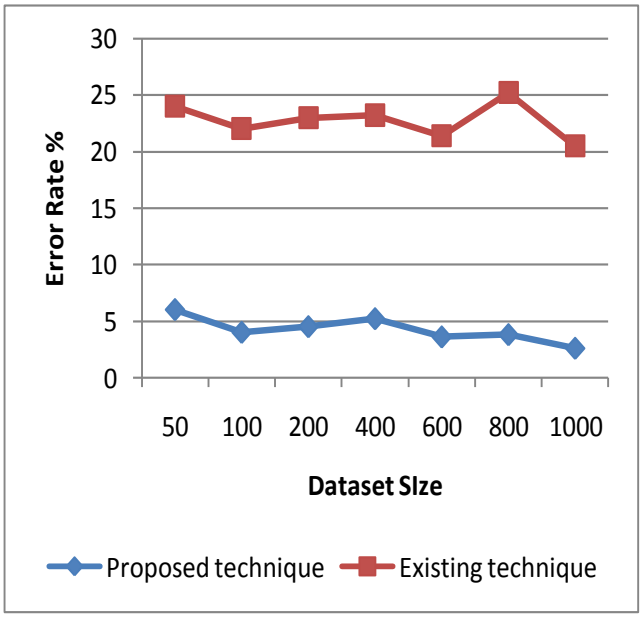

Figure 3.2 error rate %

The performance of implemented algorithms namely proposed ARM based phishing URL classification and traditional apriori algorithm based technique is compared using the figure 3.2 and table 3.2. The computed error rate in terms of percentage is reported using table 3.2 and the

graphical representation is given in figure 3.2. The X axis of the diagram includes the size of dataset in terms of number of instances and the Y axis shows the obtained error rate during classification. The obtained results demonstrate a huge gap between both the implemented methods. The error rate of the proposed technique is varying fewer than the traditional technique. Thus the proposed technique is acceptable as compared to the traditional Apriori algorithm based technique.

C. Memory usages

The memory usage is also known as the space complexity for the algorithms. That can be defined as the part of main memory acquired by the given process is termed as the space complexity or memory usages of the algorithms. The following formula can be used for measuring the main memory usages of the algorithms for JAVA technology.

$$memory\ usages = total\ allocated\ memory - total\ free\ space$$

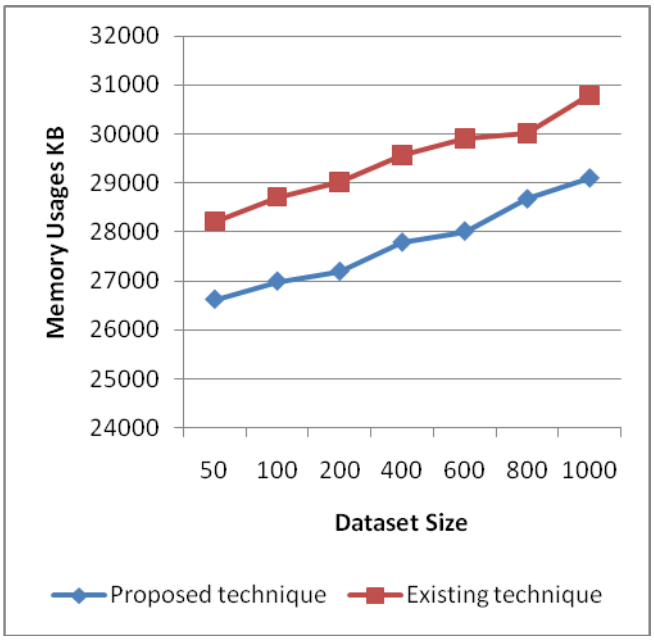

Figure 3.3 memory usages

Table 3.3 memory usages

S. No.	Proposed technique	Existing technique
50	26618	28217
100	26991	28719
200	27193	29011
400	27791	29571
600	28019	29910
800	28681	30018
1000	29109	30791

The consumed memory is measured here in terms of kilobytes (KB) for both the implemented algorithms. The table 3.3 shows the experimentally obtained memory usages values and the line graph for the given values are demonstrated in figure 3.3. The X axis of figure 3.3 shows the total number of data instances used for experiments and the Y axis show the amount of main memory consumed for processing the given data. According to the show results the proposed technique requires less amount of main memory as compared to the traditional apriori based URL classification technique.

D. Time Consumption

The processing of an algorithm with the given amount of data needs an amount of time. That amount of time is known as time consumption or time complexity of algorithm. The following formula can be used for computing the time requirements of the algorithms.

$$time\ consumption = algorithm\ end\ time - start\ time$$

Table 3.4 time consumption

Dataset Size	Proposed technique	Existing technique
50	105	126
100	156	172
200	171	231
400	284	345
600	367	418
800	417	469
1000	481	547

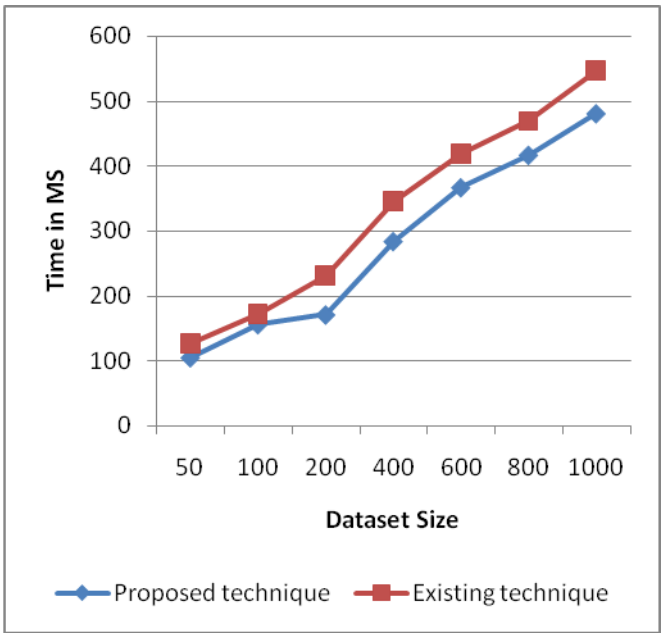

Figure 3.4 time consumption

The required time for processing data using both the algorithms namely ARM based URL classification and apriori based classification techniques are reported using table 3.4 and table 3.4. The line graph 3.4 is visualized using the reported time values in table 3.4. The X axis of diagram shows the total number of instances produced for processing with the algorithm and Y axis shows the amount of time consumed to process the data in terms of milliseconds (MS). The reported results shows the proposed technique consumes less amount of time as compared to the traditional apriori based technique. Because the traditional apriori based technique contains more number of rules as compared to the proposed ARM based technique. Thus the time for invoking the rules of the traditional approach is higher as compared to the proposed technique.

IV. CONCLUSION

The main aim of the proposed work to enhance the traditional apriori based phishing URL classification is accomplished successfully. The part contains the summary of the effort performed with the domain of URL classification as conclusion and the future extension of the proposed work is also included.

A. Conclusion

In cyber security the phishing is one of the classical problems. A number of different kinds of techniques and algorithms are presented to accurately recognize the phishing URL patterns and identification of phishing attack, but most of them are not much effective. The only solution of this critical issue is the awareness of the end internet user. The new and innocent users are not aware about the cheaters and phish-ers and they update their tricks every data therefore it is possibility to trap any new user by these attackers. Meanwhile, beyond the awareness the techniques based on white list and black list is also much time consuming and having additional overhead to store and compare each pattern to the target URLs. Thus the machine learning and data mining techniques are claimed to learn on the specified patterns and can recognize the similar possible patterns. in this context the proposed work is aimed to work with the data mining technique to handle such kind of serious issue.

The proposed work is intended to deal with a classification problem. Thus here we need to find an approach that successfully obtain the features from the raw URLs and then use some classification technique to differentiate the URL classes as phishing or legitimate. Basically the URLs are unstructured kind of data additionally the length of each URLs are different from each other. Thus first we need to estimate the meaningful features from individual URLs. Additionally after computing the features we can classify the URL features using some kinds of classification technique. Thus we applied 14 heuristics to compute the URL features as described in [1], additionally we utilize a ARM algorithm to classify the features using the rule based classification technique. The ARM algorithm is an extension of apriori algorithm which prunes the association rules using the outlier detection manner. This technique reduces the amount of rules and improves the accuracy and time consumption of the traditional apriori algorithm.

The implementation of the proposed ARM based phishing URL classification is performed using JAVA technology. Additionally, during the experimentation with the phish tank dataset we found improved results as compared to the traditional apriori based URL classification technique. To compare the proposed technique and traditional technique different performance parameters are measured and their values are summarized in table 4.1.

Table 4.1 performance summary

S. No.	Parameters	Proposed technique	Traditional technique
1	Accuracy	94 - 97.4 %	76 - 79.5 %
2	Error rate	2.6 – 6 %	20.5 - 25.2 %
3	Memory usages	26618 – 29109 KB	28217 – 30791 KB
4	Time consumption	105 – 481 MS	128 – 547 MS

According to the obtained results as shown in table 4.1 the proposed algorithm is accurate and less time and memory resource consuming. That is become fruitful for the real world applications for classifying the malicious URLs.

B. Future Work

The aim of the proposed work is to enhance the traditional apriori based technique to improve their accuracy of URL classification and improve the amount of time taken to process URL data is accomplished successfully. The following future work is proposed for extending the proposed ARM based classification model.

1. Involve more features to impure the existing ARM based URL classification problem

2. Incorporate an supervised learning algorithm which first train with the model and then classify it to reduce the time consumption of classification

3. Rule based classification need a significant amount of time and memory resources thus need to implement the technique with the opaque model

REFERENCES

[1] S. Carolin Jeeva and Elijah Blessing Rajsingh, "Intelligent phishing url detection using association rule mining", Hum. Cent. Comput. Inf. Sci. (2016) 6:10, DOI 10.1186/s13673-016-0064-3

[2] Gorunescu, F, Data Mining: Concepts, Models, and Techniques, Springer, 2011.

[3] Han, J., and Kamber, M., Data mining: Concepts and techniques, Morgan-Kaufman Series of Data Management Systems San Diego: Academic Press, 2001.

[4] Neelam adhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", vol.2, no.3, June

[5] Manoj and Jatinder Singh, "Applications of Data Mining for Intrusion Detection", International Journal of Educational Planning & Administration. Volume 1, Number 1 (2011), pp. 37-42

[6] M. Rajalakshmi, M. Sakthi, "Max-Miner Algorithm Using Knowledge Discovery Process in Data Mining", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 11, November 2015

[7] Smriti Srivastava & Anchal Garg, "Data Mining For Credit Card Risk Analysis: A Review", International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 2, Jun 2013, 193-200

[8] Dipti Verma and Rakesh Nashine, "Data Mining: Next Generation Challenges and Future Directions", International Journal of Modeling and Optimization, Vol. 2, No. 5, October 2012

[9] Neelamadhab Padhy, Dr. Pragnyaban Mishra, "The Survey of Data Mining Applications and Feature Scope", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), PP. 43-58 Vol.2, No.3, June 2012.

[10] Watson, D., Holz, T., and Mueller, S. (2005) Know your enemy: Phishing, behind the scenes of Phishing attacks, The Honeynet Project & Research Alliance

[11] Jagatic, T., Johnson, N., Jakobsson, M., and Menczer, F. (2007) Social Phishing, Community. ACM, Vol. 50, No. 10 (pp. 94-100).

Authors Profile

Mr. Rahul Patel pursed Bachelor of Engineering from RGPV Bhopal. He is currently pursuing M.tech in Computer science Engineering from SVVV indore.

Dr. Anand Rajavat currently Director of Shri Vaishnav Institute of Information Technology, Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore. Dr. Anand Rajavat does research in software engineering, Reengineering, Risk Engineering, Information Systems (Business Informatics), Artificial Intelligence and Software Engineering. He has published more than 35 research papers in reputed international journals. He has more than 20 years of teaching experience.

