ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Special Issue
National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

Cyber Crime Prediction using Data Mining Techniques

[1] K Chitra Lekha, [2] S.Prakasam, M.C.A, Ph.D.,
[1] Ph.D Research Scholar(Full-Time), [2] Associate Professor,
[1][2] Department of Computer Science and Applications
SCSVMV University, Enathur.
Kanchipuram.

Abstract:- **The cyber crime is a criminal activity in which Information technology systems are the means used for the commission of the crime. The purpose of applying data mining techniques in the field of cybercrime detection can produce significant results. The main objective of this work is to find which age group of respondents are affected by the common cyber crime threats using data mining techniques. The purpose of this work is to create an awareness of common cyber crime threats and provide some prevention measures. A survey has been conducted during September 2016 with different category respondents of 135. The questionnaire was designed to predict the factors about common cyber crime threats among the various sectors respondents of the society. In and around Kanchipuram , the respondents' survey was distributed from face to face contact. The WEKA tool is used for the study implementation for data analysis because it contains a collection of visualization tools and algorithms. In this study author used Classification algorithm (J48) that predict respondents' age group who are most affected by common cyber crime threats. The SPSS tool is used to Cross tabulate between the respondents' age and some common cyber crime attacks.**

Keywords:-- **Cybercrime, Data Mining, J48, SPSS, WEKA.**

I. INTRODUCTION

Cyber crime is a criminal activity involving computers or computer networks that include everything from electronic cracking to denial of service attacks. Daily five billion phone messages and 294 billions of emails are exchanged and most people around the Universe now depend on reliable access and accuracy of these communication channels [1]. The security mechanisms for preventing measures of the cyber crime like updating the computer, by choosing strong passwords, by protecting computer with security software, Be Social-Media Savvy[2]. Although cyber crime cells have been setup in major cities, Cyber crime is omnipresent and most cases remain unreported due to lack of awareness. The print media has a duty to educate unwary parents and adults about the dangers inherent in trending dangerous areas in the cyber world. Prediction of cybercrimes is a difficult task that needs human intelligence and experience and data mining is a technique that can assist them with cybercrime detection problems. A survey has been conducted during September 2016 with respondents of various categories of 135. The WEKA software is used in this study for prediction techniques that predict which age group of respondents are affected by the common cyber crime threats. In this study, author used Classification algorithm(J48 Decision tree algorithm) in which the classifier model builds J48 pruned tree. This analysis identifies by which cybercrime threat a respondent is affected. The some of common cybercrime threats are credit/debit card fraud, Phishing, Cyber bullying, Denial of service attack, Cyber stalking.

II. RELATED WORK

K.Sridharan and Saktheeswari(2013) have said that cyber crime differs from physical or "terrestrial" crime in four main ways: being easy to commit, requiring minimal resources for great potential damage, being committable in a jurisdiction in which the perpetrator is not physically present, and often, not being entirely clearly illegal.

Dr. Ajeet S. Poonia(2014) have said that Cybercrime has high potential and easy to commit without any physical existence required as it was global in nature due to this it has become a challenge and risk to the crime fighter and vice versa.

Janhavi J Deshmukh and Surabhi R Chaudhari(2014) have said that the global spam rate, malware rate, and phishing rate is increasing rapidly and there is a potential impact of cyber crime on economics, consumer trust and production time.

P. Arokia Vasantha Rani(2015) have said that the cyber offense as a whole refers to offenses that are committed

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Special Issue
National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm to the victim directly or indirectly, using modern telecommunication networks such as Internet (Chat rooms, emails, notice boards and groups) and mobile phones(SMS/MMS).

Umesh R. Gadhave, Dr. Sandeep R. Sirsat(2016) have said that the attacks those are processed knowingly can be considered as the cybercrime and they have serious impacts over the society in the form of economical disrupt, psychological disorder, threat to National defense, social nuance etc.

K.Chitra Lekha and S.Prakasam(2016) have said that Classification algorithms in data mining builds decision trees using pruning method to predict categorical class labels.

Malathi. A and Dr. S.Santosh Baboo(2011) have said that semi-supervised learning techniques can be used for knowledge discovery from the crime records and to help in increasing the predictive accuracy.

Shiju Sathyadevan and Devan M.S (2014) have explained an approach between computer science and criminal justice to develop a data mining procedure that helps solving crimes faster by focusing mainly on crime factors of each day.

Chun-long Yao,Cui-Cui Sun,Xu Li and Kejun Lee (2014) have said that Classification analysis can find the factors affecting the crime and help the police officers to strengthen crime preventions.

III. METHODOLOGY & TOOLS

In our study, the data mining technique used is Classification algorithm (J48 Decision tree algorithm). The J48 algorithm is used to predict the most common cybercrime threat that affects the respondents. In this work, the respondents are categorized in to different groups based on their age. The data analysis helps in better understanding of large set of data.

3.1 Classification Algorithm (J48 Decision tree algorithm) in WEKA

In WEKA tool, select the Explorer in Applications panel in WEKA GUI Chooser and open WEKA Explorer. Open the saved .csv file and convert it in to .arff file and then preprocess the file. Then select the Classify tab and click Choose button. From classifier models, choose tree and then click J48 algorithm ; choose full training set as test option and then click the start button to run the algorithm. Thus the classifier model builds J48 pruned tree.

3.2 Cross Tabulation using SPSS

SPSS is capable of handling large amounts of data and can perform all of the statistical analyses. The SPSS GUI has two views namely Data view and Variable view which can be toggled by clicking on the tab in the bottom of the left in SPSS window. The Data view shows the spread sheet view of the cases (rows) and the variable view shows variables(columns). The variable view displays the metadata dictionary each row represents a variable and shows variable name, variable label, measurement type and variety of other characteristics.

IV. DATA ANALYSIS

The questionnaire has been designed to collect the impacts of common cyber crime among the various respondents in and around Kanchipuram. The purpose of this work is : (i) To find which age group of respondents are affected by the impacts of cyber crime. This paper presents how the collected data are analyzed through appropriate data mining techniques and the results of data analysis.

4.1 Population and sample

The author collected 135 samples from the data among which there are 63 male respondents and 72 female respondents.

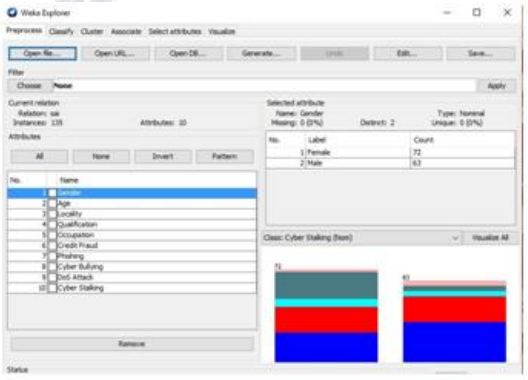

Fig 4.1: Screenshot for Sample in WEKA Explorer.

The author collected 135 samples from the data among which there are 66 respondents of age group between 18 and 30; 33 respondents of age group between 31 and 45; 21 respondents of age group between 46 and 60; 10 respondents below 18 and 5 respondents of above 60.

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Special Issue
National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

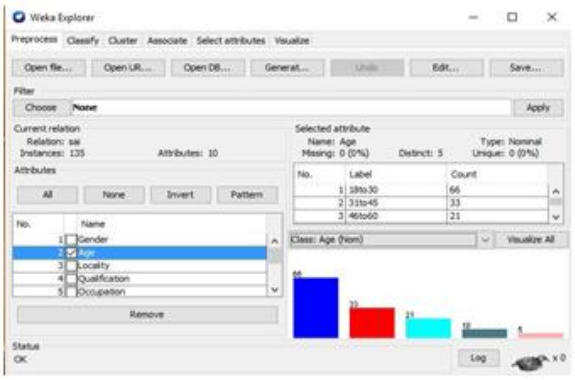

Fig 4.2: Screenshot for Respondents' age group in WEKA Explorer

4.2 Hypothesis tested

4.2.1 Research Hypothesis (H1)

There is an association between the respondents' age and the common cyber crime attacks on the respondents.

Null Hypothesis (H0)

There is no association between the respondents' age and the common cyber crime attacks on the respondents. Classification Algorithm (J48 Decision tree algorithm) in WEKA for Prediction

For classification problems involving prediction, the more powerful approach is Decision tree. This precedes building a tree and applying the tree to the dataset. J48 (modification of which uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. On training data, this technique provides balance of flexibility and maximum accuracy. To know the efficiency of prediction system, we have used J48 algorithm.

The confusion matrix showed the different cyber crime threats affecting the various age group of respondents. Here the prediction is done by the attributes credit/debit card fraud, phishing, cyber bullying, DoS attacks, Cyber stalking and the predicted age result is described below. The classification on the test data is done based on the decision tree that is created.

The confusion matrix (contingency table) has five classes in our study, and so a 5*5 confusion matrix. In the matrix, the sum of diagonals is the number of correctly classified instances; all others are incorrectly classified instances.

=== Confusion Matrix ===

a b c d e <-- classified as

52 2 0 0 0 | a = 18-30

5 35 0 0 0 | b = 31-45

2 0 8 0 0 | c = Below 18

2 0 0 23 0 | d = 45-60

0 2 0 0 4 | e = Above 60

The correctly classified instances are 122(90.3704%) which is the sum of diagonals of confusion matrix (52+35+8+23+4) and the incorrectly classified instances are 13(9.6296%).

Table 1: Prediction based on respondents' age

		Predicted					%of correctly predicted
		a	b	c	D	e	
Observed	18-30	52 (54)	2	0	0	0	96.29
	31-45	5	35 (40)	0	0	0	87.5
	Below 18	2	0	8 (10)	0	0	80.0
	45-60	2	0	0	23 (25)	0	92.0
	Above 60	0	2	0	0	4 (6)	66.66

Among 122 correctly classified instances, 96.29% of respondents of age group between 18 and 30 strongly agree; 87.5% of respondents of age group between 31 and 45 agree; 80.0% of respondents of age below are neutral; 92.0% of respondents of age group between 45 and 60 strongly disagree and 66.66% of respondents of age above 60 disagree.

Table 2: Respondents' age group Vs Phishing

Respondents' age group Vs Phishing		Phishing					Total
		SA	A	N	SD	D	
Age (in years)	<18	7	1	1	0	0	9
	18-30	23	31	5	0	8	67
	31-45	6	19	3	2	3	33
	46-60	7	12	2	0	0	21
	>60	1	3	0	0	1	5
Total		44	66	11	2	12	135

From the above cross tabulation, receiving spoofed email that contains links to fake websites (Phishing) is the most common cyber crime threat among respondents.

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Special Issue
National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

Among 9 respondents of age below 18 have high influencing percent is 7 of them strongly agree. Among 67 respondents of age group between 18 and 30 have high influencing percent is 31 of them agree. Among 33 respondents of age group between 31 and 45 have high influencing percent is 19 of them agree. Among 5 respondents of age above 60 have the high influencing percent is 3 of them agree.

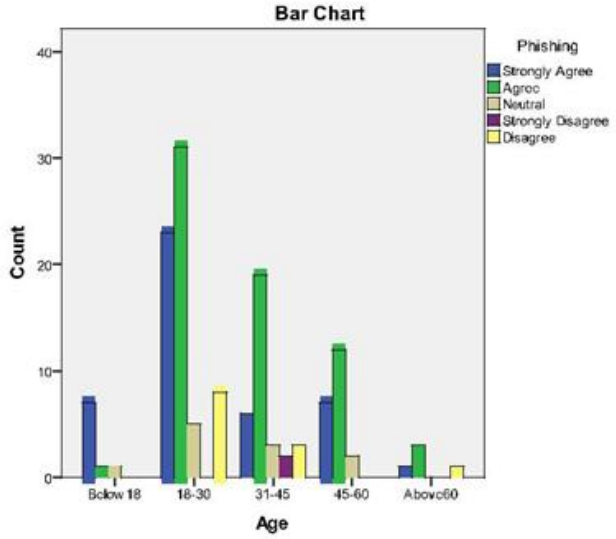

Fig 4.3: Screenshot for Respondents' age Vs Phishing

Table 3: Outcome of study Respondents' age Vs Phishing

	Description	Outcome of study
Age	To find which age group of respondents are affected by common cyber crime threats	Positive
Phishing	Respondents of age group between 18 and 30	Positive
	Respondents of age group above 60	Negative

V. CONCLUSION

Cyber crimes are varying in its nature due to enhancement in technologies. Due to diversified nature it is difficult to identify the cyber security problems which leads to unawareness on security issues. The WEKA tool is used for the study implementation since it contains a collection of visualization tools and algorithms for data analysis. The respondents' age group who are most affected by common cyber crime threats are predicted using classification algorithm (J48 decision tree algorithm). The some of common cyber crime threats are Credit/Debit card theft, Phishing, Cyber bullying, Denial of service attacks, Cyber stalking. The J48 decision tree algorithm shows the result obtained is most common cyber threat is receiving spoofed email that contains links to fake websites (Phishing). Our prediction analysis has been categorized in two different views based on respondents' age group. The respondents of age group between 18 and 30 are affected mostly by common cybercrime threats that is by phishing. Protecting our email address and not responding to the messages from spammers and phishers is one of the best way to avoid such crimes. Because responding to those messages or even downloading images may add us to their lists.

REFERENCES

[1] KPMG INTERNATIONAL Issues Monitor "Cyber Crime-A Growing Challenge for Governments, Vol.8, July 2011.

[2]Vineet Kandpal and R.K. singh, "Latst Face of Cybercrime and its Prevention in India", International Journal of Basic and
Applied Sciences, Vol.2 No.4, 2013.

[3]K.Sridharan and Saktheeswari(2013), A Case study on Cyber crime in India, International Journal of Power Control Signal and Computation, Vol. 4, Nov.2013.

[4]Dr. Ajeet S. Poonia , Cyber Crime: Challenges and its Classification, International Journal of Emerging Trends & Technology in Computer Science, ISSN 2278-6856, Volume 3, Issue 6, Nov-Dec 2014.

[5]Janhavi J Deshmukh and Surabhi R Chaudhari(2014), Cyber crime in Indian scenario-a literature snapshot, International Journal of Conceptions on Computing and Information Technology, ISSN:235-9808,Vol.2, April 2014.

[6]P. Arokia Vasantha Rani(2015), Cyber Crime – An Overview of Security Measures, Seclusion Fortification and

Suggestions, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN:2277 12BX, Vol. 5, Issue 3, March 2015.

[7]K. Chitra Lekha and S. Prakasam, An analysis on Finding the Influencing factors of supporting for the "GiveitUp" LPG

Subsidy for the Government using Data Mining Techniques, International Journal of Computer Applications, ISSN: 0975-8887),Vol. 143-No.5, June 2016.

[8]Malathi. A and Dr. S. Santhosh Baboo, An Enhanced Algorithm to Predict a Future crime using Data Mining techniques, International Journal of Computer Applications, ISSN: 0975-8887, Vol.21-No.1, May 2011.

[9]S.Yamuna and N.Sudha Bhuvaneswari, Data mining techniques to Analyze and Predict Crimes, The International Journal of Engineering and Science, ISSN:2319-1813, Vol.1, Issue 2, 2012.

[10]Arunima S. Kumar and Raju K. Gopal, Data mining based Crime Investigation Systems: Taxonomy and Relevance, Proceedings of 2015 Global Conference on Communication Technologies, 2015.

[11]Mohammad Reza Keyvanpour, Mostafa Javideh, Mohammad Reza Ebrahimi, "Detecting and investigating crime by means of data mining: a general crime matching framework, Elsevier, ISSN: 1877-0509, 2010.

[12]Deepti Gaur and Neha Aggarwal, Cybercrime Analysis and Data mining Methodologies, International Journal on Advanced Computer Theory and Engineering, ISSN: 2319-2526, Vol.3, Issue.4, 2014.

[13]K.K. Sindhu and B.B. Meshram, Digital Forensics and Cyber crime Data mining, Journal of Information Security,2012.

[14]Shijiu Sathyadevan and Devan M.S, Crime analysis and Prediction using Data mining, First International Conference on Networks & Soft Computing, 2014.

[15]Dr. B. Muthukumaran, Cyber crime Scenario in India, Criminal Investigation Department Review, January 2008.

[16]Chun-long Yao,Cui-Cui Sun,Xu Li, Kejun Lee, Detecting Crime types using Classification Algorithms, Journal of Digital Information Management, Vol.12, No.5, October 2014.