RIGA TECHNICAL UNIVERSITY

Information Technology and Management Science

ISSN 2255-9094 (online)
ISSN 2255-9086 (print)
December 2018, vol. 21, pp. 45–48
doi: 10.7250/itms-2018-0006
https://itms-journals.rtu.lv

Application of Machine Learning Classification Algorithm to Cybersecurity Awareness

Shilpa Balan[1], Sanchita Gawand[2], Priyanka Purushu[3]
[1–3]*California State University, Los Angeles, USA*

***Abstract* – Cybersecurity plays a vital role in protecting the privacy and data of people. In the recent times, there have been several issues relating to cyber fraud, data breach and cyber theft. Many people in the United States have been a victim of identity theft. Thus, understanding of cybersecurity plays an important role in protecting their information and devices. As the adoption of smart devices and social networking are increasing, cybersecurity awareness needs to be spread. The research aims at building a classification machine learning algorithm to determine the awareness of cybersecurity by the common masses in the United States. We were able to attain a good F-measure score when evaluating the performance of the classification model built for this study.**

***Keywords* – Big data, cybersecurity, classification, machine learning.**

I. Introduction

In today's digital age where every smart device and computer is connected to the internet, attention to cybersecurity has become of vital importance. However, people fail to understand its implications in daily life, for instance, in an online shopping transaction [1].

According to a survey conducted by Pew Research Center [2], it was found that many individuals in the United States are unclear about some key cybersecurity topics, terms and concepts. The need to increase measures in cybersecurity has led to an increased emphasis on cybersecurity awareness. This paper examines the extent of cybersecurity awareness among the Americans. In the paper, we attempt to build a classification model to determine how well American users are aware of cybersecurity concepts, such as phishing attack, authentication and others. Phishing is defined as "an act of attempting a victim to fraudulently acquire sensitive information by impersonating a trustworthy third party, which could be a person or a reputed business in an electronic communication" [3, p. 700]. Phishing attack includes tricking a user into revealing sensitive information, such as bank account numbers, passwords and credit card details.

In the survey conducted by Pew Research Center [2], a survey question presented to the internet users is on awareness of authentication. Users are surveyed to determine their understanding of the importance to employ a 'two-factor' or 'multi-factor' authentication on any account where it is available. Two-factor authentication generally requires users to log in to a site with credentials that the user knows. Only 10 % of online adults were able to correctly identify the example of a multi-factor authentication process. Approximately half of the internet users were able to correctly answer several questions in the survey on cybersecurity awareness. The survey consisted of other questions on awareness of phishing attacks, private browsing and internet accessibility. 54 % of the survey users were able to identify examples of phishing attacks. It is seen that a users' knowledge of cybersecurity varies by educational attainment. More than half of the questions in the survey were answered correctly with those having college degrees or higher. About 27 % of those with college degrees answered more questions correctly [2].

In this paper, we apply the classification algorithm of machine learning to build a model on the survey data. Machine learning is the practice of using algorithms to parse data, learn from it, and to then make a determination about something in the world [4].

The model is an aim to improve the existing literature on machine learning and classification. Using the classification model, we determine how many instances for each of the attributes of the data are classified correctly.

In the case of cybersecurity, machine learns, helps better analyse previous cyber-attacks and develop respective defense responses. This approach enables an automated cyber defense system with a minimum-skilled cybersecurity force [5].

In the present paper, we use Weka, an open source machine learning software, to analyse the data on cybersecurity awareness. WEKA (Waikato Environment for Knowledge Analysis) is an open source library for machine learning and is a powerful tool that can implement machine learning algorithms [6].

The background section further details the increasing types of cyber-attacks and the importance of cybersecurity.

II. Background

The average American's understanding about cybersecurity plays a vital role in protecting their information and devices at home and work. It has been seen that several individuals in America have reported a victim of identity thefts and nearly half of them have been a victim of some sort of a phishing attack [2]. In the United States, internet users are less aware of recognising phishing links or determining an encrypted website [7].

A large number of US personnel allow family members and trusted friends to check and reply to their email, and view posts on social media. Many internet users in the United States assume that a hotel or a tourist venue indicates a trusted hotspot with secured network [7]. A typical scenario of a cybercrime is the hacking of a credit card [7]. It is crucial to be aware of cybersecurity attacks and vulnerabilities. Many common US employees are unaware of security threats and how to combat them [7].

A. Cybersecurity Crimes in Recent US History

Equifax is one of the three major credit bureaus in the United States, along with Experian and TransUnion. It is a private company that collects financial information on consumers. Regarding a recent breach in Equifax, the hackers had access to information of the customers' database and managed to steal information pertaining to names, social security numbers, birth dates and addresses for up to 143 million people. The hackers were also able to steal driver's license numbers and credit card numbers for 209,000 U.S. consumers. This impacted several people in the United States [8].

A similar cyber-attack occurred in the year 2015 at Anthem. Anthem is the second largest health insurance company in the United States. The attacker was able to obtain personal information about the clients, such as names, social security numbers, birth dates, addresses and employment information. It was found that the breach had impacted many Anthem product lines. This was reported to be the largest data breach in the health care history (Ragan, Feb 4, 2015) [9].

In December 2013, Target Stores reported that credit/debit card information and contact information of several millions of people were compromised. The attackers got access to data stored on the magnetic stripe on the back of the credit and debit cards through card swiping devices [10].

All of these examples suggest the importance of cybersecurity awareness. In this paper, we attempt to demonstrate an application of a supervised learning algorithm, the classification algorithm to train the data to determine cybersecurity awareness among internet users in America. While writing this paper, we did not find earlier applications of classification algorithm on cyber security. However, there are several applications of classification algorithm in healthcare and other fields. In the previous research, an example of a classification algorithm was on anomaly detection. The classification algorithm was approached using both back propagation rules and fuzzy classifier rules [11].

The Methodology Section describes the data we used for analysis and the machine learning classifier algorithm we applied in the present study.

III. Methodology

A. Data

We analysed a survey data published by the Pew Research Center in the year 2016 [2]. The survey was conducted among adult internet users living in the United States.

The goal of the research is to build a classification algorithm to classify instances of each of the survey question on cybersecurity. The survey attributes selected are the following: 1) awareness of authentication; 2) awareness of private browsing; 3) awareness of phishing attack; 4) awareness of internet accessibility.

Each of these attributes had 3 to 5 multiple answer choices for a survey user to choose from. For example, the survey question on phishing attacks is shown in Table I [2].

For the survey question on awareness of phishing attack in Table 1, option d was the correct answer choice. From the previous research, it was found that 54 % of users were aware of phishing [2].

The survey question on the two-step authentication required the user to select the correct images from the given answer options.

TABLE I

Survey on Awareness of Phishing Attack

Serial No.	Survey Answer Options
a	Sending someone an email that contains a malicious link that is disguised to look like an email from someone the person knows
b	Creating a fake website that looks nearly identical to a real website, in order to trick users into entering their login information
c	Sending someone a text message that contains a malicious link that is disguised to look like a notification that the person has won a contest
d	All of the above [S]
e	Not sure

(Source: What the Public knows about Cybersecurity, Pew Research Center, Washington D.C., 2017. http://www.pewinternet.org/2017/03/22/what-the-public-knows-about-cybersecurity/)

Further, awareness of private browsing included the following survey question: "Private Browsing is a feature in many internet browsers that lets users access web pages without any information (like browsing history) being stored by the browser. Can internet service providers see the online activities of their subscribers when those subscribers are using private browsing?" (Pew Research Center, Olmstead, 2017) [2].

a) Yes
b) No
c) Not sure

The survey question on internet accessibility included the following survey question: (Pew Research Center, Olmstead, 2017, [2]) Do you access the internet on a cell phone, tablet or other mobile handheld device, at least occasionally?

a) Yes
b) No

B. Machine Learning Classifier Algorithm

We used Weka, an open source data mining software, to build a classification model. Weka is a collection of machine learning algorithms for data mining tasks [12]. It contains tools for data preparation, classification, regression, clustering, and association rule mining [12].

Figure 1 shows the architecture of the methodology used. After we input the data in Weka, we performed data pre-processing that included setting the appropriate data type and the class attribute. The attributes we selected from the data are the following: 1) awareness of authentication; 2) awareness of private browsing; 3) awareness of phishing attack; 4) awareness of internet accessibility. We then executed the LMT (Logistic Model Trees) classification algorithm and set the classifier rules. We tested with training data by splitting the data into training data as 66 % and remaining – for test data. We re-tested by splitting the data into training data as 80 % and 20 % – for test data. The step on data pre-processing is repeated for the different percentage split of training and test data. Table II in the Results Section gives the correctly classified instances using the classification algorithm.

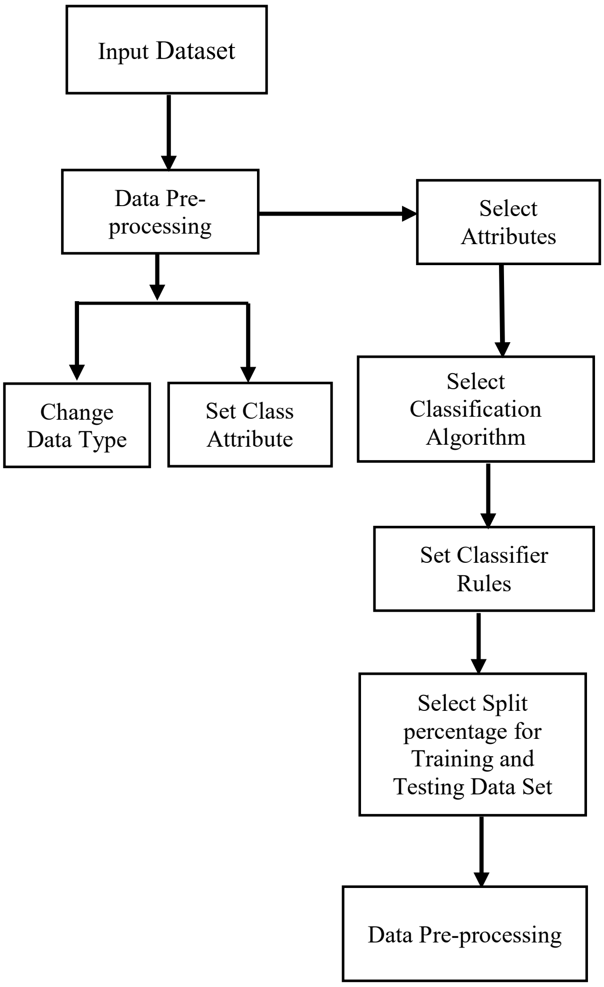

Fig. 1. Architecture diagram.

In this paper, we present a classification method, called LMT. A Logistic Model Tree is an algorithm for supervised learning tasks that is combined with linear logistic regression and tree induction [13]. We evaluate the performance of LMT on the dataset containing the survey records. LMT [13] is a recent addition to decision trees. A logistic model tree consists of a decision tree structure with logistic regression functions at the leaves. This has the benefit of producing decision trees that have higher accuracy than previous decision tree extraction algorithms. Our selected attributes from the data set were of type nominal attributes. For a nominal attribute with k values, the node has k child nodes.

A logistic model tree is made up of a set of non-terminal nodes N and a set of leaves or terminal nodes T [13]. If S denotes the instance space, spanned by all attributes that are present in the data, then the tree structure gives a disjoint subdivision into regions St, and every region is represented by a leaf in the tree [13].

$$S = \bigcup_{t\in T} S_t, S_t \cap S_{t'} = \emptyset \text{ for } t \neq t'$$

Unlike usual decision trees, the leaves t ∈ T have an associated logistic regression function f_t [13]. The model represented by the whole logistic model tree is given by:

$$\mathrm{f(x)} = \sum_{t\in T} f_t\ (\mathrm{x}) \cdot \mathrm{I}\ (\mathrm{x} \in S_t)$$

IV. Results and Analysis

Table II shows the analysis using the Logistic Model Tree with 66 % split for training data. We attained similar results with 80 % split for training data. The experiments show that LMT produces more accurate classifiers. Table II shows the correctly classified instances.

Table II shows that the correctly classified instances were 81.79 % for the attribute on awareness of authentication, and about 99 % for the other attributes on awareness of private browsing, phishing attack and internet accessibility.

Precision is the fraction of relevant instances among the retrieved instances. Recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. The F-measure is computed as follows:

$$\mathrm{F}=2\cdot\frac{precision\cdot recall}{precision+recall}$$

It is further seen that the F-measure was high for each of the four selected attributes. The F-measure for the attribute on the awareness of authentication is 70 %. In comparison, the F-measure was slightly lower in comparison to the other selected attributes.

TABLE II

Results of LMT Algorithm

	Authentication	Private browsing	Phishing attack	Internet accessibility
Correctly classified instances	81.79 %	99.1 %	99 %	99 %
Mean absolute error	0.41	0.18	0.12	0.11
Root mean squared error	0.44	0.19	0.14	0.11
Recall	0.71	0.99	0.99	0.99
Precision	0.69	0.98	0.98	0.99
F-measure	0.70	0.99	0.98	0.99

V. Limitations

For this analysis, we used the public data set published by the Pew Research Center. Due to the nature of the data collected in this survey and the type of data collected, we were unable to combine all of the selected attributes together to make a more accurate classification. Each of the survey questions collected from the user was independent from the other. For example, we were unable to create a strong model to determine that awareness of phishing attack was dependent on the awareness of authentication or any of the other attributes selected.

It has already been previously determined by the Pew Research Center that awareness of phishing attack, awareness of authentication, and awareness of private browsing are determined by education and age [2].

VI. Conclusion and Future Research

While the field of cybercrimes and cybersecurity is very popular, open data sets are still very limited for an analysis in this field of study [14].

A new survey with cybersecurity terms should be conducted in order to determine a more accurate prediction of the awareness of cybersecurity terms by users in America. For future, a more elaborated survey can be administered such that one survey question is built upon the other. However, since the younger generation is more aware of cybersecurity terms, it can be predicted that with the new generations to come, more Americans will be aware of cybersecurity terminology.

References

[1] R. von Solms and J. van Niekerk, “From Information Security to Cyber Security,” *Computers & Security*, vol. 38, pp. 97–102, Oct. 2013. https://doi.org/10.1016/j.cose.2013.04.004

[2] K. Olmstead, A. Smith, “What the Public knows about Cybersecurity”, Pew Research Center, Washington, D.C., 2017 [Online]. Available: http://www.pewinternet.org/2017/03/22/what-the-public-knows-about-cybersecurity/

[3] R. Damodaram, “Study on Phishing Attacks and Phishing Tools”, *International Research Journal of Engineering and Technology*, vol. 3, no. 1, pp. 700–705, 2016.

[4] D. Fagella, “What is Machine learning”, 2018 [Online]. Available: https://www.techemergence.com/what-is-machine-learning/

[5] R. Koppula, “Applications of machine learning in cyber security”, 2018 [Online]. Available: https://apiumhub.com/tech-blog-barcelona/applications-machine-learning-cyber-security/

[6] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, S. Cunningham, “Weka: Practical Machine Learning Tools and Techniques with Java Implementations”, *ICONIP/ANZIIS/ANNES*, 2002.

[7] Wombat Security, “Wombat Study Reveals Personal Security Habits of 2,000 US, UK Workers”, 2017 [Online]. Available: https://www.wombatsecurity.com/press/press-releases/wombat-study-reveals-personal-security-habits-2000-us-uk-workers

[8] Crediful, “Equifax Data Breach of 143 Million People: What it means for You”, 2017 [Online]. Available: https://www.crediful.com/equifax-data-breach/

[9] S. Ragan, “Anthem confirms data breach, but full extent remains unknown”, 2015 [Online]. Available: https://www.csoonline.com/article/2880352/disaster-recovery/anthem-confirms-data-breach-but-full-extent-remains-unknown.html

[10] J. Ribeiro, “Target customers' card data said to be at risk after store thefts”, 2013 [Online]. Available: https://www.csoonline.com/article/2134248/data-protection/target-customers--39--card-data-said-to-be-at-risk-after-store-thefts.html

[11] Gonzalez, F., Dasgupta, D., Kozma, R. (2002). Combining Negative Selection and Classification Techniques for Anomaly Detection. *Proceedings of the 2002 Congress on Evolutionary Computation*, *IEEE*, pp. 705–710. https://doi.org/10.1109/CEC.2002.1007012

[12] Weka, “Weka: Data Mining Software in Java”, 2018 [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/

[13] N. Landwehr, M. Hall, E. Frank, “Logistic Model Trees”, *Machine Learning*, vol. 59, pp. 161–205, 2005. https://doi.org/10.1007/s10994-005-0466-3

[14] S. Balan, J. Otto, E. Minasian, A. Aryal, “Data Analysis of Cybercrimes in Businesses”, *Information Technology and Management Science*, vol. 20, no. 1, pp. 64–68, 2017. https://doi.org/10.1515/itms-2017-0011

Dr. Shilpa Balan is an Assistant Professor at the Department of Information Systems, California State University-Los Angeles. Her research interests are in big data analytics, business intelligence, machine learning and healthcare informatics.
E-mail: sbalan@calstatela.edu
ORCID iD: https://orcid.org/0000-0002-3582-2560

Sanchita Gawand is a graduate student at the Department of Information Systems, California State University-Los Angeles. Her research interests are in business intelligence and analytics.
E-mail: sgawand@calstatela.edu
ORCID iD: https://orcid.org/0000-0003-3954-0189

Priyanka Purushu is a graduate student at the Department of Information Systems, California State University-Los Angeles. Her research interests are in business analytics and machine learning.
ORCID iD: https://orcid.org/0000-0001-7985-0400