International Journal of Multidisciplinary Research and Development
www.allsubjectjournal.com
ISSN Online: 2349-4182 ISSN Print: 2349-5979
Received: 06-10-2023, Accepted: 22-10-2023, Published: 07-11-2023
Volume 10, Issue 11, 2023, Page No. 20-23

Predicting cyber crime issues in digital environment using datamining techniques

R Sujatha[1], B Navaneetha[2]
[1] Assistant Professor, Department of Computer Science, PSG College of Arts & Science, Coimbatore, Tamil Nadu, India
[2] Assistant Professor, Department of Commerce (PA), PSG College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract
Cyber-attacks have become one of the biggest problems of the world. They cause serious financial damages to countries and people every day. The increase in cyber-attacks also brings along cyber-crime. The key factors in the fight against crime and criminals are identifying the perpetrators of cyber-crime and understanding the methods of attack. Detecting and avoiding cyber-attacks are difficult tasks. However, researchers have recently been solving these problems by developing security models and making predictions through artificial intelligence methods. A high number of methods of crime prediction are available in the literature. On the other hand, they suffer from a deficiency in predicting cyber-crime and cyber-attack methods. This problem can be tackled by identifying an attack and the perpetrator of such attack, using actual data. The data include the type of crime, gender of perpetrator, damage and methods of attack. The data can be acquired from the applications of the persons who were exposed to cyber-attacks to the forensic units.
In this paper, we analyze cyber-crimes in two different models with machine-learning methods and predict the effect of the defined features on the detection of the cyber-attack method and the perpetrator. We used eight machine-learning methods in our approach and concluded that their accuracy ratios were close. The Support Vector Machine Linear was found out to be the most successful in the cyber-attack method, with an accuracy rate of 95.02%. In the first model, we could predict the types of attacks that the victims were likely to be exposed to with a high accuracy. The Logistic Regression was the leading method in detecting attackers with an accuracy rate of 65.42%. In the second model, we predicted whether the perpetrators could be identified by comparing their characteristics. Our results have revealed that the probability of cyber-attack decreases as the education and income level of victim increases. We believe that cyber-crime units will use the proposed model. It will also facilitate the detection of cyber-attacks and make the fight against these attacks easier and more effective. In this proposed work Enhanced Fuzzy C-Means Clustering Algorithm is utilized for the datasets to predict the crime.

Keywords: Machine learning, clustering, support vector machine, logistic regression

Introduction

In today's world the challenges of crime prevention in Digital Environment are increasingly complex because of the huge amount of confidential data generated through different sources in internet. Criminal formations and various intelligent methods in illegal business use the latest technologies to the full extent for money laundering, distributing of false information, unauthorized access to information systems and other violations. The advent of e-technology has brought variety of chances to illegal opportunities.

The automation of many sectors like banks, educational institution, and railway reservation created many chances to digital crime. Cybercrime is an important issue for research as it affects many mainstream sectors such as defense, social media, government, industry, private, military and scientific sectors etc. Internet criminals use distorted or hacked data to capture their actions.

Cybercrime consists of online identity theft, financial fraud, stalking, bullying, hacking, email spoofing, information piracy and forgery and intellectual property crime. Cybercrime can lead to financial ruin and potentially threaten a victim's reputation and personal safety. In such matters, the complainant alleges that some unknown person had withdrawn money/ made transactions through his/her credit/debit cards through online purchasing. In most of these cases purchasing is done by using following crucial information of the credit/debit card.

Types of computer fraud include

- Distributing hoax emails
- Accessing unauthorized computers
- Hacking into computer systems to illegally access personal information, such as credit cards or Social Security numbers
- Sending computer viruses or worms with the intent to destroy or ruin another party's computer or system.

There are many words used to describe fraud: Scam, con, swindle, extortion, sham, double-cross, hoax, cheat, ploy, ruse, hoodwink, confidence trick. Thus in our project we use the various intelligent methods from data mining to predict and analyze the Cybercrime issues in digital environment and R Programming tool is utilized to analyze the result. This creates awareness towards Social Responsibility and Environmental Consciousness.

Literature Survey

The importance of the fight against such cyber-attacks, cyber-crimes and cyber security is highlighted in various studies. Cyber security is the protection of physical-digital data, networks, and technological systems from cyber-attacks, unauthorized accesses disruptions, modifications, destructions and damages through various processes, applications and applied technologies (Fischer, 2009).

Cyber-attacks such as distributed denial of service attacks by sending malicious packets (Kaur Chahal, Bhandari &

Behal 2019, phishing attacks to banking and shopping sites that deceive the user (Sahingoz *et al*., 2019) have increased significantly. In addition, attackers have been using malicious attack software (virus, worms, trojans, spyware and ransomware) that is installed into the user's computer without any consent of the user (Biju, Gopal & Prakash, 2019) increasingly. Again, the most common of these attacks and one of the attacks that are most difficult to be prevented is the social engineering attacks. They are based on technical skill, cunning and persuasion, made by taking advantage of the weakness of the victim.

Kevin Mitnick, one of the world's famous hackers in social engineering attacks, penetrated most systems he attacked with this method (Mitnick & Simon, 2009). In the work by Breda, Barbosa & Morais (2017) this attack is mentioned as one of the biggest security vulnerabilities in the system no matter how secure a technical system is. Likewise, attacks against IoT devices, which have increased rapidly in recent years, affect the society considerably. Thus, attacks and threats to the IoT structure should be understood for security purposes (Kagita *et al*., 2020). Studies conducted to understand and cyber-attacks reveal the importance of crime prediction as discussed in this study.

The attacks described above are defined as prohibited criminal acts within the legal framework of many governments. The duty of fighting against crime and criminals is given to law enforcement departments. Researchers assist the institutions conducting the investigation with various analysis and prediction methods.

For example, big data (Rewari & Singh, 2017) and machine-learning (Lin, Chen & Yu, 2017) methods have been used to analyze crimes in many studies. They have contributed to crime and crime fighting institutions with artificial intelligence models. Among these are determining the areas where the crime can be committed and its story (Hassan & Rahman, 2017), predicting the crime using spatial, temporal and demographic data (Zhao & Tang, 2017), and analyzing crime with literacy, unemployment and development index data (Vineeth, Pandey & Pradhan, 2016).

Time series of crime data in San Francisco, Chicago and Philadelphia were used for predicting crimes in the following years. Decision Tree (DT) classification model performed better than K-nearest neighbors (KNN) and Naive Bayes (NB) (Feng *et al*., 2018).

Proposed Work

The objective of this research is to predict and examine various crime issues in digital environment. The Predictions of crime issues in digital environment focuses and forecasts to decrease the Crime rate in the society and create awareness towards Social Responsibility and Environmental Consciousness. To predict the issues various intelligent methods from Datamining is utilized. The analysis is done using R Programming and statistical tools.

The following are the few objectives of the study

1. To understand the socio-economic profile of the respondents
2. To analyze the awareness of cybercrime among the respondents
3. To explore the issues faced by respondents in digital environment
4. To predict and provide the analysis to the respondents

Significance of study

In this digitalized environment all the industries utilizes the Internet for their working environment. By every second huge amount of data is generated in internet through social media, public profiles, and by using more applications in mobile environment like gpay, foodie applications (zomato), banking application and various ticket reservation application and websites. Hence handling the huge volume of data and providing privacy to our data is a big issue. Due to this theft of data, various illegal cybercrime issues in digital environment increases. Fraud rate and crime rate is increased due to technological manipulations, hence our research focuses to create an awareness to the society and to predict the crime issues in digital environment by applying various intelligent methods from data mining, tools for statistical methods are utilized for predicting and analyzing the datasets.

Methodology

The present study is based on both primary and secondary data. The primary data will be collected from the questionnaires.

Step 1: The Dataset is collected

Step 2: Preprocessing is applied to the dataset.

Step 3: Clustering is applied to group the dataset

Step 4: Enhanced Fuzzy C-Means Clustering Algorithm is used to predict the issues in cybercrime from digital environment.

Step 5: Accuracy is found by comparing with various benchmark algorithms.

Step 6: Based on the result the analysis is done.

Framework of Analysis

An intelligent method (algorithms) from Datamining is applied to predict the dataset and the statistical tools namely R Programming & SPSS is applied for analyzing the dataset.

1. Enhanced Fuzzy C-Means Clustering Algorithm

Enhanced Fuzzy C-Means Clustering (EFCM) Algorithm is a data clustering algorithm in which every data point has a place with a group to a degree validated by a participation grade. EFCM uses fuzzy assigning to such a level that a given data point can have a multiple clusters, which falls in the level either 0 or 1. Fuzzy C Means uses a technique to separate the cluster. The data points are permitted to have fragments with values somewhere in the range of 0 and 1. Regardless, the degree of all data points belongings of a particular data highlight all groups equally.

Fuzzy C-Means clustering incorporates two cycles: the count of collecting centers and the task of assigning centers using a sort of Euclidian distance. This cycle is repeated until the grouping is balanced out. The estimation resembles K-Means clustering from different perspectives and joins fuzzy set's thoughts the mid data points.

The algorithm needs a fuzzification boundary m in the reach $^{[1, n]}$ which decides the level of fuzziness in the clusters. At the point when m arrives at the estimation of 1 the calculation works for a bigger estimations of m the covering of cluster is generally more. Iterations is done in EFCM based on the conditions. The point of EFCM is to discover multiple data points focuses (centroids) that limit a work.

Furthermore the algorithm restricts a data point in all the groupings and it should be compared with one another. This constraint is addressed by clarification 1.

$$\sum_{j1=1}^{p1} \mu_{j1}(x_{i1}) = 1 \quad (1)$$

The new cluster communities are determined with the fuzzy enrollment along with the condition.

$$c_i = \frac{\sum_i [\mu_{j1}(x_{i1})]^{mm} x_{i1}}{\sum_i [\mu_{j1}(x_{i1})]^{mm}} \quad (2)$$

where
Ci: is the center of the the cluster
Xi1: is the ith data point
µj1: the function which returns the membership value
mm: is the fuzzification parameter
We change the degree of fuzziness in xi1's current position and begin it by xi1. The data points gained is isolated by the whole of the fuzzified cooperation.
The algorithm figures the participation esteem µ1 with the below equation 3,

$$\mu_{j1}(x_{i1}) = \frac{\left(\frac{1}{d_{ji1}}\right)^{\frac{1}{mm-1}}}{\sum_{k=1}^{p1}\left(\frac{1}{d_{ki1}}\right) m^{\frac{1}{m-1}}} \quad (3)$$

where
µj(xi1): is the membership of xi1 in the j1th cluster
dji1: is the distance of xi1 in cluster ci1
mm: is the fuzzification parameter
p1: is the number of specified clusters
8dki1: is the distance of xi1 in cluster Ck1
Therefore the algorithm is,
Step 1. Initialize the membership matrix (µ1) that has restrictions in Equation 1.
Step 2. Calculate centroids (ci1) by using Equation 2
Step 3. Calculate difference between centroids and data points.
Step 4 Calculate a new µ using Equation 3.
Step 5. Go back to Step 2 unless the centroids are unvarying.

Results and Discussion

The proposed Enhanced Fuzzy C-Means (EFCM) algorithm improves the efficiency in predicting cybercrime issues based on the protocol. In this work cybercrime data are taken from validated dataset. The proposed research strategy is implemented using R Programming and SPSS.
The accuracy of the algorithm is compared with various existing techniques such as Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO). The proposed model calculates the Accuracy, Precision and Recall. These measures demonstrate that the proposed algorithm produces good accuracy of results.

1. Prediction accuracy

The accuracy is calculated using Enhanced Fuzzy C-Means (EFCM) algorithm.

Table 1: Comparison of Accuracy with other methods

Algorithm	ACCURACY Rate
ACO	80
PSO	88
EFCM	93

Table 1 The proposed model exhibits higher exactness compared to other existing techniques.

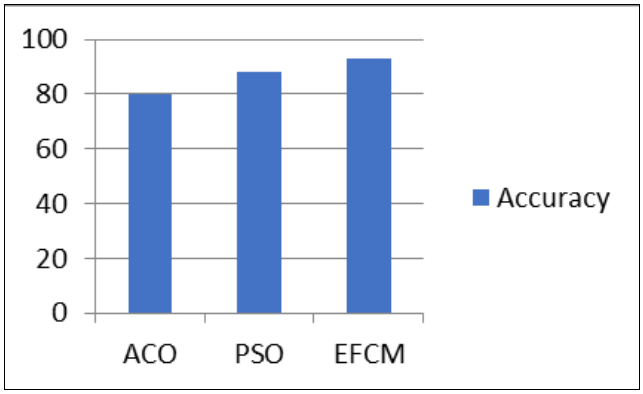

Fig 1: Comparison on prediction accuracy

Fig. 1. shows the examination of Enhanced Fuzzy C-Means (EFCM) comparison with various techniques and it proves that the accuracy is higher than other techniques.

2. Precision, Recall and Accuracy

Comparison on prediction method

From the above outcomes, it is evident that the proposed framework gives the preeminent outcome while anticipating the information. Assessment measures like precision and recall are examined underneath for efficiency assessment.

Table 2: Comparison of algorithms with various measures and accuracy

Algorithm	Precision	Recall	Accuracy
ACO	74	78	81
PSO	81	86	90
EFCM	92.5	94	93

Conclusion

The proposed model Enhanced Fuzzy C-Means (EFCM) algorithm groups the dataset to predict cyber crime attacks. This work discloses how to identify cyber crime issues in a successful strategy for the given dataset. Accuracy of the proposed algorithm is compared with various existing techniques such as ACO [10] and PSO [10] to prove the efficiency of Enhanced Fuzzy C-Mean (EFCM) algorithm. The dataset are taken from UCI Machine Learning Repository and kaggle. Implementation of the work proposed is done with R Programming and SPSS.

References

1. Adebowale MA, Lwin KT, Hossain MA. Deep learning with convolutional neural network and long short-term memory for phishing detection. In 2019 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2019, 1–8. IEEE.
2. Balakrishnan V, Khan S, Arabnia HR. Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security,2020:90:101710.
3. Basit A, Zafar M, Liu X, Javed AR, Jalil Z, Kifayat K. A comprehensive survey of ai-enabled phishing attacks detection techniques. Telecommunication Systems,2021:76:139–154.

4. Chen YC, Chen JL, Ma YW. Ai@ tss-intelligent technical support scam detection system. Journal of Information Security and Applications,2021:61:102921.
5. Dasgupta D, Akhtar Z, Sen S. Machine learning in cybersecurity: a comprehensive survey. The Journal of Defense Modeling and Simulation,2022:19:57–106.
6. Go JH, Jan T, Mohanty M, Patel OP, Puthal D, Prasad M. Visualization approach for malware classification with resnext. In 2020 IEEE Congress on Evolutionary Computation (CEC), 2020, 1–7. IEEE.
7. Hou Y, Wang H, Wang H. Identification of chinese dark jargons in telegram underground markets using context oriented and linguistic features. Information Processing & Management,2022:59:103033.
8. Jha S, Prashar D, Long HV, Taniar D. Recurrent neural network for detecting malware. computers & security,2020:99:102037.
9. Li Z, Chen J, Zhang J, Cheng X, Chen B. Detecting advanced persistent threat in edge computing via federated learning. In Security and Privacy in Digital Economy: First International Conference, SPDE 2020, Quzhou, China, October 30–November 1, Proceedings,2020:1:518–532. Springer.
10. Ravi V, Chaganti R, Alazab M. Recurrent deep learning-based feature fusion ensemble meta-classifier approach for intelligent network intrusion detection system. Computers and Electrical Engineering,2022:102:108156.
11. Zaman S, Iqbal MM. Tauqeer H, Shahzad M, Akbar G. Trustworthy communication channel for the iot sensor nodes using reinforcement learning. In 2022 International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE), 2022, 1–6. IEEE.