JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CYBER CRIME DETECTION

RAVEESHA N[1], DR. SANJAY KUMAR TIWARI[2]

[1]Research Scholar, Sunrise University, Alwar

[2]Associate Professor, Sunrise University, Alwar

ABSTRACT

Crimes when computers or other forms of electronic communication are used in any way as instruments of the crime, as means to conceal the crime, as part of the crime, or as a means to commit the crime—are collectively referred to as cybercrimes. Examining Detection Algorithms, Boat Algorithms, Selected Attribute Optimization, and Selected Attribute Optimization Applications The k-means clustering method is a common application of evolutionary algorithms. Here, we take in a parameter (k) that controls how many sections the data will be split into. The distance of an observation to the cluster mean is then used to classify the observation into one of the clusters.

KEYWORDS: Cyber Crime, Detection, Hacking, Algorithm and Cyber Stalking

INTRODUCTION

Concerns about privacy, national security, social decency, intellectual property rights, child safety, and the prevention, investigation, and punishment of cybercrime have been both aided and hindered by the expansion of the Internet. Victims in every corner of the world are often clueless about their rights online and the tools available to track down and apprehend the perpetrators of attacks like identity theft, financial loss, forgery, threats, sexual abuses, etc., even though these crimes are on the rise and affect everyone who uses the Internet. It's common knowledge that computers and the Internet have had a major influence on our culture, ushering in a new era of convenience and productivity.

It's also no secret that there's been an alarming uptick in crime because to the security holes made possible by these innovations. It is now more apparent than ever that conventional approaches to policing cybercrime, including detection, investigation, and punishment, are insufficient. For this reason, there is an immediate need for a preventive strategy, swift international coordination, and efficient public private partnerships to establish command over these criminals. For cybercriminals, a computer or network might be a tool for committing other crimes or the crime itself. Many unlawful activities are made feasible via the usage of global computer networks.

Crime rates and socioeconomic conditions tend to go hand in hand. No matter how much we try, we can never live in a world devoid of cybercrime. If we haven't been successful in reducing crime in the actual world, why should we think we can do it in the virtual one? The virtual world is more difficult to legally control since it is more abstract, permanent, and unchanging than the physical world. The nature, scope, and significance of crime, however, change during the course of a culture's development. In reality, there is no such thing as a crime-free neighborhood. As a result, criminal behavior is indicative of a society's values and norms. A culture's level of complexity may be inferred by looking at how closely its crime rate tracks that of the culture at large. If one is serious in understanding crime rates, they must first confirm all of the factors that affect and contribute to criminal activity in a given society.

The criminal justice system and its possible remedies must be understood by the political and economic structures of society. Equally crucial to consider the procedures put in place to prevent lawlessness in a specific community while analyzing the causes and consequences of crime. The introduction of new technology has not helped the state manage the problem, but rather created complex scenarios that are difficult to understand and much more difficult to deal with the laws that are now in place. The government, however, does not have the manpower or knowledge to adequately deal with modern criminality. The largest influence of computers on modern society has occurred in the last three to four decades. There has been a significant increase in understanding and friendship amongst persons of different socioeconomic backgrounds and cultural traditions.

LITERTURE REVIEW

Neha K Bhatt et al (2021) Patriarchal Indian culture has consistently and historically treated women terribly, and that hasn't changed even though we're well into the twenty-first century. 2 Statistics show a steady increase in crimes committed against women. In this day of rapid technological advancement, women still do not feel safe in cyberspace. Despite the fact that India has laws meant to protect women from cybercrime, the incidence of such crimes is increasing. It was essential that the court maintain the law and protect people's rights as outlined in the constitution. The court's principal role is to defend personal freedoms and provide relief to those who have been wronged. 3 This research aims to analyse the part that the Indian judicial system plays in the battle against cybercrime directed against women. Cybercrime is defined and surveyed in this article's introductory part. In addition, it gives a brief summary of cyber legislation in India and describes the many kinds of cybercrime that impact women in that country.

Sang Ni et.al., (2017) suggested the GCCG local generalisation approach for K-anonymity, which is based on clustering. The results showed that when compared to KACA and other classifiable local generalisation algorithms, GCCG performed better and lost less information. When compared to KACA, GCCG has a performance gain of 10 times and an information loss of just around 50%. In GCCG, we lost around a third as much information while maintaining the same level of speed. It was also shown that the parallel GCCG shows better performance with the help of 4-threads parallelization, and that it can accelerate enormous data sets by a factor of 10 with little loss of information.

MuhammetBaykara et.al (2018), the use of honeypots as a tool for intrusion detection and prevention systems was addressed and suggested. IDS was combined with the honeypot technique for analysing live data and running efficient processes. In addition, a cutting-edge hybrid honeypot system was developed by combining the best features of high- and low-interaction honeypots. Honeypots were found to be deployed on corporate networks with the intention of using virtualization technologies to reduce the setup, maintenance, and management overhead involved. The suggested and realised honeypot-based IDPS might provide an animated representation of server network activity in real time.

Khraisat et al (2019) Many of the suggested SIDS, KDD, and MD methods involve pattern matching to identify known attacks. It used pattern-matching techniques to find evidence of a breach in security in the past. That is, an alarm signal is generated whenever an intrusion signature matches a previously stored intrusion signature in the signature database. Logs from hosts were examined in SIDS to identify patterns of behaviour indicative of malware.

AshkaAshani et al (2018) better honeypots are offered here. They had very high-level interactions with the invasive system. Attackers had a realistic experience with this honeypot and were able to gather a wealth of data about their assaults, which may raise the difficulty of capturing the complete honeypot. The infrastructure and administration of this kind of honeypot was complex. A high-interactivity honeypot may help uncover previously hidden weaknesses. When it comes to so-called "zero-day assaults," honeypots of this kind are indispensable. Take the case of Honeynets, which are used in scientific studies.

DETECTION ALGORITHM

K-means Clustering Algorithm

One popular use of evolutionary algorithms is the k-means technique. Here, we take a parameter (k) as input and divide the information into that many groups. Then, we determine each observation's cluster based on its distance from the mean of that cluster. After a new generated mean has been determined, the process begins again. This is how the algorithm works:

The procedure begins by picking k random cluster centres ("means") to use as a starting point.

Second, the dataset utilises the ED of each point relative to the centre of each cluster to classify the point into one of many closed clusters.

The average of the nodes inside a cluster is used to determine a new centre for each cluster.

Steps 2 and 3 are performed once cluster convergence has commenced.

In general, when steps 2 and 3 are repeated, either no new observations affect the clusters or the changes do not noticeably alter the definition of the clusters (a phrase known as "convergence").

Decision Tree Algorithm

Decision trees may be a useful analytical technique for data classification. The ideal tree model starts with a careful selection of the most relevant association attributes. Important aspects of tree building include selecting test qualities and determining how to partition a sample set. Several decision tree systems use numerous methods to deal with such issues. Common algorithms include CART, ID3, C4.5, CHAID, etc., with each iteration improving upon the last. The final decision trees produced by C4.5 may be reused in other classification schemes. In this sense, C4.5 is likewise a statistical classifier. C4.5 uses an information entropy approach and a training data set that are functionally identical to ID3 in order to create its decision trees. In the training data, samples that have previously been labelled are represented by set S = s1,s2. The features of a given sample are represented by the elements of a vector, si, where si=x1, x2, ... To get a more in-depth understanding, we augment the training data set with a vector C = c1, c2, where c1, c2, denote the sample's class.

In C4.5, each node of the tree is in charge of picking a particular feature of the data, thereby dividing the sample set into subsets enriched in a single class. The increase in normalized data is a popular statistic used to assess the efficacy of attribute selection for data splitting (difference in entropy). If you have many valuable attributes, the one with the largest normalized information gain should be used as a decision criteria. Shorter sub lists are generated and used in the C4.5 procedure. Following the standard scenarios generated by the algorithm.

- The aforementioned cases all belong to the same category. A leaf node is then added to the branch of the decision tree that leads to the selected class.

- There is no additional value added by any of the features. The projected class value is used to create a decision node, which is then output by C4.5.

BOAT Algorithm

To keep a decision tree up-to-date while a training dataset changes, BOAT provides a scalable method. Instead of starting from scratch every time new training data becomes available, BOAT may update the existing tree. In contrast, the remaining steps of the decision tree approach include repeat database scans at successive tree nodes. With the BOAT approach, on the other hand, the whole tree is built in a single pass over the data. The remaining decision tree approaches no longer have the speed and scalability constraints.

When figuring out what to do, BOAT considers example D1 from the training set D, which could be kept in RAM. The bootstrapping technique is then used to swap out D1 and the decision trees ST1, ST2, ST3, Sn to generate a set of very small samples S1, S2, S3, Sn. STm is constructed for every sample using any of the several conventional tree-building methods now available. To do bootstrapping, a new sample of N transactions is generated at random from the pool of m already sampled transactions, with the same transaction being selected no more than t times. By iteratively repeating these procedures, several copies of a dataset may be built. Taking additional samples could help lower the sampling error. If node n's splitting properties vary across bootstrap trees, then node n and its children are omitted.

After combining the sample trees, we verify that the node splits are within the confidence range. The whole training database D is then used to build the tree T in the traditional method. The next step is to examine the differences between the starting tree (T) and the tree we're using as an example (T1) in order to find and correct any inconsistencies. Changing the level of assurance requires adjusting the number of bootstrap samples.

Dataset Used

Dataset used for ML techniques: The dataset utilized is a longitudinal study of 1,144,740 individuals accused of committing either a cybercrime (N=870) or a traditional crime (N=1,144,740) in the Netherlands between the years of 2000 and 2012. The purpose of this research is to make inferences about the personal and professional life of cybercasters based on how long it takes for them to get caught. The dataset utilized in cybercrime detection can be obtained from various places, including the CBS open data stat line (https://www.cbs.nl/en-gb/our- services/open-data) and the CBS customized services microdata (https://www.cbs.nl/en-gb/our-services/customized- services-microdata/microdata-conducting-your-own-research) websites.

Freely accessible in a machine-readable format, open data is the norm. Statistics Netherlands' Open Data is split into three distinct types. The O Data interfaces APIs developed by Statistics Netherlands provide automated data processing. The O Data protocol facilitates data collection. The data is safeguarded according to European guidelines set out in the EU's General Data Protection Regulation (GDPR).

EU's General Data Protection Regulation (GDPR) is a standard that must be met by the CBS. CBS follows not only its own code of conduct on the protection of personal information, but also the European Statistics code of practise and the Statistics Netherlands Act.

Stat Line open data: Open data also includes the tables required to view and download Stat Line. The Stat Line data portal provides access to all of the datasets in the Dutch government's open data collection (referred to as data.overheid.nl Dutch). The data utilised is from 1997-2011 and was made public in 2017. There are a total of one thousand different user IDs that may be used in conjunction with demographic information on crime victims. Statistics on the ages, genders, and locations of victims of crime are compiled. Information like age distribution, sex distribution, greatest level of education, and the density of addresses per square kilometre are included. Years 1997–2004 are used as a benchmark for analysis. Totaling 824 rows and 28 columns, the file is rather large.

One of the attributes is the proportion of total victims that may be attributed to that ID.

The number of people impacted by a certain crime,

How many victims have been subjected to multiple forms of violence, how many victims have been subjected to sexual violence, how many victims have been subjected to sexual violence against only females, how many victims have been subjected to physical abuse, how many have been subjected to threats, how many have been subjected to telephone harassment, and how many have been subjected to other forms of violence.

Set of data including 25 user attributes representing 1000 users.

Optimization of Attributes Selected

Hacking: The term "hacking," which means "the unauthorised incursion into a computer," refers to a crime that anybody, or "hacker," is capable of committing. Criminal charges might be brought when a victim loses money due to hacking. There's a big difference between hacking and ethical hacking. There are several entry points for the theft of private information.

Theft: Data theft is the illegal acquisition and use of an individual's or organization's computerised or networked personal information. This is an invasion of the user's personal space. Password theft, hacking, and the theft of sensitive financial data like credit card numbers are all too prevalent.

Cyber Stalking: If someone you don't know, or someone you don't know very well, sends you unsolicited emails or SMS, they may be engaging in online stalking. The stalker becomes very obsessive with their target and follows them wherever they go, bringing them extreme emotional and mental anguish.

Identity Theft: Theft of one's identity refers to the illegal use of one's own or another's private information. A malicious third party might, for instance, make unauthorised charges to or withdrawals from the victim's bank account. The victim might have devastating monetary repercussions as a result of this.

Malicious: Malware like this spreads across a network and wreaks havoc in order to pave the way for hackers to get in and steal information.

Child Soliciting: This offence includes any sexual or violent interaction with a juvenile that takes place via the internet. The government is taking aggressive action in response to the growth in this kind of crime.

Abuse: Institutionalized child abuse occurs when minors are forced to participate in internet pornography. Despite the fact that chat rooms are less likely to be used for these kinds of crimes because of modern monitoring, they nevertheless occur.

Assault by Threat: The victim may be threatened in written, verbal, electronic, and/or visual forms.

Child Pornography: This involves extensive online sexual exploitation of youngsters with the purpose of disrupting their nascent worldviews.

Cyber Illegal Imports: The term "illegal imports" describes the practice of transferring illegal items through the Internet while employing encryption technology."

Cyber Laundering The electronic movement of stolen money from one location to another is an example of this kind of internet crime. During the transfer, the identity of the receiver is concealed.

Cyber Terrorism: Here, politicians utilise advanced technologies to attack regular citizens.

Cyber Theft: This is a case of computer-facilitated crime, since the necessary data is taken in this way. The DNS cache may be accessed by criminal actors for a variety of fraudulent purposes, including but not limited to identity theft, piracy, malicious hacking, and plagiarism.

Advertising through the internet: The site encourages online prostitution. These communications occur between governments and people.

Soliciting harlotry through the internet: Internet marketing is spreading the word about a service or something through the World Wide Web.

Drug Sales: Only those in possession of a valid licence from the appropriate agency may lawfully sell drugs through the internet. Despite the fact that it is illegal to prescribe these drugs without a prescription or medical licence, it is becoming more common for people to do so over the internet.

Number of times the proxy server is used: Proxy helps cache data that is often accessed by consumers, which decreases wait times. The proxy maintains a log of its communications to aid in debugging issues. If this proxy server is hacked, the data stored on it might be leaked.

Malicious Code Presence: When malicious code is inserted into a programme or script, it may have a variety of negative effects, including as causing the programme to act unexpectedly, crashing the system, or compromising security. Antivirus software is insufficient in the face of this enormous danger to cyber security. Dangerous software takes numerous forms, such as but not limited to viruses, backdoors, script attacks, worms, Trojan horses, and malicious active content.

Password Violations: Strong passwords should be selected by users whenever they are asked to do so during account creation. Forms should be designed in a manner that discourages the usage of weak passwords for user accounts.

Excess Privileges: Malicious activities like data theft and illegal access may take root in the soil of excessive Internet usage. They can not only read, write, and delete the file, but also get illegal access to the socket communication rights.

Computer related offences: Unauthorized access to someone else's data is a serious computer crime. Information that can be accessed, downloaded, and removed is kept here.

Data Forwarding: To "forward" information is to transmit it to a remote server or supported storage provider.

Publication of irrelevant content: The dissemination of information that is not specifically related to any person or subject.

Transmission of obscene content: Unauthorized dissemination of private, sexually explicit information about another person is an example of cyberstalking.

Sexually explicit content: If the victim's image is utilised for financial advantage, whether in a photograph, a video, or any other medium, it is a violation of their dignity.

Note on Trolling: When you "troll" someone, you write hurtful comments about them online in an attempt to make them feel bad about themselves. It is a kind of cybercrime to utilise Twitter for criminal ends. The "right to freedom of speech and expression" is a fundamental right for all Indian citizens, as guaranteed by Article 19(1)(a) of the Indian Constitution. The right to "Freedom of Expression" is enshrined in Article 19 of the Universal Declaration of Human Rights (UDHR). Article 19 of the International Covenant on Civil and Political Rights makes the claim that these freedoms are recognised by the general public (ICCPR). Using machine learning to identify troll factories Reference: [Fornacciari, P. et al., 2018]. As stated by [Anqi Liu et al.], the troll factory cannot identify a troll who seems to be serious in their comments. The Trolls and Troll Factories have not been discussed in this research.

Applications amends

Researchers, analysts, and investigators may benefit from cybercrime detection in a number of ways. It's quite useful for both the NMC and the forensics lab while investigating cybercrime. Large companies often use cybercrime detection systems to hunt down hackers who have stolen sensitive information. The suggested method may be employed if user profile data is accessible and there is no way to negatively affect the victim's emotions.

CONCLUSION

Cybercrime is one of the fastest growing criminal industries today. The majority of cybercrime happens when a system, network, or device is compromised. Because of this, uncovering cybercrime is crucial. Up until now, cybercrime detection has solely been the focus of studies and surveys. Most of these techniques, however, have not been directly applied to the problem of cybercrime detection. In the conventional approach, the whole training database D is utilised to construct the tree T. Next, we'll look for discrepancies between T, our seed tree, and T1, our reference tree, by comparing the two and noting any discrepancies. The number of bootstrap samples must be changed in order to change the confidence level.

REFERENCE

1. Neha k bhatt, dr pareshkumar d. Dobariya (2021) a challenging role of indian judiciary at cyber space to curb cybercrime against women. A global journal of interdisciplinary studies (issn – 2581-5628) impact factor: sjif - 5.047, iifs - 4.875 globally peer-reviewed and open access journal

2. Sang ni, mengboxie, quanqian (2017), “clustering-based k-anonymity algorithm for privacy preservation”, international journal of network security, vol.19, no.6, pp.1062-1071.

3. Muhammetbaykara and resul das (2018), “a novel honeypot-based security approach for real-time intrusion detection and prevention systems”, journal of information security and applications, elsevier, vol. 41, pp. 103–116.

4. Khraisat a, gondal i, and vamplew p (2019), “an anomaly intrusion detection system using c5 decision tree classifier”, trends and applications in knowledge discovery and data mining, springer, pp 149–155.

5. Ashkaashani, deeshanirmal, viral doshi, and nikita pati (2018), “survey on security using honeypot”, international organization of scientific research, vol. 12, pp. 41-44.

6. Hamad al-mohannadi, irfanawan, jassim al hamary, and andrea cullen (2018), “cyber threat intelligence from honeypot data using elasticsearch”, ieee, international conference on advanced information networking and applications.

7. Hidemasanaruoka, masafumimatsuta, watarumachii, tomomi aoyama, masahito koike, ichiro koshijima, yoshihiro hashimoto (2014), “ics honeypot system (camouflage net) based on attackers’ human factors’, international conference on applies human factors and ergonomics, pp. 1074 -1081.

8. Rupinder kauri, sunil nagpaer, saurabhchamotra (2015), “malicious traffic detection in a private organizational network using honeynet system”, ieee india conference (indicon).

9. Chunhui yuan and haitao yang (2019), “research on k-value selection method of k-means clustering algorithm”, vol. 2, no. 16, pp. 226 - 235.

10. Nyamugudza .t, rajasekar. v, sen. p, nirmala .m, and viswanatham. v.m (2017), “network traffic intelligence using a low interaction honeypot”, iop conference series: materials science and engineering, pp. 1 - 10.

11. Tendainyamugudza (2017), “network traffic intelligence using a low interaction honeypot”, icset, vol. 263, pp. 1 – 10.

12. Huang .c, han. j, zhang .x, and Liu. j (2019), “automatic identification of honeypot server using machine learning techniques”, security and communication networks, pp. 1 - 8.

13. Al-shaer e.s, wei. j, hamlen k.w, and wang .c (2019), “autonomous cyber deception: reasoning, adaptive planning, and evaluation of honey things”, springer, pp. 1 - 235.

14. Manjeetrege and raymond blanch k. Mbah (2018), “machine learning for cyber defense and attack”, data analytics: the seventh international conference on data analytics, pp. 73 - 78.

15. Alkeshbharati and sarvanaguru ra (2018), “crime prediction and analysis using machine learning”, international research journal of engineering and technology, vol. 5, issue. 9, pp. 1037 - 1042.