CYBER ATTACKS DETECTING ON WEBAPPLICATIONS USING PERL COMPATIBLE REGULAR EXPRESSIONS (PCRE) IN ML

Yalamanchili Sai Deepthi Goipika [1], BTech , Department of CSE , 19L11A0515@gmail.com
Makkena Suvarna Kumar [2], BTech , Department of CSE , 19L11A0507@gmail.com
Gunti Jyosthna [3], BTech , Department of CSE , 19L11A0505@gmail.com
Gorrela Gangadhar Sai Subhash [4], BTech , Department of CSE , 19L11A0504@gmail.com
A.Satish Kumar [5], M.Tech , Associate professor ,Department of CSE
VRS & YRN COLLEGE OF ENGINEERING AND TECHNOLOGY

ABSTRACT: People are using more cloud services and more web applications, and the way devices connect to the internet is changing. This makes it harder to keep the internet safe from bad guys. To protect people from these threats, the tools used for internet security must also change and get better. This article is about one way to make the internet safer. The idea is to use a machine learning model to figure out what is normal online behavior and what is not. This will help detect cyber attacks.The model consists of patterns (in form of Perl Compatible Regular Expressions (PCRE) regular expressions) that are obtained using graph-based segmentation technique and dynamic programming. The model is based on information obtained from HTTP requests generated by client to a web server. We have evaluated our method on CSIC 2010 HTTP Dataset achieving satisfactory results.

Keywords:*Cyberattacksdetection,cybersecurity,applicationlayer, anomaly detection.*

1. INTRODUCTION

Almansob and Lomte used Blameless Bayes and PCA with the KDD99 dataset for their research. Chithik and Rabbani also used PCA, SVM, and KDD99 for their IDS study. Aljawarneh et al used the NSLKDD dataset for their IDS research. The KDD99 dataset is often used for IDS research, but it's old (created in 1999) and doesn't include information about newer types of attacks. So, InthismannerweutilizedanewCICIDS2017 datasetinourinvestigation.

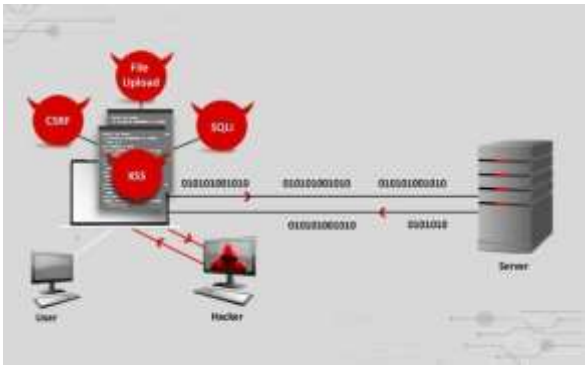

Fig.1:Cyberattacking

Network security is about keeping computer systems, networks, and all the information safe from harm.The goal is to stop bad guysfrom getting in and changing things they shouldn't. But as more things are connected to the internet, they become targets for

attacks.So, we need to find ways to protect these systems and detect when an attack is happening. The biggest challenge is knowing how to recognize different kinds of attacks, especially new ones that we haven't seen before.Theaimofprojectisnecessarytoprovideeffectivestrategiestodetectanddefendattacks and maintain network security.

Nowadays, the increasing number of security incidents being reported around the world. The National Computer Emergency Response Team (CERT) in Poland has reported a significant increase in the number of attacks, Accordingtothereport[1] in2012 there were 1082 incidents, which is an increaseof nearly 80% in comparison to the previous year,mainly due to malware and phishing. The rise in the number of incidents is linked to the growing number of people using mobile devices and connecting to the internet from anywhere, which challenges traditional network security. The trend of "bring your own device[4]" (BYOD) also presents new and evolving threats to the security ofmany businesses. Today's malware, such as ZITMO (Zeus In The Mobile), targets users' private information and access to remote services like banks and web services. The widespread use of social media has also led to a significant number of reported incidents and the quick spread of various types of malware and viruses. As reported by Sophoslabs[2]in 2013, botnets (networks of infected devices) have become more widespread, harder to detect, and more dangerous, targeting new targets.

2. LITERATUREREVIEW

2.1 Defensiveprogramming:Usinganannotationtool kitto builddos-resistantsoftware

This paper describes a toolkit to helpimprove the robustness of code against DoS attacks.Weobservethatwhendevelopingsoftware,programmersprimarilyfocusonfunctionality.Protecting code from attacks is often considered theresponsibilityoftheOS,firewallsandintrusiondetection systems.Asaresult,manyDoSvulnerabilities are not discovered until the system isattacked and the damage is done. Instead of reactingto attacks after the fact, this paper argues that a bettersolutionistomakesoftwaredefensivebysystematically injecting protection mechanisms intothecodeitself.OurtoolkitprovidesanAPIthatprogrammers use to annotate their code. At runtime,these annotations serve as both sensors and actuators:watchingforresourceabuseandtakingtheappropriateactionshouldabusebedetected.Thispaper presents the design and implementation of thetoolkit, as well as evaluation of its effectiveness withthreewidely-deployed networkservices.

2.2 Aclassificationofsql-injectionattacksandcountermeasures

SQL injection attacks pose a serioussecuritythreattoWebappli-cations:theyallow attackerstoobtainunrestrictedaccesstothedatabasesunderlyingtheapplicationsandtothepotentiallysensi-tiveinformationthesedatabasescontain. Although researchers and practitioners haveproposedvariousmethodstoaddresstheSQLinjecti

on problem, current approaches either fail toaddressthefullscopeoftheproblem.

Manyresearchers and practitioners are familiar with only asubset of the wide range of techniques available toattackers who are trying to take advantage of SQLinjectionvulnerabilities.Asaconsequence,manysolutions proposed in the literature address only someof the issues related to SQL injection. To address thisproblem,wepresentanextensivereviewofthedifferenttypesofSQLinjectionattacksknowntodate. For each type of attack, we provide descriptionsand examples of how attacks of that type could beperformed.Wealsopresentandanalyzeexistingdetectionandpreventiontechniquesagainst SQLinjection attacks. For each tech- nique, we discuss itsstrengthsandweaknessesinaddressingtheentirerangeofSQLinjectionattacks.

2.3 SAS:semanticsawaresignaturegenerationforpolymorphicwormdetection.

Stringextractionandmatchingtechniqueshavebeenwidelyusedingeneratingsignatures for worm detection, but how to generateeffectivewormsignaturesinanadversarialenvironment still remains challenging. For example,attackerscanfreelymanipulatebytedistributionswithin the attack payloads and also can inject well-crafted noisy packets to contaminate the suspiciousflow pool. To address these attacks, we propose SAS,anovelSemanticsAwareStatisticalalgorithmforautomaticsignaturegeneration.WhenSASprocessespackets in a suspicious flow pool, it uses data flowanalysis techniques to remove non-critical bytes. Wethen apply a Hidden Markov Model (HMM) to therefined data to generate state-transition-graph basedsignatures. To our best knowledge, this is the firstworkcombiningsemanticanalysiswithstatisticalanalysis to automatically generate worm signatures.Our experiments show that the proposed techniquecan accurately detect worms with concise signatures.Moreover,ourresultsindicatethatSASismorerobustto thebytedistributionchangesandnoiseinjectionattackscomparingtoPolygraphandHamsa.

2.4A novel model for detecting application layerDDoSattacks.

Counteringdistributeddenialofservice(DDoS)attacksisbecomingevermorechallenging with the vast resources and techniquesincreasingly available to attackers. DDoS attacks aretypically carried out at the network layer.

However,thereisevidencetosuggestthatapplicationlayer DDoSattackscanbemoreeffectivethanthetraditionalones. Inthispaper,weconsidersophisticatedattacksthatutilizelegitimateapplicationlayerHTTPrequestsfromlegitimately connectednetworkmachinestooverwhelmWebserver.

Since the attack signature of each applicationlayer DDoS is represented in abnormal user behavior,we propose a counter-mechanism based on Web userbrowsing behavior to protect the servers from theseattacks. In contrast to prior works, we explore hiddensemi-Markovmodeltodescribethebrowsingbehaviors of Web users and apply it to implement theanomalydetectionfortheapplicationlayerDDoSattacks which simulate the Web request behaviors ofbrowser and use HTTP requests to launch attacks. Byconductinganexperimentwitharealtrafficdata,the modelshowsthatitiseffectiveinmeasuringtheuser behaviorsanddetectingtheapplicationlayerDDoS attacks.

2.5 Robustanomalydetectionusingsupportvectormachines.

Using the 1998 DARPA BSM data setcollected at MIT's Lincoln Labs to study intrusiondetection systems, the performance of robust support -

vector machines (RSVMs) was compared with that ofconventionalsupportvectormachinesandnearestneighborclassifiersinseparatingnormalusageprofilesfromintrusiveprofilesofcomputerprograms.Theresultsindicatethesuperiorityof RSVMs not only in terms of high intrusion detectionaccuracy and low false positives but also in terms oftheir generalization ability in the presence of noiseand runningtime.

3. IMPLEMENTATION

Almansob and Lomte used Blameless Bayes and PCA with the KDD99 dataset for their research. Chithik and Rabbani also used PCA, SVM, and KDD99 for their IDS study. Aljawarneh et al used the NSLKDD dataset for their IDS research. The KDD99 dataset is often used for IDS research, but it's old (created in 1999) and doesn't include information about newer types of attacks. So, InthismannerweutilizedanewCICIDS2017 datasetinourinvestigation.

Disadvantages:

- Inmanycasesfalsepositivesaremorefrequentthanactual threats.
- They don’t take caretomonitor thefalsepositives,realattackscanslipthroughorbeignored.

This paper explains a new method called Mutation Based ABC (MABC) that helps find unused servers in data centers. People use the Internet to request cloud resources, and the Cloud Service Provider makes sure each request is given to a virtual machine (VM) to execute. Sometimes the request

needs to move between data centers to find an unused server. The data centers may also break the request into smaller tasks and search for unused servers to allocate those tasks to.

Advantages:

The proposed algorithm is able to minimize the makespan time of the jobs by assigning it to the availableunder-utilized data centers.

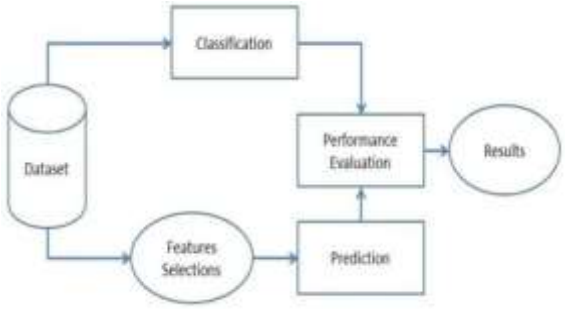

Fig.2:Systemarchitecture

MODULES:

- UploadTrainDataset
- UploadTestDatase
- PreprocessDataset
- ModelGeneration
- RunNeedleman-WunschDissimilarities
- TrainingSamplesVsTPRate

4. MACHINELEARNING

Let's first understand what machine learning is and what it is not before we learn about different types of machine learning methods. Machine learning helps us build mathematical models to understand data. By using tunable parameters, these models can learn from observed data and make predictions about new data. This type of learning is different from how the human brain learns, and it is used in data science to help us understand and predict patterns in data. To use machine learning effectively, it's important to understand the different approaches used in solving problems with machine learning.The study of machinelearning certainly arose from research in this context,butinthedatascienceapplicationofmachinelearningmethods,it'smorehelpfultothinkofmachine learning as a means of building models ofdata.Fundamentally,machinelearninginvolvesbuildingmathematicalmodelstohelpunderstanddata.

"Learning" enters the fray when we give thesemodelstunableparametersthatcanbeadaptedtoobserveddata;inthiswaytheprogramcanbeconsidered to be "learning" from the data. Once thesemodels have been fit to previously seen data, they canbe used to predict and understand aspects of newlyobserveddata.I'llleavetothereaderthemorephilosophicaldigressionregardingtheextenttowhichthistypeofmathematical,model-based"learning" is similar to the "learning" exhibited by thehumanbrain.Understandingtheproblemsettinginmachinelearningisessentialtousingthesetoolseffectively,andsowewillstartwithsomebroadcategorizationsofthetypesofapproacheswe'lldiscusshere.

CategoriesOfMachineLeaning:-

At the most fundamental level, machine learning canbecategorizedintothreemaintypes:supervisedlearning,unsupervised learning and Reinforcement learning.

Supervised Learning: This type of machine learning involves training a model on labeled data, where the input data is paired with the corresponding output or target variable. The model learns to predict the target variable for new input data based on the patterns it has learned from the training data. Common examples of supervised learning include classification and regression problems.

Unsupervised Learning: In this type of machine learning, the input data is not labeled, and the model must identify patterns and structures within the data on its own. Clustering is a common example of unsupervised learning, where the model groups similar data points together.

Reinforcement Learning: In this type of machine learning, an agent learns to make decisions based on a reward system. The agent takes actions in an environment and receives feedback in the form of rewards or penalties. The goal is to learn a policy that maximizes the cumulative reward over time. Examples of reinforcement learning include game playing and robotics.

ApplicationsofMachinesLearning:-

Machine learning has a wide range of applications in various fields, including:

- Image and speech recognition
- Natural Language Processing (NLP)
- Fraud detection and cybersecurity
- Recommendation systems

- Stockmarketanalysisandforecasting
- Speechsynthesis
- Speechrecognition
- Customersegmentation
- Objectrecognition
- Frauddetection
- Fraudprevention
- Recommendationofproductstocustomerinonlineshopping.

5. EXPERIMENTALRESULTS

Fig.3:Homescreen

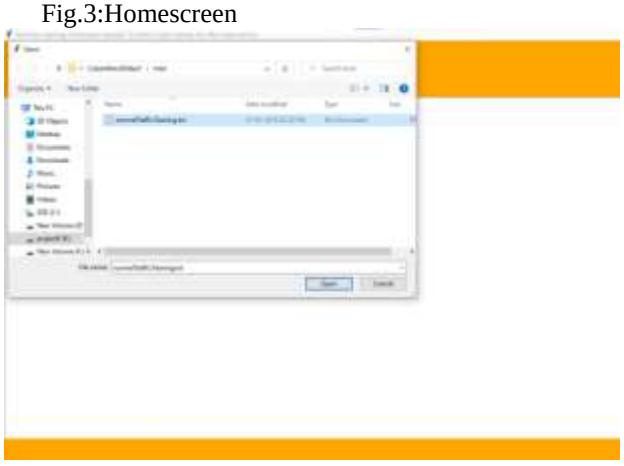

Fig.4:Uploadingscreen

Fig.5:Traindata

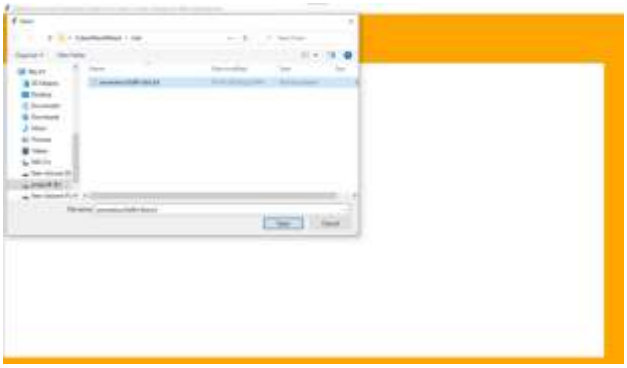

Fig.6:Uploading Test data

Fig 7:Test data

Fig.7:RunNeedleman-WunschDissimilarities

Fig.8:Detectionresult

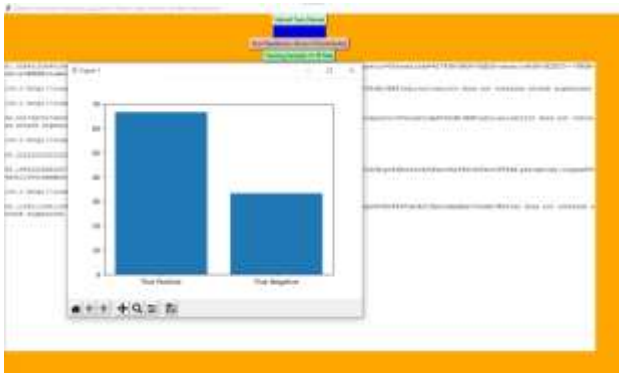

Fig.9:graph

6. CONCLUSION

In this article, the method for application layer attackdetection based on machine learning was proposed.Themodelconsistsofpatterns(inform ofPCREregular expressions) that are obtained using graph-basedsegmentationtechniqueanddynamicprogramming. The regular expressions are used formodelling the genuine behaviour of the applicationsand detecting cyber attacks. We also presented theresultsthatprovetheefficiencyoftheproposedalgorithm that can be effectively used for applicationlayer attack detection. The experiments on CSIC'10show that the proposed approach can achieve 94.46%ofdetectionratiowhilehaving<4.5%offalsepositive s.

7. FUTURESCOPE

Researchers are working on turning their models into real-time systems so they can be used to detect and prevent attacks in real-life situations. Real-time ML has two levels: online prediction, which means making predictions in real-time, and online learning, which allows the system to update itself with new data in real-time. Many researchers plan to work on turning their models into real-time systems so that they can be used in real-life situations.

8. REFERENCES

[1] ThailandMotorVehicleRegistered.CEICDataGlobal Database.
[2] Linda,S.,Canpublictransportcompetewiththeprivatec ar?IatssResearch, 2003. 27(2):p. 27-35.
[3] CO2emissions(metrictonspercapita)-Thailand.THEWORLD BANK.
[4] Cohen, A.J., et al., Estimates and 25-year trends ofthe global burden of disease attributable to ambient airpollution: an analysis of data from the Global BurdenofDiseasesStudy2015.TheLancet,2017.389(1008 2):p. 1907-1918.

[5] Litman, T. and D. Burwell, Issues in sustainabletransportation.InternationalJournalofGlobalEnvironmentalIssues, 2006.6(4):p.331-347.

[6] Liu, M. and N. Choosri.. A technical solution toimprovetheredcabfortouringinChiangMai:Chinese tourists' perspective. in 2016 Chinese Controland Decision Conference(CCDC). 2016.IEEE

[7] Farooq,M.U.,A.Shakoor,andA.B.Siddique.GPS based Public Transport Arrival Time Prediction.in2017InternationalConferenceonFrontiers ofInformationTechnology(FIT).2017. IEEE.

[8] Bin,Y.,Y.Zhongzhen,andY.Baozhen,Busarrival time prediction using support vector machines.Journal of Intelligent Transportation Systems, 2006.10(4):p. 151-158.

[9] Maiti,S.,et al.Historical data base real timeprediction of vehicle arrival time. In 17th InternationalIEEE Conference on Intelligent Transportation Systems (ITSC). 2014 IEEE

[10] Fan, W . and Z. Gurmu, Dynamic travel time prediction models for buses using only GPS data. International Journa of Transportation Science and Technology, 2015.4(4): p.353-366