International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1384-1388 www.ijaem.net ISSN: 2395-5252

Machinelearning Solution for Detecting network Attacks

V. Stella mary, M. Vasantha

Submitted: 10-08-2021 Revised: 22-08-2021 Accepted: 25-08-2021

ABSTRACT
Developing IntrusionDetectionSystem(IDS) by setting the real working environment to explore all the possibilities of attacks isexpensive. Software to detect network intrusions protects a computer network fromunauthorized users, including perhaps insiders. The intrusion detector learning task isto build a predictive model (i.e. a classifier) capable of distinguishing between "bad"connections, called intrusions or attacks, and "good" normal connections. To preventthis problem in network sectors have to predict whether the connection is attacked ornotfromKDDCup99datasetusingmachinelearning techniques.Theaimistoinvestigate machine learning based techniques for better packet connection transfersforecasting by prediction results in best accuracy from ensemble learning votingclassifier technique.To propose a machine learning-based method to accuratelypredicttheDOS,R2L,U2R,Probeandoverallattacksbypredictionresultsintheformofbestaccuracyfromcomparingsuperviseclassificationmachinelearningalgorithmswith voting classifiers.Additionally, to compare and discuss the performance ofvariousmachinelearningalgorithmsfromthegivendatasetwithevaluationclassification report, identify the confusion matrix and to categorizing data frompriority and the result shows that the effectiveness of the proposed machine learningalgorithmtechniquecanbecomparedwithbest accuracywithprecision,RecallandF1Score.

Key words: Developing IntrusionDetectionSystem, DOS, Network attacks, machine learning algorithm

I. INTRODUCTION

Thisanalysisaimstoobservewhichfeaturesaremosthelpfulinpredictingthenetworkattacks of DOS, R2L, U2R, Probe and combination of attacks or not and to see the generaltrends that may help us in model selection and hyper parameter selection. To achieve usedmachine learning classification methods to fit a function that can predict the discrete class ofnew input.

Lately, an internet network company in Japan has been facing huge losses due tomalicious server attacks. They've encountered breach in data security, reduced data transferspeed and intermittent breakdowns in user-user & user-network connections. When asked, acompanyofficialsaid,“there’sasignificantdipinthenumberofactiveusersonournetwork". The company is looking are some predictive analytics solution to help them understand,detectandcountertheattacksandmaketheirnetworkconnectionsecure.Thinkofaconnection as a sequence of TCP packets starting and ending at some well-defined times,between which data flows to and from a source IP address to a target IP address under somewell-defined protocol. In total, there are 3 major type of attacks to which their network isvulnerableto.But,3ofthemcausethemaximumdamage.Inthischallenge,youaregivenananonymised sample dataset of server connections. The type of attack(s)likeDos, R2L, U2R, Probe have to be predicted.

In the existing system, development of connected devices and their daily use is presently at the origin ofthe omnipresence of Wi-Fi wireless networks. However, these Wi-Fi networks are oftenvulnerable, and can be used by malicious people to disturb services, intercept sensitive data,ortogainaccesstothesystem.Inrailways,trainsarenowequippedwithwirelesscommunication systems for operational purposes or for passenger services. In both cases,defense strategies have to be developed to prevent the misuses of the networks. The firstobjective of this study is to propose a monitoring solution, which is independent of thecommunicationnetworks,todetectthe occurrenceofattacks.

Our proposed system isnotmeanttobeprovidingafinalconclusiononthereasons leadingtonetworksectorasitdoesn'tinvolveusinganyinferentialstatisticstechniques/machinelearning algorithms.Machine learning supervised

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1384-1388 www.ijaem.net ISSN: 2395-5252

classification algorithms will be used togive the network connection dataset and extract patterns, which would help in predicting thelikelypatientaffectedornot,therebyhelpingtheattackofavoidsformakingbetterdecisionsinthefuture.Multipledatasetsfromdifferentsourceswouldbecombinedtoformageneralized dataset, and then different machine learning algorithms would be applied toextractpatternsandto obtain resultswith maximumaccuracy.

II. LITERATURE SURVEY

In an early warning system,accurate prediction of DoS attacks is the prime aim in the network offence and defense task.Detectionbasedonabnormityiseffectivetodetect DoSattacks.AvariousstudiesfocusedonDoSattacksfromdifferentrespects[1].Socio-technical attack is an organized approach which is defined by the interaction amongpeople through maltreatment of technology with some of the malicious intent to attack thesocialstructurebasedontrustandfaith.[2][3].Intrusion detection systems (IDS) are used to detect the occurrence of malicious activitiesagainst IT system. Alertsrelationsaredifferentiatedfromduplicationrelationtosameattackscenariorelation[4].The prediction results reflect the security situation of the target network in the future, andsecurityadministratorscantakecorrespondingmeasurestoenhancenetworksecurityaccording to the results..Manymodelshavebeen proposedforperformingevaluation ofnetwork security[5].Meanwhile,withtheBayesianmethod,the calculationoftheoutputprobabilitycorrespondingtoeachsub-modelisdeducedandthenthedistributionoftheamount ofDoSattackinsomerangeinfuture isobtained[6][7].

This paper attempted to cover the state-of-the-art studies for adversarial examples in the DLdomain.Comparedwithrecentworkonadversarial examples,weanalyzedanddiscussedthecurrentchallengesandpotentialsolutionsinadversarial examples[8].Thispaperhasinvestigatedthedistributedsecurecontrolofmultiagentsystemsunder DoS attacks. We focus on the investigation of a jointly adverse impact of distributed DoSattacksfrommultipleadversaries[9].Fortheevent-triggeredcase, two effective dynamical event conditions have been designed and implemented in afully distributed way, and both of them have excluded Zeno behavior. Finally, a simulationexample has been provided to verify the effectiveness of theoretical analysis[10].

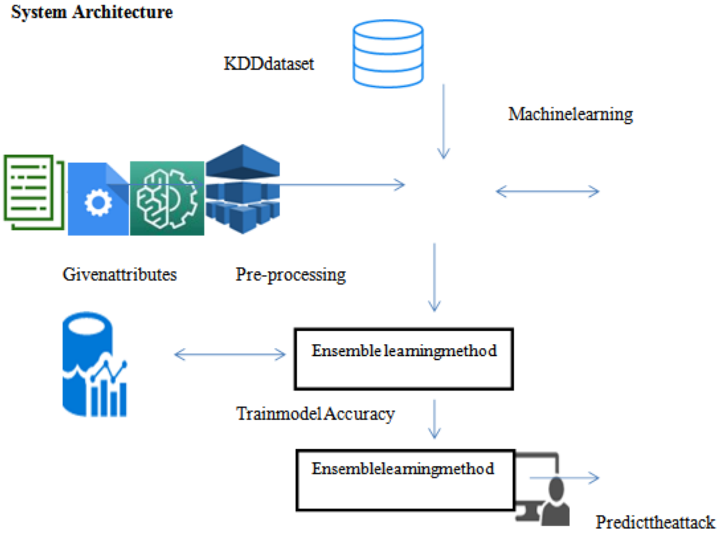
III. METHODOLOGY

As networked systems become more and more pervasive and businesses continue tomove more and more of their sensitive data online, the number and sophistication of cyber-attacks and network security breaches has risen dramatically. In earlier year, there are twokindsofbigcompaniesintheUnitedStates.Therearethosewho’vebeenhacked...andthosewho don’t yet know they’ve been hacked.” In order to secure their infrastructure and protectsensitiveassets,organizationsareincreasinglyrelyingonnetworkintrusiondetectionsystems(NIDS) to automatically monitor their network traffic and report suspicious or anomalousbehavior.Historically,mostNIDSoperateinoneoftwostyles:misusedetectionandanomalydetection.Misusedetectionsearchesforprecisesignaturesof knownmaliciousbehavior,while anomaly detection tries to build a model for what constitutes “normal” network trafficpatternsandthenflagdeviationsfromthosepatterns.Forallthesamereasons thatsignature-based antivirus software is becoming obsolete (the ease of spoofing signatures and theincreasing diversity and sophistication of new attacks), misuse-detection is struggling toremain relevant in today’s threat landscape. Anomaly-based intrusion detection offers theenticing prospect of being able to detect novel attacks even before they’ve been studied andcharacterizedbysecurityanalysts,aswellasbeingabletodetectvariationsonexistingattackmethods.Inour projectwefocusonclassifyinganomaliesusingbothsupervisedandunsupervisedlearning techniques.

In order to create data for the IDS, it is necessary to set the real working environmenttoexploreallthepossibilitiesofattacks,whichisexpensive.Dataanalysisphasesystematically identifies the patterns in the gathered information, and narrates them to thedefined issue. It is a process of examining, transforming and modeling of data and decidinghow to organize, classify, interrelate, compareand displayit. Data quality focuses thecorrectnessandreliabilityofinformationgatheredandutilizedinanevaluation.Dataquantitydeals with the quantity of information gathered for the evaluation. This task requires variousground truth databases in its region and the experimentation would be completed effectivelyifthequality andfeaturesofdataforthespecificregion aregood.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 8 Aug 2021, pp: 1384-1388 www.ijaem.net ISSN: 2395-5252

System Architecture



DATASET

The KDDCup99 data set stems from data gathered at MIT Lincoln Laboratory undersponsorship of the Defense Advanced Research Projects Agency (DARPA) to evaluateIntrusion Detection Systems (IDSs) in 1998 and 1999. These two data sets are referred to asDARPA98 and DARPA99, which consist of raw TCP dump data from a simulated mediumsizedUSairforcebase.TheKDDCup99datasetwasprovidedfortheKnowledgeDiscoveryandDataMiningToolscompetition(andassociatedconference)in 1999.ThisisatransformedversionoftheDARPATCPdumpdata,consistingofasetoffeaturesconsideredsuitable for classification with machine learning algorithms. The data set consists of 41features,someofwhichareintrinsictothenetworkconnections,whilstotherarecreatedusingdomainknowledge.

ClassificationofAttacks:

The data set in KDD Cup99 have normal and 22 attack type data with 41 features andall generated traffic patterns end with a label either as ‘normal’ or any type of ‘attack’ forupcoming analysis. There are varieties of attacks which are entering into the network over aperiodoftimeandtheattacksareclassifiedintothefollowingfourmainclasses.

- DenialofService(DoS)
- Userto Root(U2R)
- RemotetoUser (R2L)
- Probing

IV. ALGORITHM

Logisticregressionalgorithmalsousesalinear equationwithindependentpredictorstopredict a value. The predicted value can be anywhere between negative infinity to positiveinfinity. We need the output of the algorithm to be classified variable data. Higher accuracypredictingresultislogisticregressionmodelbycomparingthebestaccuracy.

TruePositiveRate(TPR)=TP/(TP+FN)FalsePositiverate(FPR)=FP/(FP+TN)

Accuracy: The Proportion of the total number of predictions that is correct otherwise overallhowoftenthe modelpredictscorrectlydefaultersandnon-defaulters.

Accuracycalculation:

Accuracy=(TP +TN)/(TP +TN+FP +FN)

Accuracy is the most intuitive performance measure and it is simply a ratio of correctlypredictedobservationtothetotalobservations.Onemaythinkthat,ifwehavehighaccuracythen our model is best. Yes, accuracy is a great measure but only when you have symmetricdatasetswherevaluesoffalse positiveand falsenegativesarealmostsame.

Precision: The proportion of positive predictions that are actually correct. (When the modelpredictsdefault:howoften iscorrect?)

Precision=TP/(TP+FP)

Precisionistheratioofcorrectlypredictedpositiveobservationstothetotalpredictedpositiveobservations.Thequestionthatthismetricanswerisofallpassengersthatlabeledassurvived, how many actually survived? High precision relates to the low false positive rate. We havegot0.788 precision whichispretty good.

Recall: The proportion of positive observed values

correctly predicted. (The proportion ofactualdefaultersthatthe modelwillcorrectly predict)
Recall=TP/(TP+FN)

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the allobservationsin actualclass-yes.
F1ScoreistheweightedaverageofPrecisionandRecall. Therefore,thisscoretakesbothfalsepositives and false negatives into account. Intuitively it is not as easy to understand asaccuracy,butF1isusuallymoreusefulthanaccuracy, especiallyifyouhaveanunevenclassdistribution.
Accuracy works best if false positives and false negatives have similar cost. Ifthe cost of false positives and false negatives are very different, it’s better to look at bothPrecisionand Recall.
GeneralFormula:F-Measure=2TP/(2TP+FP+FN)

F1-ScoreFormula:F1Score=2*(Recall*Precision)/(Recall+Precision)

V. CONCLUSION

The analytical process started from data cleaning and processing, missing value,exploratory analysis and finally model building and evaluation. The best accuracy on publictest set is higher accuracy scoreis will be find out by comparing each algorithm with type ofallnetworkattacksforfuturepredictionresultsbyfindingbestconnections.Thisbringssomeof the following insights about diagnose the network attack of each new connection. Topresentedapredictionmodelwith theaidofartificialintelligencetoimproveoverhumanaccuracyandprovidewiththescopeofearlydetection.Itcanbeinferredfromthis model that, area analysis and use of machine learning technique is useful in developingprediction models that can helps to network sectors reduce the long process of diagnosis anderadicateany human error. Our future work can be Networksectorwanttoautomatethedetectingtheattacksofpackettransfersfromeligibilityprocess(realtime)based onthe connectiondetail.Automatingthisprocessbyshowthe predictionresultinwebapplicationordesktopapplication.Optimizingthework toimplementinArtificialIntelligence environment.

Authors’ Profile

Mrs. V. Stella mary pursued Master of Engineering from Anna University, India. Currently she is working as an Assistant Professor in Sri Muthukumaran Institute of Technology. She has worked as an Assistant Professor in reputed engineering colleges under VTU university and Anna University. She has 8 years of teaching experience and 2 years of research experience.

Mrs. M. Vasantha pursed Master of Computer Application from Alagappa University, Master of Engineering from Anna University, India and Doctorate in Computer Science from Mother Teresa University, India in the year 2015. She is currently working as Associate Professor in PG Department of Computer Sciences, Bhaktavatsalm Memorial College For Women, Chennai affiliated with the University of Madras, India since 2016. She has published more than 15 research papers in reputed international journals.. Her main research work focuses on Big Data Analytics, Data Mining, and Machine learning. She has 25 years of teaching experience and 10 years of Research Experience.

REFERENCES

[1] C.H.Rowland,“Intrusiondetectionsystem,”U.S.Patent6405318,Jun.11,2002.
[2] M.SunandT.Chen,“Networkintrusiondetectionsystem,”U.S.PatentAppl.12/411916, Sep. 30, 2010.
[3] L. Vokorokos and A. Balaz, “Host-based intrusion detection system,” in Proc. 14th Int.Conf.Intell. Eng. Syst., 2010, pp. 43–47.
[4] P.VanAubel,K.Papagiannopoulos,L.Chmielewski,andC.Doerr,“Sidechannelbasedintrusiondetectionforindustrialcontrolsystems,” 2017,arXiv:1712.05745.
[5] W. Xu, W. Trappe, Y. Zhang, and T. Wood, “The feasibility of launching and detectingjamming attacks in wireless networks,” in Proc. 6th ACM Int. Symp. Mobile Ad Hoc Netw.Comput.,2005
[6] R.BhojaniandR.Joshi,“Anintegratedapproachforjammerdetectionusingsoftwaredefinedradio,” ProcediaComput. Sci.,vol. 79, pp. 809–

816, 2016.
[7] V. Deniau, C. Gransart, G. L. Romero, E. P. Simon, and J. Farah, "IEEE 802.11ncommunicationsinthepresenceoffrequency-sweepinginterferencesignals,"IEEETrans.Electromagn.Compat., vol. 59, no.5, pp. 1625–1633, Oct.2017.
[8] S.Grimaldi,A.Mahmood,andM.Gidlund,"An SVM-basedmethodforclassificationofexternalinterferenceinindustrialwirelesssensorandactuator networks,"J.SensorActuatorNetw., vol. 6, no. 2, p. 9, 2017.
[9] ElectromagneticCompatibility(EMC)—Part2-13:Environment—HighPowerElectromagnetic(HPEM)Environments—RadiatedandConducted,IECStandard61000-2-13Ed. 1, 2005.
[10] R.Vinek,BackTrack5WirelessPenetrationTestingBeginner'sGuide,PacktPublishingLtd.,Birmingham, U.K., 2011, ISBN:978-1-849515-58-0.