REVIEW OF VARIOUS DATA MINING AND MACHINE LEARNING METHOD FOR INTRUSION DETECTION

[1]Prakash Chandra, [2]Prof. Umesh Kumar Lilhore

[1]M.Tech Research Scholar, [2]Associate Professor and PG Coordinator
Department of Computer Science & Engineering, NIIST Bhopal MP, India

***ABSTRACT*-Due to rapid growth and development in digital world data are easily available for attackers. Easy availability of data create loops in security, an attacker can access and modified private data of an authorized user. Data Security is a crucial and open issue for researchers. Intrusions detections systems from point of view of security policy are a second line of defense; they have a supervisory role to observe the activities of our network or hosts to identify attacks in real time. In these days, electronics attacks can cause a very destructive damage for nations which make necessary the use of completed security policy to minimize the potential threats. IDS it is a very important element to resist against this vulnerability.KDD cup 99, N-KDD Cup and Kyoto data sets are to detect various network based IDS by using different machine learning and data mining methods. In this survey paper we are presenting review of various data mining and machine learning methods for IDS detection used in WEKA tool. Lastly in this survey we tend to explain the mostly used dataset in network security research KDDCUP 99 data sets and also present a complete study of its various components. Finally we conclude our survey with few real research proposals which will be open issues for searchers.**

***Keywords*- Intrusion Detection, Machine Learning, Network Security, WEKA**

1. INTRODUCTION

Intrusion detection System (IDS) is a type ofsecurity management system for computers andnetworks. An intrusion detection system (IDS)inspects all outbound and inbound network action andfind out the doubtful patterns that may point to anetwork or system intrusion or attack from someonetrying to crack into or conciliation of a system [7]. An ID gathers and observed information from different areasinside a network of systems to find out probablesafety breaches, which contain together calledintrusions (attacks exterior from the association) andmisuse (attacks from inside the association). IDS usesusceptibility assessment, it is an expertise which isdesign and developed to appraise the security of anetwork [12].

Data mining techniques can be used todetect intrusions. Applications of data mining havepresented a collection of research efforts on the use ofdata mining in computer security. In the context ofsecurity of the data we are looking for theinformation whether an information security breachhas been experienced [3]. This data could becollected in the perspective of discovering attacks orintrusions that aim to break the privacy and securityof services, information in a system or alternatively,in the context of discovering evidence left in acomputer system as part of criminal activity. Thereare four major categories of networking attacks:Denial of Service, Probing, and User to Root and Remoteto Local [1].

Intrusion detection system is the area where datamining concentrate heavily. There are two foldreasons for this first an IDS is very common and verypopular and extremely critical activity. Second, largevolume of the data on the network is dealing so this isan ideal condition for the data mining to use it. Thedata mining technology has the enormous benefits inthe data extracting attributes and the rule, so it issignificant to use data mining methods in theintrusion detection [4].

A significant problem of IDSis how to efficiently divide the normal behavior andthe abnormal behavior from a huge number of rawinformation's attributes, and how to effectivelygenerate automatic intrusion rules followingcomposed raw data of the network. To accomplishthis, different data mining methods must be studied,like classification, correlation analysis of data miningmethods and so on [14]. The ever rising new intrusionor attacks type poses severe difficulties for theirdetection. The human labeling of the accessiblenetwork audit information instances is generallytedious, expensive as well as time consuming. In this survey paper we are presenting review of various data mining and machine learning methods for IDS detection used in WEKA tool [7]. Lastly in this survey we tend to explain the mostly used dataset in network security research KDD CUP 99 data sets and also present a complete study of its various components. Finally we conclude our survey with few real research proposals which will be open issues for searchers [11].

2. INTRUSION DETECTION

An intrusion detection system (IDS) inspects alloutbound and inbound network action and find outthe doubtful patterns that may point to a network orsystem intrusion or attack from someone trying tocrack into or conciliation a system [2].

An ID gathers andobserved information from different areas inside anetwork of systems to find out probable safetybreaches, which contain together called intrusions(attacks exterior from the association) and misuse(attacks from inside the association) [6]. IDS usesusceptibility assessment, it is an expertise which

isdesign and developed to appraise the security of anetwork [12].

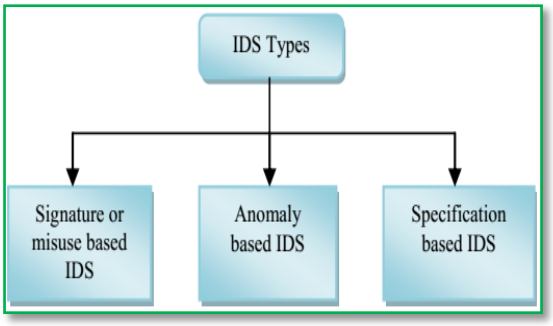

Figure 2.1 Types of IDS [2]

Data mining techniques can be used todetect intrusions. Applications of data mining havepresented a collection of research efforts on the use ofdata mining in computer security. In the context ofsecurity of the data we are looking for theinformation whether an information security breachhas been experienced [13]. This data could becollected in the perspective of discovering attacks or intrusions that aim to break the privacy and securityof services, information in a system or alternatively,in the context of discovering evidence left in acomputer system as part of criminal activity. Thereare four major categories of networking attacks:Denial of Service, Probing, and User to Root and Remoteto Local.

2.1 HOW IDS WORKS?

Working of Intrusion detection systems is based on four step approaches for thegeneralized working of IDS-

- **Collection of Data-** It involves collecting networktraffic using particular software and thus helps toget the information about the traffic like types ofpackets, hosts and protocol details.
- **Selection of Features-** The collected data issubstantially large because of the huge network traffic; we generate feature vectors that containonly necessary information. In network-basedintrusion detection, it can be IP headerinformation, which consists of source anddestination IP address, packet type, layer 4protocol type and other flags.
- **Analysis-** The collected data is analyzed in thisstep to determine whether data is anomalous ornot.
- **Action-**IDS alarm the system administrator thatan attack has happened and it tells about thenature of the attack. IDS also participate incontrolling the attacks by closing the network portor killing the processes.

3. DATA MIMING &MACHINE LEARNING METHODS FOR IDS

The term data mining is used to describe the process ofextracting useful information from the large databases.Data mining analyses the observed sets to discover theunknown relation and sum up the results of data analysisto make the owner of data to understand [3].

Hence datamining problems are considered as a data analysis problem.Data mining framework automatically detect patterns inour data set and use these patterns to find a set of malicious entries. Data mining techniques can detect patterns inlarge amount of data, such as byte code and use thesepatterns to detect future instances in similar data.
There is a significant overlap between machine learning and data mining. These two terms are commonly confused because they often employ the same methods and therefore overlap significantly. The pioneer of machine learning, Arthur samuel, defined machine learning as a [8] “field of studyThat gives computers the ability to learn without being explicitly programmed”.

Machine learning focuses on classification andPrediction, based on known properties previously learned from the training data. Machine learning algorithms need a goal (problem formulation) from the domain (e.g., dependent variable to predict). Data mining focuses on the discovery of previously unknown properties in the data. It does not need a specific goal from the domain, but instead focuses on finding new and interesting knowledge.

4. KDD CUP 99 DATASET

From1999, KDD’99 is the mainly frequent used datasetfor the assessment of anomaly detection techniques [10]. Thisdataset is made by Stolfo et al. and is built based on thedata taken in DARPA’98 IDS assessment program [7].

DARPA’98 is about 4 GB of compacted unrefined (binary)TCP dump data of seven weeks of internet network traffic,which can be developed into about five million linkrecords, each with about hundred bytes. The two weeks oftest data have around 2 million connection records. KDDtraining dataset consists of just about 4,900,000 singleconnection vectors every of which encloses 41 features andis labeled as either an attack or normal, with precisely onedefinite attack type.

The simulated attacks plunge in one ofthe following four categories-

a) **Denial of Service Attack (DoS)-** DoS is an attack inwhich the attacker creates some memory orcomputing resource too full or too busy to handlegenuine requests, or denies genuine users entranceto a machine.

b) **User to Root Attack (U2R)**- U2R is a class ofexploit in which the attacker creates entrance to astandard user account on the system (instead ofgained by sniffing passwords, social engineering,or a dictionary attack) and is capable to exploitseveral weaknesses to achieve root access to thesystem.

c) **Remote to Local Attack (R2L)**- R2L attack occurswhen an attacker whom the ability to launchpackets to a machine over a network but who doesnot have an account on that machine developsseveral weakness to achieve local entrance as auser of that machine.

d) Probing Attack- Prob is an effort to collectinformation about a network of computers for theperceptible reason of circumventing its safetycontrols.

5. WEKA TOOL FOR IDS

WEKA is a Tool for Data Mining and Machine Learningwhich was implemented at the University of Waikato, in New Zealand in the year 1997. WEKA is a set of Machine Learning and Data Mining algorithms [10]. This WEKA software is programmed in JAVA language and it has a GUI Interface to interact with data Files. With 49 datapre-processing tools WEKA tool contains 76 classification algorithms, 15 attribute evaluators and ten search algorithms for feature selection. There are threealgorithms to find association rules.

Figure 5.1 WEKA Tool

It also has three Graphical User Interfaces: "The Explorer", "TheExperimenter" and "The Knowledge Flow." The file formatto store data in WEKA is ARFF. Meaning of ARFF isAttribute Relation File Format. It also includes tools forvisualization.

6. EXISTING DM & ML FOR IDS

Bayes Classifier: They are also known as Belief Networks,belongs to the family of probabilistic Graphical Models (GM'S), These graphical models are used to representknowledge about uncertain domains, Random variables aredenoted by nodes in the graph and probabilisticdependencies are assigned as weights to the edgesconnecting corresponding random variable nodes [5].

Thesetypes of classifiers are based upon the idea of predicting theclass on the basis of value of members of the features. Thiscategory has 13 classifiers out of which 3 classifiers (BayesNet, NaiVeBayes and NaiVeBayes Updateable) arecompatible with the chosen dataset.

- **Functions Classifier**- Functional Classifier uses the conceptof neural network and regression. They maps inputdata to output. There are eighteen classifiers under thiscategory out of which only RBF Network and SMOclassifiers are compatible with our dataset RBF classifierscan model any nonlinear functions easily [5]. It does not useraw input data. The processing of RBF Networks is likeneural networks i.e. iterative in nature. The problem withRBF is the tendency to over train the model.

- **Lazy Classifier**- To construct the classification model lazyclassifiers demand to store complete training data i.e. suchclassifiers do not support inclusion of new samples intraining set while building the model. These types ofclassifiers are simple and effective [4]. Lazy classifiers aremainly used for classification on data streams, there arefive classifiers under this category out of which two arecompatible with our dataset that are: IB1 and IBK.

- **Meta Classifier**- These types of classifiers find the optimalset of attributes to train the base classifier with these parameters;this trained base classifier will be used forfurther predictions [8]. There are 26 classifiers under thiscategory out of which 21 are compatible with our dataset:AdaBoost M 1, LogistBoot, Attribute Selection Classifier,Bagging,Dagging Classification via Clustering,Classification via regression, End Multiclass Multischeme,Grading, Vote, Ordinal Class Classifier, Rotation Forest ,Random Subspace, CV Parameter Selection, RacedIncremental Logi Boost, Random Committee, Stacking,Stacking C.

- **Mi Classifier**- Mi stands for Multi- Instance Classifiers.This category of classifier consists of 12 classifiers out ofwhich no classifier is compatible with our dataset. Miclassifier is variant of supervised learning technique. It hasmultiple instances in an example but can only observe one class. These types of classifiers are originally madeavailable through a separate software package.

- **Misc Classifier**- There are three classifiers under thiscategory out of which two are compatible with our dataset.These compatible classifiers are Hyper pipes and VFI [10].

- **Rules Classifier**- In this category of classifier, associationrules are used for correct prediction of class among all theattributes and those correct predictions are called ascoverage and it is expressed in terms of percentage ofaccuracy [5]. They may predict more than one conclusion.Rules are mutually exclusive. These are learned one at a time, there are 11 classifiers under this category out ofwhich 8 are compatible with our dataset that are:Conjunctive Rule, Decision Table, DTNB, JRip, OneR, ZeroR, Part, Ridor.

- **Trees**-These are popular classification techniques in which at low chart like tree structure is produced as a result inwhich each node denotes a test on attribute value and eachbranch represents an outcome of the test [9]. They are alsoknown as Decision Trees. The tree leaves represents theclasses that are predicted. They design a model that is bothpredictive and descriptive. There are 16 classifiers underthis category out of which 10 are compatible with our chosedataset that are: Decision Stump, j48, j48 graft, LAD Tree,NB Tree, REP Tree, Random Forest, Simple Cart, Random tree.

7. CONCLUSIONS AND FUTURE WORKS

In this survey we have introduced an overview of differentdetection methodologies, approaches and techniques forIntrusion Detection System (IDS) used in WEKA using Datamining approaches. Each technique has its own superiorityand limitation. WEKA is a powerful instrument that offers several data Pre-processing facilities as well as facilities for their analysis through classification, regression, clustering, Association rules techniques, etc.For basicKnowledge of Machine Learning Approaches WEKA tooland various classification algorithms have been discussed.At last the KDD cup-99 data set which is wildly used inanomaly detection and some real reason for research scopein this field is given.

In future work we will propose an efficient IDS detection method by using weka tool and KDD cup-99 over existing methods. Existing method and proposed method will compare over weka and java netbeans and various performance comparisons parameters compare.

REFERENCES

[1] Nabila Farnaaz∗ and M. A. Jabbar,"Random Forest Modeling for Network Intrusion Detection System", Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016). ScienceDirect, Procedia Computer Science 89 (2016)PP 213-217

[2] Lata, InduKashyap, " Study and Analysis of Networkbased Intrusion Detection System", International Journalof Advanced Research in Computer and CommunicationEngineering Vol. 2, Issue 5, May 2013

[3] N.S.CHANDOLIKAR, V.D.NANDAVADEKAR,"Comparative Analysis of two Algorithms for intrusionattack classification using KDDCUP DataSet",International Journal of Computer Science andEngineering (IJCSE) Vol.1, Issue 1 Aug 2012.

[4] Devikrishna K, Ramakrishna B, "An Artificial NeuralNetwork based Intrusion Detection System andClassification of Attacks ", International Journal ofScientific & Engineering Research, Volume 6, Issue 1,January-2015

[5] Vaishali Kosamkar, Sangita S Chaudhari," ImprovedIntrusion Detection System using C4.5 Decision Treeand Support Vector Machine",International Journal ofComputer Science and Information Technologies, Vol. 5(2) , 2014.

[6] Rajendra V. Boppana, Senior Member, IEEE, and XuSu, Member, IEEE A Distributed ID for Ad HocNetworks, 26th InternationalConference on Advanced Information Networking andApplications 2012.

[7] Leila Mechtri, Fatiha Djemili Tolba, Salim Ghanemi,MASID,"Multi agent based intrusion detection inMANET", IEEE 2012.

[8] Monita waghengbam and ningrila marchang,"Intrusiondetection in MANET using fuzzy logic", IEEE 2012.

[9] X. Zhang, L. Jia, H. Shi, Z. Tang and X. Wang, "TheApplication of Machine Learning Methods to IntrusionDetection", Engineering and Technology (S-CET), 2012Spring Congress on, (2012), pp. 1-4.

[10] Robert Mitchell and Ing-Ray Chen, "Behavior RuleSpecification-based Intrusion Detection for SafetyCritical Medical Cyber Physical Systems", IEEETransactions on Dependable and Secure Computing(Volume:12 , Issue: 1), 2014

[11] Devikrishna K S and Ramakrishna B B, "An ArtificialNeural Network based Intrusion Detection System andClassification of Attacks", International Journal ofEngineering Research and Applications (IJERA) ISSN:2248-9622, Vol. 3, Issue 4, Jul-Aug 2013, pp. 1959-1964

[12] Vaishali Kosamkar, Sangita S Chaudhari, "ImprovedIntrusion Detection System using C4.5 Decision Treeand Support Vector Machine", (IJCSIT) InternationalJournal of Computer Science and InformationTechnologies, Vol. 5 (2) , 2014, 1463-1467