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

1Prakash Chandra, 2Prof. Umesh Kumar Lilhore

1M.Tech Research Scholar, 2Associate 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 is it 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 KDPCUP 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 of security management system for computers and networks. An intrusion detection system (IDS) inspects all outbound and inbound network action and find out the doubtful patterns that may point to network or system intrusion or attack from someone trying to crack into or conciliation of a system [7]. An ID gathers and observed information from different areas inside a network of systems to find out probable safety breaches, which contain together called intrusions (attacks exterior from the association) and misuse (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 have presenteda collection of research efforts on the use of data mining in computer security. In the context of security the data we are looking for the information whether an information security breach has been experienced [3]. This data could be collected in the perspective of discovering attacks or intrusions that aim to break the privacy and security of services, information in a system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity. There are 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 data mining concentrate heavily. There are two foldreasons for this first an IDS is very common and very popular and extremely critical activity. Second, large volume of the data on the network is dealing so this is an ideal condition for the data mining to use it. Thedata mining technology has the enormous benefits in the data extracting attributes and the rule, so it is significant to use data mining methods in the intrusion detection [4].

A significant problem of IDSs is how to efficiently divide the normal behavior and the abnormal behavior from a huge number of raw information’s attributes, and how to effectively generate automatic intrusion rules following composed raw data of the network. To accomplish this, different data mining methods must be studied, like classification, correlation analysis of data mining methods and so on [14]. The ever rising new intrusionor attacks type poses severe difficulties for their detection. The human labeling of the accessible network audit information instances is generally tedious, 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 out the doubtful patterns that may point to a network or system intrusion or attack from someone trying to crack into or conciliation a system [2].

An ID gathers and observed information from different areas inside a network of systems to find out probable safety breaches, 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
Data mining techniques can be used to detect intrusions. Applications of data mining have represented a collection of research efforts on the use of data mining in computer security. In the context of security of the data we are looking for the information whether an information security breach has been experienced [13]. This data could be collected in the perspective of discovering attacks or intrusions that aim to break the privacy and security of information in a system or alternatively, in the context of discovering evidence left in a computer system as part of criminal activity. There are four major categories of networking attacks: Denial of Service, Probing, and User to Root and Remote to Local.

2.1 HOW IDS WORKS?

Working of Intrusion detection systems is based on four-step approaches for the generalized working of IDS:

- **Collection of Data**: It involves collecting network traffic using particular software and thus helps to get the information about the traffic like types of packets, hosts and protocol details.

- **Selection of Features**: The collected data is substantially large because of the huge network traffic; we generate feature vectors that contain necessary information. In network-based intrusion detection, it can be IP header information, which consists of source and destination IP address, packet type, layer 4 protocol type and other flags.

- **Analysis**: The collected data is analyzed in this step to determine whether data is anomalous or not.

- **Action-IDS** alarm the system administrator that an attack has happened and it tells about the nature of the attack. IDS also participate in controlling the attacks by closing the network port and killing the processes.

3. DATA MINING & MACHINE LEARNING METHODS FOR IDS

The term data mining is used to describe the process of extracting useful information from the large databases. Data mining analyses the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand [3]. Hence datamining problems are considered as a data analysis problem. Data mining framework automatically detect patterns in our data set and use these patterns to find a set of malicious entries. Data mining techniques can detect patterns in large amounts of data, such as byte code and use these patterns 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 study that gives computers the ability to learn without being explicitly programmed.”

Machine learning focuses on classification and prediction, 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

From 1999, KDD’99 is the mainly frequent used dataset for the assessment of anomaly detection techniques [10]. This dataset is made by Stolfo et al. and is built based on the data 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 link records, each with about hundred bytes. The two weeks of the data have around 2 million connection records. KDD training dataset consists of just about 4,900,000 single connection vectors every of which encloses 41 features and is labeled as either an attack or normal, with precisely one definite attack type.

The simulated attacks plunge in one of the following four categories:

a) **Denial of Service Attack (DoS)**- DoS is an attack in which the attacker creates some memory or computing resource too full or too busy to handle genuine requests, or denies genuine users entrance to a machine.

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

c) **Remote to Local Attack (R2L)** - R2L attack occurs when an attacker whom the ability to launch packets to a machine over a network but who does not have an account on that machine develops several weaknesses to achieve local entrance as auser of that machine.
d) Probing Attack- Prob is an effort to collect information about a network of computers for the perceptible reason of circumventing its safety controls.

5. WEKA TOOL FOR IDS

WEKA is a Tool for Data Mining and Machine Learning which 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 three algorithms to find association rules.

![Figure 5.1 WEKA Tool](image)

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 for visualization.

6. EXISTING DM & ML FOR IDS

Bayes Classifier: They are also known as Belief Networks,belong to the family of probabilistic Graphical Models (GMS). These graphical models are used to represent knowledge about uncertain domains, Random variables are denoted by nodes in the graph and probabilistic dependencies are assigned as weights to the edges connecting corresponding random variable nodes [5].

The types of classifiers are based upon the idea of predicting the class on the basis of value of members of the features. This category has 13 classifiers out of which 3 classifiers (BayesNet, NaïVeBayes and NaïVeBayes Updateable) are incompatible with the chosen dataset.

- **Functions Classifier**: Functional Classifier uses the concept of neural network and regression. They maps input data to output. There are eighteen classifiers under this category out of which only RBF Network and SMO classifiers are compatible with our dataset RBF classifiers can model any nonlinear functions easily [5]. It does not use raw input data. The processing of RBF Networks is likeneural networks i.e. iterative in nature. The problem with RBF is the tendency to over train the model.

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

- **Meta Classifier**: These types of classifiers find the optimal set of attributes to train the base classifier with these parameters; this trained base classifier will be used for further predictions [8]. There are 26 classifiers under this category out of which 21 are compatible with our dataset: AdaBoost M 1, LogistBoost, Attribute Selection Classifier, Bagging, Daging 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 of which no classifier is compatible with our dataset. Mi classifier is variant of supervised learning technique. It has multiple instances in an example but can only observe one class. These types of classifiers are originally made available through a separate software package.

- **Misc Classifier**: There are three classifiers under this category 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, association rules are used for correct prediction of class among all the attributes and those correct predictions are called coverage and it is expressed in terms of percentage of accuracy [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 of which 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 in which each node denotes a test on attribute value and each branch represents an outcome of the test [9]. They are also known as Decision Trees. The tree leaves represent the class that is predicted. They design a model that is both predictive and descriptive. There are 16 classifiers under this category out of which 10 are compatible with our chosen dataset 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 different detection methodologies, approaches and techniques for Intrusion Detection System (IDS) used in WEKA using Data mining approaches. Each technique has its own superiority and 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 rule techniques, etc. For basic knowledge of Machine Learning Approaches WEKA tool and various classification algorithms have been discussed. At last the KDD cup-99 data set which is wildly used in anomaly detection and some real reason for research scope in 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


[6] Rajendra V. Boppana, Senior Member, IEEE, and XuSu, Member, IEEE A Distributed ID for Ad Hoc Networks, 26th International Conference on Advanced Information Networking and Applications 2012.


