International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

A survey of IDS classification using KDD CUP 99 dataset in WEKA

Ms. Urvashi Modi Prof. Anurag Jain

Abstract— Intrusion detection systems (IDSs) are based on two fundamental approaches first the recognition of anomalous activities as it turns from usual behavior and second misuse detection by observing those "signatures" of those recognized malicious assaults and classification vulnerabilities. Anomaly (behavior-based) IDSs presume the difference of normal behavior beneath attacks and achieve abnormal recognition evaluated with predefined system or user behavior reference model. This paper is to provide a detailed survey of intrusion detection techniques. It represents a study of Intrusion Detection and data mining techniques to classify different Intrusion attacks. This survey also focuses on WEKA (Waikato Environment for Knowledge Analysis) Tool and its various algorithms of classification. Lastly In this survey we tend to explain the mostly used dataset in network security research KDDCUP 99 and its various components. Finally we conclude our survey with few real research proposals which will be open issues for searchers.

Key words: IDS, Classification, Clustering, Machine Learning, WEKA, KDD Cup 99

1 INTRODUCTION

Computer based information Systems are becoming an important part of so many organizations. By connecting our computer to the Internet, we increase the risk that someone may install malicious programs and use it to attack other machines on the Internet by controlling it remotely. Computer security is the ability to protect a computer system and its resources in reference to Confidentiality, Integrity and Availability [1]. In order to protect from the computer threats, various protocols and firewalls are used. Confidentiality requires that information can be accessible only to those authorized for it; integrity requires that information remain unchanged without any modification by malicious attempts, and availability means the computer system and its resources always available to authorized users when they need it. By this definition, a computer system is said to be reliable if confidentiality, integrity and availability is a part of its security requirements [2].

A computer system is said to be secure if it can protects its data and resources from unauthorized access, modification, and denial of use. Intrusion is a type of attack that tries to deny the security aspects of a computer system. It is normally considered that intrusions illustrate something diverges from the ordinary pattern, and that any unknown intrusion will present patterns more similar to known intrusion [3]. Intrusion Detection is a major focus of research in the security of computer systems and networking. An intrusion detection system (IDS) is used to detect unauthorized intrusions i.e. attacks into computer systems and networks. These systems are known to generate alarms (alerts).the following general terms used for detection and identification of attack and non-attack behavior.

- True positive (TP): The amount of attack detected when it is actually attack;
- True negative (TN): The amount of normal detected when it is actually normal;
- False positive (FP):The amount of attack detected when it is actually normal called as false alarm;
- False negative (FN): The amount of normal detected when it is actually attack, namely the attacks which can be detected by intrusion detection system.

Based on the above assumption intrusion can be defined as a data analysis problem. Patterns of the intrusions and patterns of the normal behavior can be computed using data mining. . Since a large volume of network traffic that requires processing, we use data mining techniques. To apply data mining techniques in intrusion detection, preprocessing is the first step to be done on the collected data. Then convert the data into a particular format for the mining process. Next, the formatted data is used for classification and clustering.

The classification model can be rule based, decision tree based, Bayesian network based or neural network based. Data mining technique provides the guarantee that no intrusion will be missed while checking the real time data in the network, thus ensuring the accuracy and efficiency in the detection process. Data mining techniques also helps in intrusion prevention mechanisms. They can detect both known and previous unknown patterns of attacks.

This paper is organized as follows: Section 2 gives the details of intrusion detection and the general working strategy of Intrusion Detection Systems .Section 3 gives the details of data mining concepts and the system design based on Data Mining Intrusion Detection Pattern. Section 4 gives the details of different data mining techniques and explains how each technique helps in detecting intrusions. WEKA tool and various classification algorithms have been discussed in section 5. Section 6 introduces the KDDCUP99 data set which is wildly used in anomaly detection. In

- *Urvashi Modi is currently pursuing masters degree program in Computer Science & Engineering in RITS in RGPV University, M.P. INDIA, E-mail: urvashimodi90@gmail.com*
- *Anurag Jain is head of department in Computer Science & Engineering in RITS in RGPV University, M.P. INDIA E-mail: anurag.akjain@gmail.com*

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

section 7 we give some real reason for research scope in this field. At last section we conclude our paper.

2 INTRUSION DETECTION

Intrusion is a set of actions that attempt to compromise the integrity, confidentiality, or availability of any resource on a computing platform [4]. An intrusion detection system (IDS) is a combination of hardware and software that detect intrusions in the network. IDS can monitor all the network activities and hence can detect the signs of intrusions. The main objective of IDS is to alarm the system administrator that any suspicious activity happened. There are two types of Intrusion detection techniques:

- Anomaly Detection: Detecting malicious activities based on deviations from the normal behavior are considered as attacks. Although it can detect unknown intrusions, rate of missing report is low.
- Misuse Detection: Detecting intrusions based on a pattern for the malicious activity [5]. It can be very helpful for known attack patterns. Also rate of missing report is high.

One disadvantage of Misuse Detection over Anomaly Detection is that it can only detect intrusions which contain known patterns of attack.

An intrusion detection system (IDS) monitors the activities of a given environment and decides whether these activities are malicious or normal based on system integrity, confidentiality and the availability of information resources [6]. When building IDS one needs to consider many issues, such as data collection, data pre-processing, intrusion recognition, reporting, and response.

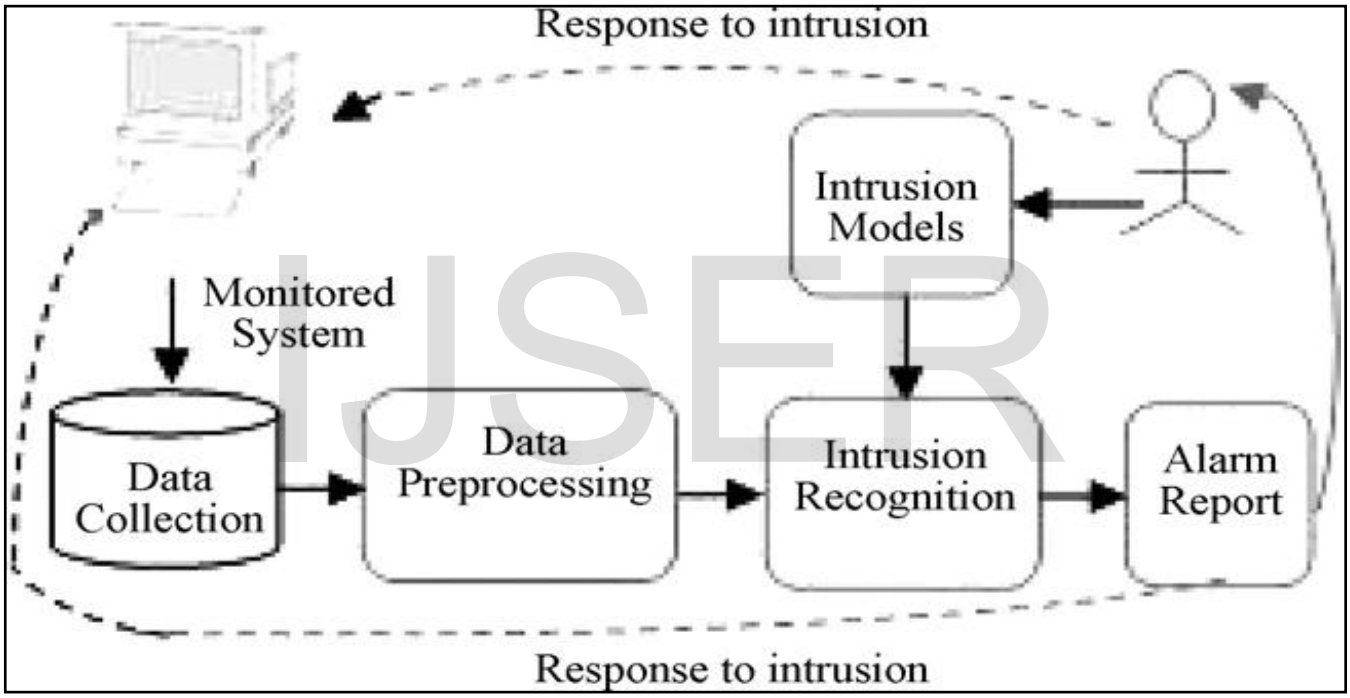

Fig.1: Organization of a generalized intrusion detection system.

Among them, intrusion recognition is most vital. Audit data is compared with detection models, which describe the patterns of intrusive behavior, so that both successful and unsuccessful intrusion attempts can be identified [7]. Fig. 1 depicts the organization of IDS where solid lines indicate data/control flow while dashed lines indicate responses to intrusive activities [7].

2.1 Working of Intrusion detection systems

Authors of [4] presented a four step approach for the generalized working of IDS:

- **Data collection:-** 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.
- **Feature Selection:-** The collected data is substantially large because of the huge network traffic; we generate feature vectors that contain only 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. Here we use various methods for detecting intrusions.
- **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 or killing the processes.

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

3 DATA MINING TECHNOLOGY

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 [8]. Hence data mining 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 binaries.ie, Data mining techniques can detect patterns in large amount of data, such as byte code and use these patterns to detect future instances in similar data.

In intrusion detection system, information comes from various sources like host data, network log data, alarm messages etc. Since the variety of different data sources is too complex, the complexity of the operating system also increases. Also network traffic is huge, so the data analysis is very hard. The data mining technology have the capability of extracting large databases; it is of great importance to use data mining techniques in intrusion detection. By applying data mining technology, intrusion detection system can widely verify the data to obtain a model, thus helps to obtain a comparison between the abnormal pattern and the normal behavior pattern. Manual analysis is not required for this method. One of the main advantages is that same data mining tool can be applied to different data sources. Main problem in intrusion detection is effective separation of the attack patterns and normal data patterns from a large number of network data and effective generation of the automatic intrusion rules after collected raw network data.

For this purpose several methods of data mining are used in such type of classification, clustering and association rule mining etc. Some Data Mining based Misuse detection model of Intrusion Detection Systems are

- Java Agents for Meta learning (JAM)
- Mining Audit Data for Automated Models for Intrusion Detection (MADAM ID)
- Automated Discovery of Concise Predictive Rules for Intrusion Detection

4 DATA MINING TECHNIQUES AND INTRUSION DETECTION

Data Mining is used in variety of applications that requires data analysis. Now a day's data mining techniques plays an important role in intrusion detection systems. Different data mining techniques like Classification, Clustering and Association rules are frequently used to acquire information about intrusions by observing network data. This section describes different data mining techniques that help in detecting intrusions.

Classification: Classification is a form of data analysis which takes each instance of a dataset and assigns it to a particular class. It extracts models defining important data classes. Such models are called classifiers [9]. A classification based IDS will classify all the network traffic into either normal or malicious. Data classification consists of two steps – learning and classification. A classifier is formed in the learning step and that model is used to predict the class labels for a given data in the classification step. In analysis of Classification the end-user/analyst requires to know ahead of time how the classes are defined [1]. Each record in the dataset already has assessment for the attribute used to define the classes. The main aim of a classifier is not only to explore the data to discover different classes, but also to find how new records should be arranged into classes. Classification helps us to categorize the data records in a predetermined set .It can be used as attribute to label each record and for distinguishing elements belonging to the normal or malicious class [1]. Different types of classification techniques are decision tree induction, Bayesian networks-nearest neighbor classifier, genetic algorithm and fuzzy logic.

As compared to the clustering technique, classification technique is less efficient in the field of intrusion detection. The main reason for this is the enormous amount of data needed to be collected to use classification. To classify the dataset into normal and abnormal, large amount of data is required to analyze its proximity. Classification method can be useful for both misuse detection and anomaly detection, but it is more commonly used for misuse detection.

Clustering Since the amount of available network data is too large, human labeling is time-consuming, and expensive. Clustering is the process of labeling data and assigning into groups.ie, Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of members from the same cluster are quite similar and members from the different clusters are different from each other. Hence clustering methods can be useful for classifying network data for detecting intrusions. Clustering algorithms can be classified into four groups: partitioning algorithm, hierarchical algorithm, density-based algorithm and grid based algorithm [10].

Clustering techniques can discovers complex intrusions over a different time period. Clustering is an unsupervised machine learning mechanism for discovering patterns in unlabeled data with many dimensions. Clustering is the collection of patterns based on similarity. Patterns within a cluster are equivalent to each other, but they are different with other clusters. Therefore patterns that are far from any of these clusters indicate that an unusual activity happened. That can be part of a new attack. Clustering can be applied on both Anomaly detection and Misuse detection.

5 WEKA (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS)

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 [11]. WEKA is a set of Machine Learning and Data Mining algorithms. This WEKA software is programmed in JAVA language and it has a GUI Interface to interact with data Files. With 49 data pre-processing tools WEKA tool contains 76 classification algorithms, 15 attribute evaluators and ten search algorithms for feature selection [12]. There are three algorithms to find association rules. It also has three

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

Graphical User Interfaces: "The Explorer", "The Experimenter" and "The Knowledge Flow." The file format to store data in WEKA is ARFF. Meaning of ARFF is Attribute Relation File Format. It also includes tools for visualization. It has a several panels that can be used to perform precise tasks. WEKA has the ability to expand and contain the new algorithms for Machine Learning in it. These expanded algorithms can directly be applied to dataset.

5.1 CLASSIFICATION ALGORITHMS USED IN WEKA

Classification algorithms also known as classifiers are used to classify the network traffic as normal or an intrusion. There are basically eight categories of classifiers and each category contains different machine learning algorithms. In this section these categories have been briefly introduced.

Bayes Classifier: They are also known as Belief Networks, belongs to the family of probabilistic Graphical Models (GM'S) [13], 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. These 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 (Bayes Net, NaiVeBayes and NaiVeBayes Updateable) are compatible with the chosen dataset.

Function Classifier: Functional Classifier uses the concept of neural network and regression [11]. 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. It does not use raw input data. The processing of RBF Networks is like neural networks i.e. iterative in nature. The problem with RBF is the tendency to over train the model [14].

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. Lazy classifiers are mainly used for classification on data streams [15], 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 [16], This trained base classifier will be used for further predictions. There are 26 classifiers under this category 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 , Raced Incremental Logi Boost , Random Committee , Stacking , Stacking C.

Mi Classifier: Mi stands for Multi- Instance Classifiers. This category of classifier consists of 12 classifiers out 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 [17]. 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 Hyperpipes and VFI.

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 as coverage and it is expressed in terms of percentage of accuracy. They may predict more than one conclusion. Rules are mutually exclusive. These are learned one at a time [11], 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, Zero R, 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. They are also known as Decision Trees. The tree leaves represents the classes that are 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 chose dataset that are: Decision Stump, j48, j48 graft, LAD Tree, NB Tree, REP Tree, Random Forest, Simple Cart, Random Tree, User Classifier.

6 KDD CUP 99 DATA SET DESCRIPTION

From1999, KDD'99 [18] is the mainly frequent used dataset for the assessment of anomaly detection techniques. This dataset is made by Stolfo et al. [19] and is built based on the data taken in DARPA'98 IDS assessment program [20]. 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 test 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:

1) 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.
2) User to Root Attack (U2R): U2R is a class of exploit in which the attacker creates entrance to a standard 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.

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

3) 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 weakness to achieve local entrance as a user of that machine.
4) Probing Attack: Prob is an effort to collect information about a network of computers for the perceptible reason of circumventing its safety controls.

It is necessary to maintain it in awareness that the experiment data is not from the same likelihood division as the training data, and it also contains precise attack categories not in the training data which build the task more realistic. Various intrusion specialists consider that most novel attacks are alternatives of known attacks and the signature of known attacks can be adequate to grab novel alternatives. The datasets include a total 24 training attack types, with an additional 14 types in the experiment data only.

KDD'99 features can be classified into three groups:

1) Basic features: this class encapsulates all the features that can be extracted from a TCP/IP connection. The majority of these feature foremost to an understood delay in recognition.

2) Traffic features: this class contains features that are computed with deference to a window interval and is separated into two groups:

a) "Same host" features: look at only the connections in the previous two seconds that have the same target host as the present connection, and compute statistics related to protocol activities, service, etc.
b) "Same service" features: observe only the connections in the previous two seconds that have the similar service as the present connection.

The two aforesaid types of "traffic" features are described as time-based. However, there are a number of slow probing attacks that examine the hosts (or ports) using a much superior time interval than two seconds, for example, one in every minute. As a consequence, these attacks do not create intrusion patterns with an occasion window of two seconds. To solve this difficulty, the "same host" and "same service" features are re-computed but based on the connection window of 100 connections rather than a time window of two seconds. These features are described connection-based traffic features

3) Content features: unlike the majority of the Probing and DoS attacks, the U2R and R2L attacks do not have several intrusion common sequential patterns. This is because the DoS and Probing attacks engage a lot of connections to several host(s) in a extremely short interval of time, though the U2R and R2L attacks are embedded in the data portions of the packets, and usually involves only a solitary connection. To identify these kinds of attacks we require several features to be capable to appear for suspicious behavior in the data portion, e.g., number of failed login attempts. These features are called content features

Evaluation of KDD99 dataset

The major aim of this survey paper is to verify the most excellent technique to organize and evaluate the KDD99 dataset to obtain highest precision in the classification of attacks and in training time, and to explore some other better technique to recognize each type of four attacks (Probe, Dos, U2R, R2L) in order to help the job of alternative for researchers in the future.

Most excellent performing occurrences of all the 20 algorithms used in WEKA were assessed on the KDD99 dataset by authors of [22]. Simulation outcomes are specified in the Table 1 to evaluate all the classifiers, they used TP and FP for every algorithm. These constraints will be the majority significant criteria for the classifier to be measured as the best algorithm for the specified attack category. In addition, it is also at equivalent significance to record Percentage of Successful (PSP) and Training Time (TT) of every algorithm in the Table 2. In the selection process, one algorithm will be disqualified if its PSP is too less, regardless of its outstanding presentation in one exact attack category. TT on the other hand, will give them the thought about which algorithm can be implemented in a real-time network intrusion detection system. Graphical representation of table 2 is shown in fig. 2.

TABLE 1: PERFORMANCE COMPARISON OF ALGORITHM USED IN WEKA

Seq.	Classifier	Percentage of successful Prediction (PSP) %	Trainig Time (TT)
1	K-Means	78.7	70.7
2	NEA	92.22	10.63
3	FCC	89.2	56.2
4	ID3	72.22	120
5	J48	92.06	15.85
6	PART	45.67	169
7	NBTree	92.28	25.88
8	SVM	81.38	222.28
9	Fuzzy logic	91.8	873.9
10	naïve Bayes	78.32	5.57
11	BayesNet	90.62	6.28
12	Decision Table	91.66	66.24
13	Random Forest Classifier	92.81	491
14	Jrip	92.30	207.47
15	OneR	89.31	3.75
16	MLP	92.03	350.15
17	SOM	91.65	192.16
18	GAU	69.9	177.4
19	MARS	**96.5**	67.9
20	Apriori	87.5	18

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

TABLE 2: EVALUATION OF 20 ALGORITHMS USED IN WEKA

Seq.	Classifier	Metric	DoS	Probe	U2R	R2L	Training Set Size
1	K-Means (Qiang W.V.,2004)	TP	87.6	97.3	29.8	6.4	2,776
		FP	2.6	0.4	0.4	0.1	
2	NEA (Maheshkumar S., 2002)	TP	96.7	72.4	22.3	7.8	1,074,991
		FP	0.8	0.2	0.1	0.6	
3	FCC (Qiang W.V.,2004)	TP	91.6	77.8	12.7	27.8	2,776
		FP	**0.03**	0.023	0.13	0	
4	ID3 (Amanpreet C., 2011)	TP	74.4	57.14	20	6.25	145,586
		FP	1.71	2.5	3.1	1.1	
5	J48 (Huy A.N., 2008)	TP	96.8	75.2	12.2	0.1	49,596
		FP	1	0.2	0.1	0.5	
6	PART (Mohammed M.M., 2009)	TP	97.0	80.8	1.8	4.6	444,458
		FP	0.7	0.3	0.5	0.01	
7	NBTree (Huy A.N., 2008)	TP	97.4	73.3	1.2	0.1	49,596
		FP	1.2	1.1	0.1	0.5	
8	SVM (Huy A.N., 2008)	TP	96.8	70.1	15.7	2.2	49,596
		FP	1.11	0.5	**0.01**	**0**	
9	Fuzzy logic (Shanmugaradtru R, 2011)	TP	94.8	**98.4**	69.6	92.1	54,226
		FP	5.5	1.8	6.7	10.7	
10	naïve Bayes (Huy A.N.,2008)	TP	79.2	94.8	12.2	0.1	49,596
		FP	1.7	13.3	0.9	0.3	
11	BayesNet (Huy A.N., 2008)	TP	94.6	83.8	30.3	5.2	49,596
		FP	0.2	0.13	0.3	0.6	
12	Decision Table (Yeung d.Y., 2002)	TP	97.0	57.6	32.8	0.3	15,919
		FP	10.7	0.4	0.3	0.1	
13	Random Forest classifier (Yeung D.Y., 2002)	TP	**99.2**	98.2	86.2	54.0	15,919
		FP	0.05	**0.01**	0.02	0.09	
14	Jrip (Huy A.N., 2008)	TP	97.4	83.8	12.8	0.1	49,596
		FP	0.3	0.1	0.1	0.4	
15	OneR (Huy A.N.,2008)	TP	94.2	12.9	10.7	10.7	49,596
		FP	6.8	0.1	2	0.1	
16	MLP (Huy A.N.,2008)	TP	96.9	74.3	20.1	0.3	49,596
		FP	1.4	0.1	0.1	0.5	
17	SOM (Huy A.N.,2008)	TP	96.4	74.3	13.3	0.1	49,596
		FP	0.8	0.3	0.1	0.4	
18	GAU (Maheshkumar S., 2002)	TP	82.4	90.2	22.8	9.6	1,074,991
		FP	0.9	11.3	0.05	0.1	
19	MARS (Sriniras M.,2002)	TP	94.7	92.32	**99.7**	**99.5**	11,982
		FP	8.9	12.2	22.4	17.9	
20	Apriori (Mohammed M.M.,2009)	TP	87.9	76.23	12.3	30.6	444,458
		FP	0.67	1.7	8.9	23.8	

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

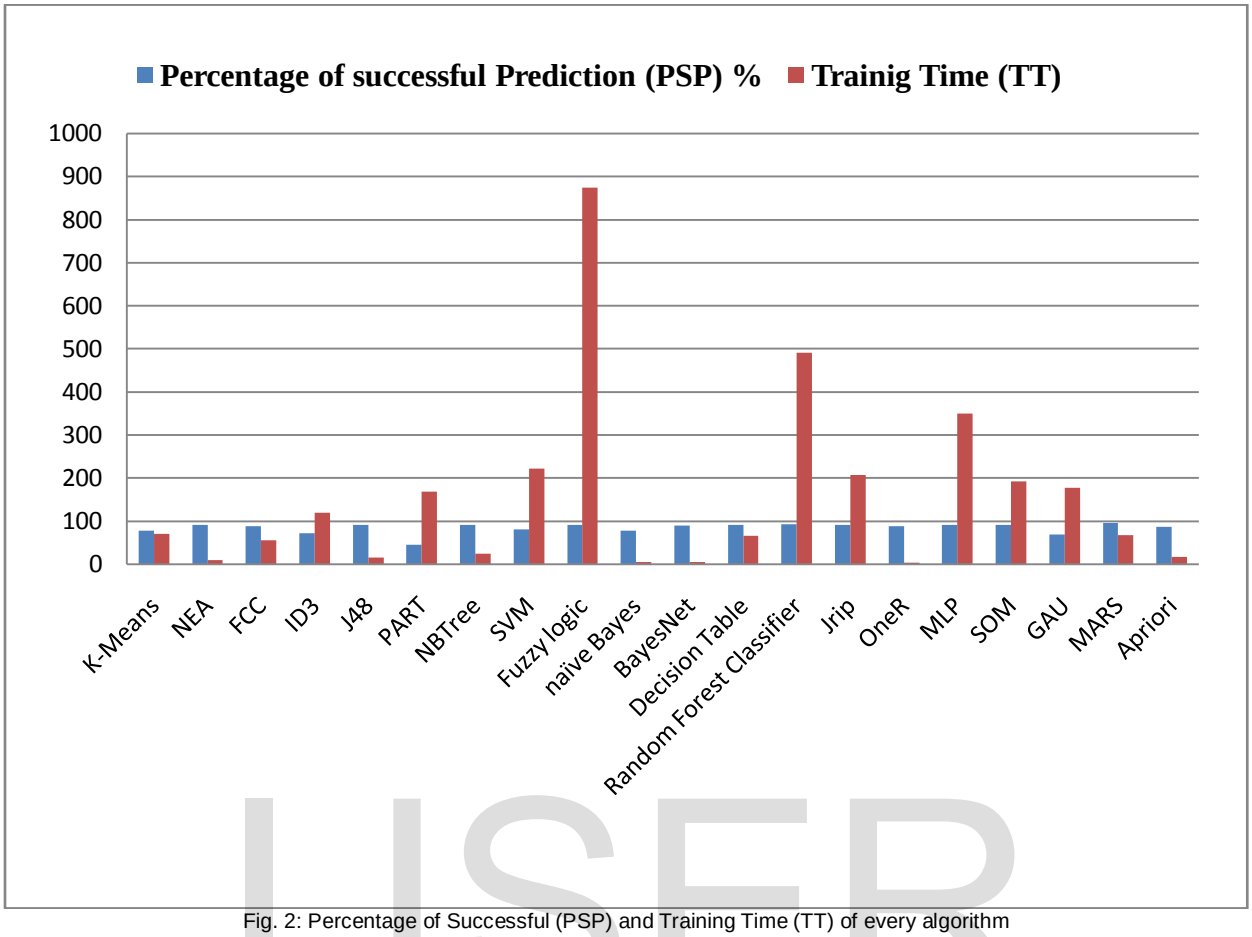

Fig. 2: Percentage of Successful (PSP) and Training Time (TT) of every algorithm

Although, not any of the assessed to most machine learning classifier algorithms was capable to perform detection of U2R and R2L attack categories appreciably (no more than averagely 27% detection for U2R and 18% for R2L category). It is logical to declare that machine learning algorithms employed as classifiers for the KDD CUP 1999 dataset don't offer a lot assure for detecting U2R and R2L attacks within the misuse detection context. But Multivariate Adaptive Regression Splines (MARS) give improved results in the detection of U2R and L2R attacks. The decision tree to obtain an improved intrusion detection rates up higher than the 96% level and less false alerts from the rest of classifier data mining algorithms

7 RESEARCH SCOPE

The effectiveness of IDS depends on the capability to detect any abnormal activity in the target system, which is called the sensitivity of IDS. If the IDS are more sensitive, the security of the system would be tighter. For Making the IDS more sensitive means to apply tighter signature rules or to be less tolerant to anomalies. As a result, the IDS become more sensitive to its input and generate a lot of alarms each day, even though most of the examined events are not illegal events.

Due to large volumes of IDS false alarms, it is a quite tough task for the security officers to investigate manually which are the real suspicious alarms and thereafter take proper action against them. Even sometimes, some real suspicious alarms are ignored mistakenly by the security officer due to large volumes of false alarms and thereby mistakenly interpret a real alarm to be a false alarm. This is the most dangerous situation when a real instance of an attack is ignored by the security officer and thus the IDS become useless though its functionality remains the same.

Presently used clustering techniques have some major limitations. First, they require mechanisms to validate the validity of the raised alerts. Unverified alerts often disgrace the superiority of alerts hence giving unreliable results. Secondly, the existing alert clustering techniques completely based on the basic information provided by the alert features. As we are familiar with the fact that unverified alerts have undesirable characteristics of noisy alerts such as incomplete information.

8 CONCLUSION

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. In This paper we give the details of intrusion detection and the general working strategy of Intrusion Detection Systems. We represent in depth of data mining concepts and the system design based on Data Mining Intrusion Detection Pattern. We did Study of

International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015
ISSN 2229-5518

different data mining techniques and explains how each technique helps in detecting intrusions. For basic Knowledge of Machine Learning Approaches WEKA tool and various classification algorithms have been discussed. At last the KDDCUP99 data set which is wildly used in anomaly detection and some real reason for research scope in this field is given.

REFERENCES

[1] Chang-Tien Lu,Arnold P.Boedihardjo,Prajwal Manalwar, "Exploiting efficient data mining techniques to enhance Intrusion Detection Systems" 0-7803-9093-8/05/$20.00 2005 IEEE,512-517.

[2] Sandeep Kumar, "Classifications and Detections of computer intrusions", A Thesis Submitted to the Faculty of Purdue University,1995.

[3] Christine Dartigue, Hyun Ik jang,and Wenjun Zeng, "A New Data-Mining Based Approach for Network Intrusion Detection",2009 Seventh Annual Communication Networks and Services Research Conference IEEE,pg 372-377.

[4] Labib, Khaled. "Computer security and intrusion detection." Crossroads TheACM students magazine.11, no. 1 (2004): 2-2.

[5] Ektefa, Mohammadreza, Sara Memar, Fatimah Sidi, and Lilly Suriani Affendey. "Intrusion detection using data mining techniques." In Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on, pp. 200-203. IEEE, 2010.

[6] A.N. Toosi, M. Kahani, A new approach to intrusion detection based on an evolutionary soft computing model using neuro-fuzzy classifiers, Computer Communications 30 (2007) 2201–2212.

[7] S.X. Wu, W. Banzhaf, The use of computational intelligence in intrusion detection systems: a review, Applied Soft Computing Journal 10 (2010) 1–35.

[8] Wang, Xiao-bin, Guang-yuan Yang, Yi-chao Li, and Dan Liu. "Review on the application of artificial intelligence in antivirus detection system i." InCybernetics and Intelligent Systems, 2008 IEEE Conference on, pp. 506-509. IEEE, 2008.

[9] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining, southeast asia edition: Concepts and techniques. Morgan kaufmann, 2006.

[10] Jianliang, Meng, Shang Haikun, and Bian Ling. "The application on intrusion detection based on k-means cluster algorithm." In Information Technology and Applications, 2009. IFITA'09. International Forum on, vol. 1, pp. 150-152. IEEE, 2009.

[11] Dash, Ranjita Kumari. "Selection of the Best Classifier from Different Datasets Using WEKA." In International Journal of Engineering Research and Technology, vol. 2, no. 3 (March-2013). ESRSA Publications, 2013.

[12] Nguyen, Huy Anh, and Deokjai Choi. "Application of data mining to network intrusion detection: classifier selection model." In Challenges for Next Generation Network Operations and Service Management, pp. 399-408. Springer Berlin Heidelberg, 2008.

[13] Ben-Gal, I., F. Ruggeri, F. Faltin, and R. Kenett. "Bayesian Networks, Encyclopedia of Statistics in Quality and Reliability." (2007).

[14] Panda, Mrutyunjaya, and Manas Ranjan Patra. "A comparative study of data mining algorithms for network intrusion detection." In Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on, pp. 504-507. IEEE, 2008.

[15] Panda, Mrutyunjaya, and Manas Ranjan Patra. "Ensembling rule based classifiers for detecting network intrusions." In Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on, pp. 19-22. IEEE, 2009.

[16] Neethu, B. "Classification of intrusion detection dataset using machine learning approaches." International Journal of Electronics and Computer Science Engineering (2012): 1044-1051.

[17] Multi-instance Classifiers at http://weka.wikispaces.comlMultiinstance+ classification.

[18] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html

[19] Stolfo, Salvatore J., Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip K. Chan. "Cost-based modeling for fraud and intrusion detection: Results from the JAM project." In DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings, vol. 2, pp. 130-144. IEEE, 2000.

[20] Lippmann, Richard P., David J. Fried, Isaac Graf, Joshua W. Haines, Kristopher R. Kendall, David McClung, Dan Weber et al. "Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation." InDARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings, vol. 2, pp. 12-26. IEEE, 2000.

[21] MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corp ora/ideval/ index.html,

[22] Al-mamory, Safaa O., and Firas S. Jassim. "Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set." Journal of Babylon University/Pure and Applied Sciences/ No.(8)/ Vol.(21): 2013 pp 2663-2681.