Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

IDS Approach Using Data Mining Tool WEKA

Aakshi Choudhary*
Research Scholar MIET Department of Computer Sc & Engineering
Kurukshetra University
Haryana, India

Sarbjit Kaur
Assistant Professor MIET Department of Computer Sc & Engineering
Kurukshetra University
Haryana, India

Abstract*— The demand of the Intrusion Detection methods and algorithms have also been asked to improve. By analyzing the technology of Intrusion Detection System and Data mining in this paper, the author uses K-mean algorithm which is the classic of association rules in Web-based Intrusion Detection System and applies the rule base generated by the K-mean algorithm to identify a variety of attacks, improves the overall performance of the detection system. Hence intrusion is used as a key to compromise the integrity, availability and confidentiality of a computer resource. The Intrusion Detection System (IDS) plays a vital role in detecting anomalies and attacks in the network. In this work, data mining concept is integrated with an IDS to identify the relevant, hidden data of interest for the user effectively and with less execution time.

***Keywords*— IDS, Clustring , Weka ,Data set**

I. INTRODUCTION

Data Mining is the use of algorithms to extract the information and patterns derived by the knowledge discovery in databases process. Data mining is being used to clean, classify, and examine large amount of network data to correlate common infringement for intrusion detection. The main reason for using Data Mining Techniques for Intrusion Detection Systems is due to the enormous volume of existing and newly appearing network data that require processing. The amount of data accumulated each day by a network is huge. Several Data Mining techniques such as clustering, classification, and association rules are proving to be useful for gathering different knowledge for Intrusion Detection. Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other. Therefore clustering methods can be useful for classifying log data and detecting intrusions. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. In many data mining applications that address classification problems, feature and model selection are considered as key tasks. That is, appropriate input features of the classifier must be selected from a given set of possible features and structure parameters of the classifier must be adapted with respect to these features and a given data set.

II. IDS

A. Intrusion Detection System (IDS):-

is an important detection used as a countermeasure to preserve data integrity and system availability from attacks. Intrusion Detection Systems (IDS) is a combination of software and hardware that attempts to perform intrusion detection. It is a process of gathering intrusion related knowledge occurring in the process of monitoring the events and analyzing them for sign or intrusion. It raises the alarm when a possible intrusion occurs in the system. The network data source of intrusion detection consists of large amount of textual information, which is difficult to comprehend and analyze. The main motivation behind using intrusion detection in data mining is automation. Pattern of the normal behavior and pattern of the intrusion can be computed using data mining. To apply data mining techniques in intrusion detection, first, the collected monitoring data needs to be preprocessed and converted to the format suitable for mining processing. Next, the reformatted data will be used to develop a clustering or classification model.

The classification model can be rule-based, decision-tree based, association-rule based, Bayesian-network based, or neural network based. Intrusion Detection mechanism based on IDS are not only automated but also provides for a significantly elevated accuracy and efficiency. Unlike manual techniques, Data Mining ensures that no intrusion will be missed while checking real time records on the network. Credibility is important in every system. IDS are now becoming important part of our security system, and its credibility also adds value to the whole system. Data mining techniques can be applied to gain insightful knowledge of intrusion prevention mechanisms. They can help detect new vulnerabilities and intrusions, discover previous unknown patterns of attacker behaviors, and provide decision support for intrusion management. Intrusion detection is detection of intrusion behavior, it collects information of the key part of

computer network and system, then analyzes them to detect whether occur the action of disobey security strategy. Intrusion Detection System (IDS) is the software or combination of software and hardware to detect intrusion behavior. IDS can examine intrusion attack before system is damaged, and make use of alerting and defense system to deport the intrusion attack. In the process of intrusion attack,

B. Methods of Intrusion Detection:-

1.Anomaly detection:-

Anomaly detection defines or summarizes pattern of user normal behavior It assumes that an intrusion will always reflect some deviations from normal patterns. When there is major difference between user's operation and normal behavior pattern, user's behavior is regarded as intrusion attack.

2 -Anomaly detection is divided into two types:-

2.1- Static Anomaly Detection

Static anomaly detector is based on the assumption that there is a portion of the system being monitored that does not change. Usually, static detectors only address the software portion of a system and are based on the assumption that the hardware need not be checked. The static portion of a system is the code for the system and the constant portion of data upon which the correct functioning of the system depends. For example, operating systems software and data to bootstrap a computer never change. If the static portion of the system ever deviates from its original form, an error has occurred or an intruder has altered the static portion of the system. Therefore static anomaly detectors focus on integrity checking.

2.2- Dynamic Anomaly detection

Dynamic Anomaly detector typically operates on audit records or on monitored networked traffic data. Audit records of operating systems do not record all events that is recorded in the audit will be observed and these events may occur in a sequence. In distributed systems, partial ordering of events is sufficient for detection. In other cases, the order is not directly represented: only cumulative information, such as cumulative processor resources used during a time interval, is maintained. In this case, thresholds are defined to separate normal resources consumption from anomalous resources consumption.

III. CLUSTERING ALGORITHMS:

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar, and the members from different clusters are quite different from each other These groups can be used to increase the performance of existing classifiers. High quality clusters can also assist human expert with labeling. A cluster is 100% pure if it contains only data instances from one category Until now,the clustering algorithms can be categorized into four main groups:partitioning algorithm,hierarchical algorithm, density-based algorithm and grid-based algorithm.Partitioning algorithms construct a partition of a database of N objects into a set of K clusters.Usually they start with an initial partition and then use an iterative control strategy to optimize an objective function.

A. Clustering techniques can be categorized into the following classes: -

- **Pairwise clustering** (i.e., similarity based clustering) unifies similar data instances based on a data-pairwise distance measure.

- **Central clustering**, also called centroid-based or model-based clustering, models each cluster by its "centroid". In terms of runtime complexity, centroid-based clustering algorithms are more efficient than similarity-based clustering algorithm.Clustering discovers complex intrusions occurred over extended periods of time and different spaces, correlating independent network events. The sets of data belonging to the cluster are modeled according to pre-defined metrics and their common features. It is used to detect hybrids of attack in the cluster. Clustering is an unsupervised machine learning mechanism for finding patterns in unlabeled data with many dimensions. K-means clustering is used to find natural groupings of similar alarm records. Records that are far from any of these clusters indicate unusual activity that may be part of a new attack. The network data available for intrusion detection is primarily categorical with attributes having a small number of unordered values

IV. IV.RELATED WORK

A. Su-Yun Wua et al:-

With popularization of internet, internet attack cases are increasing, and attack methods differs each day, thus information safety problem has became a significant issue all over the world. Nowadays, it is an urgent need to detect, identify and hold up such attacks effectively. The research intends to compare efficiency of machine learning methods in intrusion detection system, including classification tree and support vector machine, with the hope of providing reference for establishing intrusion detection system in future. Compared with other related works in data mining-based intrusion detectors, we proposed to calculate the mean value via sampling different ratios of normal data for each measurement, which lead us to reach a better accuracy rate for observation data in real world. We compared the accuracy, detection rate, false alarm rate for four attack types. More over, it shows better performance than KDD Winner, especially for U2R type and R2L type attacks.

B. Li Hanguang et al:-

With rapidly development of Internet, especially the wide open of the Internet, more and more systems encounter the invasion threatens. Therefore, safeguards of the computer system, the network system as well as the entire information infrastructure security have become an urgent question. Besides the traditional firewall isolation technology, in the security domain another important technology and the research direction is Intrusion Detection [1]. At present, the method of the Intrusion Detection technology is mainly paused at anomaly detection
and misuse detection. It is based on the foundation that any kind of invasion could be detected for deviating from normal state and the expected system and user's activities regulation. Misuse detection technology is an Intrusion Detection technology based on the knowledge, it mainly established on
accumulation of past invasion methods and system flaw knowledge. First, a database that contains the above knowledge needs to be established, describes the characters of the correspondence invasions, then match it with the current user behavior and the system mode. When the database discovers active clue conforming to the conditions, it will issue a warning, that is to say, an action that does not match the specific condition is legal.

C. G.V. Nadiammai et al:-

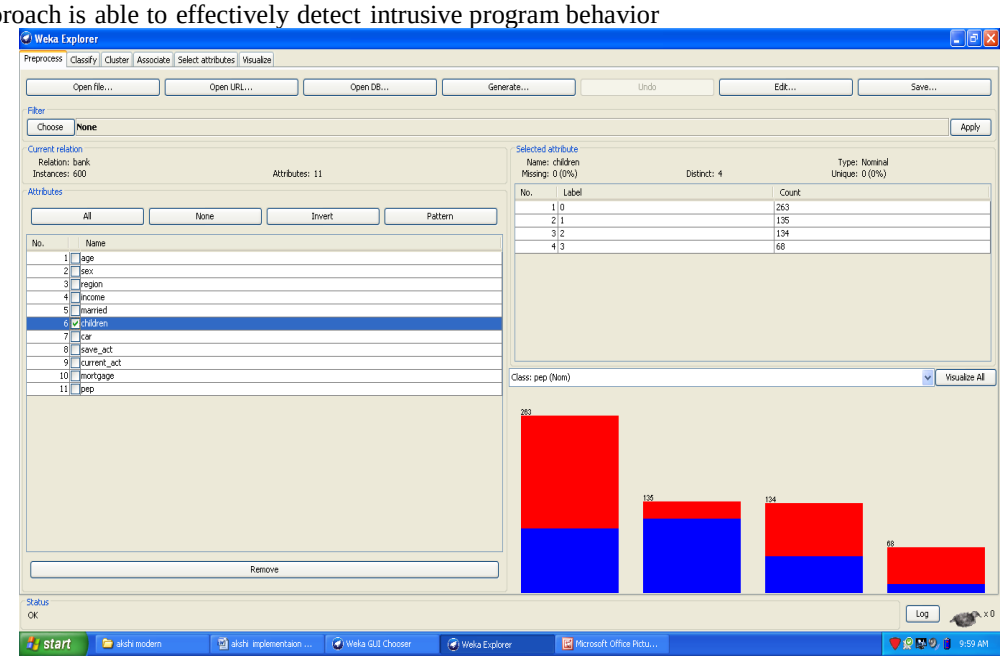
Data mining based IDS can efficiently identify these data of user interest and also predicts the results that can be utilized in the future. Data mining or knowledge discovery in databases has gained a great deal of attention in IT industry as well as in the society. Data mining has been involved to analyze the useful information from large volumes of data that are noisy, fuzzy and dynamic. illustrates the overall architecture of IDS. It has been placed centrally to capture all the incoming packets that are transmitted over the network. Data are collected and send for pre-processing to remove the noise; irrelevant and missing attributes are replaced. Then the preprocessed data are analyzed and classified according to their severity measures. If the record is normal, then it does not require any more change or else it send for report generation to raise alarms. Based on the state of the data, alarms are raised to make the administrator to handle the situation in advance. The attack is modeled so as to enable the classification of network data. All the above process continues as soon as the transmission starts.

V. PURPOSED WORK

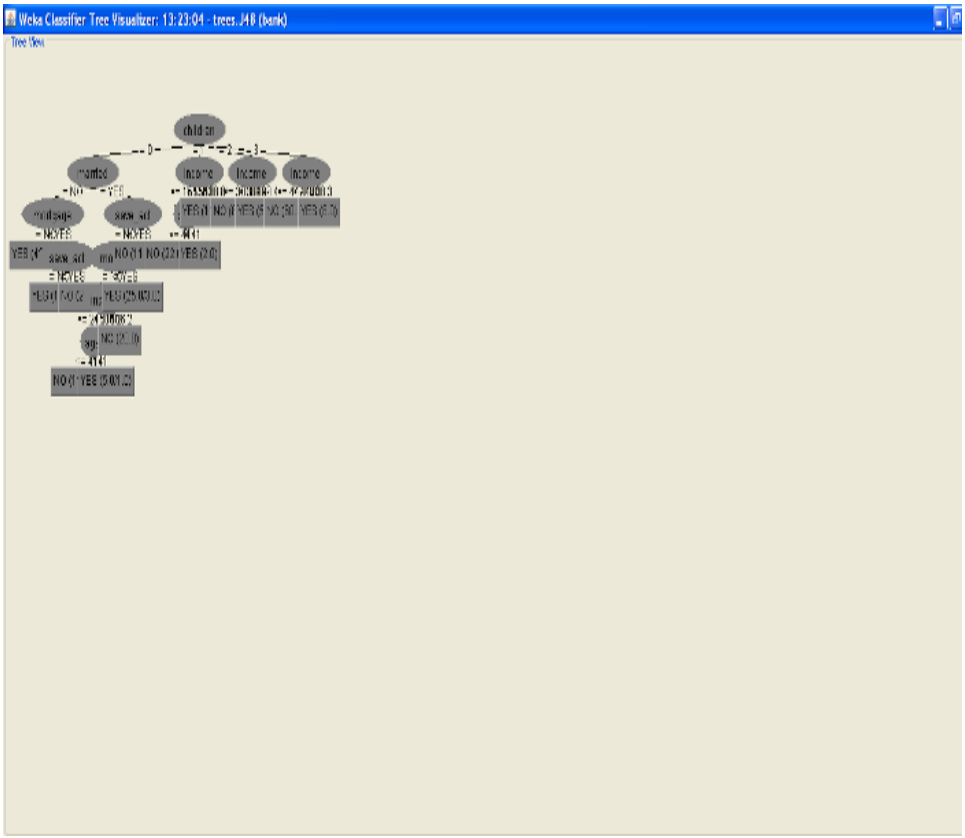
To address the core problem of an intrusion detection system (IDS) the accuracy of the detection results, in this Dissertation a Data Clustering Using K-Mean Algorithm for Network Intrusion Detection that possess highly accurate intrusion detection capability.

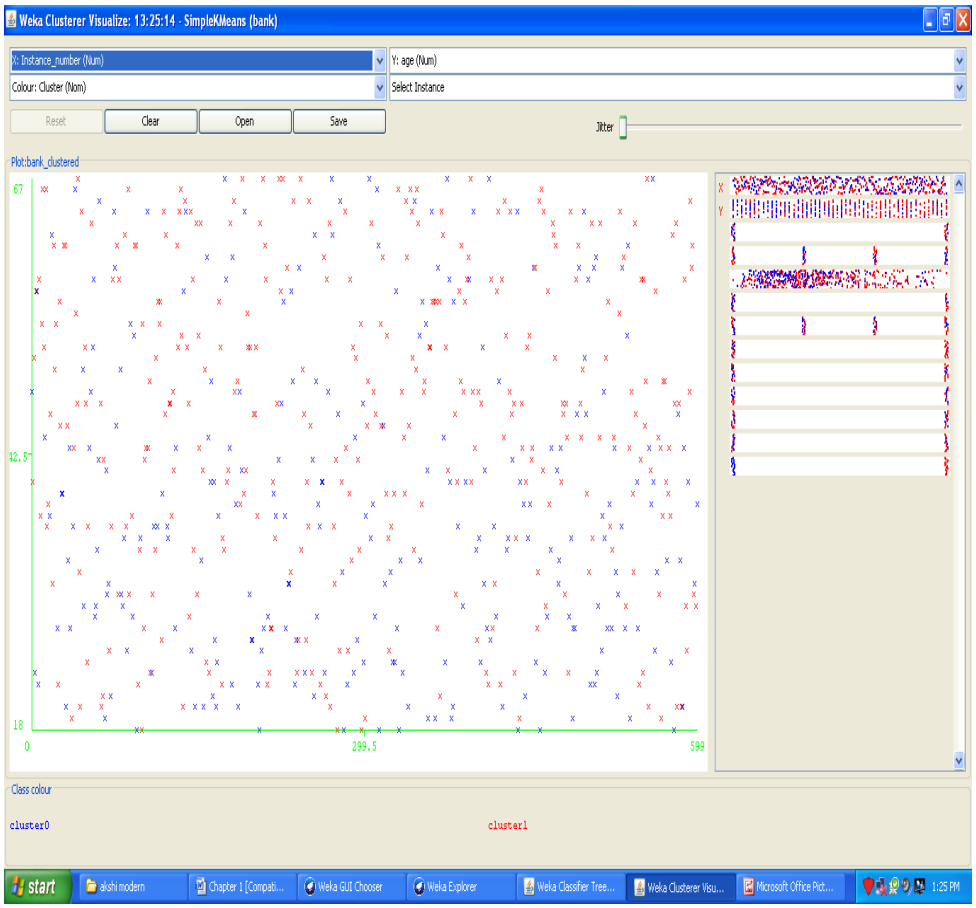
- It first describes the data set used and IDS architecture, followed by a detailed description of the IDS components. Then, it illustrates the K-mean algorithm
- the dataset used for the evaluation is described and then a large training data set for intrusion detection is presented
- Describes the intrusion-tolerant framework, and how the TC-K-Mean based IDS can be made intrusion-tolerant by introducing an intrusion-tolerant mechanism. The mechanism improves the dependability of the IDS since it prevents discontinuity in the detection service.
- An algorithm based on the *k*-mean clustering for analyzing program behavior in intrusion detection is evaluated by experiments.
- The preliminary experiments with the Bank data set audit data have shown.

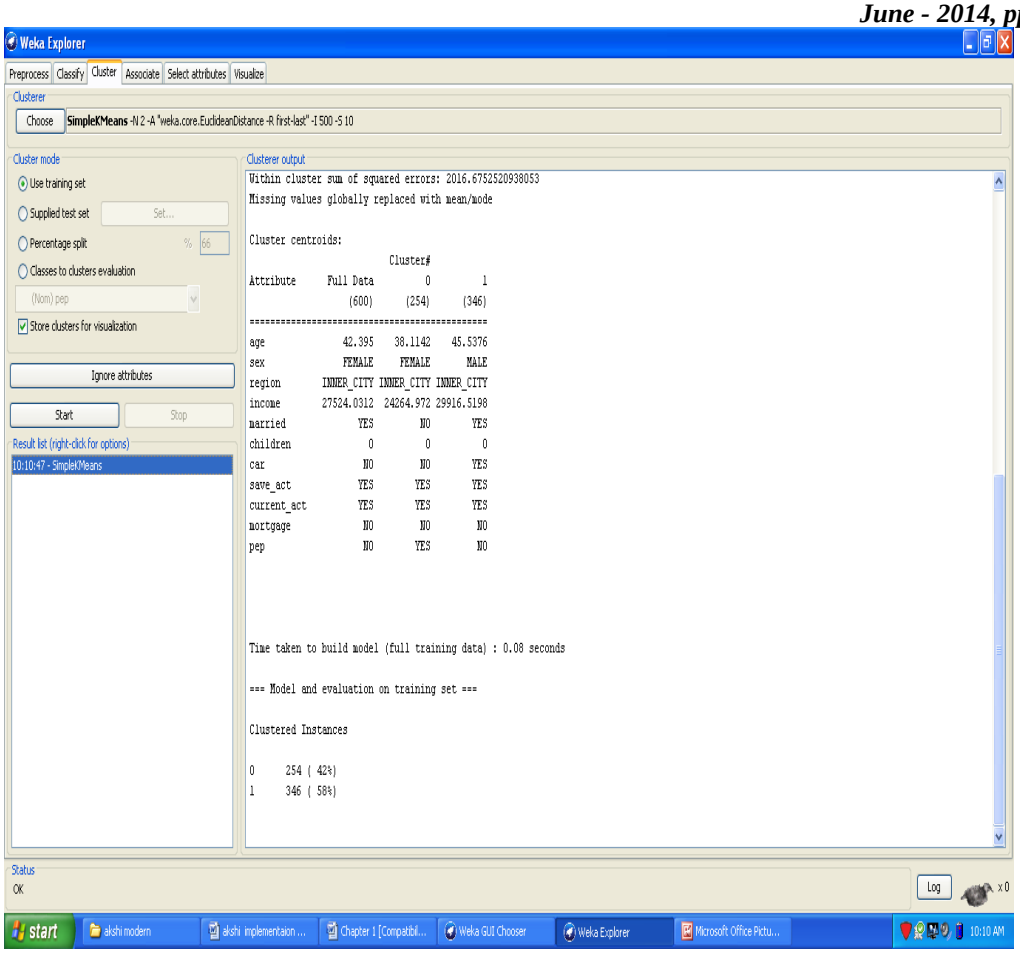
That this approach is able to effectively detect intrusive program behavior



To convert the data set into the standard spreadsheet format and convert categorical attributes to binary. It may also be necessary to normalize the values of attributes that are measured on substantially different scales (e.g., "age" and "income"). While WEKA provides filters to accomplish all of these preprocessing tasks, they are not necessary for clustering in WEKA . This is because WEKA SimpleKMeans algorithm automatically handles a mixture of categorical and numerical attributes.







VI. CONCLUSION

A new training process could be easily added to the training data set without changing the weights of the existing training samples. The performance of the k-mean clustering algorithm depends on the value of k, for k=2, the detection rate reached 96.6% rapidly and the low false positive rate. This could make the k mean clustering method more suitable for dynamic environments that require frequent updates of the training data.

REFERENCES

[1] SU-YUN WUA, ESTER YEN""ELSEVIER 2009 DATA MINING-BASED INTRUSION DETECTORS

[2] Li Hanguang, Ni Yu" Intrusion Detection Technology Research Based on Apriori Algorithm"Elsevier 2012

[3] G.V. Nadiammai, M. Hemalatha "Effective approach toward Intrusion Detection System using data mining techniques" Elsevier 2013

[4] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

[5] R. Ramanujan, S. Kudige, T. Nguyen "TECHNIQUES FOR INTRUSION-RESISTANT AD HOC ROUTING ALGORITHMS (TIARA)" IEEE Volume 2, Issue 8, November 2003.

[6] Wenke Lee Salvatore J. Stolfo Kui W. Mok "A Data Mining Framework for Building Intrusion Detection Models "Published by IEEE Computer Society,2005

[7] Sang-Hyun Oh, Jin-Suk Kang Yung-Cheol Byun "Anomaly Intrusion Detection Based on Clustering a Data Stream" Springer-Verlag Berlin Heidelberg 2006.

[8] Surasit Songma, Witcha Chimphlee "Classification via k-Means Clustering and Distance- Based Outlier Detection" 2012 Tenth International Conference on ICT and Knowledge Engineering.

[9] Min Sun, Yuanzhi Wang, Yun Luo" E-Alarm: An Anomaly Detection System on Large Network" 2009 International Joint Conference on Artificial Intelligence.

[10] N. Srinivasan' and V. Vaidehil "Reduction of False Alarm Rate in Detecting Network Anomaly using Mahalanobis Distance and Similarity Measure" IEEE - ICSCN 2007, MIT Campus, Anna University, Chennai, India. Feb. 22-24, 2007. pp.366-371.

