ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Review on Data Mining Techniques for Intrusion Detection System

Sandeep D[1], M. S. Chaudhari[2]

Research Scholar, Dept. of Computer Science, P.B.C.E, Nagpur, India[1]

HoD, Dept. of Computer Science, P.B.C.E, Nagpur, India[2]

ABSTRACT: With the rapid growth of computer networks during the past few years, security has become a crucial issue for modern computer systems. A good way to detect illegitimate use is through monitoring unusual user activity. This can be achieved with an Intrusion Detection System, which identifies attacks and reacts by generating alerts or by blocking the unwanted data/traffic. These systems are mainly classified as Anomaly based Intrusion Detection Systems and Misuse based Intrusion Detection Systems. Anomaly based Intrusion Detection System has the benefit of detecting novel attacks but has a high false positive rate. On the other hand, Misuse based systems are signature based having higher accuracy. Misuse based Intrusion Detection System fails to detect novel attacks. To overcome these limitations, both Anomaly based and Misuse based Intrusion Detection Systems should be combined to form a new Hybrid Intrusion Detection System. A new Hybrid Intrusion Detection System is proposed. In this system, fuzzy data-mining concept based on genetic algorithm is used as an intrusion detection system. KDD dataset is used to train the system and test the system.

Keywords: Intrusion Detection System (IDS), Network Security, Fuzzy Logic, Data Mining, Genetic Algorithm (GA).

I. INTRODUCTION

In recent years, internet and computers have been utilized by many people all over the world in several fields. In order to come up with efficiency and up to date issues, most organizations rest their applications and service items on internet. On the other hand, network intrusion and information safety problems are ramifications of using internet. For instance, on February 7th, 2000 the first DoS attacks of great volume where launched, targeting the computer systems of large companies like Yahoo!, eBay, Amazon, CNN, ZDnet and Dadet. In other words, network intrusion is considered as new weapon of world war. Therefore, it has become the general concern of the computer society to detect and to prevent intrusions efficiently. There are many methods to strengthen the network security at the moment, such as encryption, VPN, firewall, etc., but all of these are too static to give an effective protection. However, intrusion detection is a dynamic one, which can give dynamic protection to the network security in monitoring, attack and counter-attack. Thus, Intrusion Detection System (IDS) has been applied to detect intrusion network. Intrusion Detection technology can be defined as a system that identifies and deals with the malicious use of computer and network resources. In the case of detecting data target, intrusion detecting system can be classified as host-based and network-based [2].

- HOST-BASED IDS: Its data come from the records of various host activities, including audit record of operation system, system logs, application programs information, and so on.
- NETWORK-BASED IDS: Its data is mainly collected network generic stream going through network segments, such as: Internet packets.

Intrusion Detection techniques fall into two categories [2]:

- ANOMALY DETECTION: is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns.
- MISUSE DETECTION: is the ability to identify intrusions based on a known pattern for the malicious activity.

Data mining techniques have taken beneficial steps towards solution of various problems in different issues; we decided to utilize data mining for solving the problem of network intrusion because of following reasons:

- It can process large amount of data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

- User's subjective evaluation is not necessary, and it is more suitable to discover the ignored and hidden information.

Furthermore, data mining systems make it possible easily perform data summarization and visualization that help the security analysis in various areas [2].

A. *Data Mining*

Data mining generally refers to the process of extracting or mining knowledge from large amount of data. This process, first understand the existing data and then predicts the new data. It is the core of Knowledge discovery from Databases (KDD). Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions. The recent rapid development in data mining contributes to developing wide variety of algorithms suitable for network-intrusion-detection problems. As one of the most popular data mining methods for wide range of applications, association-rule mining is used to discover association rules or correlations among a set of attributes in a dataset [9].

1) Association Rule Mining

Association rule mining finds interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. The relationship between datasets can be represented as association rules. An association rule is expressed by $X \Rightarrow Y$, where X and Y contain a set of attributes. This means that if a tuple satisfies X, it is also likely to satisfy Y [9].

B. *Genetic Algorithm (GA)*

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics, introduced by J Holland in the 1970's and inspired by the biological evolution of living beings. Genetic algorithms abstract the problem space as a population of individuals, and try to explore the fittest individual by producing generations iteratively. GA evolves a population of initial individuals to a population of high quality individuals, where each individual represents a solution of the problem to be solved. The quality of each rule is measured by a fitness function as the quantitative representation of each rule's adaptation to a certain environment. The procedure starts from an initial population of randomly generated individuals. During each generation, three basic genetic operators are sequentially applied to each individual with certain probabilities, i.e. selection, crossover and mutation. The algorithm flow is presented in Fig.1 [8],[10].

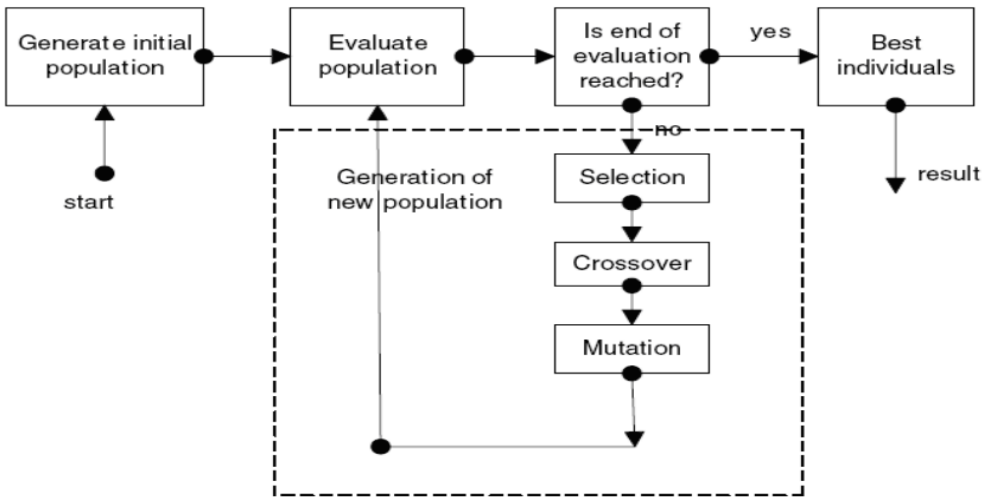

Fig.1. Flowchart of GA system [8]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

C. ***Fuzzy Set Theory***

Georg Cantor, who was a main inventor of the set theory, defined a set as follows: "By a set we mean an aggregation *M* of certain unequal objects *m* in our opinion or in our thought (which are called "elements" of *M*) to a whole." Such crisp sets do not always satisfy the needs of real world applications, because they only allow a membership of 1 or 0, i.e. member or non-member. In the real world, it is not at all times possible to assign an object clearly to a certain group of objects. Rather, it might lie in between two different sets [1].

Crisp sets are discriminating between members and non-members of a set by assigning 0 or 1 to each object of the universal set. Fuzzy sets generalize this function by assigning values that fall in a specified range, typically 0 to 1, to the elements. Let *X* be the universal set. The function $\mu_A$is the membership function which defines set *A*, where μ_A*: X* → [0, 1]. Fig. 2 shows difference between non-fuzzy sets and fuzzy sets.

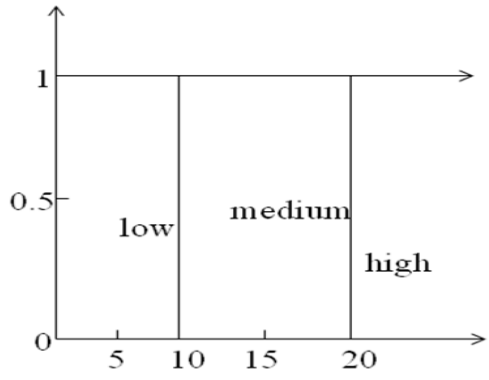

Fig. 2(a) Non-fuzzy sets [1]

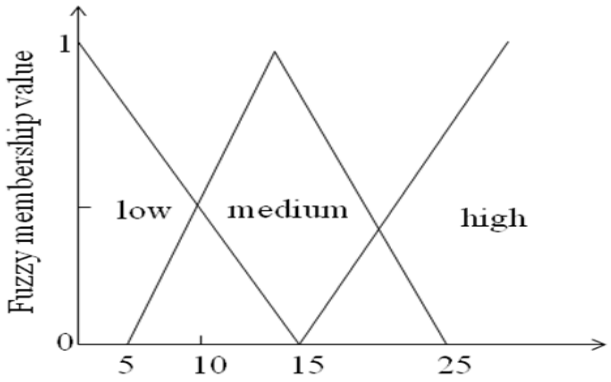

Fig. 2(b) Fuzzy sets [1]

D. ***DATA SET***

The data applied in the research comes from KDD Cup 99dataset, which was initially used for The Third International Knowledge Discovery and Data Mining Tools Competition. There are approximately 4,940,000 kinds of data in training dataset, 10% of which is provided, there are 3,110,291 kinds of data in test dataset, and there are totally 41 types of network connection characteristic (characterized by continuous data and discrete data) in each kind of network connection record. And its property can be divided into three major types: Basic characteristic of network connection, characteristic of network connection content, network transmission characteristic. Data pattern include nominal, binary and numeric. The data set contains a total of 24 attack types (connections) that fall into 4 major categories: Denial of service (Dos), Probe, User to Root (U2R), Remote to User (R2L). Each record is labeled either as normal, or as an attack, with exactly one specific attack type [7].

Probe: Attackers usually apply probe to get information, to determine the targets and the type of operating system.

Dos (Denial of service): Such attack may cause the stop of server operation, and the server cannot provide services. The attack usually occupies all system source of server, or occupies the Band-width and disables system resource and makes operation stop. Common attacks are SYN Flooding, Ping Flooding, and so on.

U2R (User gain root): In the attack, users take advantage of system leak to get access to legal purview or administrator's purview, such as: Buffer Overflow is among them.

R2L (Remote file access): The attack is to apply the advantage of server providing services, to get related safety setting or user's encrypted files, such as: Unicode leak, SQL Injection, and so on.

II. LITERATURE REVIEW

In this section a survey of data mining techniques that have been applied to IDSs by various research groups is presented [2],[3].

A. **Feature Selection:** Feature selection, also known as subset selection or variable selection. It is a process commonly used in machine learning. Feature selection is necessary because it is computationally infeasible to use all available features, or because of problems of estimation when limited data samples (but a large number of features) are present.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

B. **Machine Learning:** Machine Learning is defined as the study of computer algorithms that improve automatically through experience. Applications varies from data mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests. As compared to statistical techniques, machine learning techniques are well suited to learning patterns with no a priori knowledge of what those patterns may be. Classification and Clustering are the two most popular machine learning problems.

1) **Classification Techniques:** In a classification task in machine learning, the task is to take each instance of a dataset and assign it to a specific class. IDS based on classification, attempts to classify all traffic as either normal or malicious. The challenge in this method is to minimize the number of false positives and false negatives. Five general types of techniques have been tried to perform classification for intrusion detection purposes:
   - Inductive Rule Generation
   - Genetic Algorithms
   - Neural Networks
   - Immunological based techniques
   - Support Vector Machine

2) **Clustering Techniques:** Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics and many more. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets, so that the data in each subset share some common trait - often proximity according to some defined distance measure. Machine learning typically regards data clustering as a form of unsupervised learning. It is useful in intrusion detection as malicious activity should cluster together, separating itself from non-malicious activity [3].

Comparison table for some techniques used for Intrusion Detection System as follows:

Technique	Advantages	Disadvantages
ADAM [5]	Uses association and classification algorithm.	Uses only for anomaly detection, requires large no. of events in short time, sharp boundary problem.
Random Forests Algorithm [6]	Overcomes the problem of rule based system, uses feature extraction algorithm, also detect minority intrusions.	Sharp boundary problem, not able to work for mixed database.
Fuzzy Association Rule Mining [3]	Avoids sharp boundary problem.	Required more complex algorithm for rule generation, not able to work for mixed database, not able to generate more rules.
Genetic Algorithm [4]	Uses string as a chromosome, due to genetic operators generate more rules.	Overhead for more numbers of rules generation due to string as a chromosome, loss of information in crossover of strings, required more time for rules extraction, not used for mixed database.

Table I: Comparison table for different techniques used for IDS

III. PROPOSED SYSTEM

For extracting the rules with attributes of continuous value, fuzzy set theory is combined with association rule mining algorithm. Fuzzy Class-association-rule mining based on GA method for intrusion detection system overcomes many problems like sharp boundary problem, deals with mixed database, and increases rule pool [1]. Therefore, extraction of many rules as compared to other is possible. Support and fitness factors are calculated for each rule. Fitness function contributes to mining more rules with higher accuracy.Block diagram for proposed work is given in fig.3. Figure shows the exact flow of proposed system.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

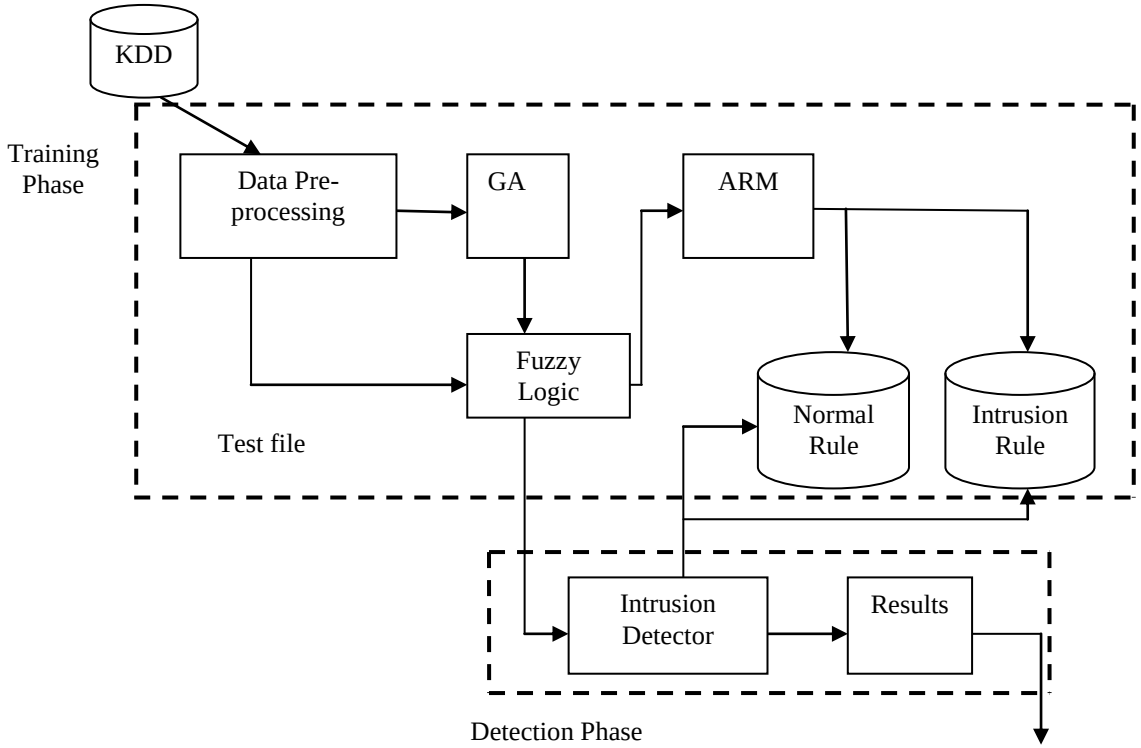

Fig. 3: Proposed System Overview

KDD- Knowledge Discovery and Data Mining
ARM- Association Rule Mining
GA- Genetic Algorithm
Proposed system objectives are as follows:

- Avoiding the sharp boundary problem by using fuzzy set theory.
- Use of mixed database, increases the detection rate and increases accuracy.
- Increases the size of rule pool by using the genetic operators.
- The proposed framework for intrusion detection can be flexibly applied to both misuse and anomaly detection with specific designed classifiers.

IV. CONCLUSION

Intrusion detection is an important but complex task for a computer system. Here, various methods for intrusion detection are studied and compared. Crisp data mining methods such as ADAM, Random Forest algorithm are used for intrusion detection but suffer from sharp boundary problem which gives less accurate results. In this proposed method use of fuzzy logic overcomes the sharp boundary problem. In this paper, we have proposed a GA-based fuzzy Class Association Rule Mining with Sub-Attribute Utilization and its application to classification, which can deal with discrete and continuous attributes at the same time. In addition, this method was applied them to both misuse detection and anomaly detection and performed experiments with practical data provided by KDD99 Cup.

REFERENCES

1. Mabu S., Chen C., Shimada K., "An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming," IEEE Transactions Systems, Man, Cybernetics C, Application and Reviews, volume 41, number 1, pp. 130–139, January 2011.
2. Ektefa M., Memar S., "Intrusion Detection Using Data Mining Techniques," IEEE Trans., 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

3. Harshna, NavneetKaur, "Survey paper on Data Mining Techniques of Intrusion Detection", IJSETR, vol-II, issue-4, April 2013.
4. Z. Bankovic, D. Stepanovic, S. Bojanic, "Improving Network Security using Genetic Algorithm Approach," Computer and Electrical Engineering, pp. 438-451, 2007.
5. Barbara, D., Couto, J., Jajodia, S., & Wu, N., "ADAM: A testbed for exploring the use of data mining in intrusion detection", ACM SIGMOD Record, 30 (4), 15—24, 2001.
6. J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," IEEE Transactions Systems, Man, Cybernetics C, Applications and Reviews, volume 38, no. 5, pp. 649–659, September 2008.
7. Kddcup 1999data [Online]. Available: kdd.ics.uci.edu/ databases/kddcup99/kddcup99 .html.
8. B. Abdullah, I. Abd-alghafar, "Performance Evaluation of a Genetic Algorithm Based Approach to Network Intrusion Detection System", 13th International Conference on AEROSPACE SCIENCES & AVIATION TECHNOLOGY, ASAT- 13, 2009.
9. J. Han, M. Kamber, "Data Mining", Morgan Kaufmann Publishers, 2001.
10. D. E. Goldberg, "Genetic Algorithm in Search, Optimization and Machine Learning", Reading, MA: Addison-Wesley, 1989.