Survey Paper of Fuzzy Data Mining using Genetic Algorithm for Intrusion Detection

Harshna
M.Tech(CSE)
harshna_attri@yahoo.com

Navneet kaur
Assistant Professor(CSE)
naiv_sekhon@yahoo.co.in

Abstract— In spite of growing information technology widely, security has remained one challenging area for computers and networks. Recently many researchers have focused on intrusion detection system based on data mining techniques as an efficient strategy. Intrusion Detection is one of the important and essential area of research. This work has explored the possibility of integrating the fuzzy logic with Data Mining methods using Genetic Algorithms for intrusion detection. Due to the use of fuzzy logic, the proposed system can deal with mixed type of attributes and also avoid the sharp boundary problem as the reasons for introducing fuzzy logic is two fold, the first being the involvement of many quantitative features where there is no separation between normal operations and anomalies. Thus fuzzy association rules can be mined to find the abstract correlation among different security features rather than to extract all the rules meeting the criteria which are useful for misuse detection. Genetic algorithm is used to extract many rules which are required for anomaly detection systems. So, proposed architecture for Intrusion Detection methods by using Data Mining algorithms to mine fuzzy association rules by extracting the best possible rules using Genetic Algorithms.

Index Terms — KDD, Data Mining, Security, Intrusion Detection System (IDS), Association Rules, Genetic Algorithm (GA), Fuzzy Logic.

1 INTRODUCTION

As the Internet services spread all over the world, several kinds and a large number of security threats are increasing. Many kinds of systems over the Internet such as Internet banking, online shopping, trading stocks and foreign exchange, and online auction have been developed. However, due to the open culture of the Internet, the security of our computer systems and data is always at risk. Computer security is defined as the protection of computing systems against threats to integrity, confidentiality, and availability [7]. Security threats come from different sources such as natural forces (flood), accidents (fire), failure of services (power) and people known as intruders.

Two types of intruders are: the external intruders who are unauthorized users of the machines who attack by using various penetration techniques, and internal intruders, refers to those with access permission who wish to perform unauthorized activities [2]. When an unauthorized user attempts to break into an information system or performs an action not legally allowed, this activity is referred to as an intrusion. An intrusion is a deliberate, unauthorized attempt to access or manipulate information or system and to render them unreliable or unusable. If a doubtful activity is from our internal network or system it will also be classified as intrusion. Intrusive activities may include password cracking, exploiting software bugs and system misconfiguration, sniffing unsecured traffic, or exploiting the design flaw of particular protocols. An Intrusion Detection System is a system for detecting intrusions and reporting to the proper authority.[11]

Some of the objectives of the IDS are described as following as:

- Detect wide variety of attacks.
- Detect intrusions in timely fashion.
- Present analysis in simple, easy-to-understand format.
- Minimize false positives, false negatives:
  a. False positive: An event, incorrectly recognized by the IDS as being an intrusion when none has occurred.
  b. False negative: An event that the IDS fails to recognize as an intrusion when one has in fact occurred.

Intrusion Detection techniques have been investigated since the mid 80s and depending on the various approaches and source of the information used to identify security breaches, they are classified as

Fig1: Different approaches of IDS
1.1 Approaches based upon the type of attacks

1.1.1 Host-based IDS: This method is based on data source category; therefore its data comes from the records of various activities of hosts, including system logs, audit operation system information, file integrity etc.

- Advantages
  - Due to the fact that HIDS monitor only the host, it can determined intrude more precise
  - It does not need to install extra hardware or software because everything is available on the host.
  - Encrypted messages are not serious problem because they received in the host and can be decrypted more easily.

- Disadvantages
  - It can not detect some types of attacks that they need to monitor traffic of network e.g. DOS and DDOS.
  - Duplicity is an important problem in HIDS especially when we want to install this system for a network, because we should have a HIDS for each host.
  - Because of the fact that HIDS should be installed in each host, it is clear that expense of system will be increased.
  - Efficiency in terms of speed up is going to be decreased, due to having a monitoring system for each host.

1.1.2 Network based IDS: Network-based IDS refers to systems that identify intrusions by monitoring traffic through network devices (e.g. Network Interface Card, NIC).

- Advantages
  - Detection of some attacks such as DOS and DDOS need to monitor traffic of whole network and it is possible by NIDS.
  - Low expense is brilliant advantage for NIDS because it is not necessary to install many monitoring systems.

- Disadvantages
  - Accuracy is a challenging problem due to losing some data during the process of detection.
  - Encrypted data are problematic in NIDS because of the fact that it is not possible to decrypt data in level of network.
  - In large-scale network more facility is required to monitor network; thus, scalability is another significant problem in NIDS.

1.2 Based upon the detection techniques :

1.2.1 Misuse Detection Technique: The misuse detection approach attempts to recognize attacks that follow intrusion patterns that have been identified and reported by experts. Misuse detection systems are vulnerable to intruders who use new patterns of behavior or who mask their illegal behaviour to deceive the detection system.

- Advantages
  - Specifying exact class of attacks.
  - Efficiency is high and complexity is low.

- Disadvantages
  - Known intrusion patterns have to be hand-coded
  - Many false positives: prone to generating alerts when there is no problem in fact.
  - Cannot detect unknown intrusions.

1.2.2 Anomaly Detection Technique: With the anomaly detection approach, one represents patterns of normal behaviour, with the assumption that an intrusion can be identified based on some deviation from this normal behaviour. When such a deviation is identified, an intrusion alarm is raised.

- Advantages
  - Anomaly detection can detect novel and unknown attacks to increase the detection rate.

- Disadvantages
  - Selecting the right set of system features to be measured is ad hoc and based on experience
  - Possible high false alarm rate.

1.3 Approaches based upon the behaviour of IDS:

1.3.1 Passive IDS: Passive IDS simply detects and alerts the administrator.

1.3.2 Active IDS: Active IDS will not only detect suspicious or malicious traffic and alert the administrator, but will take predefined proactive actions to respond to the threat.

2. Data Mining in Intrusion Detection

2.1. Classification

Classification is one of the topics of data mining. Its goal is to build the classification attribute model based on attribute. Data classification has two steps. The first, a data set is selected. The class label of each set (training samples) for training data set is known. The class label of each training samples is provided, so the first step also is the supervised learning process. Usually, the learning model is described by the classification rules, decision tree or mathematical formula. The second step, the model is classified. First the prediction accuracy of the model (classification rules) is evaluated. For each test sample, the known class label and the prediction label of the sample are compared. If the model’s exactness rate can be accepted, it will be used to classify the data set that the class label is unknown.

2.2. Clustering

Clustering is to identify the internal rules of the data object. The objects are grouped to form a class of similar objects, and export the data distribution. Similar or dissimilar measure is based on the values of the property defined by the data object. Usually, it is defined by the distance. When the mining task is confronted with the lack of domain knowledge or incomplete data set, clustering is used to divide the unidentified data object into different classes automatically. It can not be restricted and interfered by a priori knowledge, to obtain the information of the original data set. Distinction between classification and clustering is that classification is applied to the data object, and clustering is to find the classification rules implied in the mixed data objects.

2.3. Association rules

Association rule induction is one of the most well-known approaches in data mining techniques. In association, a pattern is discovered based on a relationship of a particular item on oth-
er items in the same transaction. It is to find the exciting con-
nexions between items of a given data set. Database T is a
collection of n transactions, (T1, T2, . . . , Tn) and I is the set of
all items, {i1, i2, . . . , im}, where each of the transactions Tj(1 ≤ j
≤ n) in the database T represents a set of items (Tj ⊆ I). An item
set is defined as a non-empty subset of I. An association rule
can be represented as: X→Y(c, s), where X ⊆ I, Y ⊆ I and X∩Y
= φ. In this association rule, s is called support and c is confi-
dence of the association rule. The support is the percentage
of the transactions in which both X and Y appear in the same
transaction and the confidence is the ratio of the number of
transactions that contain both X and Y to the number of trans-
actions that contain only X. It can be described as follows:
Support (X→Y) = P(X∪Y)
Confidence (X→Y) = P(Y |X)
Association rule is that support and confidence to meet the
thresholds given by the user. The basic Association rule works
in two steps. First, it finds frequent itemsets. In the second
step the minimum confidence rules are generated from the
frequent itemsets found in the first step.

3. Fuzzy Logic

Fuzzy logic is appropriate for the intrusion detection problem
for two major reasons. First, many quantitative features are
involved in intrusion detection. Security-related data catego-
rizes the statistical measurements into four types:
- ordinal,
- categorical,
- binary categorical,
- and linear categorical.
Both ordinal and linear categorical measurements are quan-
titative features that can potentially be viewed as fuzzy
variables. Two examples of ordinal measurements are the
CPU usage time and the connection duration. An example of a
linear categorical measurement is the number of different
TCP/UDP services initiated by the same source host. The se-
cond motivation for using fuzzy logic to address the intrusion
detection problem is that security itself includes fuzziness.
Given a quantitative measurement, an interval can be used to
denote a normal value. Then, any values falling outside the
interval will be considered anomalous to the same degree
regardless of their distance to the interval. The same applies to
values inside the interval, i.e., all will be viewed as normal to
the same degree. The use of fuzziness in representing these
quantitative features helps to smooth the abrupt separation
of normality and abnormality and provides a measure of the de-
gree of normality or abnormality of a particular measure.

4. Fuzzy Association Rules

4.1 Limitations of association rules

According to the different quantitative attributes, association
rule is divided into Boolean association rules and quantitative
association rules. In reality, the data are quantitative in most
cases, so t quantitative association rules mining research is
very important. The general method to solve quantitative as-

sociation rules is that the value of the property is divided into
several regions by a certain criteria and then is converted to a
sequence-<attribute, interval>. Thus quantitative association
rule will be transformed into Boolean association rules. How-
ever, there are some problems. On the one hand, if the interval
division is too large, confidence of the rules included in the
interval will be very low. So that it will cause a small number
of rules, and will be a corresponding reduction in the amount
of information. If the interval division is too small, support of
the rules included in the interval will be very low. So that it
will cause a small number of rules. On other hand, if the do-
main of property is divided into the non-overlapping interval,
the discrete data in the database is mapped to the interval. As
potential elements near the interval are excluded by clear divi-
sion, it will lead to some significant interval is ignored. If the
domain of property is divided into overlapping intervals, the
elements in the border may be in two intervals at the same
time. These elements will contribute to the two intervals, re-
sulting in some intervals are overemphasized. In order to
solve the problem of sharp boundary, fuzzy theory is prop-
osed. The membership function is used to define data set in
fuzzy sets of the attribute domain, in order to achieve the pur-
pose of softening the border.

4.2 Fuzzy Association Rules

Given a database T with attributes I and the definitions of
fuzzy sets associated with attributes in I, the objective is to
find out some interesting regularities between attribute values
in a guided way. Any fuzzy association rule is in the following
form:
If X is A then Y is B. (1)
In the above rule, X = {x1, x2, . . . , xp} and Y = {y1, y2, . . . , yq}
are attribute sets. X and Y are disjoint subsets of I. A ={f x1 , f
dx2, . . . , fxp} fuzzy sets associated with the cor-
responding attributes in X and Y. For example f xi is a fuzzy
set, defined on xi domain. Each pair of (xi , f xi) is called
an item, and each pair of (X, A) or (Y, B) is called an itemset.
The first part of the rule ‘X is A’ is called the antecedent and ‘Y
is B’ is called the consequent of the rule. The semantics of
the rule is when ‘X is A’ is satisfied, we can imply that ‘Y is B’
is also satisfied. Here the word “satisfied” means there are suffi-
cient amount of records which contribute their votes to the
attribute/fuzzy set pairs and the sum of these votes is greater
than a user specified threshold. An appropriate rule should
have enough significance and a high certainty factor. Signifi-
cance and certainty factor are two concepts, equivalent to sup-
port and confidence.

5. Genetic Algorithm

Genetic algorithms (GA) are search algorithms based on the
principles of natural selection and genetics, introduced by
John Holland in the 1970s and inspired by the biological ev-
olution of living beings. Genetic algorithms abstract the prob-
lem space as a population of individuals, and try to explore
the fittest individual by producing generations iteratively.
Individuals are represented by a string of symbols. Each indi-
vidual is called a chromosome, and is composed of a prede-
termined number of genes [4]. The generation of new off-
springs includes the operations such as crossover, mutation and
selection operations [10], [1].
Working steps of Genetic Algorithm are:
1. [START] Generate random population of n chromosomes i.e. suitable for the problem.
2. [FITNESS] Evaluate the fitness f(x) of each chromosome x in the population.
3. [NEW POPULATION] Create a new population by repeating following steps until the new population is complete.
   a) [SELECTION]: Reproduction (or selection) is an operator that makes more copies of better strings in a new population. Reproduction is usually the first operator applied on a population [9].
   b) [CROSSOVER]: A crossover operator is used to recombine two strings/parents to get better new two strings/children. It is important to note that no new strings are formed in the reproduction phase. In the crossover operator, new strings are created by exchanging information among strings of the mating pool. Types of crossover are explained in [2], [5].
   c) [MUTATION]: Mutation adds new information in a random way to the genetic search process [5], [7]. It is an operator that introduces diversity in the population whenever the population tends to become homogeneous due to repeated use of reproduction and crossover operators.
   d) [ACCEPTING] place new offspring in the new population.
4. [REPLACE] use new generated population for the further run of the algorithm.
5. [TEST] if the end condition is satisfied then stops and returns the best solution in current population.

6. CONCLUSION

Intrusion Detection is one of the major concerns in any computer networks environment. Various methods related to intrusion detection system are studied and compared. Crisp data mining methods such as ADAM method etc. are used for intrusion detection but suffer from sharp boundary problem which gives less accurate results. Use of fuzzy logic overcomes the sharp boundary problem. The reasons for introducing fuzzy logic is two fold, the first being the involvement of many quantitative features where there is no separation between normal operations and anomalies. Thus fuzzy association rules can be mined to find the abstract correlation among different security features. Using genetic algorithms with the fuzzy data mining method may result in the tune of the fuzzy membership functions to improve the performance and select the set of features available from the audit data that provide the most information to the data mining component. These algorithms are often used for optimization problems. Therefore, integration of fuzzy logic with class-association rules and GA generates more abstract and flexible patterns for intrusion detection.

7. ACKNOWLEDGEMENT

I would like to thanks CSE department of RIMT-IET , Mandi Gobindgarh, Punjab.

8. REFERENCES

School (AIS), University Technology Malaysia International Campus, Kuala Lumpur, Malaysia.

