An Intrusion-Detection Model based on Fuzzy Frequent Episodes using Probability Distribution

Vivek P. Kshirsagar\(^1\) and Madhuri S. Joshi\(^2\)

\(^1\)Comp. Science and Engg. Department, Govt. College of Engg., Aurangabad, India
vkshirsagar@gmail.com

\(^2\)Associated with Govt. College of Engg., Aurangabad, India
madhuris.joshi@gmail.com

Abstract— Intrusion detection (IDS) is a computer based information system designed to collect information about malicious activities in a set of targeted IT resources, analyze the information and respond to a predefined security policy. This paper describes a fuzzy frequency episodes and probabilistic classification for detecting intrusions in a network. There are two main reasons for introducing fuzzy logic for intrusion detection. First, many quantitative attributes are involved in intrusion detection. Fuzzy set theory provides an efficient way to categorize these quantitative attributes in order to form high-level patterns. Second, Fuzzy itself is secure. We use fuzzy frequent episodes to extract patterns for temporal statistical measurements at a higher level data level. A fuzzy frequent episode reduces the error rate as compared to the non-fuzzy frequent episodes. Proposed method efficiently extracts many patterns that are significantly and flexibly applied to both misuse and anomaly detections.

Index Terms— Intrusion Detection System (IDS), Fuzzy frequency episodes, Malicious activity, Probabilistic Classification, High level patterns, Temporal Statistical Measurement, Anomaly detections.

I. INTRODUCTION

In information Security, IDS is the act of detecting actions that attempt to compromise the integrity, availability or confidentiality of a resource. When an IDS takes preventive action without direct intervention, then it becomes an Intrusion-prevention system [1]. There are two types of intrusion detection: misuse detection and anomaly detection. Misuse detection can be applied to attacks that follow some fixed patterns and are usually constructed to examine intrusion patterns that have been recognized and reported by experts. The anomaly detection system looks for any evidence of activities that deviate from what is considered normal system use [1] [2].

Association rule algorithms find correlations between features or attributes used to describe a data set. On the other hand, frequency episode techniques are effective at detecting sequential patterns of occurrences in a sequence of events. This paper explores the use of frequency episode to produce rules suitable for anomaly-based and misuse-based detection.

There are two main reasons for introducing fuzzy logic for intrusion detection. First, many quantitative attributes are involved in intrusion detection. Fuzzy set theory provides an efficient way to categorize these quantitative attributes in order to form high-level patterns. Second, fuzzy is secure. Fuzzy frequent episodes...
are used to extract patterns for temporal statistical measurements at a higher level than the data level. A fuzzy frequent episode reduces the error rate as compared to the non-fuzzy frequent episodes [4] [5] [6].

II. OVERVIEW OF FREQUENT EPISODES RULES

Mannila and Toivonen’s[3] describe algorithm for mining frequent episodes similar to the Apriori algorithm except for the difference between calculation episode frequencies and calculating itemset supports. Episodes rules can be directly established from frequent episodes. Let $P(e_1,e_2...e_k)$ be a frequent episode and $P'(e_1,e_2...e_k) \subset P$ be k-1 non-empty ordered sub episodes where $1 \leq i \leq k - 1$. In order to find frequent episodes rule we need infrequency (representing minimum frequency), the threshold of the window and threshold minconfidence. A serial episode rule of the form $P_i \rightarrow Q_i.c.s.w$ is as follows [5].

$$P(e_1,e_2...e_l) \subset P, Q(e_{l+1},e_{l+2}...e_k) = P - P_i \subset P$$

\(s=\text{frequency}(P) \geq \text{minfrequency},\)

\(c=(\text{frequency}(P)/\text{frequency}(P)) \geq \text{minconfidence}, &\)

\(w=\text{window} \text{(representing timestamp bounds)}\).

Shingo Mabu et al., [14] describes a novel fuzzy class-association rule mining method based on genetic network programming (GNP) for detecting network intrusions. GNP is an evolutionary optimization technique, which uses directed graph structures instead of strings in genetic algorithm or trees in genetic programming, that leads to enhancing therepresentation ability with compact programs derived from the reusability of nodes in a graph structure. The integration of fuzzy logic with Data Mining methods using Genetic Network Programming (GNP) for intrusion detection help to deal with the mixed database that contains both discrete and continuous attributes and also extract many important class-association rules that contribute to enhancing detection ability. Therefore, the method can be flexibly applied to both misuse and anomaly detection in network-intrusion-detection problems. Such mixed database is normal in real-world applications and GNP can extract rules that include both discrete and continuous attributes consistently. The initiative of combining association rule mining with fuzzy set theory has been applied more frequently in recent years. There are two main reasons for introducing fuzzy logic for intrusion detection. First, many quantitative attributes are involved in intrusion detection, where discretization of the quantitative attributes into intervals would lead to under- or overestimate the values that are near the borders. This is called the sharp boundary problem. Fuzzy sets can help us to overcome this problem by allowing different degrees of fuzzy membership’s functions. Second, security itself is fuzzy [7].

III. MINING FUZZY FREQUENT EPISODES

Integrating fuzzy logic with frequency episodes allows one to extract more abstract patterns at a higher level. Many quantitative attributes are involved in intrusion detection. Fuzzy set theory provides an efficient way to categorize these quantitative attributes in order to form high-level patterns. Let $A = \{a_1, a_2, ..., a_n\}$ be the set of event attributes and each attribute a in A may be categorical or quantitative (fuzzy) wherever $1 \leq j \leq m$. Suppose $f(a)$ represents the maximum number of categories or the maximum number of fuzzy sets and $m_a(j,x,y)$ represents the membership degree. If $a_j$ is categorical, $m_a(j,x,y) \subseteq \{0,1\}$. If $a_j$ is fuzzy, $0 \leq m_a(j,x,y) \leq 1$. For an event attribute $A$, its categories or fuzzy sets can be mapped to consecutive integers. Let $e^j(\text{attr}_1=c_1, \text{attr}_2=c_2,...,\text{attr}_j=c_j)$ is event variable where $e^j(\text{attr}_1=c_1, \text{attr}_2,...,\text{attr}_j=c_j) \subseteq A$ and for all $j (1 \leq j \leq k), 1 \leq c_j \leq f(a_j)$. Consider two event variables $e^d$ and $e^l$ as $e^d(\text{attr}_1=c_1, \text{attr}_2=c_2,...,\text{attr}_d=c_d)$ and $e^l(\text{attr}_1=c_1, \text{attr}_2=c_2,...,\text{attr}_l=c_l)$. These event variables are said to be homogenous, if $e^d(\text{attr}_1,c_1,\text{attr}_2,...,\text{attr}_d,c_d)$ and $e^l(\text{attr}_1,c_1,\text{attr}_2,...,\text{attr}_l,c_l)$. The minimal occurrence of an episode is the product of the occurrences of its event variables. Due to the
introduction of fuzzy sets event E has several occurrences of an event variable. However, a side effect may arise due to very small membership values. The occurrence of an event variable with a very small membership may change the minimal occurrence of an episode in the event sequence. To overcome this problem, threshold minioccurrence is used to represent the smallest allowable occurrence of an event variable. Given an event variable $e^k$, if occurrence $(e^k, E) < \text{minioccurrence}$, it will not be considered to have occurred in E. The following normalization process will conducted to account for the deletion of these low membership occurrences [6]

If occurrence $(e^k, E) < \text{leastoccurence}$ occurrence $(e^k, E)=0$

Else

\[
\text{occurrences}(e^k) = \frac{\text{occurrences}(e^k, E)}{\sum_q \text{occurrences}(e^q, E)}
\]  

(3)

Jianxiong Luo et al.,[5] describe fuzzy frequent episodes that can provide evidence of intrusions in near real-time. It uses fuzzy frequent episodes rule mining method for detecting network intrusions. The disadvantages of these systems are:
1) Frequent episodes have been used to mine training data to establish normal patterns for only anomaly detection and not for misuse detection.
2) A fuzzy frequent episodes rule mining method is not able to extract accurate rules with attributes of both binary and continuous values [8][9].

IV. PROPOSED METHOD

This paper describes a fuzzy frequent episode rule mining method based on probabilistic classification [9], by combining fuzzy set with probabilistic classification; the proposed method can deal with the mixed database that contains both discrete and continuous attributes. Sub-attribute utilization considers all discrete and continuous attribute values as information, which contributes to avoid data loss. The proposed framework for intrusion detection can be flexibly applied to both misuse and anomaly detection with specific designed classifiers. High detection rate(DR) are obtained in both misuse and anomaly detection. Fig.1 shows system architecture diagram which consist of DARPA99 dataset, preprocessing, fuzzy membership function, rule extraction, updating fuzzy rules in rule pool and probability density function for classification[8][9].

The DARPA99 training data includes “list file,” which identifies each network connection’s time stamps, service type, type of each attack, source IP address, destination IP address, source port and destination port etc. A preprocessor is a program that processes raw input data, eliminating missing values from dataset to produce output which is used as input to another program.

A. Extract Dataset & Preprocessing

The DARPA 99 training dataset includes time stamps, service type, type of each attack, source IP address, destination IP address, source port and destination port etc. A pre-processor is a program that processes raw input data, eliminating missing values from dataset to produce output which is used as input to another program.

B. Fuzzy Membership Function

In each network connection sub attribute utilization is considers to avoid data loss and more efficient rule mining as it contain all discrete and continuous attribute values as information. This attributes are divided into sub attributes as: binary, continuous and symbolic attributes. Binary attributes are dividing into two sub
attributes. For example, if A1(=phf) is binary attribute then A11 (representing phf = 1) and A12 (representing phf = 0) is sub attribute of A1. The symbolic attribute was divided into three sub attributes which represent linguistic terms (low, middle and high) of fuzzy membership functions as shown in table 1. GNP examines the attributes of tuples at judgment nodes and calculates the measurements of association rules at processing nodes. The extracted fuzzy class-association rules are stored in a rule pool through generations. When an important rule is extracted by GNP, it is stored in the pool with its support, confidence, $\chi^2$ value, and the parameters of the fuzzy membership function. Occasionally, a fuzzy rule already stored in the pool would be extracted again. In that case, the membership function and $\chi^2$ value might be changed. If the fuzzy rule has higher $\chi^2$ value, it will replace the same old fuzzy rule in the pool along with its fuzzy parameters. Therefore, the pool is updated every generation and only important fuzzy rules with higher $\chi^2$ values and better-adapted fuzzy parameters are stored.

![Intrusion Detection Model](image)

**TABLE I. EXAMPLE OF SUBATTRIBUTE UTILIZATION**

<table>
<thead>
<tr>
<th>Binary attributes</th>
<th>Symbolic attribute</th>
<th>Continuous attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>phf (phf=0,phf=1)</td>
<td>http</td>
<td>Low</td>
</tr>
<tr>
<td>guess (guess=0,guess=1)</td>
<td>ftp-data</td>
<td>medium</td>
</tr>
<tr>
<td>rcp (rcp=0,rcp=1)</td>
<td>telnet</td>
<td>high</td>
</tr>
<tr>
<td>rlogin (rlogin=0,rlogin=1)</td>
<td>smtp</td>
<td></td>
</tr>
<tr>
<td>port-scan (port-scan=0,port scan=1)</td>
<td>pop3</td>
<td></td>
</tr>
</tbody>
</table>

**C. Rule Extraction and Updating Fuzzy Rules in Rule Pool**

To distinguish between intrusive and normal behavior, algorithms are needed to generate the frequent episode rules (FER) from audit traffic data. The FER describes the interrelationship among multiple connection records. Frequent episode rules are extracted using fuzzy membership functions. The confidence of an episode rule is usually less than 1 and provides a measure of the strength of the evidence of an anomaly. When mining episode rules, only the rules which span less than or equal to window timestamp bound intervals will be mined. Therefore Users can use this mining parameter to avoid mining rules that span across too many intervals. The extracted rules as shown in table 2 are stored in rule pool with support, confidences, occurrences and window threshold.
D. Classification Using Probability Density Function

In probability theory, a probability density function (pdf), is a function that describes the relative likelihood for this random variable to take on a given value. The probability for the random variable to fall within a particular region is defined by the integral of this variable density over the region. The probability density function takes a nonnegative value everywhere and its integral is equal to one. The terms “probability distribution function” and “probability function” have also sometimes been used to denote the probability density function. The probability distribution function may be used when it is defined as a function over general sets of values, or cumulative distribution function or a probability mass function rather than the density.

<table>
<thead>
<tr>
<th>Intrusion Rules</th>
<th>Support</th>
<th>Confidence</th>
<th>Normal Rules</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(phf=1)</td>
<td>at(1)/N</td>
<td>at(1)/a</td>
<td>(phf=1)</td>
<td>at(2)/N</td>
<td>at(2)/a</td>
</tr>
<tr>
<td>(phf=1)A(pro=http)</td>
<td>b(1)/N</td>
<td>b(1)/b</td>
<td>(phf=1)A(pro=http)</td>
<td>b(2)/N</td>
<td>b(2)/b</td>
</tr>
<tr>
<td>(phf=1)A(pro=http) (count=low)</td>
<td>c(1)/N</td>
<td>c(1)/c</td>
<td>(phf=1)A(pro=http)A (count=low)</td>
<td>c(2)/N</td>
<td>c(2)/c</td>
</tr>
<tr>
<td>(phf=1)A(pro=http)A (count=medium)</td>
<td>c(1)/N</td>
<td>c(1)/c</td>
<td>(phf=1)A(pro=http)A (count=medium)</td>
<td>c(2)/N</td>
<td>c(2)/c</td>
</tr>
<tr>
<td>(phf=1)A(pro=http)A (count=high)</td>
<td>c(1)/N</td>
<td>c(1)/c</td>
<td>(phf=1)A(pro=http)A (count=high)</td>
<td>c(2)/N</td>
<td>c(2)/c</td>
</tr>
</tbody>
</table>

Before applying the probabilistic classification, the probability density function should be generated. First, the matching degree of each data d with every rule in the rule pool is calculated. The matching degree between the continuous attribute Ai with linguistic term Qi in rule r in class k and the value ai of attribute Ai of a testing data is defined as:

\[ \text{MatchDegree}_k(Q_i, a_i) = FQ_i(a_i) \]  \hspace{1cm} (4)

Where, FQi is the membership function of linguistic term Qi and matching between rule r in class k and new unlabeled connection d is given by:

\[ \text{Match}_k(d, r) = \frac{1}{p+q} \left( \sum_{i \in CA} \text{MatchDegree}_k(Q_i, a_i) + t \right) \]  \hspace{1cm} (5)

Where

- i – Index of continuous attributes in rule r;
- CA – set of suffixes of continuous attributes in rule r;
- P – total number of continuous attributes in rule r;
- q - total number of discrete attribute in rule r;
- t – the number of matched discrete attributes with new unlabeled connection d in rule r.

After creating the probability density function \( f_k(x_k) \) of the average matching degree between training data \( d \in D_{train}(k) \) and rule \( r \in R_k \), the probability that new connection data \( d \in D_{test} \) belongs to class k is represented as follow:

\[ p_k(d) = \int_0^{f_1} \int_0^{f_2} \cdots \int_0^{f_m} f_k(x_k) dx_k \cdots f_k(x_1) dx_1 \cdots \]  \hspace{1cm} (6)

where,

- \( D_{test} \) is the set of suffix of testing data. The probability that \( d \in D_{test} \) belongs to anomaly class is defined as:

\[ p_a(d) = 1 - \sum_{k \in C} p_k(d) \]  \hspace{1cm} (7)

where, C is the set of suffix of classes having training data.

Then, the probability that a new connection data belongs to anomaly and misuse class is calculated by:

\[ p_{a\&m}(d) = 1 - \sum_{k \in C} p_k(d) \]  \hspace{1cm} (8)

Based on the probabilities, d is assigned to the class with the highest probability for anomaly and lowest for misuse detection.

V. CONCLUSION

The paper describes a Probabilistic-based Fuzzy Frequent Episodes rule mining with sub-attribute utilization which avoids data loss and the classifiers based on the extracted rules have been proposed, which consistently use and combine discrete and continuous attributes in a rule and efficiently extract many good rules for classification. After extracting a number of important frequent episodes rules including normal and
intrusion, a classifier is constructed to classify new connection data into normal, misuse and anomaly intrusion correctly. The probability density function is used to calculate probability of a new connection data belongs to anomaly and misuse class. It is then assigned to the class with the highest probability for anomaly and lower for misuse detection.

REFERENCES