Information-Theoretic Measures for Anomaly Detection

Wenke Lee and Dong Xiang

Dependable Computing Systems Laboratory

DongHoon Shin

11 March 2007

1

Contents

- Introduction
- Motivation and Objective
- Information-Theoretic Measures
- Case Study
- Conclusion

2

Introduction

- Intrusion Detection System (IDS)

  : collects system and network activity data and analyzes the information to determine whether there is an attack occurring.

- Two main techniques for intrusion detection
  - Misuse Detection
    - Misuse detection systems use the patterns of attack behavior or effects to identify a matched activity as an attack instance.
    - Misuse detection are not effective against new attacks.
  - Anomaly Detection
    - Anomaly detection systems use established normal profiles, i.e., expected behavior to identify any unacceptable deviation as possibly the result of an attack
    - Anomaly detection can be effective against new attacks.
    - New legitimate behavior can also be falsely identified as an attack.

3

Introduction (con't)

- Assessment of the state-of-the-art of ID research
  - An evaluation conducted by DARPA in 1998
    - The results showed that the best research system had detection rates below 70%
    - Most of the missed intrusions were new attacks
  - Another evaluation conducted by DARPA in 1999
    - It is conducted with improved capabilities, e.g., the added modules for detecting the attacks missed in the previous evaluation.
    - The research IDSs still had detection rates below 70% because many new attacks were missed.

Even the cutting-edge ID technology is not very effective against new attacks, and the improvement is often too slow and too little to keep up with the "innovation" by sophisticated attackers.

4

Motivation and Objective

- The basic premise for anomaly detection
  - There is intrinsic characteristic or regularity in audit data that is consistent with normal behavior and thus distinct from the abnormal behavior
- Motivation
  - Lack of theoretical understanding and useful tools for characterizing audit data
  - Most anomaly detection models are built based solely on “expert” knowledge or intuition
- Objective
  - The research aims to provide theoretical foundation as well as useful tools that can facilitate the IDS development process and improve the effectiveness of ID technologies.

5

Information-Theoretic Measures

- Information-Theoretic Measures
  - Entropy
  - Conditional Entropy
  - Relative Conditional Entropy
  - Information Gain
  - Information Cost

6

Entropy

- Entropy
  - Definition

For a dataset X where each data item belongs to a class $x \in C_X$, the entropy of X relative to this $|C_X|$-wise classification is defined as

$$H(X) = \sum_{x \in C_X} P(x) \log \frac{1}{P(x)}$$

where $P(x)$ is the probability of x in X.

  - It measures the uncertainty (or impurity) of a collection of data item.
  - The entropy value is smaller when the class distribution is skewer, i.e., when the data is "purer".

7

Entropy (con't)

- The entropy value is larger when the class distribution is more even, i.e., when the data is more "impure".
- For anomaly detection, we can use entropy as a measure of the regularity of audit data.
- The smaller the entropy, the fewer the number of different records, i.e., the more regularity the audit dataset.
- High-regularity data contains redundancies that help predicting future events because the fact that many events are repeated (or redundant) in the current dataset suggests that they will likely to appear in the future.
- Therefore, anomaly detection model constructed using dataset with smaller entropy will likely be simpler and have better detection performance.

8

Conditional Entropy

- Conditional Entropy
  - Definition

The conditional entropy of X given Y is the entropy of the probability distribution $P(x \mid y)$, that is,

$$H(X \mid Y) = \sum_{x,y \in C_X, C_Y} P(x,y) \log \frac{1}{P(x \mid y)}$$

where $P(x, y)$ is the joint probability of x and y and $P(x \mid y)$ is the conditional probability of x given y.

  - For anomaly detection, we can use conditional entropy as a measure of regularity of sequential dependencies.

Conditional Entropy (con't)

- Let X be a collection of sequences where each is denoted as $(e_1, e_2, \ldots, e_{n-1}, e_n)$, and each e_i is an audit event.
  Let Y be the collection of subsequences where each is $(e_1, e_2, \ldots, e_{k-1}, e_k)$, and k<n.
  Then, the conditional entropy H(X|Y) tells us how much uncertainty remains for the rest of audit events in a sequence x after we have seen y, i.e., the first k events of x.
  For example, let X = {aaaaa, bbbbb, ccccc, …}, then the conditional entropy is 0 and the event sequences are deterministic.
- The smaller the conditional entropy, the better
- Conversely, a large conditional entropy indicates that the sequences are less deterministic and thus much harder to model.

Relative Conditional Entropy

- Relative Entropy
  - Definition

The relative entropy between two probability distributions $p(x)$ and $q(x)$ that are defined over the same $x \in C_X$ is

$$relEntropy(p \mid q) = \sum_{x \in C_X} p(x) \log \frac{p(x)}{q(x)}$$

  - Relative entropy measures the distance of the regularities between two datasets.
  - For anomaly detection, we often build a model using a training dataset and apply the model to the test dataset.

11

Relative Conditional Entropy

  - These datasets must have the same (or very similar) regularity for the anomaly detection model to attain high performance.
  - It is obvious that the smaller relative entropy, the better.

- Relative Conditional Entropy
  - Definition

The relative conditional entropy between two probability distributions $p(x \mid y)$ and $q(x \mid y)$ that are defined over the same $x \in C_X$ and $y \in C_Y$ is

$$relCondEntropy(p \mid q) = \sum_{x,y \in C_X, C_Y} p(x,y) \log \frac{p(x \mid y)}{q(x \mid y)}$$

  - Again, for anomaly detection, the smaller the relative conditional entropy, the better

12

Information Gain and Classification

- Intrusion detection can be cast as a classification problem
  - An audit event can be classified as belonging to
    - Normal class
    - Abnormal class (in the case of anomaly detection)
    - A particular class of intrusion (in the case of misuse detection)
- Construction of Classifier
  - In a training dataset, the records are defined by a set of features and each record belongs to a class.
  - After applying a sequence of feature value tests, the dataset can be partitioned into “pure” subsets.
  - The sequence of feature value tests can be used as the conditions in the classifier to determine the class of a new record.
  - It is obvious that the purer the final subsets, the more accurate the classifier.

13

Information Gain and Classification

- Information Gain
  - Definition

The information gain of attribute (i.e., feature) A on dataset X is

$$Gain(X, A) = H(X) - \sum_{v \in Values(A)} \frac{|X_v|}{|X|} H(X_v)$$

  - It is the reduction of entropy when the dataset is partitioned according to the feature values.
  - Therefore, when constructing a classifier, a classification algorithm needs to search for features with high information gain.

14

Information Cost

- Trade-off of the detection performance versus the amount of information
  - Intuitively, the more information we have, the better the detection performance. However, there is always a cost for gain.
  - For intrusion detection, one important goal is to detect intrusions as early as possible, which makes the model simpler but causes the degradation of detection performance

Information Cost
: average time for processing an audit record and checking against the detection model

15

Application in Anomaly Detection

- Approach for anomaly detection using the information theoretic measures

1. Measure regularity of audit data and perform appropriate data transformation
   - Iterate this step if necessary so that the dataset used for modeling has high regularity.
2. Determine how the model should be built, i.e., how to achieve the best performance or the optimal performance/cost trade-off, according to the regularity measure
3. Use relative entropy to determine whether a model is suitable for a new dataset (e.g., from a new environment)

16

Case Studies

- Three case studies are provided to show how to use the information-theoretic measures defined to build anomaly detection models.
- Experiments on
  - University of New Mexico (UNM) sendmail system call data
  - MIT Lincoln Lab sendmail BSM data
  - MIT Lincoln Lab tcpdump data

17

UNM sendmail System Call Data

- A ground-breaking study by Forrest et al ("A sense of self for Unix processes")
  - The short sequences of system calls made by a program during its normal executions are very consistent.
  - More importantly, the sequences are different from the sequences of its abnormal executions as well as the executions of other programs.
  - Therefore, a very concise database containing these normal sequences can be used as the definition of the normal behavior of a program and as the basis to detect anomalies.
  - However, a means to determine the appropriate sequence length was not suggested, rather, an ad hoc trial-and-error approach was used.

18

UNM sendmail System Call Data

open, read, mmap, mmap, open, getrlimit, mmap, close

< Trace of System Calls for Training Dataset >

call	position 1	position 2	position 3
open	read, getrlimit	mmap	mmap, close
read	mmap	mmap	open
mmap	mmap, open, close	open, getrlimit	getrlimit, mmap
getrlimit	mmap	close	
close			

< Database for Normal Behavior >

open, read, mmap, open, open, getrlimit, mmap, close

< Intrusion Traces >

19

UNM sendmail System Call Data

- How to measure the data regularity and use it to determine the sequence length
  - Using a sliding window of size n, a system call trace is process into a set of length-n sequences
    - This set is used as our dataset, X
    - Each sequence is a data point, x
  - Let X represent the set of length-n sequences and Y be the set of (prefix) subsequences of the length n-1
  - The conditional entropy H(X|Y) measures the regularity of how the first n-1 system calls determines the nth system call.

20

UNM sendmail System Call Data

- Mathematically,

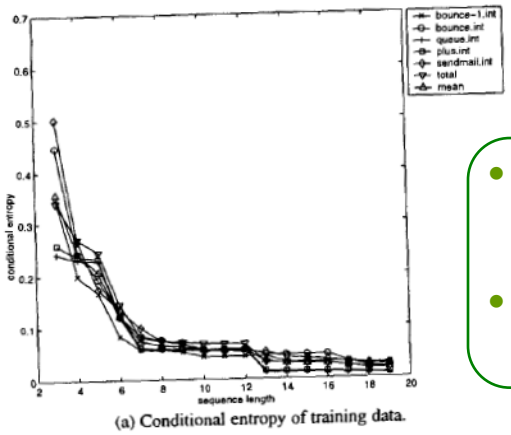
For each unique $x \in X$, $|x|$ is the number of occurrences of x in X, and $y(x)$ is the length $n-1$ subsequence of x, i.e.,
if $x = (s_1 s_2 \ldots s_{n-1} s_n)$, then $y(x) = (s_1 s_2 \ldots s_{n-2} s_{n-1})$.
Then, the conditional entropy is

$$H(X \mid Y) = \sum_{x \in X} \frac{|x|}{|X|} \log \frac{|y(x)|}{|x|}.$$

21

UNM sendmail System Call Data

- Conditional entropy of training data



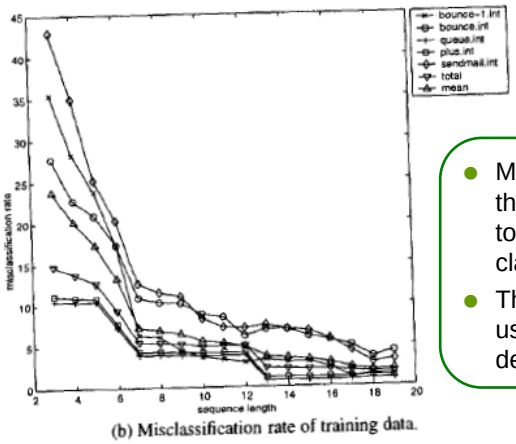
(a) Conditional entropy of training data.

- Each graph represents a particular kind (or configuration) of normal sendmail runs
- The more information, the more deterministic (i.e., regular)

22

UNM sendmail System Call Data

- Misclassification rate of training data



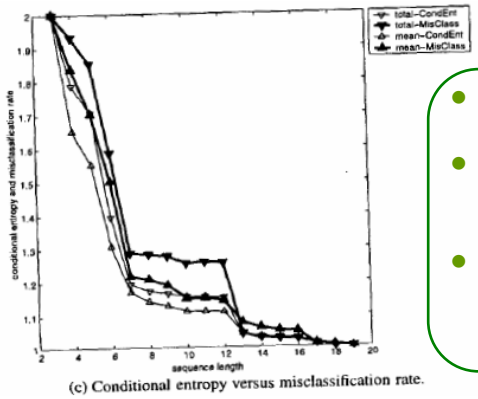
(b) Misclassification rate of training data.

- Misclassification means that the classifier predicts an item to be in class i while the actual class is j.
- The misclassification rate is used to measure anomaly detection performance.

23

UNM sendmail System Call Data

- Comparison of misclassification rate on the training data and conditional entropy



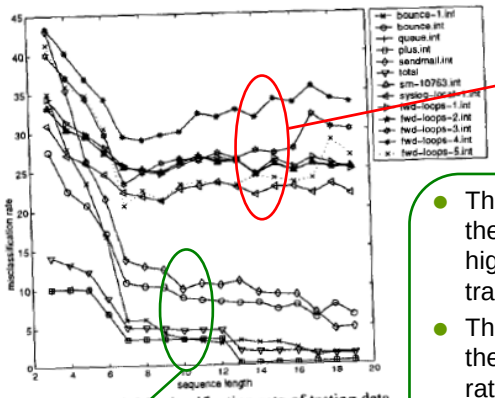
(c) Conditional entropy versus misclassification rate.

- The values are all scaled into 1 to 2 range
- The trend of misclassification rate coincides with the trend of conditional entropy
- Conditional entropy can be considered as the estimated trend of misclassification to select a sequence length

24

UNM sendmail System Call Data

- Misclassification rate on testing data



(d) Misclassification rate of testing data.

Intrusion traces

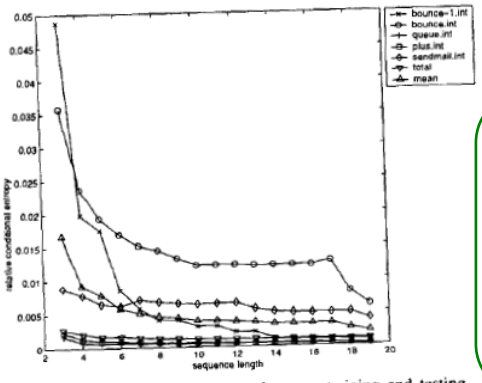
normal traces

- The misclassification rates for the intrusion traces are much higher than those for the normal traces
- This suggests that we can use the range of the misclassification rates as the indicator of whether a given trace is normal or intrusion

25

UNM sendmail System Call Data

- Relative conditional entropy between training and testing normal data



(e) Relative conditional entropy between training and testing normal data.

- When, relative conditional entropy, the misclassification rate on testing normal data is also higher
- This suggests that we can use R.C.E between the training and testing sets to discover the difference of regularity between them.

26

Conclusion

- In this paper, some information-theoretic measures for anomaly detection have been proposed.
  - Entropy
  - Conditional Entropy
  - Relative (Conditional) Entropy
  - Information Gain
- Case studies on sendmail system call data were provided to show how to use the information-theoretic measures to build anomaly detection models.

27