FUZZY DATA MINING AND GENETIC ALGORITHMS APPLIED TO INTRUSION DETECTION

Susan M. Bridges

Bridges@cs.msstate.edu

Rayford B. Vaughn

vaughn@cs.msstate.edu

23rd National Information Systems Security Conference

October 16-19, 2000

OUTLINE

- AI and Intrusion Detection
- Intrusion Detection System Design
- Fuzzy Logic and Data Mining
  - Fuzzy Association Rules
  - Fuzzy Frequency Episodes
- Intrusion Detection via Fuzzy Data Mining
- GA's for System Optimization
- Summary and Future Work

AI TECHNIQUES AND INTRUSION DETECTION

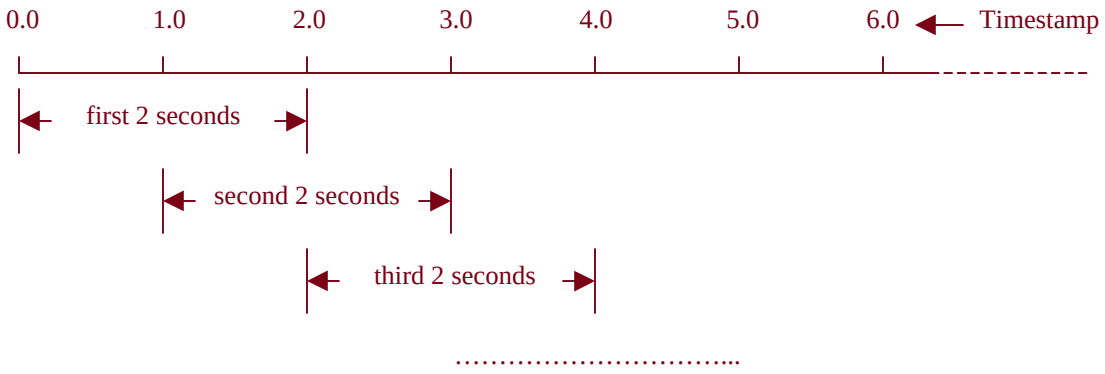
- Long history of AI techniques applied to intrusion detection. For example:
  - Rule-Based Expert Systems(Lunt and Jagannathan 1988)
  - State Transition Analysis (Ilgun and Kemmerer 1995)
  - Genetic Algorithms (Me 1998)
  - Inductive Sequential Patterns (Teng, Chen and Lu 1990)
  - Artificial Neural Networks (Debar, Becker, and Siboni 1992)
- Data mining applied to intrusion detection is an active area of research. Examples include:
  - Lee, Stolfo, and Mok (1998)
  - Barbara, Jajodia, Wu, and Speegle (2000)

UNIQUE FEATURES OF OUR WORK

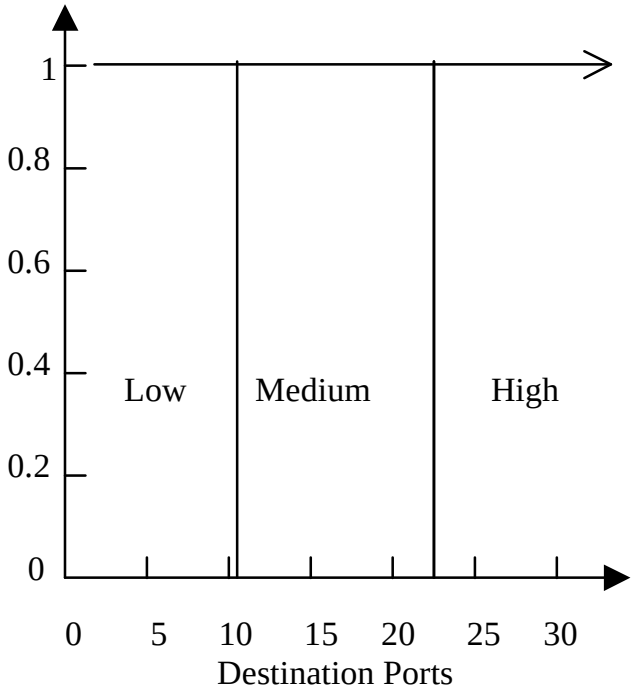
- Combines fuzzy logic with data mining
- Overcomes sharp boundary problems of many systems
- Reduces false positive errors
- Can be used for both anomaly detection and misuse detection
- Includes real-time components
- Uses genetic algorithms for system optimization

FUZZY LOGIC AND SECURITY

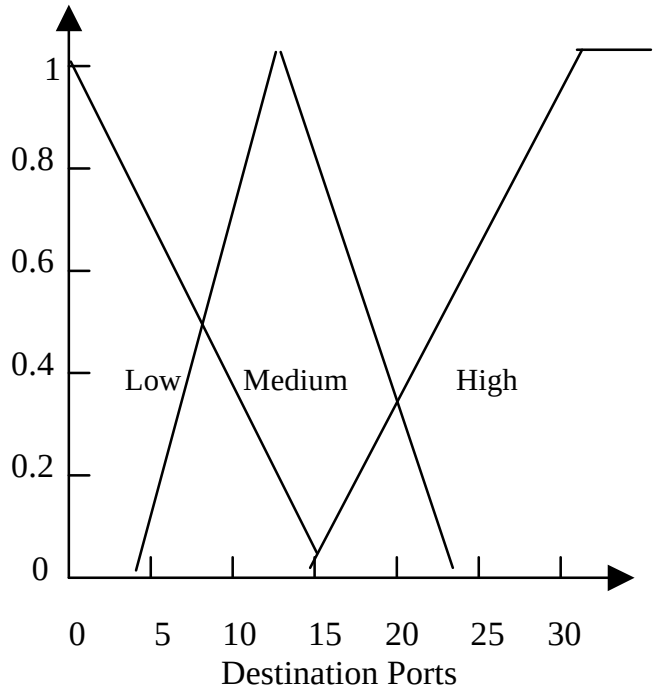
- Many security-related features are quantitative
  - e.g., temporal statistical measurements
    (Porras and Valdes 1998; Lee and Stolfo 1998)
    - » SN: number of SYN flags in TCP header during last 2s
    - » FN: number of FIN flags in TCP header during last 2s
    - » RN: number of RST flags in TCP header during last 2s
    - » PN: number of distinct destination ports during last 2s



FUZZY LOGIC ALLOWS OVERLAPPING CATEGORIES

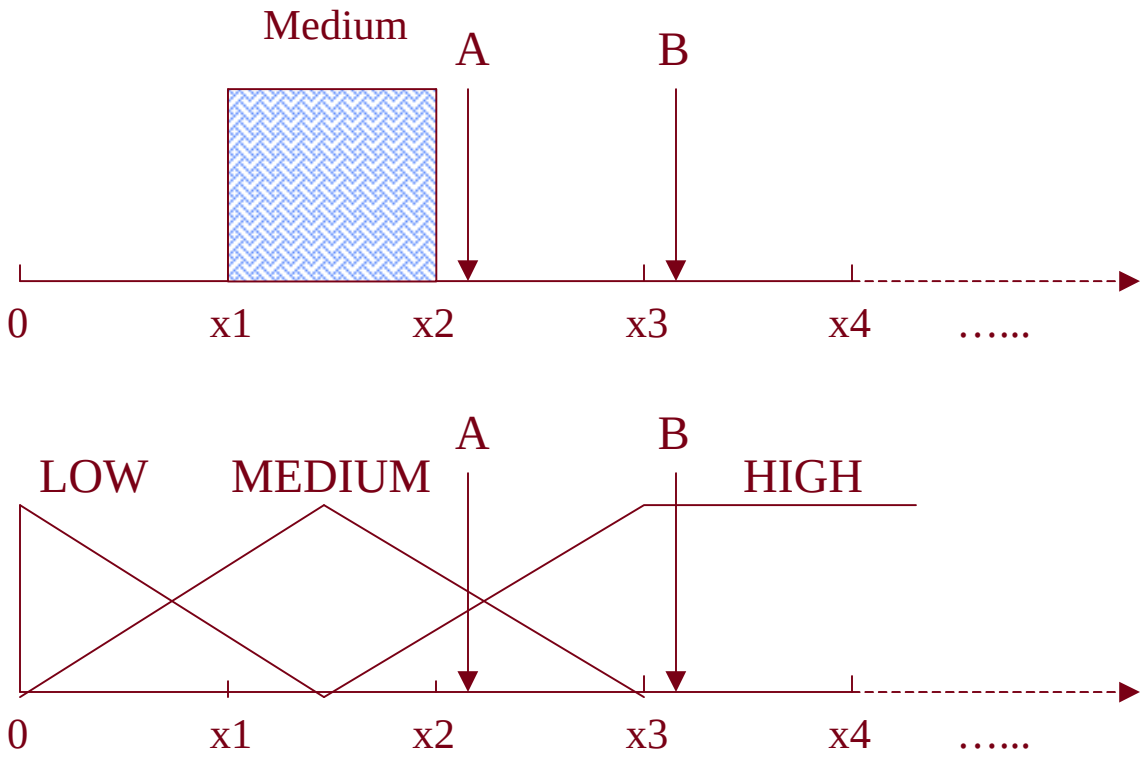


a. Non-fuzzy sets



b. Fuzzy sets

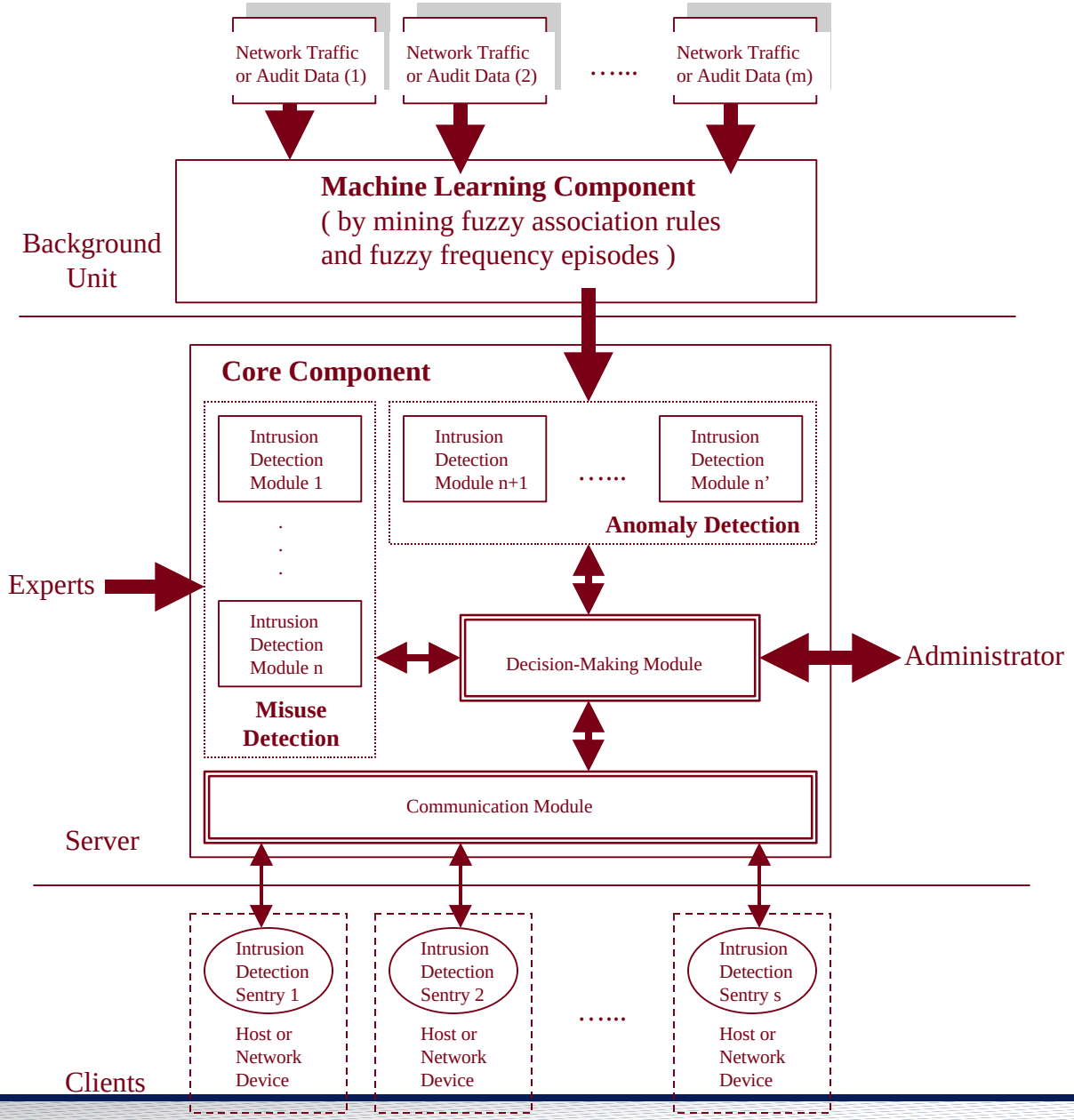
FUZZY LOGIC OVERCOMES SHARP BOUNDARY PROBLEMS



If Medium is the normal pattern, then without fuzzy sets, A & B are both “outside” of the normal pattern. Fuzzy logic allows “degrees of normality.”

INTELLIGENT INTRUSION DETECTION MODEL

- Integration of fuzzy logic with data mining
  - Fuzzy association rules
  - Fuzzy frequency episodes
- Preliminary architecture
  - Includes both misuse detection and anamoly detection
  - Integrates machine-level and network-level information
- Optimization using genetic algorithms


Network Traffic or Audit Data (1)
Network Traffic or Audit Data (2)
.....
Network Traffic or Audit Data (m)
Machine Learning Component
(by mining fuzzy association rules and fuzzy frequency episodes)
Background Unit
Core Component
Intrusion Detection Module 1
Intrusion Detection Module n+1
.....
Intrusion Detection Module n'
Anomaly Detection
Experts
Intrusion Detection Module n
Misuse Detection
Decision-Making Module
Administrator
Communication Module
Server
Intrusion Detection Sentry 1
Intrusion Detection Sentry 2
.....
Intrusion Detection Sentry s
Host or Network Device
Host or Network Device
Host or Network Device
Clients

MINING FUZZY ASSOCIATION RULES

- Association rules represent commonly found patterns in data.
- Association Rule Format: *R1: X ® Y, c, s*

  X and Y are disjoint sets of items

  s (support) tells how often X and Y co-occur in the data

  c (confidence) tells how often Y is associated with X.
- Our system is unique: X and Y are fuzzy variables that take fuzzy sets as values

FUZZY ASSOCIATION RULES

Sample Fuzzy Association Rule:

{ SN=*LOW*, FN=*LOW* } $\rightarrow$ { RN=*LOW* }, c = 0.924, s = 0.49

Interpretation:

SN, FN, and RN are fuzzy variables.

The pattern { SN=*LOW*, FN=*LOW*, RN=*LOW* } has occurred in 49% of the training cases;

When the pattern { SN=*LOW*, FN=*LOW* } occurs, there will be 92.4% probability that { RN=*LOW* } will occur at the same time.

SAMPLE FUZZY FREQUENCY EPISODE RULES

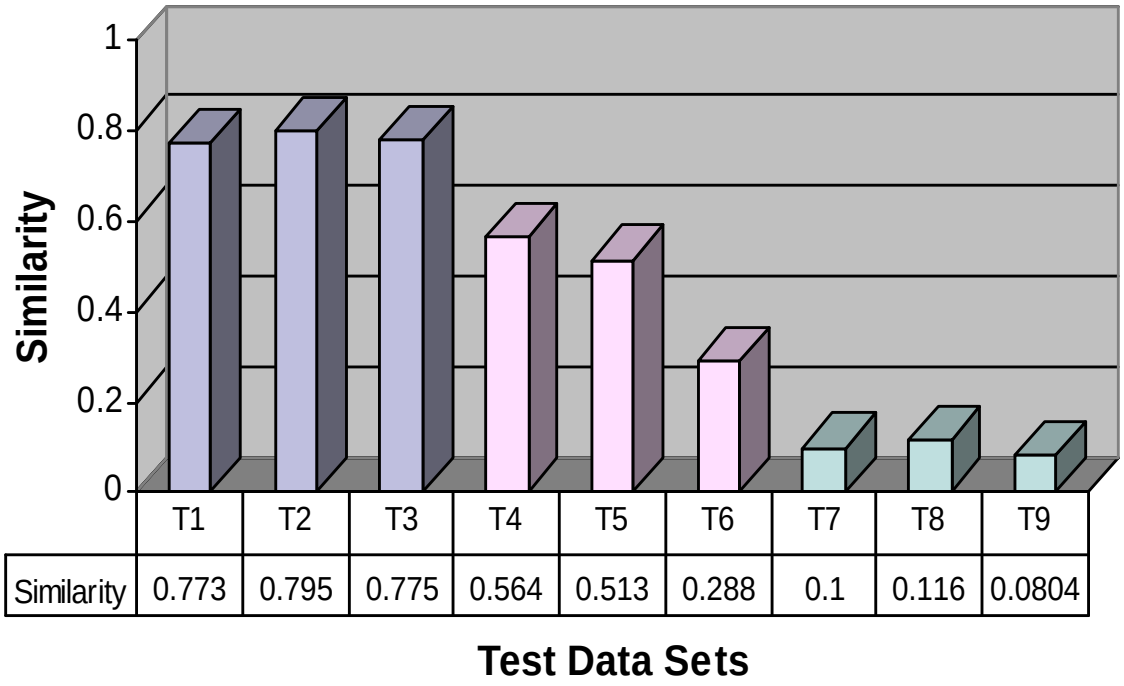
{ E1: PN=*LOW*, E2: PN=*MEDIUM* } →{ E3:PN=*MEDIUM* }

c = 0.854, *s* = 0.108, *w* = 10 seconds

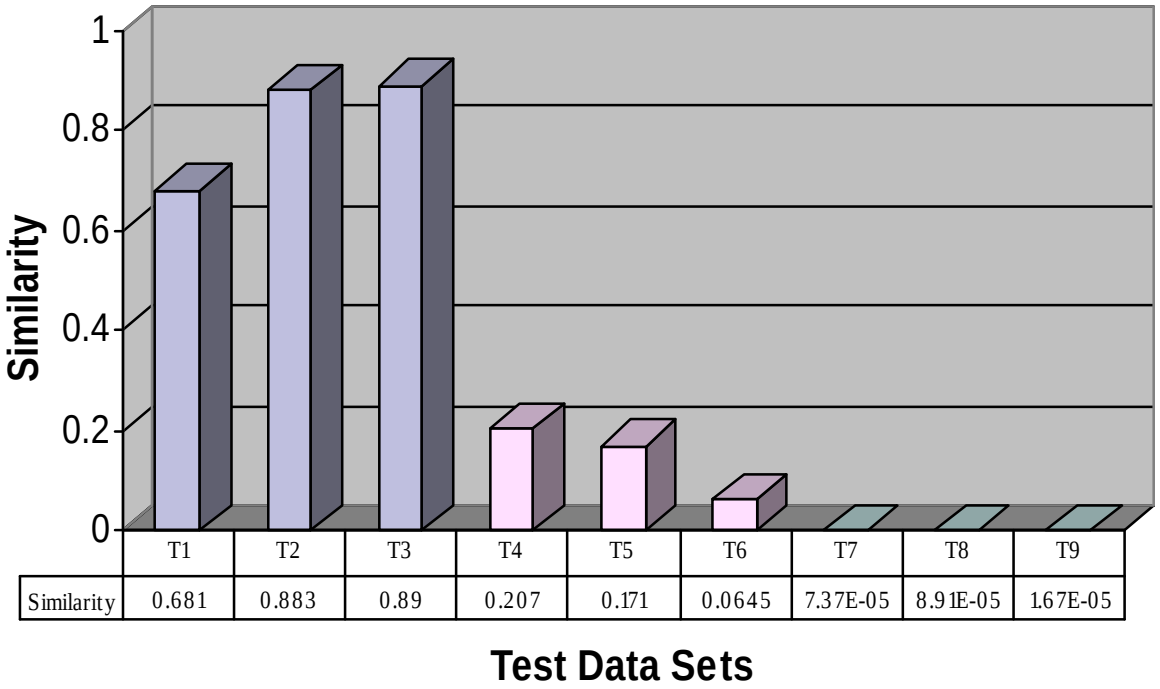
- E1, E2 and E3 are events occurring within the time window 10 seconds.
- PN is a fuzzy variable
- The events occur in the order E1, E2, E3, but there may also be intervening events
- { PN=*LOW*, PN=*MEDIUM*, PN=*MEDIUM* } has occurred 10.8% in all training cases;
- When { PN=*LOW*, PN=*MEDIUM* } occurs, { PN=*MEDIUM* } will follow with 85.4% probability.

FUZZY DATA MINING FOR INTRUSION DETECTION

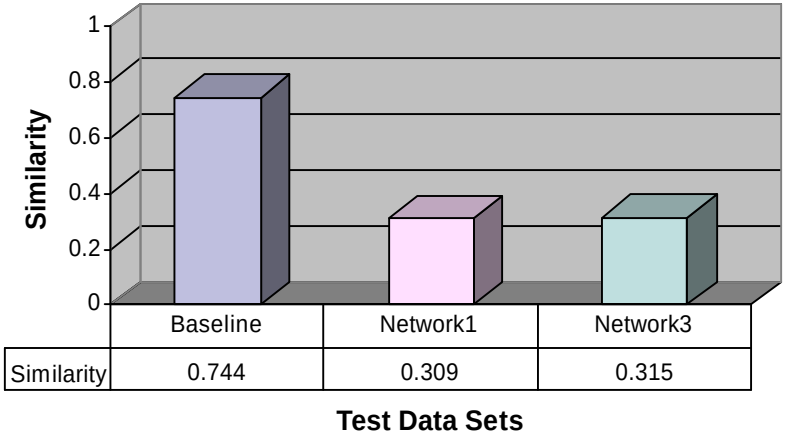
- Modification of non-fuzzy methods developed by Lee, Stolfo, and Mok (1998)
- Anomaly Detection Approach
  - Mine a set of fuzzy association rules from data with no anomalies.
  - When given new data, mine fuzzy association rules from this data.
  - Compare the similarity of the sets of rules mined from new data and “normal” data.



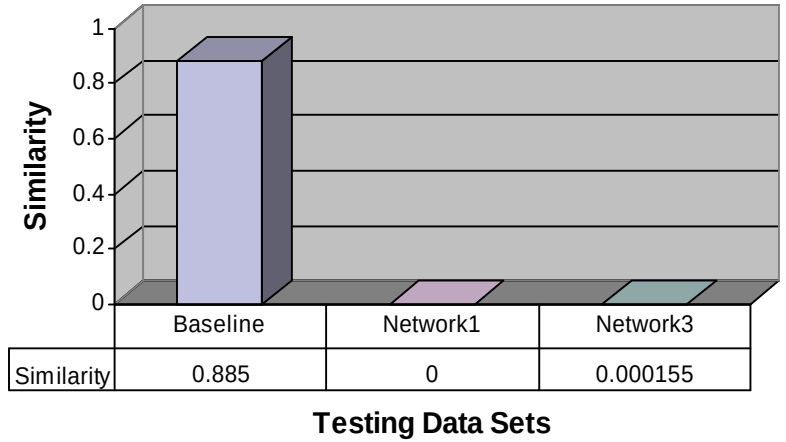
Similarities between Training Data Set and Different Test Data Sets by Mining Fuzzy Association Rules on SN, FN, and RN.
Training data collected in the afternoon.
T1-T3—afternoon T4-T6—evening T7-T9—late night
Data source: MSU CS network



Similarities between Training Data Set and Different Test Data Sets by Mining Fuzzy Frequency Episodes on PN.
Training data collected in the afternoon.
T1-T3—afternoon T4-T6—evening T7-T9—late night
Data source: MSU CS network



Similarities between Training Data Set and Different Test Data Sets by Mining Fuzzy Association Rules on SN, FN, and RN



Similarities between Training Data Set and Different Test Data Sets by Mining Fuzzy Frequency Episodes on PN

Training data: no intrusions

Test data: *baseline* (no intrusions)

network1 (includes simulated IP Spoofing intrusions)

network3 (includes simulated port scanning intrusions)

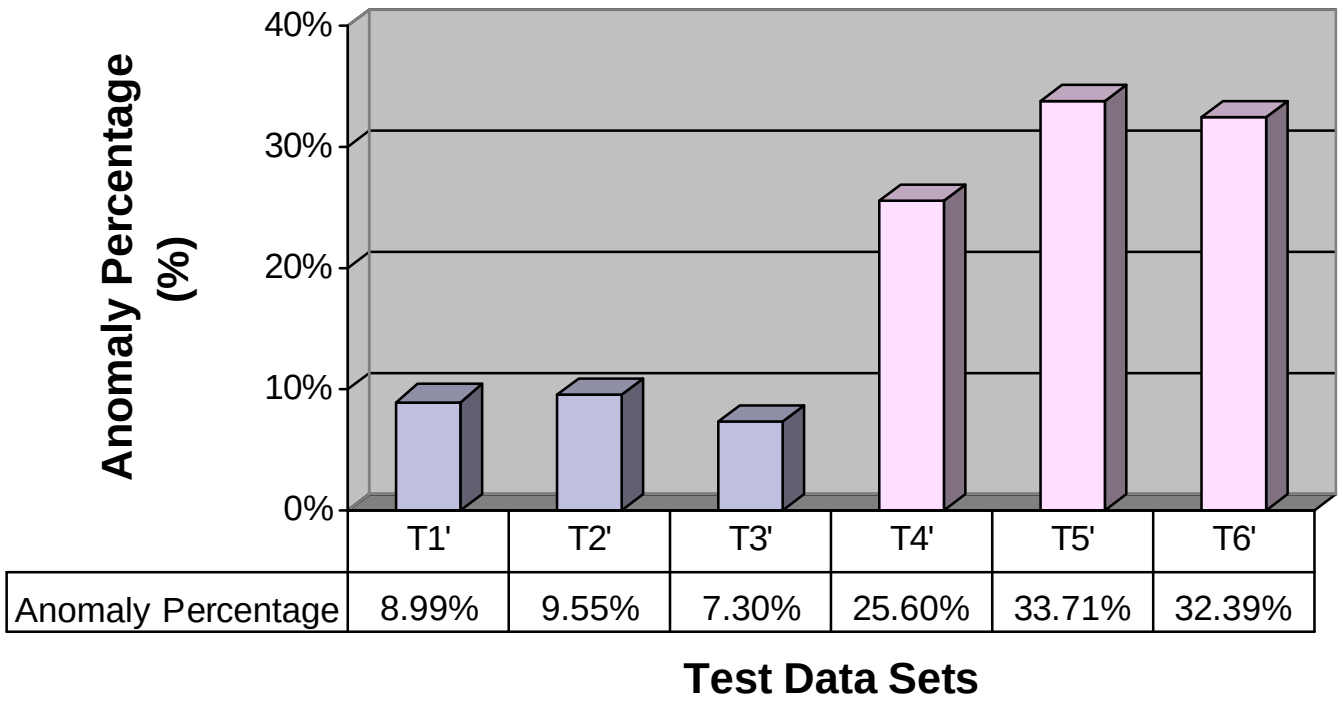
REAL-TIME INTRUSION DETECTION

Given a fuzzy episode rule R: $\{ e_1, \ldots, e_{k-1} \} \rightarrow \{ e_k \}$, c, s, w,
if $\{e_1, \ldots, e_{k-1}\}$ has occurred in the current event sequence,
then $\{ e_k \}$ can be predicted to occur next with confidence of c.

If the next event does not match any prediction from the rule set, it will be alarmed as an anomaly.

Define

anomaly percentage = number of anomalies / number of events



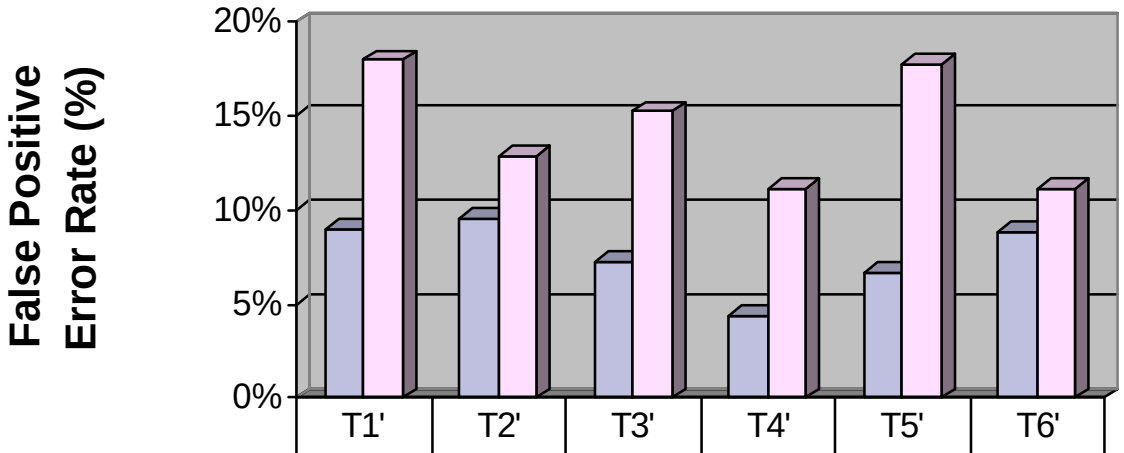
Anomaly Percentages of Different Test Data Sets in Real-time Intrusion Detection by Mining Fuzzy Frequency Episodes on PN

Training Data: No intrusions

Test Data: T1'-T3'—no intrusions T4'-T6'—simulated mscan

FUZZY VS. NON-FUZZY

Comparing the false positive error rates of fuzzy episode rules with non-fuzzy versions for real-time intrusion detection



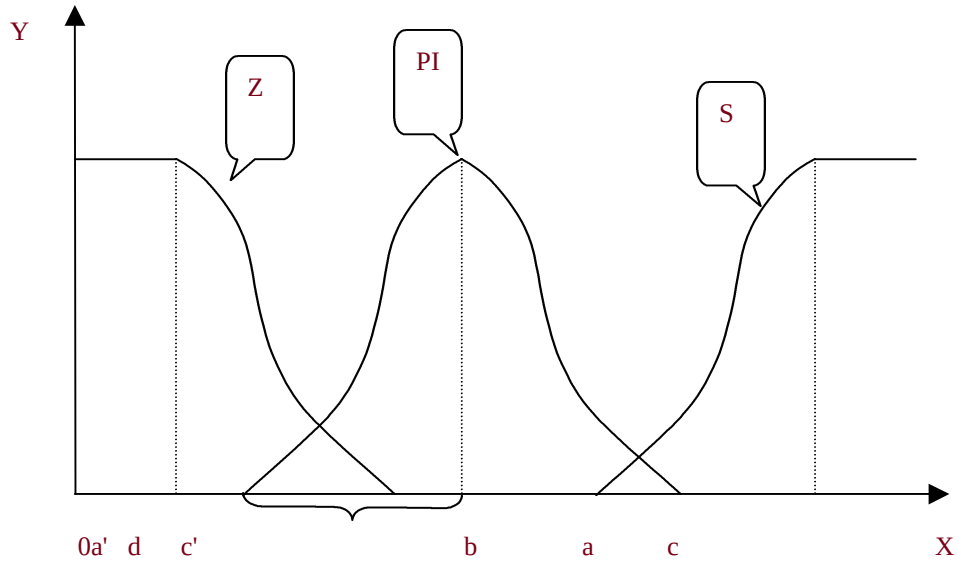
USING GENETIC ALGORITHMS FOR OPTIMIZATION

- Problem with NIDS
  - System uses a fixed set of features for all kinds of situations
  - Fuzzy membership functions must be predefined.
- Hypothesis
  - Different features may be useful for different classes of intrusion attacks and for different situations.
  - The performance of the system can be improved by using a GA to evolve an optimal set of features and fuzzy membership functions.

GENETIC ALGORITHMS APPROACH

- Optimization goals
  - Maximize the similarity of rules mined from "normal" data with baseline rule set
  - Minimize the similarity of rules mined from "abnormal" data with baseline rule set
- Parameters to change
  - Features available from audit data
  - Fuzzy membership function parameters

FUZZY SETS ARE DEFINED BY PARAMETERIZED MEMBERSHIP FUNCTIONS



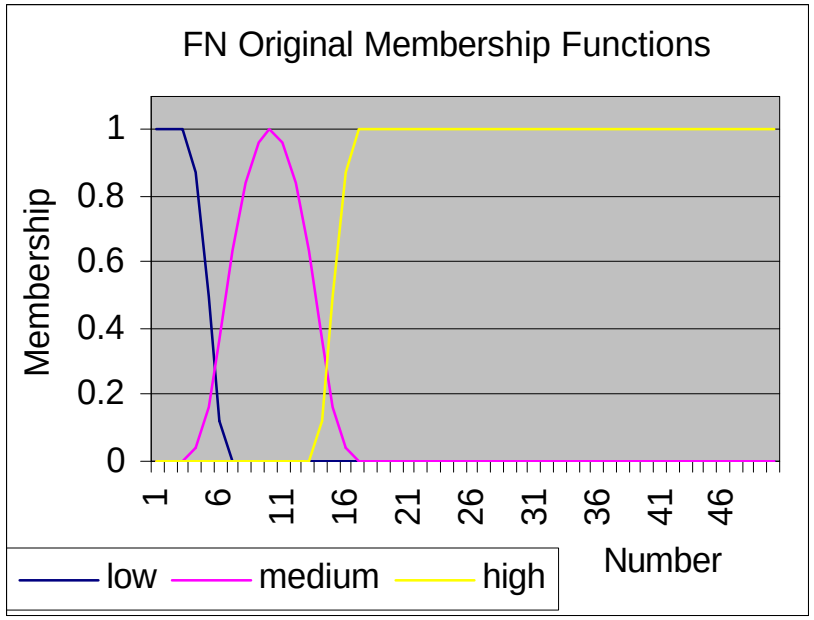
$$S(m,a,c)=\begin{cases}0 & m\le a \\ 2\left(\frac{m-a}{c-a}\right)^2 & a<m\le\frac{a+c}{2} \\ 1-2\left(\frac{c-m}{c-a}\right)^2 & \frac{a+c}{2}<m\le c \\ 1 & c<m\end{cases}$$

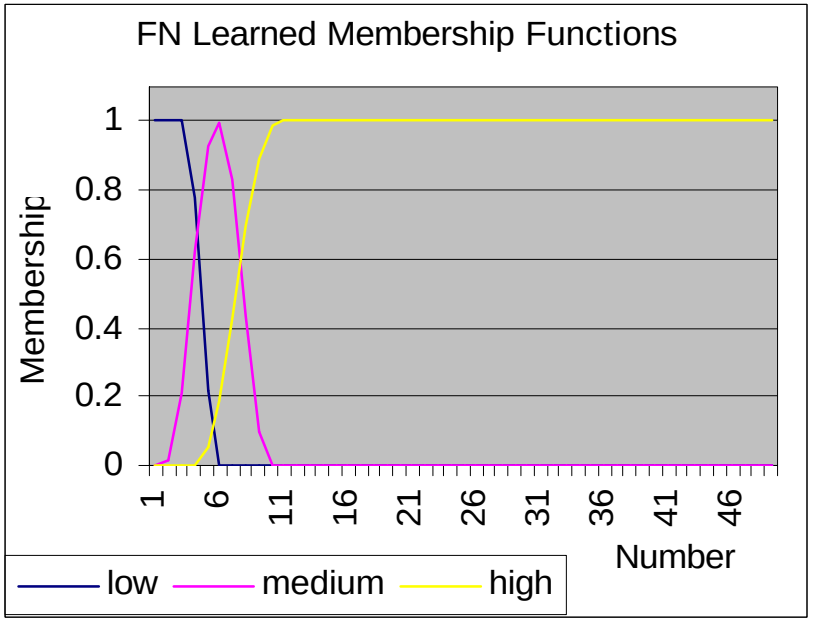
$$Z(m,a',c')=1-S(m,a',c')$$

$$PI(m,d,b)=\begin{cases}S(m,b-d,b) & m<b \\ Z(m,b,b+d) & b<m\end{cases}$$

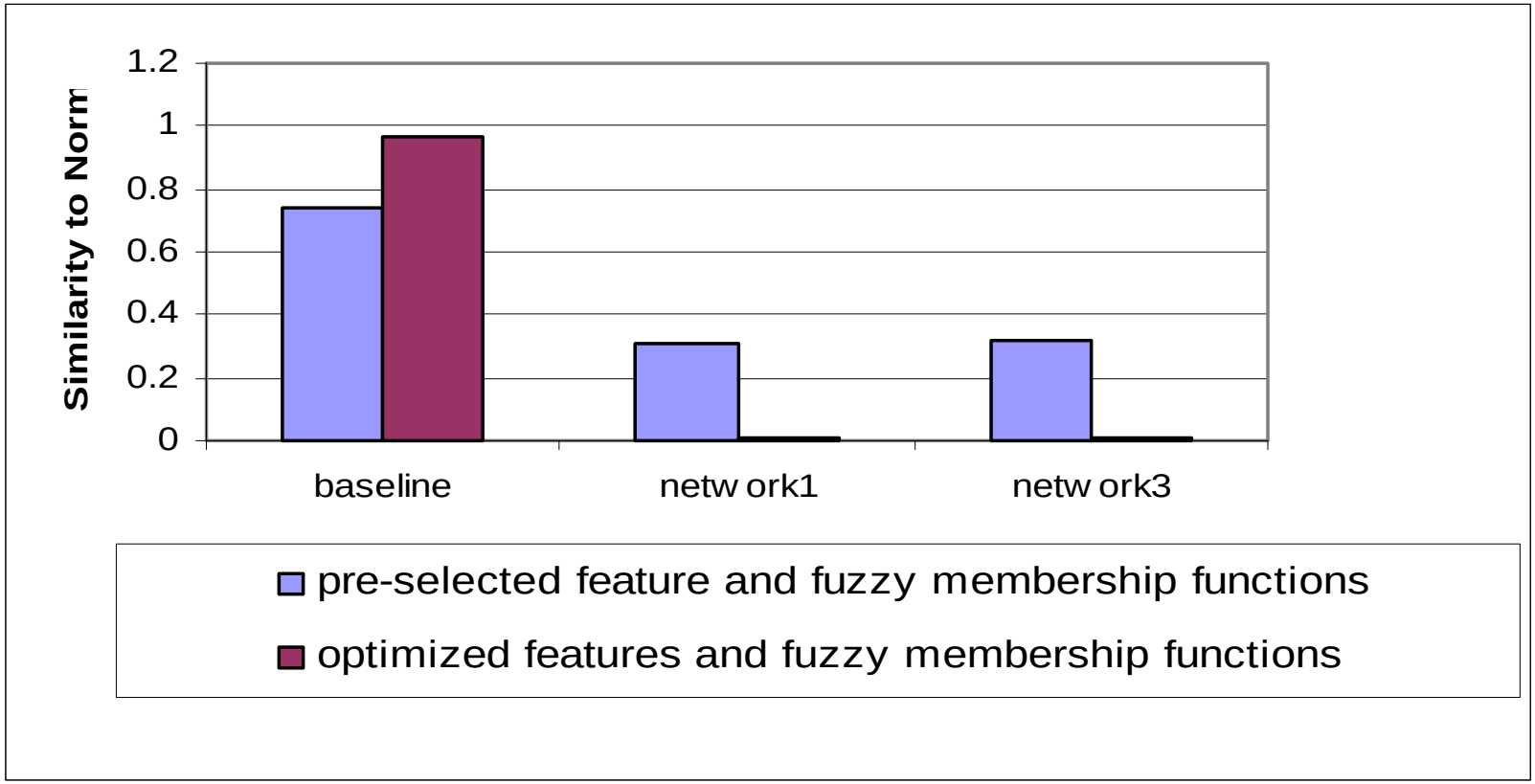
EXAMPLE RUN

Membership functions before and after the GA optimization





OPTIMIZATION RESULTS



baseline (no intrusions)

network1 (includes simulated IP Spoofing intrusions)

network3 (includes simulated port scanning intrusions)

FEATURES SELECTED FOR IP SPOOFING AND PORT SCANNING ATTACKS

	Feature Selected	**Luo's Features**
IP Spoofing	Source IP, FIN, Data Size, Port number	SYN, FIN, RN
Port Scanning	Source IP, Destination IP, Source Port, and Data size.	SYN, FIN, RN

CONCLUSIONS

- Developed an architecture for integrating machine learning methods with other intrusion detection methods.
- Extended data mining techniques by integrating fuzzy logic
- Demonstrated that these methods are superior to their non-fuzzy counterparts.
- Developed a method for real-time intrusion detection using fuzzy frequency episodes.
- Used GA's to improve the performance of the system by selecting best set of features and by tuning the fuzzy membership function parameters

Current and Future Work

- Further work with fuzzy frequency episodes and real-time intrusion detection
- Using fuzzy logic for data fusion by the decision module
- Generating misuse modules from association rules
- Using incremental data mining to deal with “drift” in “normality”
- Investigating intrusion detection in high speed clusters of workstations