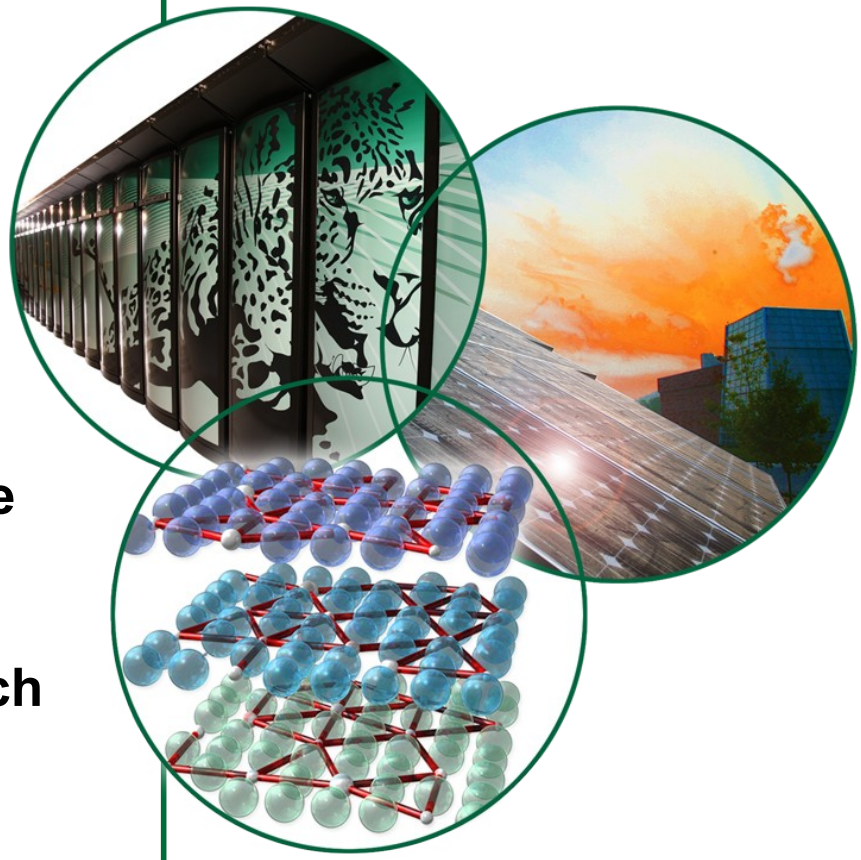
Data Clustering for Anomaly Detection in Network Intrusion Detection

Jose F. Nieves
Polytechnic University of Puerto Rico
Research Alliance in Math and Science

Dr. Yu (Cathy) Jiao
Applied Software Engineering Research
Oak Ridge National Laboratory



August 2009

U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Outline

- Introduction and motivation
- Background
  - Network intrusion detection
  - Anomaly detection
  - Data clustering
- Performance evaluation
- Summary

Introduction and motivation

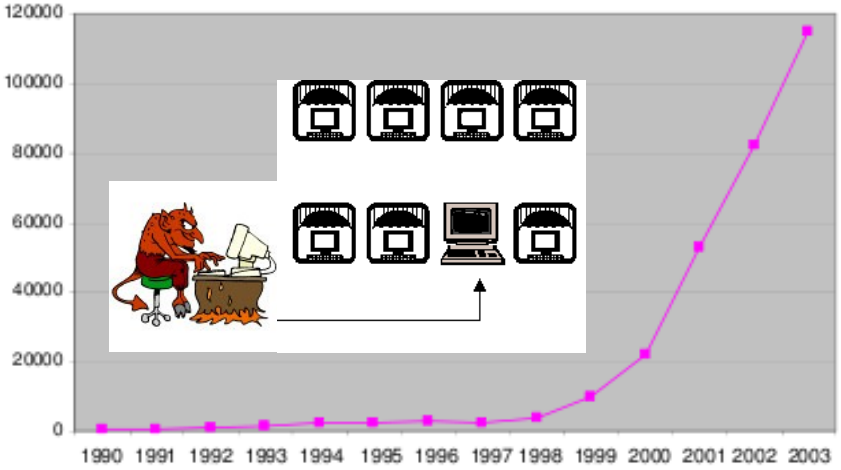


Figure 1. Computer incidents report (CERT/CC)

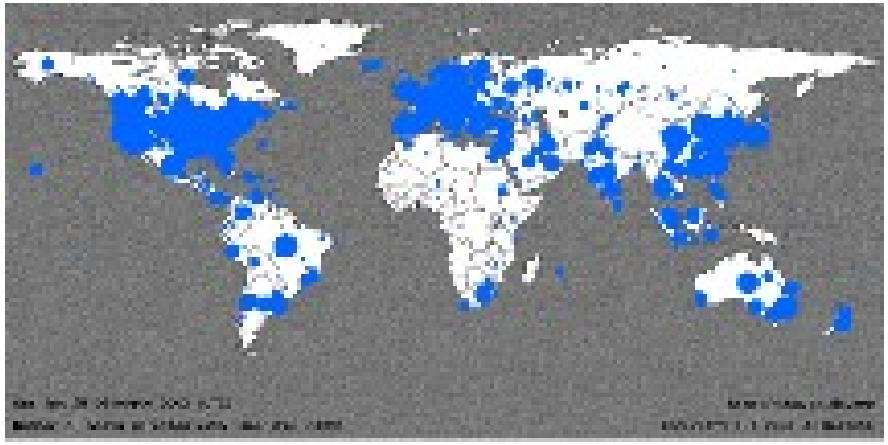


Figure 2. Worm spread after 30 min www.caida.org

- **Network intrusions pose a high security risk**
- **NIDS aim to identify intrusions**
- **NIDS capable of detecting new emerging threats**
- **Main goal**
  - **Present different methods for network intrusion detection**
  - **Select one method and evaluate its performance**

Network intrusion detection methods

- Signature-based
  - Based on signatures of known attacks
  - Signature database has to be manually updated
  - Expensive and time consuming
  - Cannot detect new emerging threats
- Misuse detection
  - Models built from labeled data sets
  - Samples labeled as 'normal' or 'attack'
  - Expensive and time consuming
  - Can not detect new threats
- Anomaly detection
  - Identify anomalies, deviations from 'normal' behavior
  - Can detect new emerging threats
  - Potential false alarm rate

Anomaly detection

- Supervised methods
  - Models built from 'normal' labeled samples
  - Expensive and time consuming
- Unsupervised method (data clustering)
  - Group unlabeled patterns into clusters based on similarities
  - Do not required labeled samples
  - Anomalies are deviation from 'normal' clusters

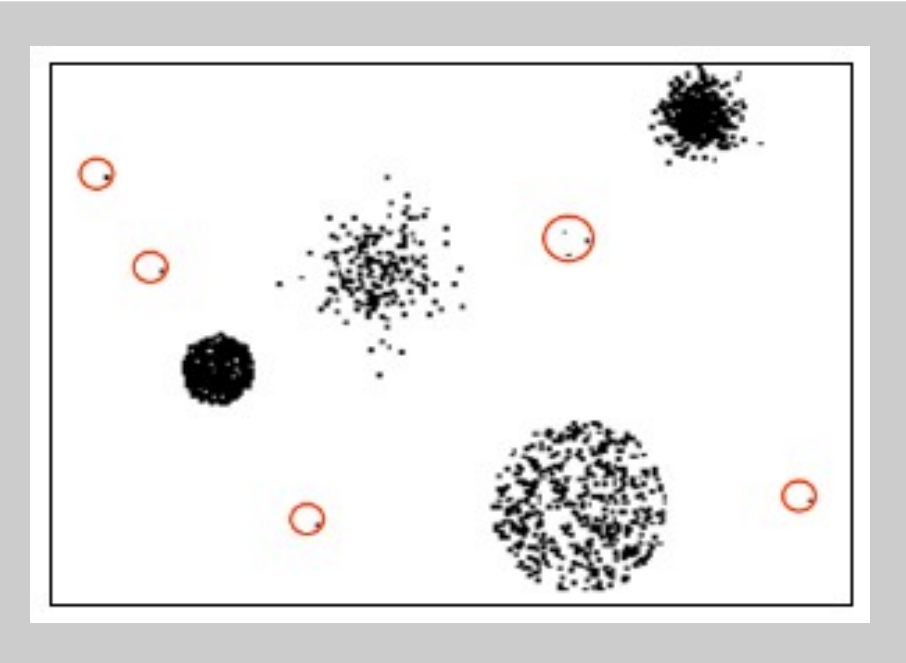


Figure 3. Clustering for anomaly detection

Basic steps for data clustering

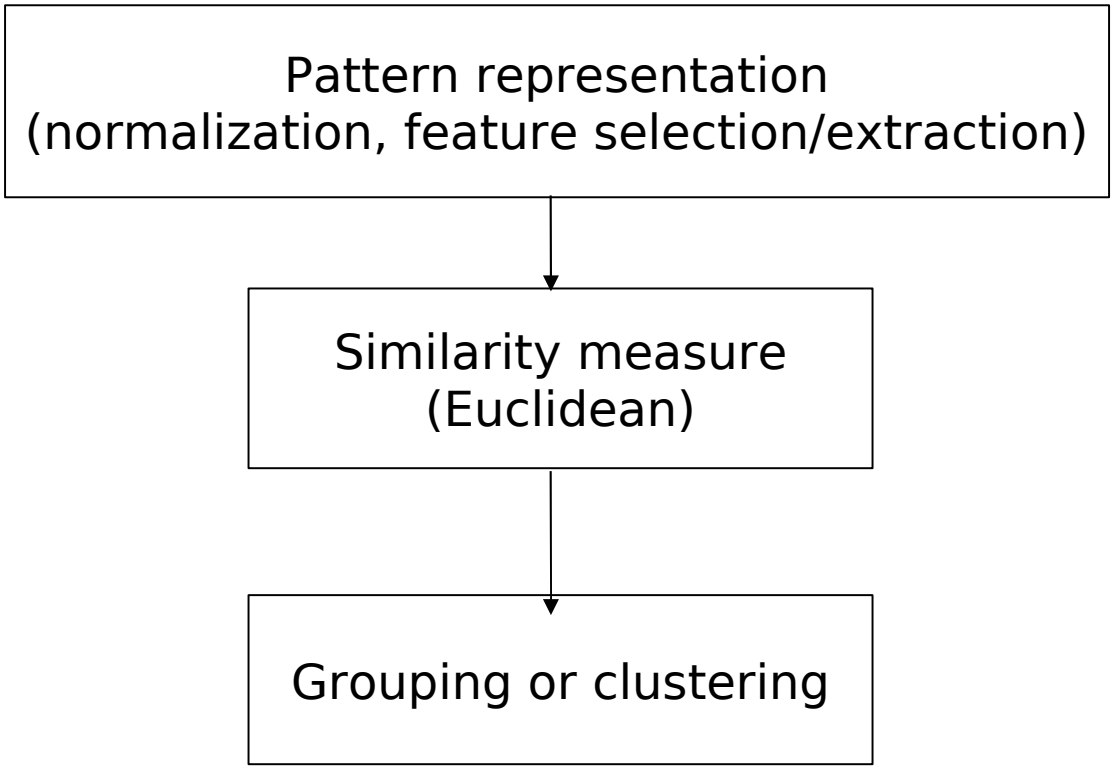


Figure 4. Data clustering steps

Algorithm

- Kmeans
  - Choose number of clusters
  - Initialize centroids or clusters centers
  - Assign each sample to nearest cluster centroid
  - Calculate means to be new centroid of each cluster

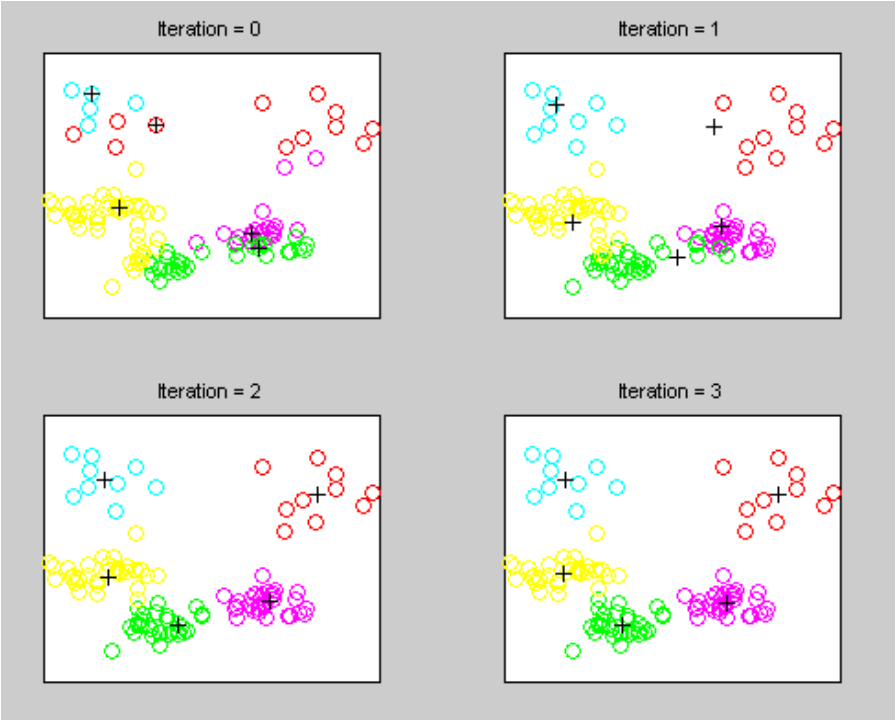
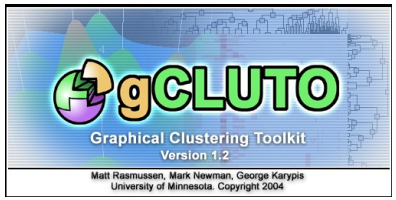
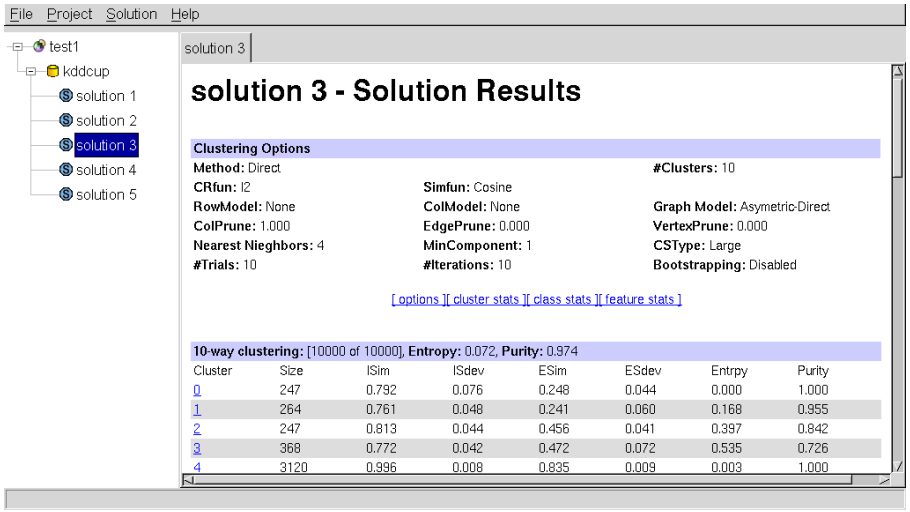


Figure 5. Kmeans clustering algorithm

Software

- **Cluto (University of Minnesota)**
  - **Clustering toolkit**
  - **Simple interface**
  - **3-D cluster output visualization**





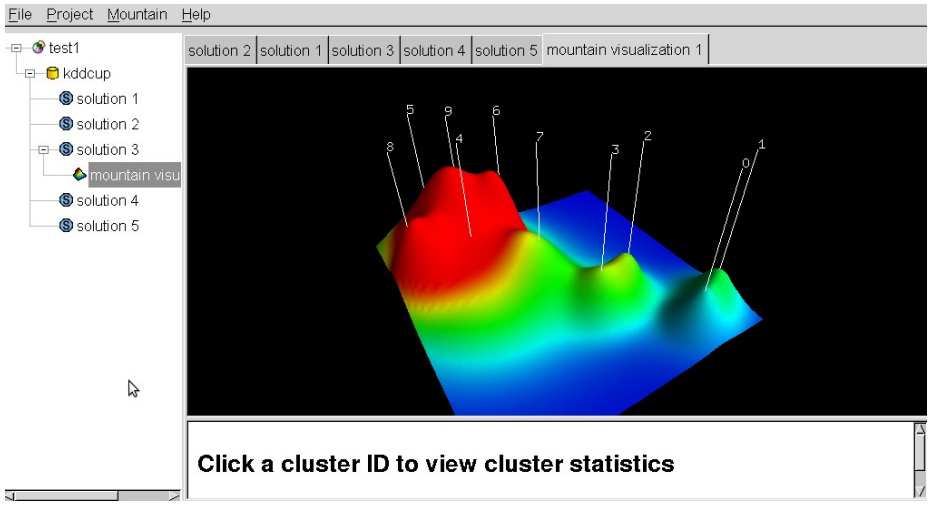


Figure 6. Cluto software for data clustering

Data set

Kddcup 1999

Based on the Darpa 1998 data set from MIT Lincoln Lab

- Original data set
  - Total samples = 400,000
  - Total features = 41 (categorical and continuous values)
  - Types of attacks: DoS, Probe, U2R, R2L
- Data set used
  - Total samples = 9,200
  - Categorical features encoded to binary values
  - Total features = 80 (continuous and binary values)
  - Attacks 2% of samples

Performance metrics

- **Detection rate:** detected attacks / total attacks
- **False alarm rate:** normal classify as attacks / total normal

Performance evaluation

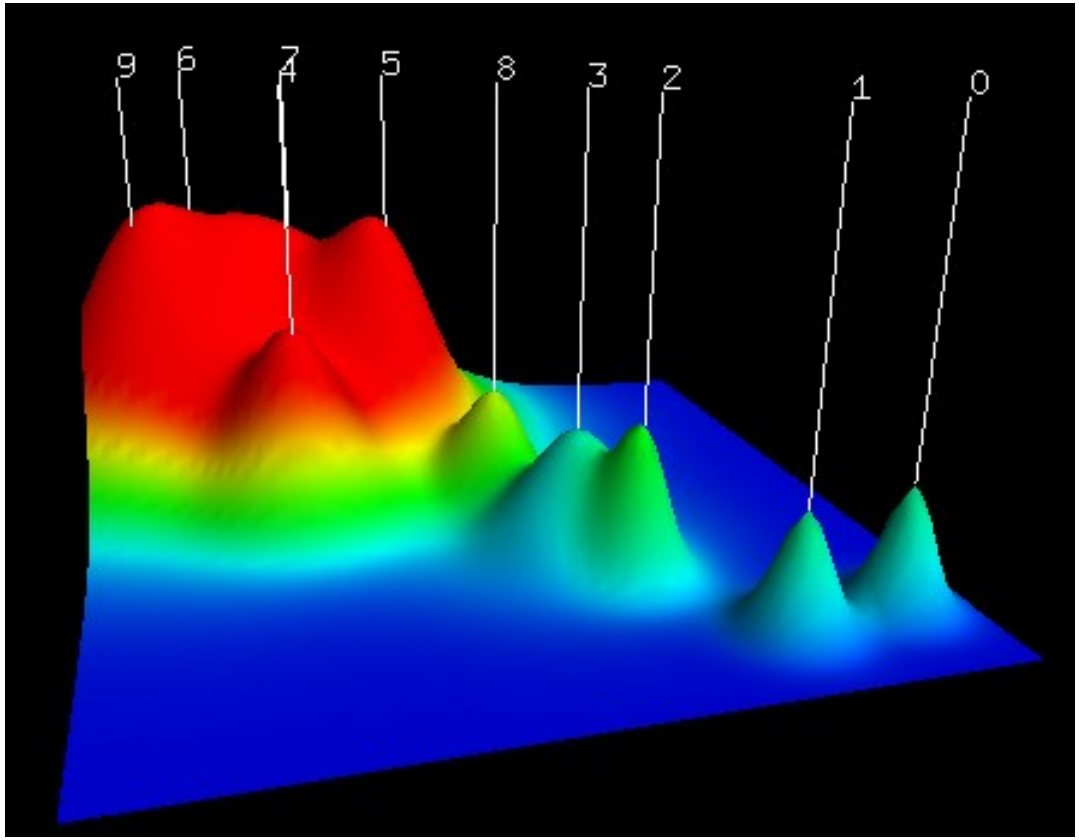


Figure 7. Clustering visualization for K = 10 using Cluto

- **Smaller clusters labeled as 'attacks'**
- **Big clusters labeled as 'normal'**

Performance evaluation

K = 10
(7 sec)

K = 20
(11 sec)

K = 30
(14 sec)

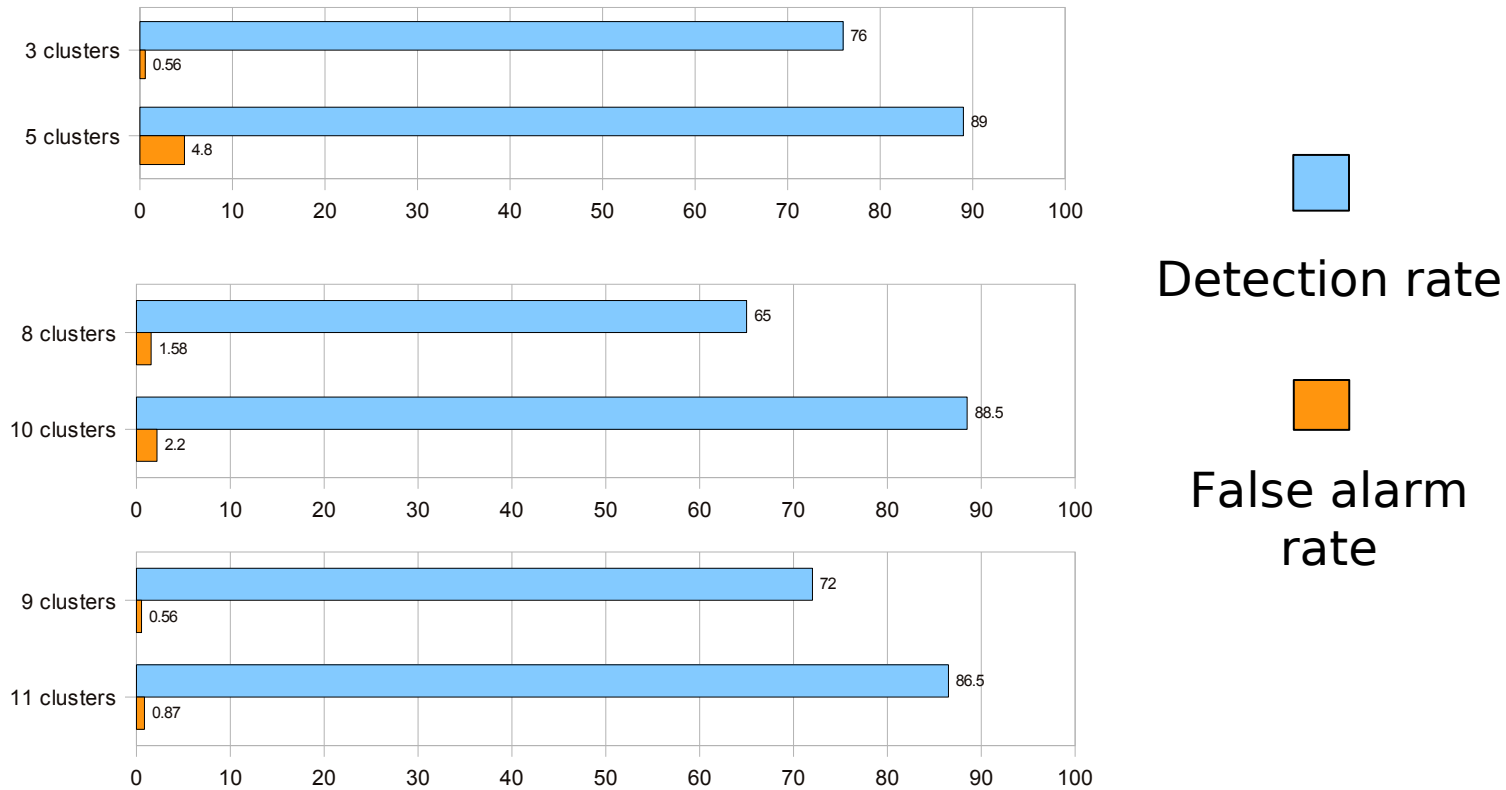


Figure 8. Bar plots of detection and false alarm rate

Clusters	Purity (0 → 1)
K=10	0.989
K=20	0.995
K=30	0.997

Summary

- Network intrusion detection
  - Methods
  - Advantages and disadvantages
- Method used
  - Data clustering for anomaly detection
  - Advantages
- Evalutation results
  - Increasing number of 'attacks' clusters affect the detection rate and false alarm rate
  - High detection rate can be achieved with a low false alarm rate

References

1. *Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly Detection: A survey. 2007*
2. *Leonid Portnoy, Eleazar Eskin, Sal Stolfo. Intrusion detection with unlabeled data using clustering. 2001*
3. *S Zhong, TM Khoshgoftaar, N Seliya. Clustering-based network intrusion detection. 2007*
4. *Richard O. Duda, Peter E. Hart, David G. Stork. Pattern Classification. 2001*
5. *Stefano Zanero, Sergio M. Savaresi. Unsupervised learning techniques for a intrusion detection system. 2004*
6. *Alexandar Lazarevic, Levent Ertoz, Vipin Kumar, Aysel Ozgur, Jaideep Srivastava. A Comparative study of anomaly detection schemes in network intrusion detection. 2003*
7. *Varun Chandola, Eric Eilertson, Leven Ertoz, Gyorgy Simon, Vipin Kumar. Data mining for cyber security. 2006*
8. *Anita K. Jones, Robert S. Sielken. Computer system intrusion detection: A survey. 2000*
9. *Shyam Boriah, Varun Chandola, Vipin Kumar. Similarity measures for categorical data. 2008*
10. *Stefan Axelsson. Intrusion detection systems: A survey and taxonomy. 2000*
11. *George Karypis. Cluto: A clustering toolkit. 2003*
12. *A.K. Jain, M.N. Murty, P.J. Flynn. Data clustering: A review. 1999*

Acknowledgements

Thanks to my mentor Cathy Jiao for her time and support.

Thanks to RAMS and Debbie Mccoy for this opportunity.

Thanks to all the interns for their friendship.

Thanks!!

Questions?