# 7. Data Mining Techniques

Simona Buchovecká, Tomáš Čejka

Faculty of Information Technology, CTU in Prague

*simona.buchovecka@fit.cvut.cz, tomas.cejka@fit.cvut.cz*

November 9, 2020

# Data Mining Fundamentals

## Definition

Data mining is the process of automatically discovering useful information in large repositories. Data mining techniques are deployed to scour large databases in order to find novel and useful patterns that might otherwise remain unknown.

**Knowledge Discovery in Databases** (KDD) — process of converting raw data into useful information or knowledge.
Data mining is a step in the KDD process which includes::

- data preparation,
- data selection,
- data cleaning,
- incorporation of appropriate prior knowledge,
- and proper interpretation of the results of mining to ensure useful knowledge is derived from the data.

Bhuyan, Monowar H., Dhruba K. Bhattacharyya, and Jugal K. Kalita. *Network traffic anomaly detection and prevention: concepts, techniques, and tools.* Springer, 2017.

# Data Mining Tasks

## Categories

- predictive (inference on the current data in order to make future predictions)
- descriptive (characterizes the general properties of the data and underlying relationships among them)

## Most important tasks

1. Classification and Regression
2. Cluster Analysis
3. Association Analysis (discovering the most important and most strongly associated feature patterns in data)
4. Evolution Analysis
5. Outlier Detection

Bhuyan, Monowar H., Dhruba K. Bhattacharyya, and Jugal K. Kalita. *Network traffic anomaly detection and prevention: concepts, techniques, and tools.* Springer, 2017.

# Data Mining in Network Security #1

Do not use signatures, but learn about usage patterns

- Also viewed as Machine Learning techniques
- or Behavioral Analysis

Two basic detection approaches:

- **Misuse detection:**
  - detection of normal vs. intrusive data flows
- **Anomaly detection:**
  - classification algorithms, rare class predictive models, association rules, cost sensitive modeling

# Data Mining in Network Security #2

**Misuse Detection**

- Models of misuse are created automatically
- Often better rules than manually created signatures
- Very good for detection of known (or modified) attacks
- But deciding normal vs. intrusive is human resource intensive

**Anomaly detection**

- Models of normal behavior are created automatically
- Deviations from normal behavior are detected automatically
- Can potentially detect unexpected and unknown attacks

# Data Mining for Anomaly Detection

**Advantages**

- Detection of unknown attacks
- Detection of unusual behavior interesting to IT managers

**Limitations**

- Possibly high false alarm rate (FAR)

**Types of "training"**

- **Supervised Detection**
  - Models for normal behavior are built from training data
- **Unsupervised Detection**
  - No training data, attempts to learn about anomalies automatically

# Unsupervised Anomaly Detection

**No training data**

- Attempts to learn about anomalies automatically

**Algorithms used**

- Statistical methods, clustering, outlier detection, state machines

# Detection Based on Association Rules

## General Information

- conceptually a simple method based on counting of co-occurrences of items in transactions databases
- unsupervised
- used for one-class anomaly detection by generating rules from data

M. V. Mahoney and P. K. Chan, *Learning rules for anomaly detection of hostile network traffic*. In Proc. 3rd IEEE International Conference on Data Mining. Washington: IEEE CS, 2003.

R. Agrawal and R. Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*. In Proc. 20th International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann, 1994, pp. 487–499.

# Summarizing Anomalous Connections Using Association Rules

Association patterns

- specified as frequent item sets or association rules

Summary of detected anomalous flows:x

- Humans can analyze the summary
- E.g., a frequent set for scanning: *srcip* = X, *dstport* = Y
- This is a candidate signature for a signature-based system

Constructing a profile of normal network traffic:

- Identify sets of features that are found in normal traffic
- Examples:
  - web browsing (HTTP request): protocol=TCP, dstPort=80, NumPackets=3. . . 6
  - if port=80 and word3=HTTP/1.0 then word1=GET or POST

# Challenges in Mining Association Rules

The most difficult and dominating part of an association rules discovery algorithm is to find the itemsets that have strong support.

- an algorithm known as LERAD by Mahoney and Chan

**Challenges**

- Imbalanced class distribution
  - Small support for attack, large for normal traffic
- Binarization and grouping of attribute values
  - Supervised or unsupervised
- Pruning the redundant patterns
  - Must be subsets with similar support
- Finding discriminating patterns
  - Patterns that identify attacks and normal traffic
- Grouping the discovered patterns

# Minnesota Intrusion Detection System #1

Abbreviated as MINDS Uses data mining techniques for

- unsupervised anomaly detection (assigns an "anomaly" score to each network connection)
- association pattern analysis (for suspected anomalous network connections)

Performance

- Detects new intrusions that escape signature-based IDSs

# Minnesota Intrusion Detection System #2

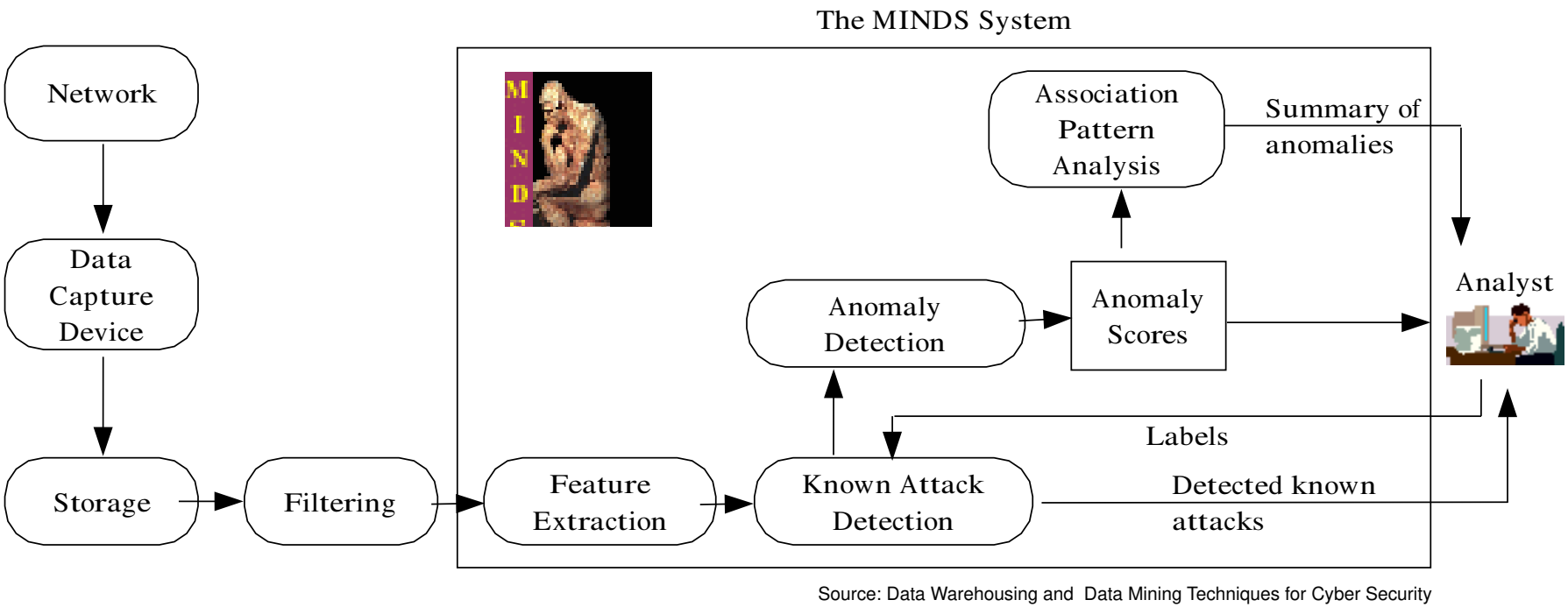
**Data capture**

- Netflow v5 (Cisco protocol, today v9; IPIFIX is standard)
- Using flow-tools (`https://code.google.com/p/flow-tools/`)
- Header information only, summarized into (one-way) flows
- Much less data storage than tcpdump (or, e.g., Wireshark)
- Flow analysis period: 10 min. (1–2 mil. flows) [today 5 min.]

**MINDS processing speed**

- Processing of 10 min. data: less than 3 min
- But unwanted traffic is filtered out before processing

# MINDS Architecture




Source: Data Warehousing and Data Mining Techniques for Cyber Security

# MINDS Operation Steps #1

1. Feature Extraction
   - Basic features: source and destination IP addresses and ports, protocol, flags, number of bytes, number of packets
   - Derived features for a time-window of last T seconds:
     - Capture connections with similar recent characteristics
     - Useful to detect sources of high volume connections per unit time (e.g., fast scanning)
   - Derived features for a window of last N connections:
     - Similar characteristics, but for the last connections from distinct sources (good for, e.g., slow scanning)

# Time-window based features

**count-dest** Number of flows to unique destination IP addresses inside the network in the last T seconds from the same source

**count-src** Number of flows from unique source IP addresses inside the network in the last T seconds to the same destination

**count-serv-src** Number of flows from the source IP to the same destination port in the last T seconds

**count-serv-dest** Number of flows to the destination IP address using same source port in the last T seconds

# Connection-window based features

count-dest-conn Number of flows to unique destination IP addresses inside the network in the last N flows from the same source

count-src-conn Number of flows from unique source IP addresses inside the network in the last N flows to the same destination

count-serv-src-conn Number of flows from the source IP to the same destination port in the last N flows

count-serv-dest-conn Number of flows to the destination IP address using same source port in the last N flows

# MINDS Operation Steps #2

2. Signature based detection for known attacks
   - Detected attacks are not further analyzed. Not our focus now.
3. Anomaly detection
   - Outlier detection assigns an anomaly score to network flows
   - Humans can analyze only the most anomalous connections
4. Association pattern analysis
   - Summarization of highly anomalous network connections
   - Humans decide whether the summaries can be used to create signatures for detection in step 2

# MINDS Anomaly Detection

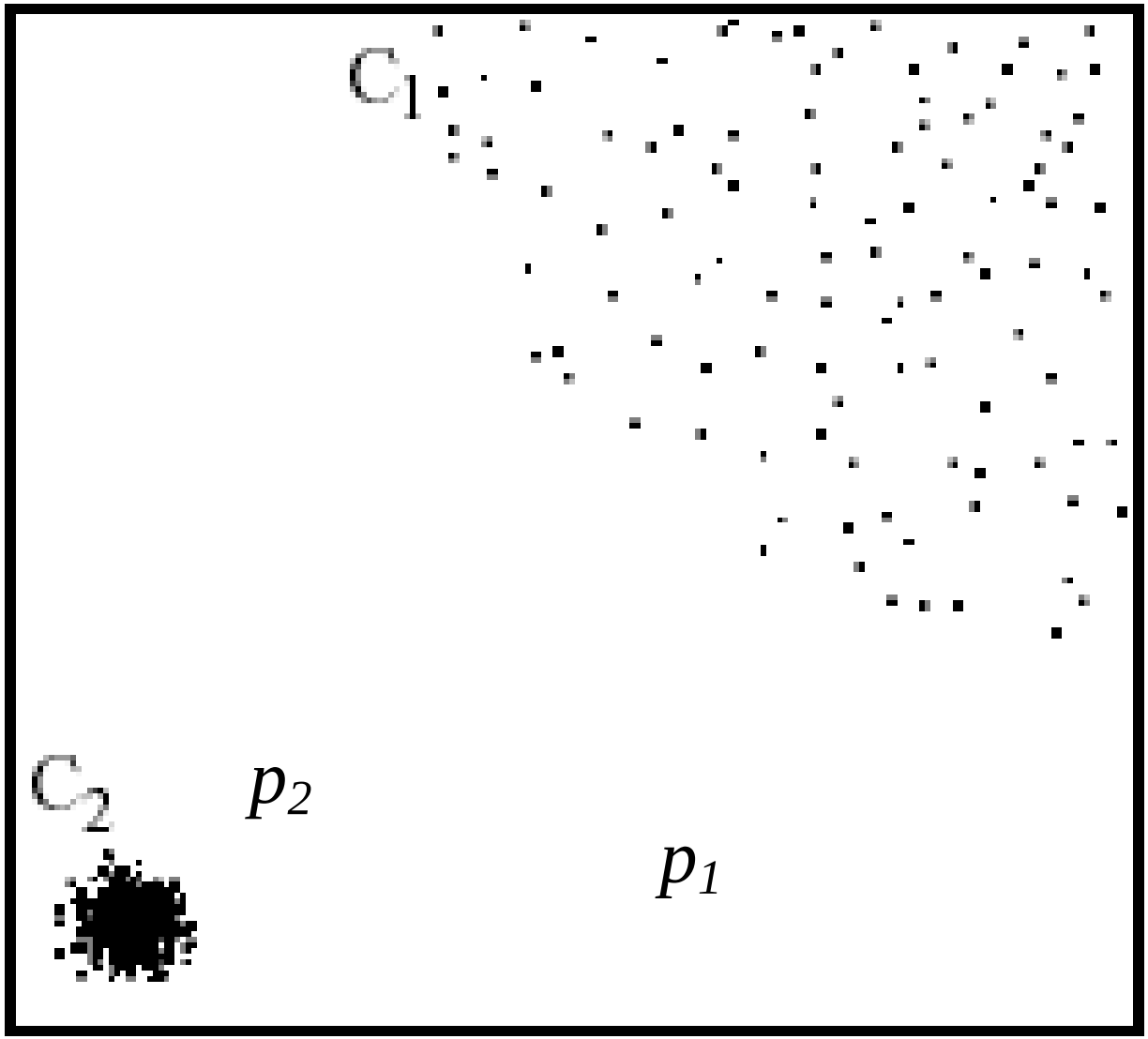
Local Outlier Factor (LOF)

- Assigned to each data point
- It is local: outliers with respect to their neighborhood

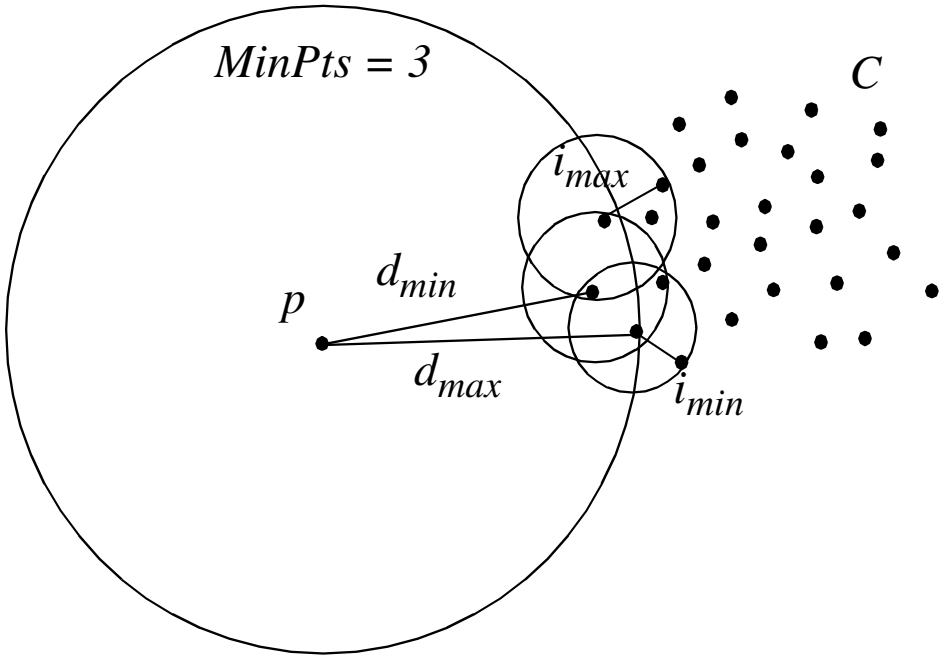
LOF Calculation

- Compute density of each point's neighborhood
- Find the average of the density ratios over all its neighbors
- Calculate the ratio of the point's density and the density of all its neighbors

# 2D Outlier Detection



# LOF Bounds




$$d_{min} = 4 * i_{max} \Rightarrow LOF_{MinPts}(p) \geq 4$$

$$d_{max} = 6 * i_{min} \Rightarrow LOF_{MinPts}(p) \leq 6$$

More details in:

Breunig, Markus M., et al. *LOF: identifying density-based local outliers*. ACM sigmod record. Vol. 29. No. 2. ACM, 2000.

# Linear and Non-Linear Classification

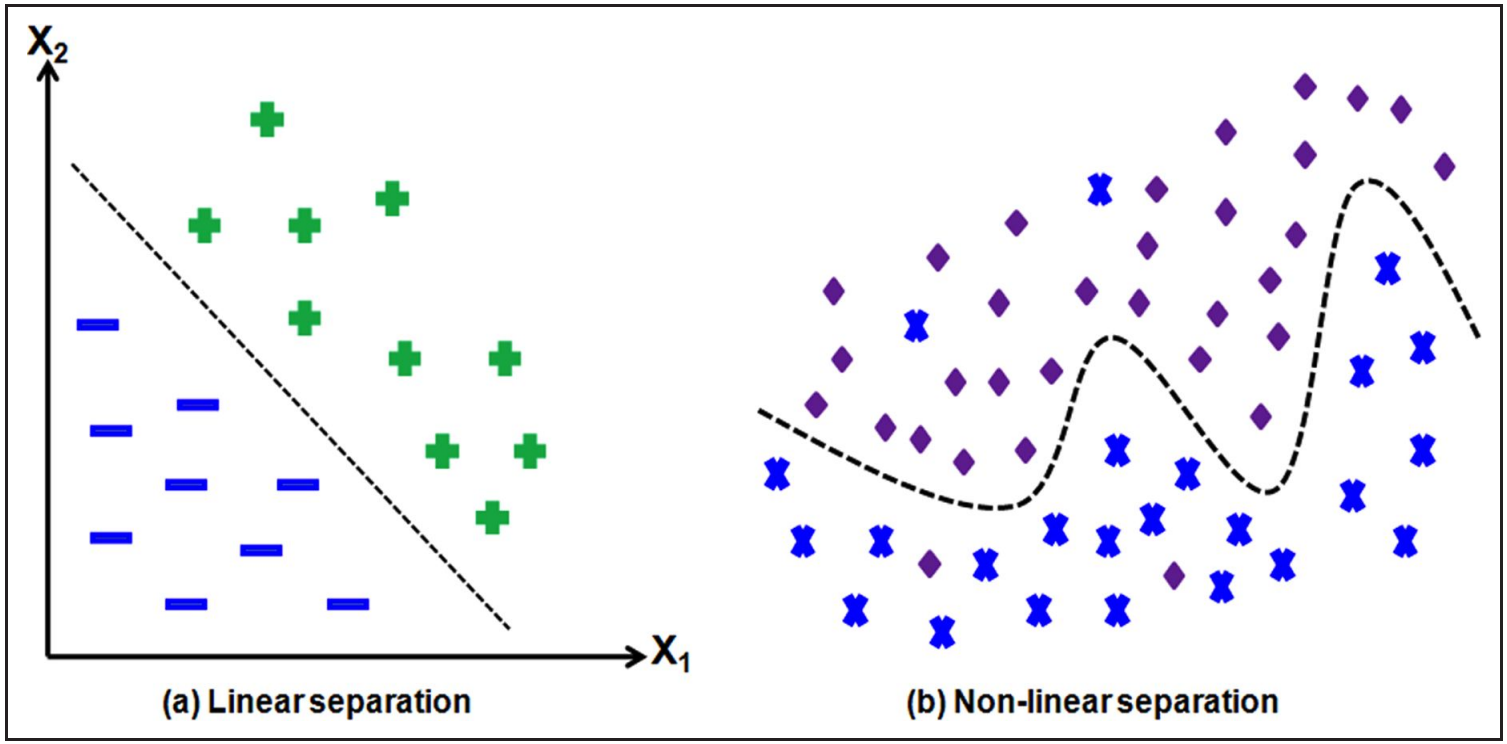



Fig. 7. Linear and non-linear classif cation in 2-D

Bhuyan, Monowar H., Dhruba Kumar Bhattacharyya, and Jugal K. Kalita. *Network anomaly detection: methods, systems and tools.* Ieee communications surveys & tutorials 16.1 (2014): 303-336.

# Comparison of Classification Algorithms

Source: http://scikit-learn.org/stable/modules/clustering.html

# Computational Complexity

Neighborhoods around all data points:

- Calculate distances between all pairs of data points
- This is $O(n^2)$ calculation – computationally infeasible for millions of data points.

Sampling a training data set from the observed data

- Compare all data points to this small set,
- Reduced complexity: $O(\text{data size} * \text{sample size})$
- Additional benefit: anomalous behavior will not be seen often in the sample, so it will not be mistaken with normal behavior

# Performance of MINDS #1

Worms

- October 10, 2002: MINDS detected two activities of the slapper worm that were not identified by SNORT since they were new variations of an existing worm code

Scanning and DoS activities

- August 9, 2002: CERT/CC issued an alert for "widespread scanning and possible denial of service activity targeted at the Microsoft-DS service on port 445/TCP" August 13, 2002: Network connections related to such scanning were the top ranked outliers in MINDS The port scan module of SNORT failed (slow scanning)

# Performance of MINDS #2

Policy Violations

- August 8 and 10, 2002: MINDS detected a machine running a Microsoft PPTP VPN server, and another one running a FTP server on non-standard ports.
- Both policy violations were the top ranked outliers since they are not allowed, and therefore very rare.
- February 6, 2003: unsolicited ICMP echo reply messages to a computer previously infected with Stacheldract worm (a DDoS agent) were detected by MINDS.
- The infected machine has been removed from the network, but other infected machines outside the local network were still trying to talk to the previously infected machine.

# Questions?