

Anomalieerkennung

UnFUG WS2011/2012

Alexander Passfall
<alex@passfall.de>

Hochschule Furtwangen

3. November 2011

Inhalt

- 1 Grundlagen
 - Typen
 - Funktionsweise
- 2 Algorithmen
 - Outlier Detection
 - Machine Learning
- 3 Anwendung



Was ist eine Anomalie?

- „Anomaly is a pattern in data that does not conform to expected behavior“
- Auch: outlier, exception, peculiarity, surprise, ...



Einsatzgebiete

- Bankenwesen
 - Aufdecken von Kreditkartenbetrug
 - z.B. eine unnatürlich hohe Abbuchung
- Medizin
 - Entdecken von Krebszellen
 - Gesundheitsüberwachung



Einsatzgebiete

- Bankenwesen
 - Aufdecken von Kreditkartenbetrug
 - z.B. eine unnatürlich hohe Abbuchung
- Medizin
 - Entdecken von Krebszellen
 - Gesundheitsüberwachung
- **Intrusion Detection**



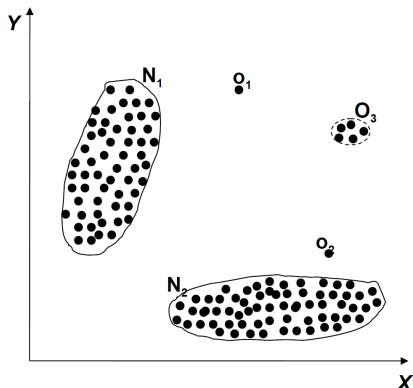
Anomalie-Typen

- Punkt
- kontextabhängig
- kollektiv



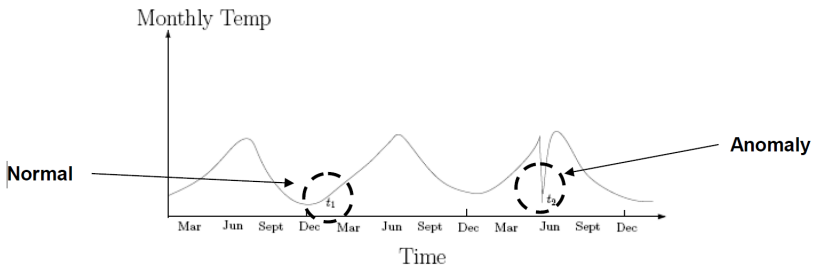
Punkt-Anomalie

- Ein einzelner Datenpunkt, der im Vergleich zu anderen anomal ist



Kontextabhängige Anomalie

- Ein einzelner Datenpunkt, der nur im Kontext anormal ist
- Kann in einem anderen Kontext völlig normal sein





Funktionsweise

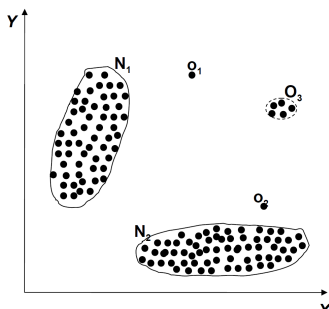
- Lernphase
 - Supervised - Normale Daten und Anomalien markiert
 - Semi-supervised - Nur normale Daten
 - Unsupervised - Keine vorherige Definition von normalen Daten/Anomalien
- Anwendungsphase
 - Daten werden entsprechend dem gelernten Verhalten gefiltert

Outlier Detection

- k-th Nearest Neighbor
- Nearest Neighbor
- Mahalanobis Distance
- Local Outlier Factor (LOF)

Gemeinsamkeiten

- Daten werden auf Punkte in n-dimensionalem Raum abgebildet
- 1 Dimension entspricht 1 Key Faktor, z.B. TCP-Port
- Kontextabhängige, kollektive Anomalien werden auf Punkt-Anomalien reduziert (z.B. Verbindungen/Minute)



k-th Nearest Neighbor

- Berechnet die Euklidische Entfernung des k. Nachbar-Punkt von einem Punkt
- k muss zuvor manuell bestimmt und angepasst werden
- Zuletzt werden die n Punkte, welche den größten Abstand haben, als Outlier betrachtet

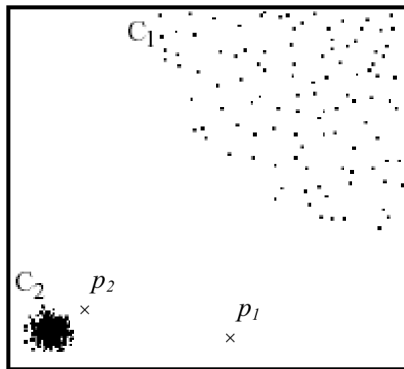


Nearest Neighbor

- k-th Nearest Neighbor mit $k=1$
- Schwach bei Sammlung von Anomalien

Local Outlier Factor

- Berücksichtigt die Punkt-Dichte in bestimmten Bereichen



Machine Learning

- Bayessche Netze
- Markov-Modelle
- Neurale Netze
- Fuzzy Logic
- Genetische Algorithmen



Vorteile

- Je nach Einsatzgebiet sehr gute Ergebnisse



Nachteile

- Hoher Ressourcen-Verbrauch
- Schlechte Nachvollziehbarkeit der Entscheidungen

SPADE

- Statistical Packet Anomaly Detection Engine
- Snort-Plugin
- Von Silicon Defence entwickelt
- 2003 DARPA-Finanzierung eingestellt → GPL

Funktionsweise

- Wahrscheinlichkeitstabellen über Auftreten verschiedener Pakete
- Normaler Traffic z.B.
 $P(\text{dest_IP} = \text{DNS_SRV}, \text{dest_port} = 53) = 10\%$
- Anomalie z.B.
 $P(\text{dest_IP} = \text{DNS_SRV}, \text{dest_port} = 80) = 0.1\%$
- Errechnen einer „relative anomaly score“ (0-1)

PHAD

- Paket Header Anomaly Detection
- Port als Snort-Preprocessor verfügbar
- <http://cs.fit.edu/~mmahoney/dist/>,
<http://seclists.org/snort/2010/q2/751>
- Untersucht den IP/TCP/UDP/ICMP-Header auf ungewöhnliches Verhalten
 - Angriffe modifizieren den Header um IDS auszutricksen



Und sonst?

- Viele akademische Systeme
- Kommerzielle Systeme?



Fragen?