MACHINE LEARNING & PUBLIC POLICY

LECTURE 4: ANOMALOUS PATTERN DETECTION & SURVEILLANCE SYSTEMS

Professor Edward McFowland III

Information Systems and Decision Sciences
Carlson School of Managment

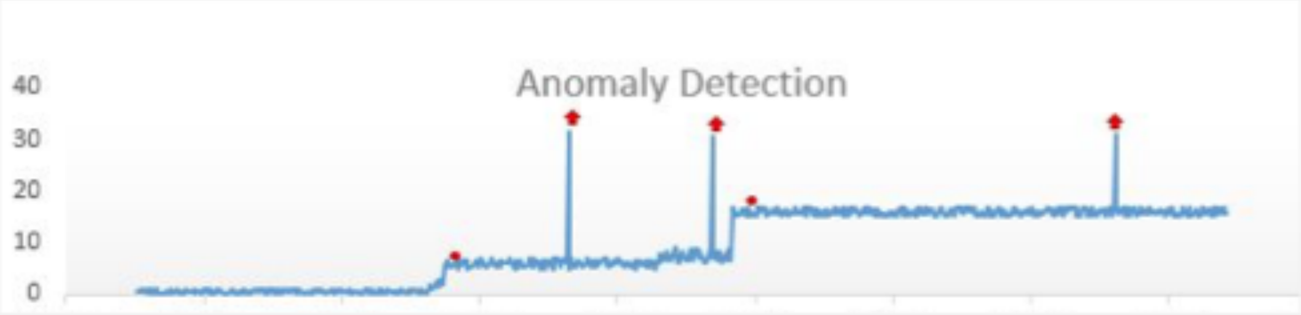
University of Minnesota

ANOMALY DETECTION

Observation II: Sometimes goal is one of detection or discovery

ANOMALY DETECTION PARADIGM

- Identifying when a "system" deviates away from its expected behavior.



ANOMALY DETECTION

- Challenges?
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging
  - Finding needle in a haystack
- Working assumption?
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

ANOMALY DETECTION

- General Steps?
  - Build a **profile** of the "normal" behavior
    - I.e., patterns or summary statistics for the overall population
  - Use the "normal" profile to detect **anomalies**
    - I.e., observations whose characteristics differ significantly from the normal profile
- Detection schemes?
  - Graphical
  - Distance/proximity - based
  - Statistical/model-based

MODEL BASED ANOMALY DETECTION

Observation II: Sometimes goal is one of detection or discovery

GENERALIZED EXTREME STUDENTIZED DEVIATE

r : upper-bound on # of anomalies

x : sample average

σ : sample standard deviation

α : significance level

H0: There are no anomalies in the data

Ha: There are up to *r* anomalies in the data

$$R_i = \frac{\max_i |x_i - x|}{\sigma}$$

$$\lambda_i = \frac{(n-1)t_{p,n-i-1}}{\sqrt{\left(n-i-1+t_{p,n-i-1}^2\right)(n-i+1)}}$$

Remove the observation *i* and recompute. Repeat until *r* observations have been removed, result in the *r* test statistics *R1*, *R2*, ..., *Rr*.

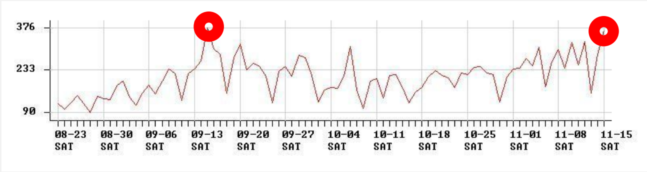
$$\mathrm{P}(\cup_i \{R_i \geq \lambda_i\}) \leq \alpha$$

- Select largest *i* such that $R_i > \lambda_i$
- Assumes Gaussian Distribution
- Seasonality unaware

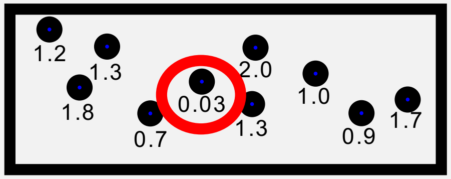
SPATIAL AND TEMPORAL ANOMALY DETECTION

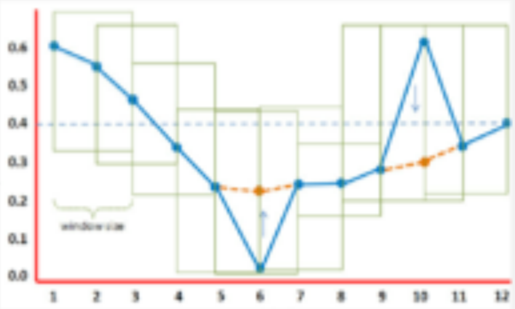
- A simple model-based anomaly detection when we monitor a single real-valued quantity over time and/or space.
- Report any observed value that is significantly above or below its expected value (e.g., using GESD).

Time series data



Spatially distributed data

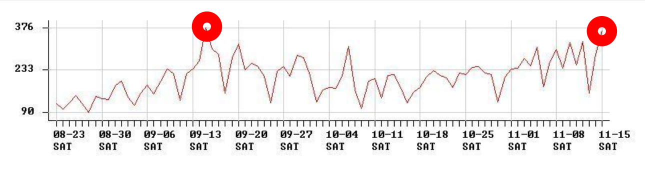




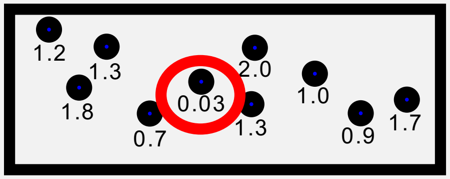
SPATIAL AND TEMPORAL ANOMALY DETECTION

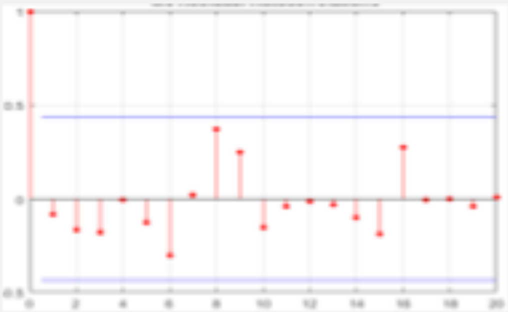
- A simple model-based anomaly detection when we monitor a single real-valued quantity over time and/or space.
- Report any observed value that is significantly above or below its expected value (e.g., using GESD).

Time series data



Spatially distributed data

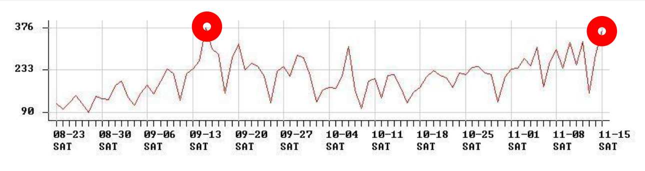




SPATIAL AND TEMPORAL ANOMALY DETECTION

- A simple model-based anomaly detection when we monitor a single real-valued quantity over time and/or space.
- Report any observed value that is significantly above or below its expected value (e.g., using GESD).

Time series data

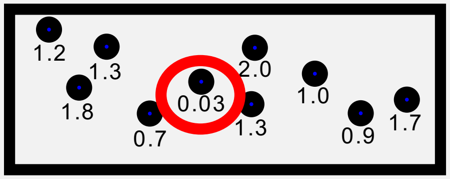


Time series analysis: the expected value for time step t is a function of the values for time steps 1 through t – 1.

Exponentially weighted averaging:

$E[x_t] = (\Sigma\ w_i x_i) / (\Sigma\ w_i)$, $w_i = e^{-(t-i)/b}$
where i = 1… t – 1.

Spatially distributed data



Spatial regression: the expected value for location s is a function of the values for all other locations.

Kernel regression, exponential kernel:

$E[x_s] = (\Sigma\ w_i x_i) / (\Sigma\ w_i)$, $w_i = e^{-d(s,\ i)/b}$
where i ≠ s and d is Euclidean distance.

ANOMALOUS PATTERN DETECTION

Observation II: Sometimes goal is one of detection or discovery

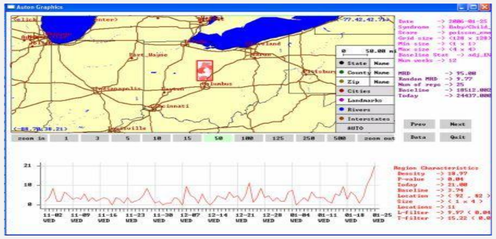
ANOMALOUS PATTERN DETECTION

Main goal of pattern detection: to **identify** and **characterize** relevant subsets of a massive dataset, i.e. groups of records that differ from the rest of the data in an interesting way.

Question 1: Are any relevant patterns present in the data, or is the entire dataset "normal"?

Example: outbreak detection

Are there any emerging outbreak on Twitter? If so, what type of outbreak, and what areas are affected?



```
Date       -> 2006-01-25
Syndrome   -> Baby/Child_
Score      -> poisson_
Grid size  -> (128 x 128)
Min size   -> (1 x 1)
Max size   -> (4 x 4)
Baseline Stat  -> adj_
Num weeks -> 12

MRD          -> 95.00
Random MRD   -> 9.77
Num of reps  -> 25
Baseline     -> 18512.00
Today        -> 24437.00
```

```
Region Characteristics
 Density    -> 18.97
 P-value    -> 0.04
 Today      -> 21.00
 Baseline   -> 3.74
 Location   -> < 92 , 82 >
 Size       -> < 1 x 4 >
 Locations  -> 11
 L-filter   -> 9.97 < 0.04
 T-filter   -> 15.22 < 0.0
```

Question 2: If there are any patterns, identify the pattern type and the affected subset of data records for each.

Example: financial analysis

Can we deduce the membership and structure of money laundering ring based on known links between suspected individuals?

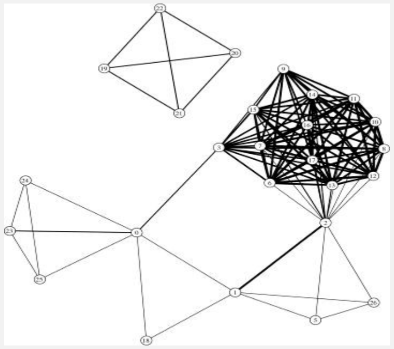
ANOMALOUS PATTERN DETECTION

Main goal of pattern detection: to **identify** and **characterize** relevant subsets of a massive dataset, i.e. groups of records that differ from the rest of the data in an interesting way.

What makes a group of records "relevant"?

1. Matching some known pattern or structure.

Group detection: given a financial network, find highly connected sets of individuals.



Many efficient algorithms have been developed to find dense subgraphs or other structures in network data.

ANOMALOUS PATTERN DETECTION

Main goal of pattern detection: to **identify** and **characterize** relevant subsets of a massive dataset, i.e. groups of records that differ from the rest of the data in an interesting way.

What makes a group of records "relevant"?

1. Matching some known pattern or structure.
2. Multiple related records that are individually anomalous.

Fraud detection: look for individuals with a history of suspicious transactions.

Network intrusion detection: look for suspicious combinations of activities (e.g. port scanning).

In these domains, multiple "slightly anomalous" behaviors may together provide evidence of a major deviation from normal.

ANOMALOUS PATTERN DETECTION

Main goal of pattern detection: to **identify** and **characterize** relevant subsets of a massive dataset, i.e. groups of records that differ from the rest of the data in an interesting way.

What makes a group of records "relevant"?

1. Matching some known pattern or structure.
2. Multiple related records that are individually anomalous.
3. Higher (or lower) than expected number of records with some combination of attributes.
4. Change in data distribution as compared to the rest of the dataset.

Cluster detection: find spatial areas or periods of time with more records than expected.

Event detection: is the recent data differently distributed than the past?

Key concept: A group of records may be highly anomalous or interesting even if none of the individual records is itself anomalous.

SUBSET SCANNING

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

There are many options for computing the score of a subset S.

WHAT'S STRANGE ABOUT RECENT EVENTS?

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

There are many options for computing the score of a subset S.

In the WSARE method ("What's Strange About Recent Events"), we consider the subsets of the data defined by a one- or two-component rule R, and find rules where the current data is significantly different than the past.

For each rule, we create a 2x2 contingency table comparing current and past data:

	Current	Past
# records satisfying R	48	45
# records satisfying ~R	86	220

Compute p-value using a statistical test (C^2 or Fisher's Exact). Lower p-value = higher score.

WHAT'S STRANGE ABOUT RECENT EVENTS?

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

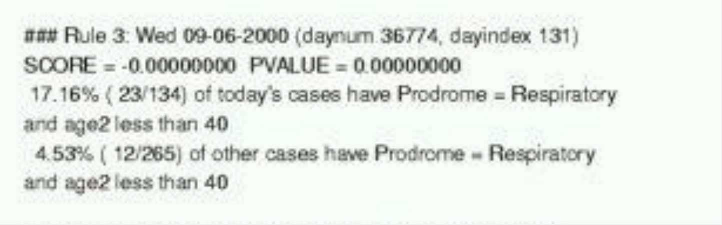
Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

There are many options for computing the score of a subset S.

In the WSARE method ("What's Strange About Recent Events"), we consider the subsets of the data defined by a one- or two-component rule R, and find rules where the current data is significantly different than the past.

For example, using WSARE for hospital Emergency Department surveillance resulted in finding the following significant rule, corresponding to an outbreak of respiratory illness on 9/6/2000.



```
### Rule 3: Wed 09-06-2000 (daynum 36774, dayindex 131)
SCORE = -0.00000000  PVALUE = 0.00000000
 17.16% ( 23/134) of today's cases have Prodrome = Respiratory
and age2 less than 40
  4.53% ( 12/265) of other cases have Prodrome = Respiratory
and age2 less than 40
```

MODEL-BASED PATTERN DETECTION

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

There are many options for computing the score of a subset S.

In the **model-based** anomalous pattern detection approach, we model the effects of each pattern type P on the affected subset of the data S.

Now we create tables comparing the numbers of anomalous and normal records:

	Anomalous	Normal
# records satisfying R	17	5
# records satisfying ~R	93	400

Compute p-value using a statistical test (C^2 or Fisher's Exact). Lower p-value = higher score.

MODEL-BASED PATTERN DETECTION

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset $S = \{x_i\}$ and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

There are many options for computing the score of a subset S.

In the **model-based** anomalous pattern detection approach, we model the effects of each pattern type P on the affected subset of the data S.

Commonly we compute the **likelihood ratio statistic** $\Pr(\text{Data} \mid H_1(S, P)) / \Pr(\text{Data} \mid H_0)$ for each (S, P).

In **event detection**, we model the null hypothesis H_0 by estimating expected counts for each data stream assuming no events.

Each pattern P is assumed to increase the counts for some data streams in the affected set of spatial locations S.

WHICH PATTERNS TO REPORT?

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

Option 1: Report the k highest scoring subsets, ordered by score.

The disadvantage of this approach is that the user is not informed whether any of the discovered patterns are likely to be relevant.

However, this may be acceptable in monitoring systems or scientific discovery applications where the user is willing to evaluate a fixed number of potential patterns.

WHICH PATTERNS TO REPORT?

We can scan over subsets of the dataset in order to find those groups of records that correspond to a pattern.

Step 1: Compute **score** F(S, P) for each subset S = {x_i} and for each pattern type P, where higher score means more likely to be a pattern.

Step 2: Consider the highest scoring potential patterns (S, P) and decide whether each actually represents a pattern.

Option 2: Perform hypothesis tests, and report all **significant** patterns (S, P).

In the hypothesis testing framework, we must adjust for the fact that we're performing so many tests. Otherwise we will report too many false positives!

In model-based approaches, one way to do this is **randomization**: we generate a large number of simulated datasets assuming the null model, and compare the scores of the potential patterns in the real dataset to the highest scoring patterns in the simulated data.

An alternative is to adjust the p-value threshold for each test based on the number of tests performed (e.g. Bonferroni threshold = .05 / # tests)

References

- A coherent text on anomalous pattern detection has yet to be written, but many methods have been proposed and are becoming common:
  - WSARE: W.-K. Wong et al., “Rule-based anomaly pattern detection for detecting disease outbreaks,” *Proc. 18th Natl. Conf. on Artificial Intelligence*, 2002.
  - APD (“Anomaly Pattern Detection”). K. Das, J. Schneider, and D.B. Neill, *Proc. KDD 2008*.
  - D.B. Neill and W.-K. Wong, “A Tutorial on Event Detection,” presented at *KDD 2009* conference.
  - Edward McFowland III, Skyler Speakman, and Daniel B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533-1561, 2013.
- Software for spatial cluster detection and for WSARE is available on the Auton Laboratory web page, http://www.autonlab.org.