Efficient Methods for Anomalous Pattern Detection in General Datasets

Event & Pattern Detection Lab
Carnegie Mellon University
Edward McFowland III (mcfowland@cmu.edu)
Daniel B. Neill (neill@cs.cmu.edu)

This work was partially supported by:
AT&T Research Labs Fellowship
NSF Graduate Research Fellowship
NSF grants IIS-0916345, IIS-0911032, and IIS-0953330
Anomalous Pattern Detection

• Two set of processes generating data
  ▫ Typical system behavior
  ▫ Anomalous system behavior

• Discover and characterize the anomalous processes
  ▫ Evaluating records in isolation may be insufficient
  ▫ Find the subsets of data that correspond to anomalous system behavior
  ▫ An anomalous subsets is self-similar and as a group different from rest of the data
Is it Useful?

- Fraud Detection
- Network Intrusion Detection
- Anomalous Patterns of Smuggling
- Disease Surveillance
- ...And many more ways to make the world a better place
The Goal!

I. Compute the anomalousness of each attribute value (for each record)

II. Discover subsets of records and attributes that are most anomalous
Fast Generalized Subset Scan (FGSS)

Attributes $A_1...A_M$

Records $R_1...R_N$

I. Compute the anomalousness of each attribute (for each record)

In order to compute the anomalousness of the data, FGSS models the data distribution under expected system behavior.
Fast Generalized Subset Scan (FGSS)

In order to compute the anomalousness of the data, FGSS models the data distribution under expected system behavior.

I. Compute the anomalousness of each attribute (for each record)

- Compute the anomalousness of each attribute (for each record)

\[ P(A_5 | A_1) \]
Fast Generalized Subset Scan (FGSS)

Attributes $A_1...A_M$

Records $R_1...R_N$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods

By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values
Fast Generalized Subset Scan (FGSS)

Attributes $A_1...A_M$

Empirical p-values are a measure, mapped onto the interval $[0,1]$, of how surprising each attribute value is given the model of normal system behavior.

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values
      i. maps each attribute distribution to same space
      ii. $p_{ij}$ in $S \sim \text{Uniform}(0,1)$ under $H_0$
Fast Generalized Subset Scan (FGSS)

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

Subsets of data with a higher than expected quantities of significantly low p-values are possibly indicative of an anomalous process
Fast Generalized Subset Scan (FGSS)

Nonparametric Scan Statistic (NPSS)

\[ F(S) = \max_{\alpha} F(\alpha) = \max_{\alpha} F_{\alpha}(N_{\alpha}, N) \]

\[ N_{\alpha} = | \{ p_{ij} \in S : p_{ij} \leq \alpha \} | \]

\[ N_{\text{tot}} = | \{ p_{ij} \in S \} | \]

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   • Evaluate subsets with NPSS

NPSS quantifies how dissimilar the distribution of empirical p-values in S are from Uniform(0,1)
Fast Generalized Subset Scan (FGSS)

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Naïve search is infeasible $O(2^{N+M})$

Attributes $A_1...A_M$

Records $R_1...R_N$

Search over all possible subsets of records’ p-value ranges and find the maximizing $F(S)$
Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff:

\[
\max_{S \subseteq D} F(S) = \max_{i=1\ldots N} F(\{R_{(1)}\ldots R_{(i)}\})
\]

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   •Naïve search is infeasible \(O(2^{N+M})\)

We can reduce the search over records from \(O(2^N)\) to \(O(N \log N)\)
Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A $F(S)$ satisfies LTSS iff:

$$\max_{S \subseteq D} F(S) = \max_{i=1 \ldots N} F\left( \{ R_{(i)} \ldots R_{(i)} \} \right)$$

We only need to consider:

- $\{ R_{(1)} \}$
- $\{ R_{(1)}, R_{(2)} \}$
- $\{ R_{(1)}, R_{(2)}, R_{(3)} \}$
  
  
  
- $\{ R_{(1)}, \ldots, R_{(n)} \}$

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize $F(S)$ over all subsets of $S$
   - Naïve search is infeasible $O(2^{N+M})$
   - NPSS satisfies LTSS with:
     $$F(S) = \max_\alpha F_\alpha (N_\alpha, N_{\text{tot}})$$

We can reduce the search over records from $O(2^N)$ to $O(N \log N)$
Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff:

\[
\max_{S \subseteq D} F(S) = \max_{j=1...M} F\left(\left\{A_{(1)}...A_{(j)}\right\}\right)
\]

We only need to consider:

\{A_{(1)}\} \\
\{A_{(1)},A_{(2)}\} \\
\{A_{(1)},A_{(2)},A_{(3)}\} \\
\vdots \\
\{A_{(1)},.............,A_{(m)}\}

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize \(F(S)\) over all subsets of \(S\)
   - Naïve search is infeasible \(O(2^{N+M})\)
   - NPSS satisfies LTSS with:
     \[
     F(S) = \max_{\alpha} F_\alpha(N_\alpha,N_{\text{tot}})
     \]

We want to maximize of subsets of records AND attributes; Observe \(F(S)\) is only a function of \(p_{ij}\), thus we can use LTSS to also maximize over the attributes.
Fast Generalized Subset Scan (FGSS)

Linear Time Subset Scanning Property (LTSS)

A F(S) satisfies LTSS iff:

\[
\max_{S \subseteq D} F(S) = \max_{j=1 \ldots M} F\left(\{A_{(1)} \ldots A_{(j)}\}\right)
\]

We only need to consider:

\{A_{(1)}\}
\{A_{(1)}, A_{(2)}\}
\{A_{(1)}, A_{(2)}, A_{(3)}\}
\vdots
\{A_{(1)}, \ldots, A_{(m)}\}

I. Compute the anomalousness of each attribute (for each record)

1. Learn Bayesian Network
2. Compute attribute value likelihoods
3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous

1. Maximize F(S) over all subsets of S
   - LTSS over records O(N log N)
   - LTSS over attributes O(M log M)

We can iterate between maximizing over the records and maximizing over the attributes.
Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure
Attributes A_1...A_M

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize F(S) over all subsets of S
      • LTSS over records O(N log N)
      • LTSS over attributes O(M log M)

1. Start with a randomly chosen subset of attributes
Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure
Attributes $A_1 \ldots A_M$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      - LTSS over records $O(N \log N)$
      - LTSS over attributes $O(M \log M)$

1. Start with a randomly chosen subset of attributes
2. Use LTSS to find the highest-scoring subset of recs for the given atts

(Score = 7.5)
Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure
Attributes $A_1...A_M$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • LTSS over records $O(N \log N)$
      • LTSS over attributes $O(M \log M)$

2. Use LTSS to find the highest-scoring subset of recs for the given atts
3. Use LTSS to find the highest-scoring subset of atts for the given recs
Fast Generalized Subset Scan (FGSS)

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

3. Use LTSS to find the highest-scoring subset of atts for the given recs
4. Iterate steps 2-3 until convergence
Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure
Attributes $A_1...A_M$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
         i. LTSS over records $O(N \log N)$
         ii. LTSS over attributes $O(M \log M)$

3. Use LTSS to find the highest-scoring subset of atts for the given recs
4. Iterate steps 2-3 until convergence
Fast Generalized Subset Scan (FGSS)

**FGSS Search Procedure**

**Attributes** $A_1...A_M$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      - Iterate between following steps
         i. LTSS over records $O(N \log N)$
         ii. LTSS over attributes $O(M \log M)$

**Good News:** Run time is (near) linear in number of recs & number of atts.

**Bad News:** Not guaranteed to find global maximum of the score function.
Fast Generalized Subset Scan (FGSS)

FGSS Search Procedure
Attributes $A_1...A_M$

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
      i. LTSS over records $O(N \log N)$
      ii. LTSS over attributes $O(M \log M)$

5. Repeat steps 1-4 for 50 random restarts
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods.
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods defined by a center record.
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize F(S) over all subsets of S
      • Iterate between following steps
        i. LTSS over records O(N \log N)
        ii. LTSS over attributes O(M \log M)

We want to enforce self-similarity, thus we create local neighborhoods defined by a center record and all other records within a max dissimilarity
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood.
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize \( F(S) \) over all subsets of \( S \)
      • Iterate between following steps
        i. LTSS over records \( O(N \log N) \)
        ii. LTSS over attributes \( O(M \log M) \)

We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood.
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
         i. LTSS over records $O(N \log N)$
         ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood.
Fast Generalized Subset Scan (FGSS)

FGSS Constrained Search Procedure

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihoods
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize $F(S)$ over all subsets of $S$
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

We want to enforce self-similarity, thus we create local neighborhoods, do the unconstrained search within each local neighborhood, and maximize $F(S)$ over all local neighborhoods.
Experiments

• Network Activity and Intrusion Data (KDDCUP ’99)
  ▫ 41 attributes representing extracted information from the raw data of the network connection

• BARD Simulated Anthrax Outbreak in ED visits
  ▫ Hospital Id
  ▫ Prodrome
  ▫ Age Decile
  ▫ Patient Home Zip-code
  ▫ Chief Complaint

• U.S. Customs and Boarder Patrol Data
  ▫ Country of origin
  ▫ Departing & Arriving ports, Shipping line
  ▫ Shipper’s & Vessel’s name
  ▫ Commodity being shipped

• We compare FGSS to other recently proposed methods
  ▫ Bayesian Network Anomaly Detector
  ▫ Anomaly Pattern Detection (APD) (Das et al. 2008)
  ▫ Anomalous Group Detection (AGD) (Das et al. 2009)
Results

Run Times (100 Records)

Run Times (1,000 Records)

Run Times (10,000 Records)

Run Times (100,000 Records)

Exhaustive FGSS (Constrained)  FGSS (Constrained)  AGD
(BARD) Simulated Anthrax ED Dataset

Precision vs. Recall

Receiver Operator Characteristic

The proportion of true anomalies detected.
Results

Pattern Characterization Accuracy

- 80.0% for 1000 Records | 1% Injected
- 60.0% for 1000 Records | 10% Injected
- 40.0% for 10000 Records | 1% Injected

Legend:
- BJ
- APD
Conclusions

- FGSS run significantly faster than methods with comparable detection power
- FGSS outperforms other methods when patterns are:
  - a small portion of the data
  - subtle (not extremely individually anomalous)
- FGSS can characterize anomalous patterns
- Extensions
  - Extend method to handle multiple anomaly detectors
  - Extend method to handle multiple models
  - Active Learning
Extensions
(Preliminary)
Fast Generalized Subset Scan (FGSS)

I. Compute the anomalousnessness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihood

By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute attribute value likelihood

By performing inference on the Bayesian Network, for each record we can determine the likelihood of each of its attribute values
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute outlier scores

Is the value sufficiently far away from the mean (outlier)?
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute duplicate scores

Is the value a duplicate?
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute missing scores

Is the value missing?
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute anomaly scores

\[
l_{ij} = \frac{\sum_{k} \alpha_k \cdot I(isAnom(l_{ij}))}{K}
\]
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute glitch scores

Now have new measure of the anomalous each record x attribute pair.
Multiple Anomaly Detectors

I. Compute the anomalousness of each attribute (for each record)
   1. Learn Bayesian Network
   2. Compute glitch scores
   3. Compute empirical p-values

II. Discover subsets of records and attributes that are most anomalous
   1. Maximize F(S) over all subsets of S
      • Iterate between following steps
        i. LTSS over records $O(N \log N)$
        ii. LTSS over attributes $O(M \log M)$

Search over all possible subsets of data and find the maximizing $F(S)$

Attributes $A_1...A_M$

Records $R_1...R_N$
Thank You...Questions/Comments?