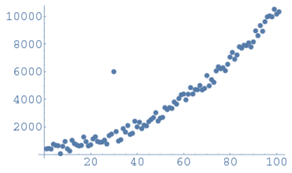
Fast Generalized Subset Scanning for Anomalous Pattern Detection

Edward McFowland III, Skyler Speakman, and Daniel B. Neill

This work was partially funded by NSF grants IIS-0916345 and IIS-0911032.

Motivation



Many techniques exist for detecting **single records** that are anomalous.

However, a more interesting problem is to find records that may appear normal when looked at individually, but are anomalous **as a group.**

Detecting anomalous patterns has a wide range of applications to business and policy:

Public Health: Patterns in electronic medical records can be used to detect disease outbreaks.

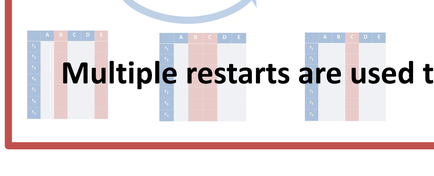
Network Security: Intrusion attempts can be detected by looking for anomalous patterns in activity.

Anti-Terrorism & Smuggling: Container shipment data can be analyzed to help customs officials inspect the most anomalous containers.

Method

A data set consists of records and categorical attributes.

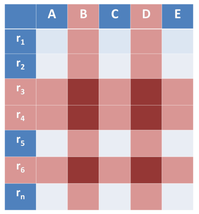
	A	B	C	D	E
r_1					
r_2					
...					
r_n					



A table of p-values is formed by learning a Bayes Net from the data.

The scoring function uses p-values to assign a score to any subset of records and attributes.

The goal is to *efficiently* find the *most anomalous* (highest scoring) subset of records and attributes.



The algorithm initializes by selecting a random subset of attributes.

For this set of attributes, we can quickly find the highest scoring set of records…

which can be used to determine the highest scoring set of attributes.

The process iterates through ordinal descent to a local optimum.

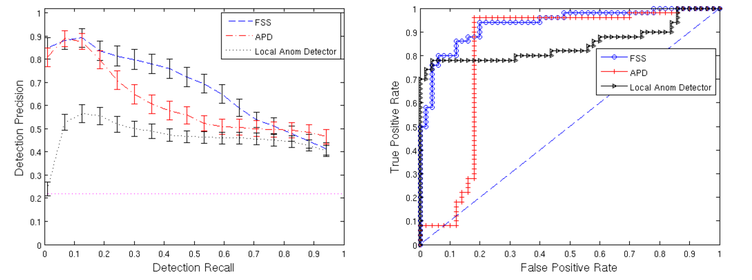
Multiple restarts are used to find the global optimum.

A distance metric, such as Hamming distance, can be used to constrain the search so that only self-similar groups are considered.

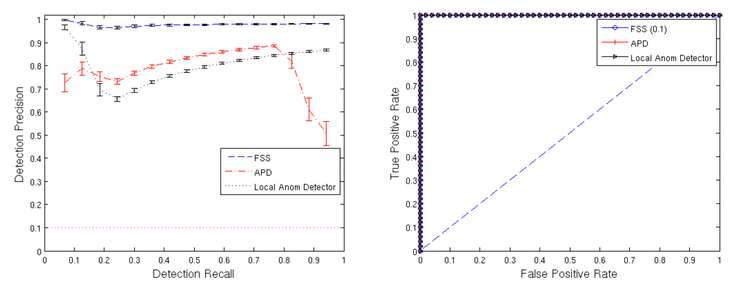
Results

The supervised version of the algorithm was performed on multiple real-world datasets:

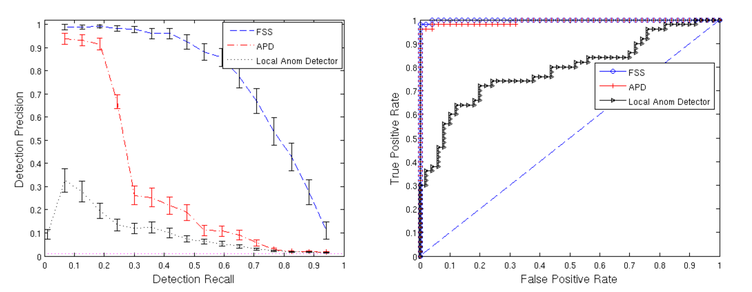
- Emergency Department data
- Network Intrusion data
- Container Shipment data



Simulated Anthrax outbreak (first 2 days)



KDDCup (1999) 'Guess Password' Attack



Semi-synthetic injects by altering attributes