Based on Tan, Steinbach, Kumar, Andrew Moore, Arindam Banerjee, Varun Chandola, Vipin Kumar, Jaideep Srivastava, Aleksandar Lazarevic
Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data

- Applications:
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection
Related problems

- Rare Class Mining
- Chance discovery
- Novelty Detection
- Exception Mining
- Noise Removal
- Black Swan*

Importance of Anomaly Detection

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels.

- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?

- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!

Sources:
Why Anomaly Detection?

- Anomalies are the focus
  - Anomaly detection is the goal
- Anomalies distort the data analysis
  - Anomaly removal as data preprocessing
Causes of Anomalies

- Data from different classes
- Natural variation
- Data measurement and collection errors
- Combination of different reasons
**Data Labels**

- **Supervised Anomaly Detection**
  - Labels available for both normal data and anomalies
  - Similar to rare class mining
- **Semi-supervised Anomaly Detection**
  - Labels available only for normal data
- **Unsupervised Anomaly Detection**
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data
Type of Anomaly

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies
Point Anomalies

An individual data instance is anomalous w.r.t. the data
Contextual Anomalies

- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies*

Collective Anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a collective anomaly are not anomalous by themselves

Anomalous Subsequence
Evaluation of Anomaly Detection – F-value

- Accuracy is not sufficient metric for evaluation
  - Example: network traffic data set with 99.9% of normal data and 0.1% of intrusions
  - Trivial classifier that labels everything with the normal class can achieve 99.9% accuracy !!!!!

<table>
<thead>
<tr>
<th>Confusion matrix</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC</td>
<td>C</td>
</tr>
<tr>
<td>Actual class</td>
<td>NC</td>
</tr>
<tr>
<td>C</td>
<td>TN</td>
</tr>
<tr>
<td>FNN</td>
<td>TP</td>
</tr>
</tbody>
</table>

- Focus on both recall and precision
  - Recall \( (R) \) = \( \frac{TP}{TP + FN} \)
  - Precision \( (P) \) = \( \frac{TP}{TP + FP} \)

- \( F \) – measure = \( \frac{2*R*P}{R+P} \)
Evaluation of Outlier Detection – ROC & AUC

- Standard measures for evaluating anomaly detection problems:
  - Recall (Detection rate) - ratio between the number of correctly detected anomalies and the total number of anomalies
  - False alarm (false positive) rate – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class
  - ROC Curve is a trade-off between detection rate and false alarm rate
  - Area under the ROC curve (AUC) is computed using a trapezoid rule
Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Industrial Damage Detection
- Image Processing / Video surveillance
- Novel Topic Detection in Text Mining
- ...
Unsupervised Anomaly Detection

- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)

- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data
Anomaly Detection Schemes

- **General Steps**
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

- **Types of anomaly detection schemes**
  - Graphical & Statistical-based
  - Distance-based
  - Model-based
Graphical/Visualization Approaches

- Boxplot (1-D), Scatter plot (2-D)

- Limitations
  - Time consuming
  - Subjective

![Boxplot and Scatter plot examples](image)
Visualization Based (2)

- Detecting Tele-communication fraud
- Display telephone call patterns as a graph
- Use colors to identify fraudulent telephone calls (anomalies)

Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
Grubbs’ Test

- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - $H_0$: There is no outlier in data
  - $H_A$: There is at least one outlier
- Grubbs’ test statistic:
  $$G = \frac{\max|X - \bar{X}|}{s}$$
- Reject $H_0$ if:
  $$G > \frac{(N - 1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N - 2 + t^2_{(\alpha/N, N-2)}}}$$
Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
  - M (majority distribution)
  - A (anomalous distribution)

- General Approach:
  - Initially, assume all the data points belong to M
  - Let $L_t(D)$ be the log likelihood of D at time $t$
  - For each point $x_t$ that belongs to M, move it to A
    - Let $L_{t+1}(D)$ be the new log likelihood.
    - Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
    - If $\Delta > c$ (some threshold), then $x_t$ is declared as an anomaly and moved permanently from M to A
Statistical-based — Likelihood Approach

- Data distribution, \( D = (1 - \lambda) M + \lambda A \)
  - \( M \) is a probability distribution estimated from data
    - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
  - \( A \) is initially assumed to be uniform distribution
- Likelihood at time \( t \):

\[
L_t(D) = \prod_{i=1}^{N} P_D(x_i) = (1 - \lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \left( \lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)
\]

\[
LL_t(D) = |M_t| \log(1 - \lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)
\]

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004
Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
Distance-based Approaches

- Data is represented as a vector of features

- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based
Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are fewer than $p$ neighboring points within a distance $D$
    - The top $n$ data points whose distance to the kth nearest neighbor is greatest
    - The top $n$ data points whose average distance to the k nearest neighbors is greatest
Outliers in Lower Dimensional Projection

- Divide each attribute into $\phi$ equal-depth intervals
  - Each interval contains a fraction $f = 1/\phi$ of the records

- Consider a $k$-dimensional cube created by picking grid ranges from $k$ different dimensions
  - If attributes are independent, we expect region to contain a fraction $f^k$ of the records
  - If there are $N$ points, we can measure sparsity of a cube $D$ as:
    \[
    S(D) = \frac{n(D) - N \cdot f^k}{\sqrt{N \cdot f^k \cdot (1 - f^k)}}
    \]
  - Negative sparsity indicates cube contains smaller number of points than expected
Example

- \( N=100, \phi = 5, f = 1/5 = 0.2, N \times f^2 = 4 \)
Density-based: LOF approach

- For each point, compute the density of its local neighborhood.
- Compute local outlier factor (LOF) of a sample \( p \) as the average of the ratios of the density of sample \( p \) and the density of its nearest neighbors.
- Outliers are points with largest LOF value.
Clustering-Based

- Cluster the data
- Points in small cluster as candidate outliers
- If candidate points are far from all other non-candidate points, they are outliers
Detecting geographic hotspots

Entire area being scanned

(Philadelphia Metro)
One Step of Spatial Scan

- Entire area being scanned
- Current region being considered
One Step of Spatial Scan

I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)
One Step of Spatial Scan

Entire area being scanned

Current region being considered

I have a population of 5300 of whom 53 are sick (1%)

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

So... is that a big deal? Evaluated with Score function (e.g. Kulldorf’s score)
One Step of Spatial Scan

Entire area being scanned

Current region being considered

I have a population of 5300 of whom 53 are sick (1%)

[Score = 1.4]

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

So... is that a big deal?
Evaluated with Score function (e.g. Kulldorf’s score)
Many Steps of Spatial Scan

I have a population of 5300 of whom 53 are sick (1%)

[Score = 1.4]

Everywhere else has a population of 2,200,000 of whom 20,000 are sick (0.9%)

[Score = 9.3]

So... is that a big deal?
Evaluated with Score function (e.g. Kulldorf’s score)
Scan Statistics

Standard scan statistic question: Given the geographical locations of occurrences of a phenomenon, is there a region with an unusually high (low) rate of these occurrences?

Standard approach:

1. Compute the likelihood of the data given the hypothesis that the rate of occurrence is uniform everywhere, $L_0$.
2. For some geographical region, $W$, compute the likelihood that the rate of occurrence is uniform at one level inside the region and uniform at another level outside the region, $L(W)$.
3. Compute the likelihood ratio, $L(W)/L_0$.
4. Repeat for all regions, and find the largest likelihood ratio. This is the scan statistic, $S^*_W$.
5. Report the region, $W$, which yielded the max, $S^*_W$.

See [Glaz and Balakrishnan, 99] for details.
Significance testing

Standard approach:

1. Generate many randomized versions of the data set by shuffling the labels (positive instance of the phenomenon or not).

2. Compute $S^*_W$ for each randomized data set. This forms a baseline distribution for $S^*_W$ if the null hypothesis holds.

3. Compare the observed value of $S^*_W$ against the baseline distribution to determine a p-value.

Given that region W is the most likely to be abnormal, is it significantly abnormal?
Fast squares speedup

- Theoretical complexity of fast squares: $O(N^2)$ (as opposed to naïve $N^3$), if maximum density region sufficiently dense.
  
  *If not, we can use several other speedup tricks.*

- In practice: 10-200x speedups on real and artificially generated datasets.

  *Emergency Dept. dataset (600K records): 20 minutes, versus 66 hours with naïve approach.*
Base Rate Fallacy

- Bayes theorem:

\[
P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}
\]

- More generally:

\[
P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}
\]
The base-rate fallacy is best described through example. Suppose that your doctor performs a test that is 99% accurate, i.e. when the test was administered to a test population all of whom had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative. Upon visiting your doctor to learn the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1/10000, i.e. only 1 in 10000 people have this ailment. What, given this information, is the probability of you having the disease? The reader is encouraged to make a quick “guesstimate” of the answer at this point.
Base Rate Fallacy

\[
P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}
\]

\[
P(S|P) = \frac{\frac{1}{10000} \cdot 0.99}{\frac{1}{10000} \cdot 0.99 + (1 - \frac{1}{10000}) \cdot 0.01} = 0.00980\ldots \approx 1\%
\]

- Even though the test is 99% certain, your chance of having the disease is 1/100, because the population of healthy people is much larger than sick people
Base Rate Fallacy in Intrusion Detection

- I: intrusive behavior,
  \(\neg I\): non-intrusive behavior
- A: alarm
  \(\neg A\): no alarm

- Detection rate (true positive rate): \(P(A|I)\)
- False alarm rate: \(P(A|\neg I)\)

- Goal is to maximize both
  - Bayesian detection rate, \(P(I|A)\)
  - \(P(\neg I|\neg A)\)
Detection Rate vs False Alarm Rate

\[ P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)} \]

- Suppose:
  \[ P(I) = \frac{1}{\frac{1 \cdot 10^6}{2 \cdot 10}} = 2 \cdot 10^{-5} \]
  \[ P(\neg I) = 1 - P(I) = 0.99998 \]

- Then:
  \[ P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)} \]

- False alarm rate becomes more dominant if \( P(I) \) is very low.
Axelsson: We need a very low false alarm rate to achieve a reasonable Bayesian detection rate
Contextual Anomaly Detection

- Detect context anomalies

- General Approach
  - Identify a context around a data instance (using a set of *contextual attributes*)
  - Determine if the data instance is anomalous w.r.t. the context (using a set of *behavioral attributes*)

- Assumption
  - All normal instances within a context will be similar (in terms of behavioral attributes), while the anomalies will be different
Conclusions

- Anomaly detection can detect critical information in data
- Highly applicable in various application domains
- Nature of anomaly detection problem is dependent on the application domain
- Need different approaches to solve a particular problem formulation
References

- W. Fan et al, Using Artificial Anomalies to Detect Unknown and Known Network Intrusions, ICDM 2001
- Barbara, D., Couto, J., Jajodia, S., and Wu, N. Adam: a testbed for exploring the use of data mining in intrusion detection. SIGMOD Rec., 2001
- Ide, T. and Kashima, H. Eigenspace-based anomaly detection in computer systems. KDD, 2004
- Sun, J. et al., Less is more: Compact matrix representation of large sparse graphs. ICDM 2007
References

- Ratsch, G., Mika, S., Scholkopf, B., and Muller, K.-R. Constructing boosting algorithms from SVMs: An application to one-class classification. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2002
- Baker, D. et al., A hierarchical probabilistic model for novelty detection in text. ICML 1999
- Das, K. and Schneider, J. Detecting anomalous records in categorical datasets. KDD 2007
References

- Caudell, T. and Newman, D. An adaptive resonance architecture to define normality and detect novelties in time series and databases. World Congress on Neural Networks, 1993
- Steinwart, I., Hush, D., and Scovel, C. A classification framework for anomaly detection. JMLR, 2005
- Li, Y., Pont et al. Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown faults may occur. Pattern Recognition Letters, 2002