Scalable Anomaly Detection with Spark and SOS
Hi there, my name is Jeroen Janssens
Today

- SOS, World!
- Anomalies and outliers
- Evaluating outlier-selection algorithms
- Various approaches to outlier selection
- Stochastic Outlier Selection
- Conclusion
SOS, World!
Implementations of SOS

Anomalies and outliers
An anomaly is an observation or event that deviates qualitatively from what is considered to be normal, according to a domain expert.
Detecting anomalies is important

- Expensive
- Dangerous
- Mess up your model
Human anomaly detection may suffer from

- Fatigue
- Information overload
- Emotional bias
Feature-vector representation

**Observations**

![Image of an apple and an orange]

**Data points**

<table>
<thead>
<tr>
<th>width</th>
<th>height</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.4</td>
<td>7.3</td>
<td>apple</td>
</tr>
<tr>
<td>6.7</td>
<td>7.1</td>
<td>orange</td>
</tr>
<tr>
<td>8.0</td>
<td>6.8</td>
<td>apple</td>
</tr>
<tr>
<td>7.4</td>
<td>7.2</td>
<td>apple</td>
</tr>
<tr>
<td>9.6</td>
<td>9.2</td>
<td>orange</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

**Visualisation**

- Chart showing scatter plot with points representing width and height, differentiated by 'apple' and 'orange' labels.

---

[Image source: data-science-workshops.com]
Dissimilarity-matrix representation

\[ d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{jk} - x_{ik})^2} \]
From anomaly to outlier
An outlier is a data point that deviates quantitatively from the majority of the data points, according to an outlier-selection algorithm.
The symbiotic relationship between the domain expert and the algorithm

<table>
<thead>
<tr>
<th>The</th>
<th>Domain expert</th>
<th>Outlier-selection algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>employs</td>
<td>Knowledge and experience</td>
<td>Mathematics and statistics</td>
</tr>
<tr>
<td>on the</td>
<td>Real world</td>
<td>Data set</td>
</tr>
<tr>
<td>to</td>
<td>Detect</td>
<td>Select</td>
</tr>
<tr>
<td>those</td>
<td>Observations</td>
<td>Data points</td>
</tr>
<tr>
<td>that are</td>
<td>Abnormal</td>
<td>Outlying</td>
</tr>
<tr>
<td>as</td>
<td>Anomalies</td>
<td>Outliers</td>
</tr>
</tbody>
</table>
Data flow diagram illustrating the relationship between the domain expert (square) and the outlier-selection algorithm (top circle).
Six Euler diagrams (1/2)

Euler diagram | Set legend | Set notation
--- | --- | ---
(a) | real-world observations | $\mathcal{X}$
(b) | labelled by expert as anomalous | $\mathcal{A}$
| | labelled by expert as normal | $\mathcal{X} \cap \overline{\mathcal{A}}$
(c) | data points | $\mathcal{D}$
| | unrecorded | $\mathcal{X} \cap \overline{\mathcal{D}}$
Six Euler diagrams (2/2)

<table>
<thead>
<tr>
<th>Euler diagram</th>
<th>Set legend</th>
<th>Set notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d)</td>
<td>anomalies represented as data points</td>
<td>$C_A = D \cap A$</td>
</tr>
<tr>
<td></td>
<td>normalities represented as data points</td>
<td>$C_N = D \cap \overline{A}$</td>
</tr>
<tr>
<td>(e)</td>
<td>classified by algorithm as an outlier</td>
<td>$C_O$</td>
</tr>
<tr>
<td></td>
<td>classified by algorithm as an inlier</td>
<td>$C_I = D \cap \overline{C_O}$</td>
</tr>
<tr>
<td>(f)</td>
<td>hits</td>
<td>$Hi = C_A \cap C_O$</td>
</tr>
<tr>
<td></td>
<td>false alarms</td>
<td>$FA = C_N \cap C_O$</td>
</tr>
<tr>
<td></td>
<td>misses</td>
<td>$Mi = C_A \cap C_I$</td>
</tr>
<tr>
<td></td>
<td>correct rejects</td>
<td>$CR = C_N \cap C_I$</td>
</tr>
</tbody>
</table>
Evaluating outlier-selection algorithms
Confusion matrix

Computer says no.

<table>
<thead>
<tr>
<th>Expert labels the observation as an</th>
<th>Anomaly ($C_A$)</th>
<th>Normality ($C_N$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier ($C_O$)</td>
<td>hit $H_i$</td>
<td>false alarm $FA$</td>
</tr>
<tr>
<td>Inlier ($C_I$)</td>
<td>miss $M_i$</td>
<td>correct reject $CR$</td>
</tr>
</tbody>
</table>
Four possible outcomes
Evaluation

Illustration of relabelling a multi-class data set into multiple one-class data sets.
Anomalies are rare

In order to evaluate the algorithm we simulate anomalies to be rare. Banana for scale.
Outlier scores

The dashed line indicates the threshold chosen by the domain expert.
An ROC curve plots the false alarm rate against the hit rate for all possible thresholds.

\[ \theta' = 0.57 \]
Various approaches to outlier selection
Distribution-based outlier selection
Distance-based outlier selection

Size does matter
Density-based outlier-selection
Stochastic Outlier Selection
Stochastic Outlier Selection

- Unsupervised outlier selection algorithm
- Employs concept of affinity (inspired by t-SNE)
- One parameter: perplexity
- Computes outlier probabilities
t-Distributed Neighbor Embedding (t-SNE; Van der Maaten, Hinton) employs affinity to perform dimensionality reduction.
A data point is selected as an outlier when all the other data points have insufficient affinity with it.
From input to output

\[ X \rightarrow D \rightarrow A \rightarrow B \rightarrow \Phi \]
From feature matrix to dissimilarity matrix

\[ d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{jk} - x_{ik})^2}, \]
From input to output
Smooth neighborhoods
Affinity between data points

\[ a_{ij} = \begin{cases} 
\exp\left(-\frac{d_{ij}^2}{2\sigma_i^2}\right) & \text{if } i \neq j \\
0 & \text{if } i = j 
\end{cases} \]
From affinity to binding probability

The binding matrix $B$ is obtained by normalising each row in the affinity matrix $A$.
Binding probabilities form a graph
Binding probabilities form a graph
Stochastic Neighbor Graph

A data point belongs to the outlier class when it is not selected by any other data points.

\[ G = (\mathcal{V}, \mathcal{E}_G) \]

\[ p(G) = \prod_{i \rightarrow j \in \mathcal{E}_G} b_{ij}. \]

\[ \mathcal{C}_0 \mid G = \{ x_i \in \mathcal{X} \mid \deg^-_G(v_i) = 0 \} \]
\[ = \{ x_i \in \mathcal{X} \mid \nexists v_j \in \mathcal{V} : j \rightarrow i \in \mathcal{E}_G \} \]
\[ = \{ x_i \in \mathcal{X} \mid \forall v_j \in \mathcal{V} : j \rightarrow i \notin \mathcal{E}_G \} \]
The three SNGs Ga, Gb, and Gc are sampled from the discrete probability distribution \( P(G) \).

\[
p(G_a) = 3.931 \cdot 10^{-4}
\]

\[
C_0|G_a = \{x_1, x_4, x_6\}
\]

\[
p(G_b) = 4.562 \cdot 10^{-5}
\]

\[
C_0|G_b = \{x_5, x_6\}
\]

\[
p(G_c) = 5.950 \cdot 10^{-7}
\]

\[
C_0|G_c = \{x_1, x_3\}
\]
Set of all SNGs
Approximating outlier probabilities by sampling SNGs

\[ p(x_i \in C_0) = \lim_{s \to \infty} \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\{x_i \in C_0 \mid G^{(s)}\}, \quad G^{(s)} \sim P(G) \]
Demo: Sampling SNGs in CoffeeScript and D3

Computing outlier probabilities through marginalisation

\[ p(x_i \in C_0) = \sum_{G \in \mathcal{G}} \mathbb{I}\{x_i \in C_0 \mid G\} \cdot p(G) \]

\[ = \sum_{G \in \mathcal{G}} \mathbb{I}\{x_i \in C_0 \mid G\} \cdot \prod_{q \rightarrow r \in \mathcal{E}_G} b_{qr} \cdot |\mathcal{G}| = (n - 1)^n \]
Computing outlier probabilities in closed form

\[ p(x_i \in C_0) = \prod_{j \neq i} (1 - b_{ji}) \]
Proof!

\[ x_i \in C_0 \mid G \leftrightarrow \deg^-(v_i) = 0 \]  
\[ p(x_i \in C_0) = \mathbb{E}_G \left[ \prod \{ j \to i \notin \mathcal{E}_G \} \right] \]  
\[ p(x_i \in C_0) = \mathbb{E}_G \left[ \prod_{j \neq i} (1 - \mathbb{I}\{ j \to i \in \mathcal{E}_G \}) \right] \]  
\[ p(x_i \in C_0) = \prod_{j \neq i} (1 - \mathbb{E}_G [\mathbb{I}\{ j \to i \in \mathcal{E}_G \}]) \]  
\[ p(x_i \in C_0) = \prod_{j \neq i} (1 - p(j \to i \in \mathcal{E}_G)) \]  
\[ p(x_i \in C_0) = \prod_{j \neq i} (1 - b_{ji}) \]
Selecting outliers

\[ f(x) = \begin{cases} 
\text{outlier} & \text{if } p(x \in \mathcal{C}_0) > \theta, \\
\text{inlier} & \text{if } p(x \in \mathcal{C}_0) \leq \theta.
\end{cases} \]
Adaptive variances via the perplexity parameter

\[ h(b_i) = 2^{H(b_i)}, \quad H(b_i) = - \sum_{j=1}^{n} b_{ij} \log_2 (b_{ij}) \]
Continuous binary search

![Graph showing variance and perplexity over binary search iterations for different variables.](image-url)
Perplexity influences outlier probabilities
Evaluation and comparison
Putlier-score plots

![Putlier-score plots](image)
Real-world datasets
Synthetic datasets
## Synthetic datasets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Determines</th>
<th>Parameter $\lambda$</th>
<th>$\lambda_{\text{start}}$</th>
<th>step size</th>
<th>$\lambda_{\text{end}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>Radius of ring</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>Cardinality of cluster and ring</td>
<td>100</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>Distance between clusters</td>
<td>4</td>
<td>0.1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>Cardinality of one cluster and ring</td>
<td>0</td>
<td>5</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>(e)</td>
<td>Density of one cluster and ring</td>
<td>1</td>
<td>0.05</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>(f)</td>
<td>Radius of square ring</td>
<td>2</td>
<td>0.05</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>(g)</td>
<td>Radius of square ring</td>
<td>2</td>
<td>0.05</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>
Synthetic datasets
Synthetic datasets

(e)  

(f)  

(g)  

AUC = 1  

AUC  

AUC = 0.6  

AUC = 0.7  

AUC = 0.8  

AUC = 0.9  

AUC = 1  

\( \lambda \)  

\( \lambda \)  

\( \lambda \)  

\( \lambda \)  

SOS  

KNNDD  

LOF  

LOCI  

LSOD
SOS performs significantly better
Spark implementation of SOS
Spark implementation of SOS

- Developed by Fokko Driesprong
- Works with DataFrame API
- Available on GitHub
- Plan is to make it part of MLLib
SOS on PySpark

92-pyspark-sos.ipynb
Summary

- Outlier selection can support the detection of anomalies
- SOS is an intuitive and probabilistic algorithm to select outliers
- SOS has a very good performance
- No free lunch
Thank you! Here are some links