Anomaly Detection in Scikit-Learn and new tools from Multivariate Extreme Value Theory

Nicolas Goix

Supervision:

Detecting Anomalies with Multivariate Extremes: Stéphane Clémençon and Anne Sabourin

Contributions to Scikit-Learn: Alexandre Gramfort

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay
1. Anomaly Detection and Scikit-Learn

2. Multivariate EVT & Representation of Extremes

3. Estimation

4. Experiments
Anomaly Detection (AD)

What is Anomaly Detection?

"Finding patterns in the data that do not conform to expected behavior"

Huge number of applications: Network intrusions, credit card fraud detection, insurance, finance, military surveillance,...
Machine Learning context

Different kind of Anomaly Detection

- **Supervised AD**
  - Labels available for both normal data and anomalies
  - Similar to rare class mining

- **Semi-supervised AD (Novelty Detection)**
  - Only normal data available to train
  - The algorithm learns on normal data only

- **Unsupervised AD (Outlier Detection)**
  - No labels, training set = normal + abnormal data
  - Assumption: anomalies are very rare
Important literature in Anomaly Detection:

- **statistical AD techniques**
  fit a statistical model for normal behavior
  ex: `EllipticEnvelope`

- **density-based**
  - ex: Local Outlier Factor (LOF) and variantes (COF ODIN LOCI)

- **Support estimation** - `OneClassSVM` - MV-set estimate

- **high-dimensional techniques**: - Spectral Techniques - Random Forest - `Isolation Forest`
Isolation Forest:

Liu Tink Zhou icdm2008

(a) Isolating $x_i$  (b) Isolating $x_o$
Anomaly Detection and Scikit-Learn

average path length

nb. of tree (log scale)
IsolationForest.fit(X)

**Inputs:** X, n_estimators, max_samples

**Output:** Forest with:
- # trees = n_estimators
- sub-sampling size = max_samples
- maximal depth $max_depth = \text{int}(\log_2 max\_samples)$

Complexity: $O(n\_estimators\ max\_samples\ \log(\max\_samples))$

default: n_estimators=100, max_samples=256
IsolationForest.predict(X)

Finding the depth in each tree

\[ \text{depth}(\text{Tree}, X): \]

\# – Finds the depth level of the leaf node for each sample \( x \) in \( X \).
\# – Add average path length \( (n \text{ samples in leaf}) \)
\# if \( x \) not isolated

\[ \text{score}(x, n) = 2^{\frac{E(\text{depth}(x))}{c(n)}} \]

Complexity: \( O( n \text{ samples} \ n \text{ estimators} \ \log(\text{max samples})) \)
Examples

- code example:

```python
>>> from sklearn.ensemble import IsolationForest
>>> IF = IsolationForest()
>>> IF.fit(X_train)  # build the trees
>>> IF.predict(X_test)  # find the average depth
```

- plotting decision function:
Anomaly Detection and Scikit-Learn

\[ n_{\text{samples\ normal}} = 150 \]
\[ n_{\text{samples\ outliers}} = 50 \]
Anomaly Detection and Scikit-Learn

\[ n_{\text{samples\_normal}} = 150 \]
\[ n_{\text{samples\_outliers}} = 50 \]
1 Anomaly Detection and Scikit-Learn

2 Multivariate EVT & Representation of Extremes

3 Estimation

4 Experiments
General idea of our work

- Extreme observations play a special role when dealing with outlying data.

- But no algorithm has **specific treatment for such multivariate extreme observations**.

- Our goal: Provide a method which can improve performance of standard AD algorithms by combining them with a **multivariate extreme analysis** of the **dependence structure**.
Multivariate EVT & Representation of Extremes

DAMEX

Classical AD Algo.
Goal:

\[ X = (X_1, \ldots, X_d) \]

Find the groups of features which can be large together

ex: \( \{X_1, X_2\}, \{X_3, X_6, X_7\}, \{X_2, X_4, X_{10}, X_{11}\} \)

\( \Leftrightarrow \) Characterize the extreme dependence structure

Anomalies = points which violate this structure
Framework

- **Context**
  - Random vector $\mathbf{X} = (X_1, \ldots, X_d)$
  - Margins: $X_j \sim F_j$ (for continuous $F_j$)

- **Preliminary step: Standardization of each marginal**
  - Standard Pareto: $V_j = \frac{1}{1 - F_j(X_j)}$ \quad \left( \mathbb{P}(V_j \geq x) = \frac{1}{x}, \quad x \geq 1 \right)$
Problematic

Joint extremes: $V$’s distribution above large thresholds?

$\mathbb{P}(V \in A)$? (A ‘far from the origin’).
Fundamental hypothesis and consequences

- Standard assumption: let $A$ extreme region,

$$\mathbb{P}[\mathbf{V} \in t\ A] \sim t^{-1}\mathbb{P}[\mathbf{V} \in A] \quad \text{(radial homogeneity)}$$

- Formally,

**regular variation** (after standardization):

$0 \notin \overline{A}$

$$t\mathbb{P}[\mathbf{V} \in t\ A] \xrightarrow{t \to \infty} \mu(A), \quad \mu: \text{exponent measure}$$

Necessarily: $\mu(tA) = t^{-1}\mu(A)$

- $\Rightarrow$ **angular measure** on sphere $S_{d-1}$: $\Phi(B) = \mu\{tB, \ t \geq 1\}$
Intuitively: \( \mathbb{P}[ \mathbf{V} \in A] \approx \mu(A) \) For a large \( r > 0 \) and a region \( B \) on the unit sphere:

\[
\mathbb{P} \left[ \| \mathbf{V} \| > r, \frac{\mathbf{V}}{\| \mathbf{V} \|} \in B \right] \sim \frac{1}{r} \Phi(B) = \mu(\{tB, t \geq r\}) , \ r \to \infty
\]

\( \Rightarrow \) \( \Phi \) (or \( \mu \)) rules the joint distribution of extremes (if margins are known).
Angular distribution

- $\Phi$ rules the joint distribution of extremes

- Asymptotic dependence: $(V_1, V_2)$ may be large together.

- Asymptotic independence: only $V_1$ or $V_2$ may be large.
General Case

- Sub-cones: $\mathcal{C}_\alpha = \{ \|v\| \geq 1, \ v_i > 0 \ (i \in \alpha), \ v_j = 0 \ (j \notin \alpha) \}$
- Corresponding sub-spheres: $\{ \Omega_\alpha, \alpha \subset \{1, \ldots, d\} \}$
  $(\Omega_\alpha = \mathcal{C}_\alpha \cap S_{d-1})$
Representation of Extreme Data

- Natural decomposition of the angular measure:

\[ \Phi = \sum_{\alpha \subset \{1, \ldots, d\}} \Phi_\alpha \quad \text{with} \quad \Phi_\alpha = \Phi|_{\Omega_\alpha} \leftrightarrow \mu|_{C_\alpha} \]

- \[ \Rightarrow \text{yields a representation} \]

\[
\mathcal{M} = \left\{ \Phi(\Omega_\alpha) : \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\} \\
= \left\{ \mu(C_\alpha) : \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\}
\]

- Assumption: \[ \frac{d\mu|_{C_\alpha}}{dv_\alpha} = O(1). \]

- Remark: Representation \( \mathcal{M} \) is linear (after non-linear transform of the data \( X \rightarrow V \)).
Sparse Representation?

Full pattern: anything may happen

Sparse pattern: \(V_1\) not large if \(V_2\) or \(V_3\) large
1 Anomaly Detection and Scikit-Learn
2 Multivariate EVT & Representation of Extremes
3 Estimation
4 Experiments
Problem: $\mathcal{M}$ is an asymptotic representation

$$\mathcal{M} = \{ \Phi(\Omega_\alpha), \alpha \} = \{ \mu(C_\alpha), \alpha \}$$

is the restriction of an asymptotic measure

$$\mu(A) = \lim_{t \to \infty} t \mathbb{P}[V \in tA]$$

to a representative class of set $\{C_\alpha, \alpha\}$, but only the central sub-cone has positive Lebesgue measure!

$\Rightarrow$ Cannot just do, for large $t$:

$$\Phi(\Omega_\alpha) = \mu(C_\alpha) \approx t \hat{\mathbb{P}}(tC_\alpha)$$
Solution

Fix $\epsilon > 0$. Affect data $\epsilon$-close to an edge, to that edge.

$$\Omega_\alpha \rightarrow \Omega^\epsilon_\alpha = \{ \mathbf{v} \in S_{d-1} : v_j > \epsilon (j \in \alpha), v_j \leq \epsilon (j \notin \alpha) \}.$$

$$C_\alpha \rightarrow C^\epsilon_\alpha = \{ t \Omega^\epsilon_\alpha, t \geq 1 \}$$

New partition of $S_{d-1}$, compatible with non asymptotic data.
\[ \hat{V}_i^j = \frac{1}{1 - \hat{F}_j(X_i^j)} \] with \[ \hat{F}_j(X_i^j) = \frac{\text{rank}(X_i^j) - 1}{n} \]

\[ \Rightarrow \text{get an natural estimate of } \Phi(\Omega_\alpha) \]

\[ \hat{\Phi}(\Omega_\alpha) := \frac{n}{k} \mathbb{P}(\hat{V} \in \frac{n}{k} C^\epsilon) \]

\[ (\frac{n}{k} \text{ large, } \epsilon \text{ small}) \]

\[ \Rightarrow \text{we obtain} \]

\[ \hat{M} := \{ \hat{\Phi}(\Omega_\alpha), \alpha \} \]
**Theorem**

*There is an absolute constant $C > 0$ such that for any $n > 0$, $k > 0$, $0 < \varepsilon < 1$, $\delta > 0$ such that $0 < \delta < e^{-k}$, with probability at least $1 - \delta$,*

$$
\| \hat{M} - M \|_{\infty} \leq Cd \left( \sqrt{\frac{1}{\varepsilon k}} \log \frac{d}{\delta} + Md \varepsilon \right) + \text{bias}(\varepsilon, k, n),
$$

**Comments:**

- $C$: depends on $M = \sup$(density on subfaces)
- Existing litterature (for spectral measure) Einmahl Segers 09, Einmahl *et.al. 01*

\[ d = 2. \]

asymptotic behaviour, rates in $1/\sqrt{k}$.

**Here:** $1/\sqrt{k} \to 1/\sqrt{\varepsilon k} + \varepsilon$. Price to pay for biasing our estimator with $\varepsilon$. 
Theorem’s proof

1. Maximal deviation on VC-class:

\[ \sup_{x \geq \epsilon} |\mu_n - \mu([x, \infty[) \leq C d \sqrt{\frac{2}{k} \log \frac{d}{\delta}} + \text{bias}(\epsilon, k, n) \]

Tools: Vapnik-Chervonenkis inequality adapted to small probability sets: bounds in \( \sqrt{p} \sqrt{\frac{1}{n} \log \frac{1}{\delta}} \)

On the VC class \( \{\lceil \frac{n}{k} x, \infty \rceil, x \geq \epsilon \} \)
Theorem’s proof

1. Maximal deviation on VC-class:
2. Decompose error:

$$|\mu_n(C^\varepsilon_\alpha) - \mu(C_\alpha)| \leq \underbrace{|\mu_n - \mu|}_{A}(C^\varepsilon_\alpha) + \underbrace{|\mu(C^\varepsilon_\alpha) - \mu(C_\alpha)|}_{B}$$

- A: First step.
- B: density on $C^\varepsilon_\alpha \times \text{Lebesgue}$: small
Algorithm

**DAMEX in O\((dn \log n)\)**

**Input:** parameters \( \epsilon > 0, \ k = k(n) \),

1. **Standardize via marginal rank-transformation:**
   \[
   \hat{V}_i := \left(1/(1 - \hat{F}_j(X^j_i))\right)_{j=1,\ldots,d}.
   \]

2. **Assign to each** \( \hat{V}_i \) **the cone** \( \frac{n}{k} \mathcal{C}_\epsilon \) **it belongs to.**

3. \( \Phi_{n,\epsilon}^{\alpha} := \hat{\Phi}(\Omega_\alpha) = \frac{n}{k} \mathbb{P}_n(\hat{V} \in \frac{n}{k} \mathcal{C}_\epsilon) \) **the estimate of the** \( \alpha \)-**mass of** \( \Phi \).**

**Output:** *(sparse)* **representation of the dependence structure**

\[
\hat{M} := (\Phi_{n,\epsilon}^{\alpha})_{\alpha \subset \{1,\ldots,d\}, \Phi_{n,\epsilon}^{\alpha} > \Phi_{\min}}
\]
Application to Anomaly Detection

After standardization of marginals: \( \mathbb{P}[R > r, W \in B] \equiv \frac{1}{r} \Phi(B) \)

\[ \rightarrow \text{scoring function} = \Phi_n^\epsilon \times 1/r : \]

\[ s_n(x) := \left( \frac{1}{\| \hat{T}(x) \|_\infty} \right) \sum_{\alpha} \Phi_{n,\epsilon}^{\alpha} \mathbb{1}_{\hat{T}(x) \in C_\alpha^\epsilon}. \]

where \( T : X \mapsto V \quad (V_j = \frac{1}{1 - F_j(x_j)}) \)
1 Anomaly Detection and Scikit-Learn

2 Multivariate EVT & Representation of Extremes

3 Estimation

4 Experiments
<table>
<thead>
<tr>
<th></th>
<th>number of samples</th>
<th>number of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>shuttle</td>
<td>85849</td>
<td>9</td>
</tr>
<tr>
<td>forestcover</td>
<td>286048</td>
<td>54</td>
</tr>
<tr>
<td>SA</td>
<td>976158</td>
<td>41</td>
</tr>
<tr>
<td>SF</td>
<td>699691</td>
<td>4</td>
</tr>
<tr>
<td>http</td>
<td>619052</td>
<td>3</td>
</tr>
<tr>
<td>smtp</td>
<td>95373</td>
<td>3</td>
</tr>
</tbody>
</table>

**Table:** Datasets characteristics
Figure: ROC and PR curve on SF dataset

$\epsilon = 0.01, \ k = n^{1/2}$

ROC curve for SF

PR curve for SF

- DAMEX+iForest (area = 0.694)
- iForest (area = 0.041)

True Positive Rate vs False Positive Rate

Precision vs Recall
Figure: ROC and PR curve on http dataset
Figure: ROC and PR curve on shuttle dataset
Experiments

$\epsilon = 0.0001, \ k = n^{1/3}$

**Figure:** ROC and PR curve on forestcover dataset

- ROC curve for forestcover
- PR curve for forestcover

- DAMEX+iForest (area = 0.363)
- iForest (area = 0.193)

**Figure:** ROC and PR curve on forestcover dataset
Figure: ROC and PR curve on SA dataset
Figure: ROC and PR curve on smtp dataset
Thank you!
Some references:

- N. Goix, A. Sabourin, S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study.
- L. de Haan, A. Ferreira. Extreme value theory, 2006
- FT Liu, Kai Ming Ting, Zhi-Hua Zhou. Isolation forest, 2008
- Y. Qi. Almost sure convergence of the stable tail empirical dependence function in multivariate extreme statistics, 1997
- S. Resnick. Extreme Values, Regular Variation, Point Processes, 1987
- S.J. Roberts. Novelty detection using extreme value statistics, Jun 1999