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

Nicolas Goix

Supervision:

Detecting Anomalies with Multivariate Extremes:
Stéphan Clémençon and Anne Sabourin

Contributions to Scikit-Learn: Alexandre Gramfort

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay

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

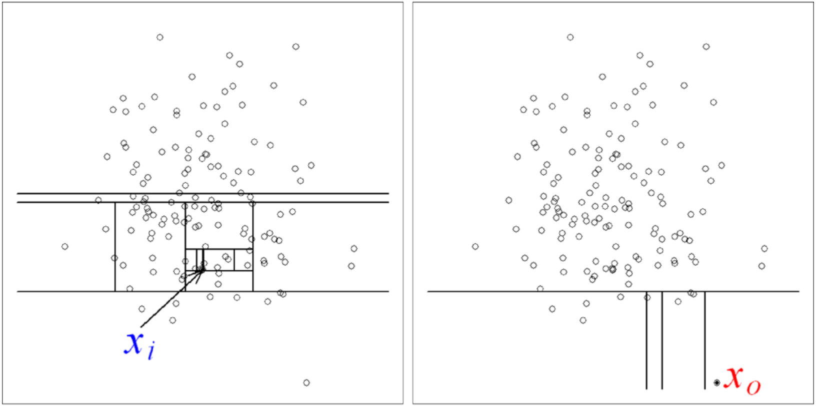
- **Unsupervised** AD (Outlier Detection)
  - no labels, training set = normal + abnormal data
  - Assumption: anomalies are very rare

Important litterature in Anomaly Detection:

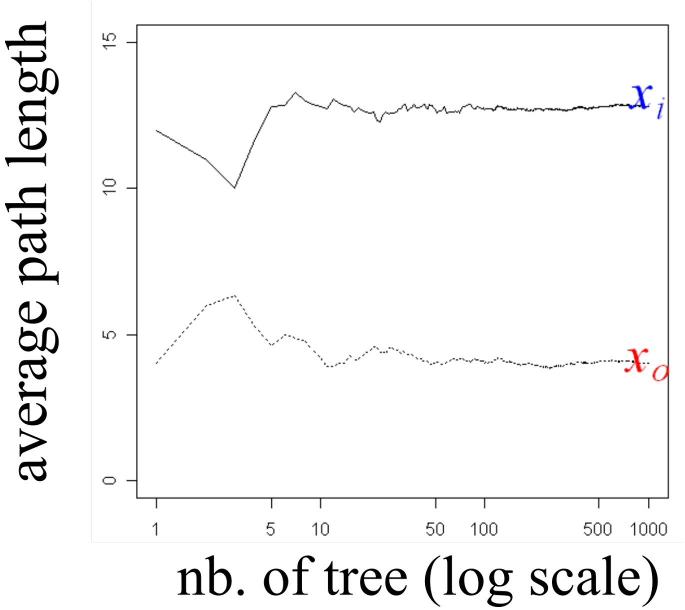
- **statistical AD techniques**
  fit a statistical model for normal behavior
  ex: EllipticEnvelope
- **density-based**
  - ex: Local Outlier Factor (LOF) and variantes (COF ODIN LOCI)
- **Support estimation** - OneClassSVM - MV-set estimate
- **high-dimensional techniques:** - Spectral Techniques - Random Forest - Isolation Forest

Isolation Forest:

Liu Tink Zhou icdm2008



(a) Isolating x_i (b) Isolating x_o


average path length
15
10
5
0
1
5
10
50
100
500
1000
x_i
x_o
nb. of tree (log scale)

IsolationForest.fit(X)

IsolationForest

Inputs: X, n_estimators, max_samples

Output: Forest with:

- # trees = n_estimators
- sub-sampling size = max_samples
- maximal depth $max\_depth = int(\log_2 max\_samples)$

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

```
depth(Tree, X):
    # - Finds the depth level of the leaf node
    #   for each sample x in X.
    # - Add average_path_length(n_samples_in_leaf)
    #   if x not isolated
```

$$score(x, n) = 2^{-\frac{E(depth(x))}{c(n)}}$$

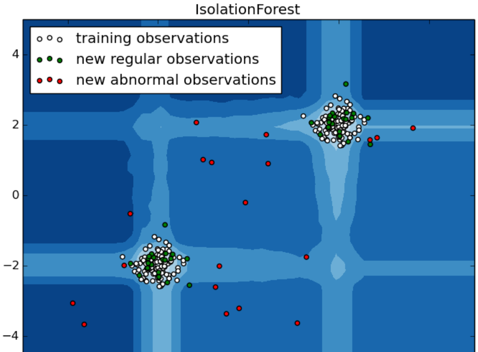
Complexity: O(n_samples n_estimators log(max_samples))

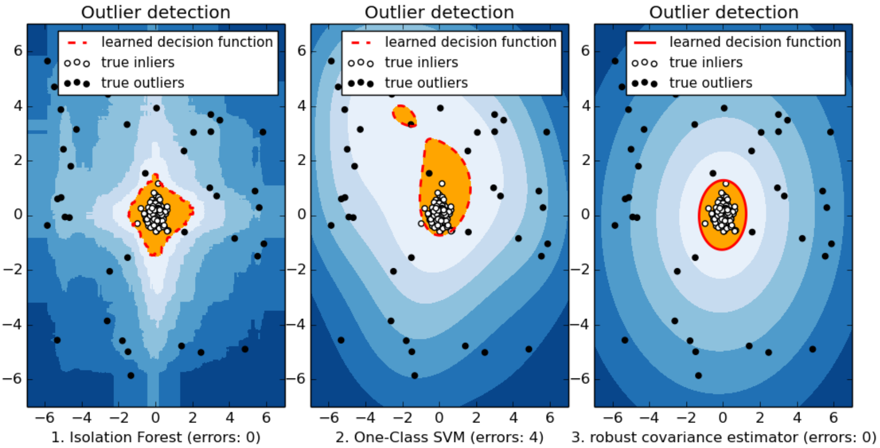
Examples

- code example:

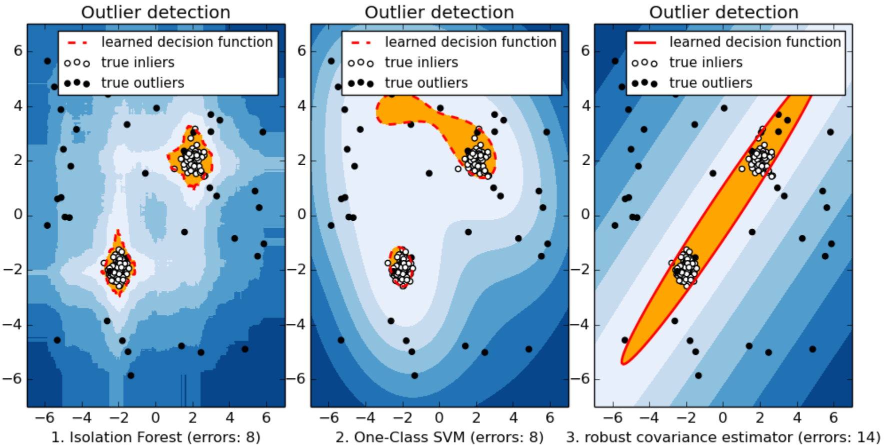
```
>> from sklearn.ensemble import IsolationForest
>> IF = IsolationForest()
>> IF.fit(X_train)   # build the trees
>> IF.predict(X_test)   # find the average depth
```

- plotting decision function:





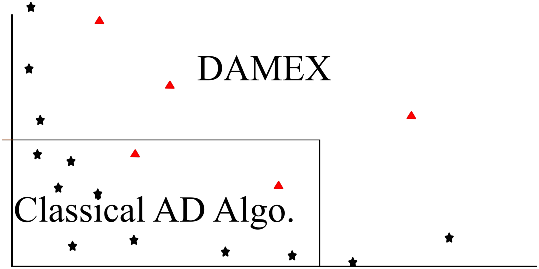

```
n_samples_normal = 150
n_samples_outiers = 50
```




```
n_samples_normal = 150
n_samples_outiers = 50
```

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**.


DAMEX
Classical AD Algo.

Goal:

$$\mathbf{X} = (X_1, \dots, X_d)$$

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

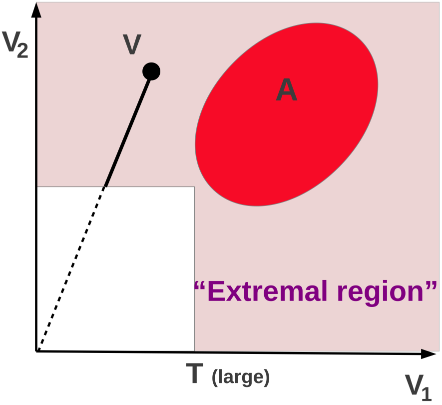
- **Context**
  - Random vector $\mathbf{X} = (X_1, \ldots, X_d)$
  - Margins: $X_j \sim F_j$ (F_j continuous)

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

Problematic

Joint extremes: $\mathbf{V}$'s distribution above large thresholds?

$$\mathbb{P}(\mathbf{V} \in A)? \quad (A \text{ 'far from the origin'}).$$



Fundamental hypothesis and consequences

- Standard assumption: let A extreme region,

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

- Formally,

regular variation (after standardization):

$0 \notin \overline{A}$

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

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

- $\Rightarrow$ **angular measure** on sphere $\mathbf{S}_{d-1}$: $\Phi(B) = \mu\{tB, t \geq 1\}$

General model in multivariate EVT

Model for excesses

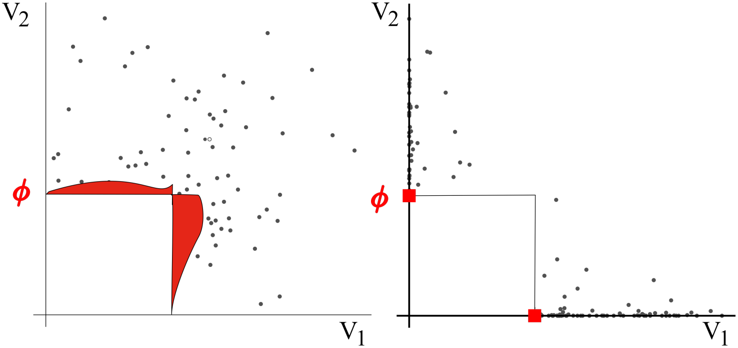
Intuitively: $\mathbb{P}[\mathbf{V} \in A] \;\; \simeq \;\; \mu(A)$ For a large $r > 0$ and a region B on the unit sphere:

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

$\Rightarrow \Phi$ (or μ) **rules the joint distribution of extremes** (if margins are known).

Angular distribution

- Φ rules the joint distribution of extremes

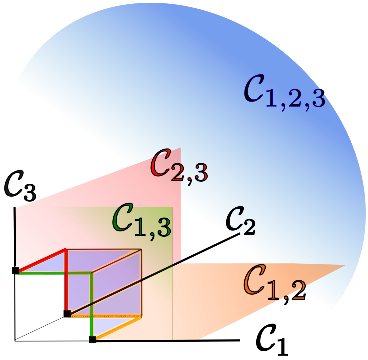


  - Asymptotic dependence: (V_1, V_2) may be large together.

  vs

  - Asymptotic independence: only V_1 *or* V_2 may be large.

General Case



- Sub-cones: $\mathcal{C}_\alpha = \left\{\|v\| \geq 1, \quad v_i > 0 \ (i \in \alpha), \quad v_j = 0 \ (j \notin \alpha)\right\}$
- Corresponding sub-spheres: $\left\{\Omega_\alpha, \alpha \subset \{1, \ldots, d\}\right\}$ $(\Omega_\alpha = \mathcal{C}_\alpha \cap \mathbf{S}_{d-1})$

Representation of extreme data

- Natural decomposition of the angular measure :

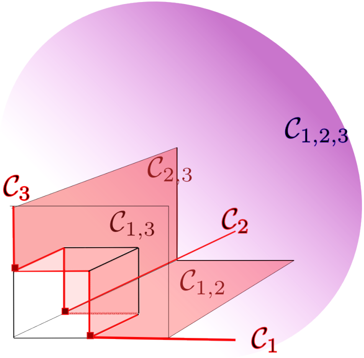
$$\Phi = \sum_{\alpha \subset \{1,\ldots,d\}} \Phi_\alpha \qquad \text{with} \quad \Phi_\alpha = \Phi_{|\Omega_\alpha} \leftrightarrow \mu_{|\mathcal{C}_\alpha}$$

- $\Rightarrow$ yields a representation

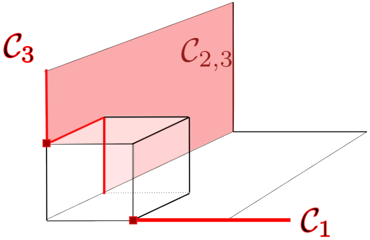
$$\begin{aligned} \mathcal{M} &= \left\{ \Phi(\Omega_\alpha) : \quad \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\} \\ &= \left\{ \mu(\mathcal{C}_\alpha) : \quad \emptyset \neq \alpha \subset \{1, \ldots, d\} \right\} \end{aligned}$$

- Assumption: $\frac{d\mu_{|\mathcal{C}_\alpha}}{dv_\alpha} = O(1)$.
- Remark: Representation $\mathcal{M}$ is linear (after non-linear transform of the data $\mathbf{X} \to \mathbf{V}$).

Sparse Representation ?



Full pattern :
anything may happen



Sparse pattern
(V_1 not large if V_2 or V_3 large)

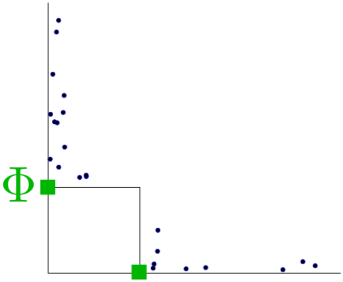
Problem: $\mathcal{M}$ is an **asymptotic** representation

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

is the restriction of an asymptotic measure

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

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

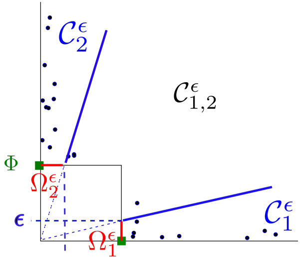


$\Rightarrow$ Cannot just do, for large t:

$$\Phi(\Omega_\alpha) = \mu(\mathcal{C}_\alpha) \simeq t\widehat{\mathbb{P}}(t\mathcal{C}_\alpha)$$

Solution

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



$$\Omega_\alpha \to \Omega_\alpha^\epsilon = \{v \in \mathbf{S}_{d-1} : v_i > \epsilon \, (j \in \alpha), \; v_j \le \epsilon \, (j \notin \alpha)\}.$$

$$\mathcal{C}_\alpha \to \mathcal{C}_\alpha^\epsilon = \{t \, \Omega_\alpha^\epsilon, t \ge 1\}$$

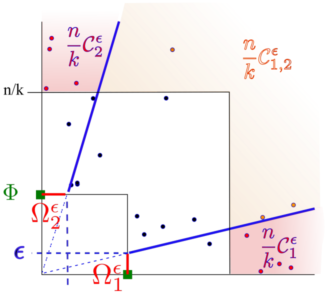
New partition of $\mathbf{S}_{d-1}$, compatible with non asymptotic data.

$$\hat{V}_i^j = \frac{1}{1-\hat{F}_j(X_i^j)} \text{ with } \hat{F}_j(X_i^j) = \frac{rank(X_i^j)-1}{n}$$

$\Rightarrow$ get an natural estimate of $\Phi(\Omega_\alpha)$

$$\widehat{\Phi}(\Omega_\alpha) := \frac{n}{k}\mathbb{P}_n(\hat{V} \in \frac{n}{k}\mathcal{C}_\alpha^\epsilon)$$

$(\frac{n}{k}$ large, ϵ small$)$



$\Rightarrow$ we obtain

$$\widehat{\mathcal{M}} := \{ \ \widehat{\Phi}(\Omega_\alpha), \ \alpha \ \}$$

Theorem

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

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

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: $1/\sqrt{k} \to 1/\sqrt{\epsilon k} + \epsilon$. Price to pay for biasing our estimator with ϵ.

Theorem's proof

1. **Maximal deviation on VC-class:**

$$\sup_{x \succeq \epsilon} |\mu_n - \mu|([x, \infty[) \leq Cd\sqrt{\frac{2}{k} \log \frac{d}{\delta}} + \mathrm{bias}(\epsilon, k, n)$$

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 $\{[\frac{n}{k}x, \infty], x \geq \epsilon\}$

Theorem's proof

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

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

- *A* : First step.
- *B* : density on $\mathcal{C}_\alpha^\epsilon \times$ *Lebesgue* : small

Algorithm

DAMEX in $O(dn\log n)$

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

1. *Standardize via marginal rank-transformation:* $\hat{V}_i := \left(1/(1-\hat{F}_j(X_i^j))\right)_{j=1,\dots,d}$.
2. *Assign to each $\hat{V}_i$ the cone $\frac{n}{k}\mathcal{C}_\alpha^\epsilon$ it belongs to.*
3. $\Phi_n^{\alpha,\epsilon} := \widehat{\Phi}(\Omega_\alpha) = \frac{n}{k}\mathbb{P}_n(\hat{V} \in \frac{n}{k}\mathcal{C}_\alpha^\epsilon)$ *the estimate of the α-mass of Φ.*

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

$$\widehat{\mathcal{M}} := (\Phi_n^{\alpha,\epsilon})_{\alpha\subset\{1,\dots,d\},\Phi_n^{\alpha,\epsilon}>\Phi_{\min}}$$

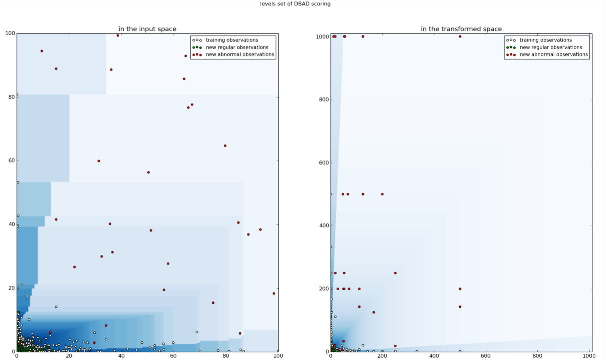
Application to Anomaly Detection

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

$\longrightarrow$ scoring function = $\Phi_n^{\epsilon} \times 1/r$:

$$s_n(\mathbf{x}) := (1/\|\hat{T}(\mathbf{x})\|_\infty) \sum_\alpha \Phi_n^{\alpha,\epsilon} \mathbb{1}_{\hat{T}(\mathbf{x}) \in \mathcal{C}_\alpha^\epsilon}.$$

where $T : \mathbf{X} \mapsto \mathbf{V}$ $\qquad$ $(V_j = \frac{1}{1 - F_j(X_j)})$



	number of samples	number of features
shuttle	85849	9
forestcover	286048	54
SA	976158	41
SF	699691	4
http	619052	3
smtp	95373	3

Table: Datasets characteristics

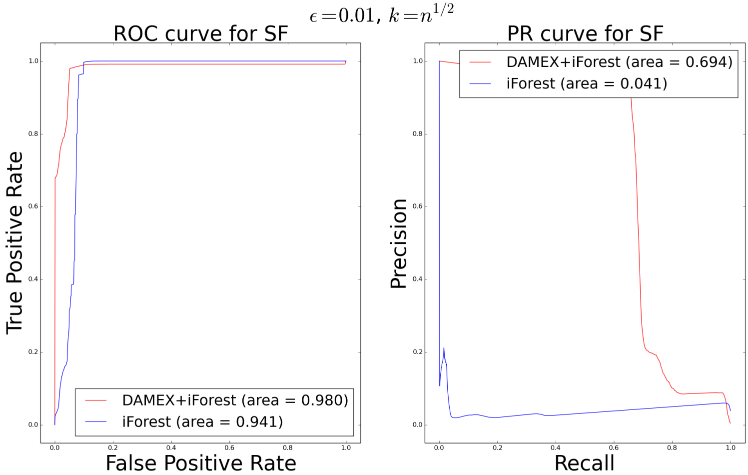


Figure: ROC and PR curve on SF dataset

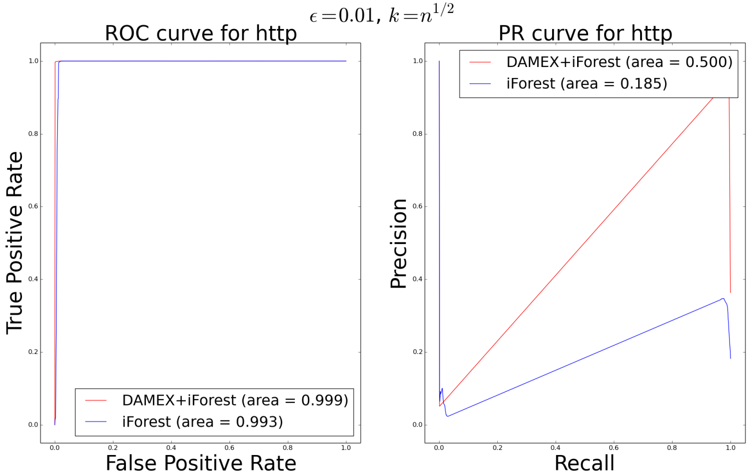


Figure: ROC and PR curve on http dataset

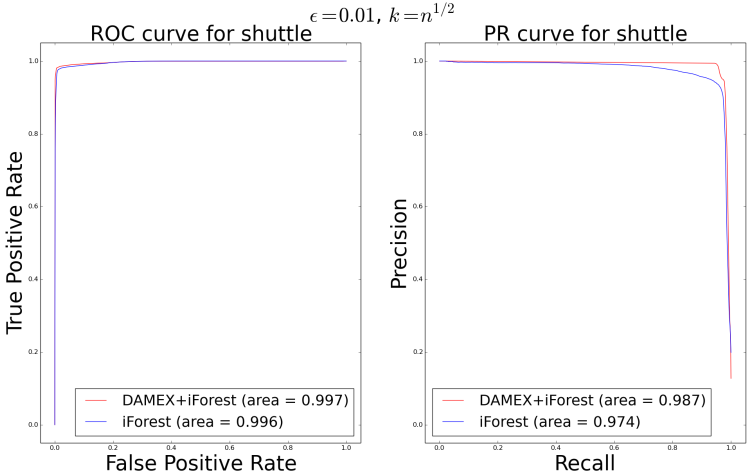


Figure: ROC and PR curve on shuttle dataset

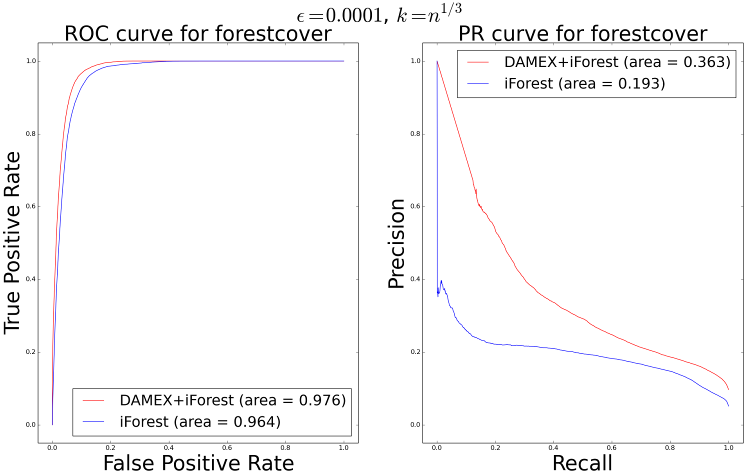


Figure: ROC and PR curve on forestcover dataset

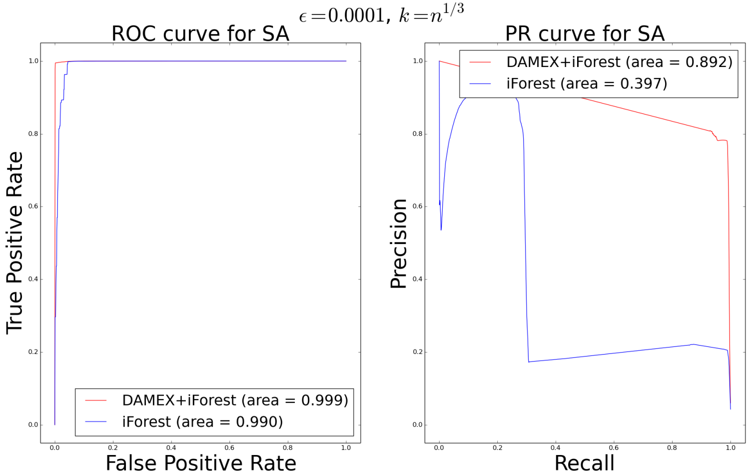


Figure: ROC and PR curve on SA dataset

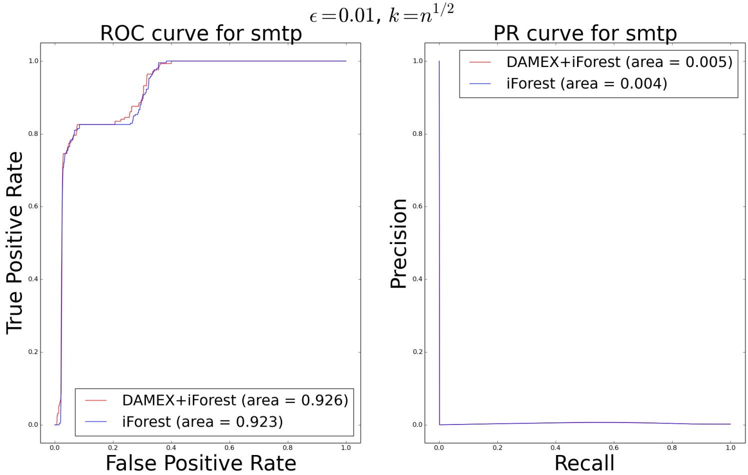


Figure: ROC and PR curve on smtp dataset

Thank you !

- Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey, 2009
- J. H. J. Einmahl , J. Segers Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution
- J. H. J. Einmahl, Andrea Krajina, J. Segers. An m-estimator for tail dependence in arbitrary dimensions, 2012.
- 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
- J. Segers. Max-stable models for multivariate extremes, 2012