

How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms?

Nicolas Goix

LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay, France

ICML Workshop on Anomaly Detection, June 2016

Motivations

Most of the time, data come without any label.

- ▶ Labeled data: we use ROC or PR curves
- ▶ No labels available: ?
 - Our particular setting: **anomaly detection**.
 - Idea: In this field, we like estimating **level sets**.

How good is an anomaly detection algorithm?



How good is it estimating the level sets?

One-Class Classification, Novelty Detection

- ▶ **Data:** i.i.d. observations in \mathbb{R}^d from the normal behavior, density f .
In practice, data can be polluted by a small proportion of anomalies.
- ▶ **Output to evaluate: scoring function.**
 - AD algorithms return a scoring function $s : \mathbb{R}^d \rightarrow \mathbb{R}$
 - s defined a pre-order on \mathbb{R}^d = 'degree of abnormality'
 - s level sets are estimates of f level sets
 - s can be interpreted as a **box which contains an infinite number of level sets estimates** (at different levels)

Remark: Perfect scoring functions: f or any increasing transform of f .

Problem reformulation

How can we know if the level sets of s are close to those of f ?

Performance criterion for a scoring function.

- ▶ **Fact:** For any strictly increasing transform T , level sets of $T \circ f$ are exactly those of f .
 \Rightarrow Criterion $\mathcal{C}(s) = \|s - f\|$ does'nt work! ($s = 2f$ is perfect)
- ▶ **We would like**
 - $\mathcal{C}^\Phi(s) = \|\Phi(s) - \Phi(f)\|$ with Φ s.t. $\Phi(T \circ s) = \Phi(s)$.
 - $\{\text{level sets of optimal } s^*\} = \{\text{level sets of } f\}$.
 - $\mathcal{C}^\Phi(s) =$ 'distance' between level sets of s and those of f .

$\Rightarrow \Phi(s) := MV_s$ or EM_s , the Mass-Volume and Excess-Mass curves of s .

Criteria satisfying these requirements: MV and EM

Mass-volume and excess-mass curves

► Definitions:

$$MV_s(\alpha) = \inf_{u \geq 0} \text{Leb}(s \geq u) \quad \text{s.t.} \quad \mathbb{P}(s(\mathbf{X}) \geq u) \geq \alpha$$

$$EM_s(t) = \sup_{u \geq 0} \{ \mathbb{P}(s(\mathbf{X}) \geq u) - t \text{Leb}(s \geq u) \}$$

► Optimal curves:

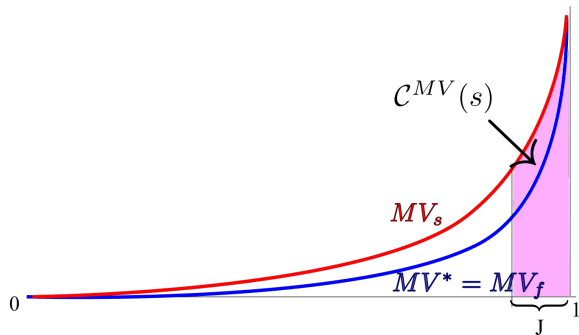
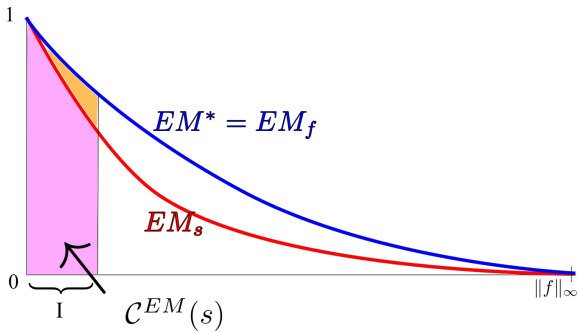
$$MV^*(\alpha) = \min_{\Omega \text{ borelian}} \text{Leb}(\Omega) \quad \text{s.t.} \quad \mathbb{P}(\mathbf{X} \in \Omega) \geq \alpha \quad = MV_f = MV_{T \circ f}$$

$$EM^*(t) = \max_{\Omega \text{ borelian}} \{ \mathbb{P}(\mathbf{X} \in \Omega) - t \text{Leb}(\Omega) \} \quad = EM_f = EM_{T \circ f}$$

► Interpretation: $(EM_s - EM_f)(t) \simeq \inf_{u > 0} \text{Leb}(\{s > u\} \Delta \{f > t\})$

$\|EM_s - EM_f\|_{L^1(I)}$: distance between t -level sets of s and f , $t \in I$.

$\|MV_s - MV_f\|_{L^1(J)}$: distance between α -level sets of s and f , $\alpha \in J$.



Estimation and Issues

► **Estimation:**

$$\widehat{MV}_s(\alpha) = \inf_{u \geq 0} \text{Leb}(s \geq u) \quad \text{s.t.} \quad \mathbb{P}_n(s \geq u) \geq \alpha$$

$$\widehat{EM}_s(t) = \sup_{u \geq 0} \mathbb{P}_n(s \geq u) - t \text{Leb}(s \geq u)$$

► **Empirical criteria:**

$$\widehat{C}^{EM}(s) = \|\widehat{EM}_s\|_{L^1(I)} \quad I = [0, \widehat{EM}^{-1}(0.9)],$$

$$\widehat{C}^{MV}(s) = \|\widehat{MV}_s\|_{L^1(J)} \quad J = [0.9, 1],$$

- **Issues:** The volume $\text{Leb}(s \geq u)$ has to be estimated (Monte-Carlo). Challenging in large dimensions.

Feature sub-sampling and Aggregating

Inputs: AD algorithm \mathcal{A} , data set X size $n \times d$, feature sub-sampling size d' , number of draws m .

for $k = 1, \dots, m$ **do**

-randomly select a sub-group F_k of d' features

-compute the associated scoring function $s_k = \mathcal{A}((x_i^j)_{1 \leq i \leq n, j \in F_k})$

-compute $\widehat{C}_k^{EM} = \|\widehat{EM}_{s_k}\|_{L^1(I)}$ or $\widehat{C}_k^{MV} = \|\widehat{MV}_{s_k}\|_{L^1(J)}$

end for

Return performance criteria:

$$\widehat{C}_{high_dim}^{EM}(\mathcal{A}) = \frac{1}{m} \sum_{k=1}^m \widehat{C}_k^{EM} \quad \text{or} \quad \widehat{C}_{high_dim}^{MV}(\mathcal{A}) = \frac{1}{m} \sum_{k=1}^m \widehat{C}_k^{MV}.$$

Benchmarks

Does performance in term of EM/MV correspond to performance in term of ROC/PR?

- ▶ **Experiments:** 12 datasets, 3 AD algorithms (LOF, OCSVM, iForest) \rightarrow 36 possible pairwise comparisons:

$$\left\{ \left(A_1 \text{ on } \mathcal{D}, A_2 \text{ on } \mathcal{D} \right), A_1, A_2 \in \{\text{iForest, LOF, OCSVM}\}, \right. \\ \left. \mathcal{D} \in \{\text{adult, http, } \dots, \text{spambase}\} \right\}.$$

- ▶ **Results:** If we only consider the pairs *s.t.* ROC and PR agree on which algorithm is the best, we are able (with EM and MV scores) to recover it in 80% of the cases.

Thank you!

Table: Original Datasets characteristics

	nb of samples	nb of features	anomaly class	
adult	48842	6	class '> 50K'	(23.9%)
http	567498	3	attack	(0.39%)
pima	768	8	pos (class 1)	(34.9%)
smtp	95156	3	attack	(0.03%)
wilt	4839	5	class 'w' (diseased trees)	(5.39%)
annthyroid	7200	6	classes $\neq 3$	(7.42%)
arrhythmia	452	164	classes $\neq 1$ (features 10-14 removed)	(45.8%)
forestcover	286048	10	class 4 (vs. class 2)	(0.96%)
ionosphere	351	32	bad	(35.9%)
pendigits	10992	16	class 4	(10.4%)
shuttle	85849	9	classes $\neq 1$ (class 4 removed)	(7.17%)
spambase	4601	57	spam	(39.4%)

Table: Results for the novelty detection setting. One can see that ROC, PR, EM, MV often do agree on which algorithm is the best (in bold), which algorithm is the worse (underlined) on some fixed datasets. When they do not agree, it is often because ROC and PR themselves do not, meaning that the ranking is not clear.

Dataset	iForest				OCSVM				LOF			
	ROC	PR	EM	MV	ROC	PR	EM	MV	ROC	PR	EM	MV
adult	0.661	0.277	1.0e-04	7.5e01	0.642	0.206	2.9e-05	4.3e02	<u>0.618</u>	<u>0.187</u>	<u>1.7e-05</u>	<u>9.0e02</u>
http	0.994	0.192	1.3e-03	9.0	0.999	0.970	6.0e-03	2.6	<u>0.946</u>	<u>0.035</u>	<u>8.0e-05</u>	<u>3.9e02</u>
pima	0.727	0.182	5.0e-07	1.2e04	0.760	0.229	5.2e-07	<u>1.3e04</u>	<u>0.705</u>	<u>0.155</u>	<u>3.2e-07</u>	2.1e04
smtp	0.907	<u>0.005</u>	<u>1.8e-04</u>	<u>9.4e01</u>	<u>0.852</u>	0.522	1.2e-03	8.2	0.922	0.189	1.1e-03	5.8
wilt	0.491	0.045	4.7e-05	<u>2.1e03</u>	<u>0.325</u>	<u>0.037</u>	5.9e-05	4.5e02	0.698	0.088	<u>2.1e-05</u>	1.6e03
annthyroid	0.913	0.456	2.0e-04	2.6e02	<u>0.699</u>	<u>0.237</u>	<u>6.3e-05</u>	2.2e02	0.823	0.432	6.3e-05	<u>1.5e03</u>
arrhythmia	0.763	0.487	1.6e-04	9.4e01	0.736	0.449	1.1e-04	1.0e02	<u>0.730</u>	<u>0.413</u>	<u>8.3e-05</u>	<u>1.6e02</u>
forestcov.	<u>0.863</u>	<u>0.046</u>	<u>3.9e-05</u>	<u>2.0e02</u>	0.958	0.110	5.2e-05	1.2e02	0.990	0.792	3.5e-04	3.9e01
ionosphere	<u>0.902</u>	<u>0.529</u>	<u>9.6e-05</u>	<u>7.5e01</u>	0.977	0.898	1.3e-04	5.4e01	0.971	0.895	1.0e-04	7.0e01
pendigits	0.811	0.197	2.8e-04	2.6e01	<u>0.606</u>	<u>0.112</u>	<u>2.7e-04</u>	<u>2.7e01</u>	0.983	0.829	4.6e-04	1.7e01
shuttle	0.996	0.973	1.8e-05	5.7e03	<u>0.992</u>	<u>0.924</u>	3.2e-05	2.0e01	0.999	0.994	<u>7.9e-06</u>	<u>2.0e06</u>
spambase	0.824	0.371	9.5e-04	4.5e01	<u>0.729</u>	0.230	4.9e-04	1.1e03	0.754	<u>0.173</u>	<u>2.2e-04</u>	<u>4.1e04</u>