The Mass Volume curve, a performance metric for unsupervised anomaly detection

Rare Events, Extremes and Machine Learning Workshop
May 24th, 2018

Albert Thomas
Huawei Technologies - Télécom ParisTech - Airbus Group Innovations

Joint work with Stephan Clémençon, Alexandre Gramfort, Vincent Feuillard and Anne Sabourin.
Outline

1. Unsupervised anomaly detection
2. The Mass Volume curve
3. Anomaly detection in extreme regions
Unsupervised anomaly detection

1. Find anomalies in an unlabeled data set \( X_1, \ldots, X_n \)

2. Anomalies are assumed to be rare events
Unsupervised anomaly detection

\[ X_1, \ldots, X_n \in \mathbb{R}^d \] unlabeled data set \( \sim P \)

- density \( f \) w.r.t. Lebesgue measure \( \lambda \)
- anomalies = rare events

Estimate a **density level set**
Minimum Volume set (Polonik, 1997)

A density level set is a Minimum Volume set, i.e. a solution of

$$\min_{\Omega \in \mathcal{B}(\mathbb{R}^d)} \lambda(\Omega) \text{ such that } P(\Omega) \geq \alpha$$
Minimum volume set $\approx$ density level set

1. A density level set is always a minimum volume set

2. If $f$ has no flat parts, a minimum volume set is a density level set.  
   (Einmahl and Mason, 1992), (Polonik, 1997), (Nunez-Garcia et al., 2003)
Unsupervised anomaly detection algorithms

Common approach for most algorithms

1. Learn a scoring function

\[ \hat{s} : \mathbb{R}^d \rightarrow \mathbb{R} \]

such that the smaller \( \hat{s}(x) \) the more abnormal is \( x \).

2. Threshold \( \hat{s} \) at an offset \( q \) such that \( \hat{\Omega}_\alpha = \{ x, \hat{s}(x) \geq q \} \) is an estimation of the Minimum Volume set with mass \( \alpha \).

→ density estimation (Cadre et al., 2013), One-Class SVM/Support Vector Data Description (Schölkopf et al., 2001), (Vert and Vert, 2006), (Tax et al., 2004), k-NN (Sricharan & Hero, 2011)

→ Isolation Forest (Liu et al., 2008) and Local Outlier Factor (Breunig et al., 2000)
Ideal scoring functions

Ideal scoring functions $s$ preserve the order induced by density $f$

$$s(x_1) < s(x_2) \iff f(x_1) < f(x_2)$$

i.e. strictly increasing transform of $f$.

$s$ does not need to be close to $f$ in the sense of a $L^p$ norm.
Scoring function of the One-Class SVM

Asymptotically constant near the modes and proportional to the density in the low density regions (Vert and Vert, 2006)
Problem

One-Class SVM with Gaussian kernel $k_{\sigma}$

- The user needs to choose the bandwidth $\sigma$ of the kernel

$\sigma = 0.5$

$\sigma = 10$

Overfitting

Underfitting

How to automatically choose $\sigma$?
Problem

Given a data set $S_n = (X_1, \ldots, X_n)$, a hyperparameter space $\Theta$ and an unsupervised anomaly detection algorithm

$$A : S_n \times \Theta \rightarrow \mathbb{R}^{R^d}$$

$$(S_n, \theta) \mapsto \hat{s}_\theta$$

How to assess the performance of $\hat{s}_\theta$ without a labeled data set?

(Anomaly Detection Workshop, Thomas, Clémençon, Feuillard, Gramfort, ICML 2016)
Mass Volume curve

\[ X \sim P, \text{ scoring function } s : \mathbb{R}^d \rightarrow \mathbb{R}, \]

\[ \text{t-level set of } s : \{x, s(x) \geq t\} \]

- \[ \alpha_s(t) = \mathbb{P}(s(X) \geq t) \text{ mass of the t-level set} \]
- \[ \lambda_s(t) = \lambda(\{x, s(x) \geq t\}) \text{ volume of } t\text{-the level set.} \]
Anomaly detection

Mass Volume curve

Mass Volume curve $MV_s$ of a scoring function $s$
(Clémençon and Jakubowicz, 2013), (Clémençon and Thomas, 2017)

$t \in \mathbb{R} \mapsto (\alpha_s(t), \lambda_s(t))$
Mass Volume curve

Easier to work with the following definition: \( MV_s \) is defined as the plot of the function

\[
MV_s : \alpha \in (0, 1) \mapsto \lambda_s(\alpha_s^{-1}(\alpha)) = \lambda(\{x, s(x) \geq \alpha_s^{-1}(\alpha)\})
\]

where \( \alpha_s^{-1} \) generalized inverse of \( \alpha_s \).

Property

(Clémentençon and Jakubowicz, 2013), (Clémentençon and Thomas, 2017)

Assume that the underlying density \( f \) has no flat parts, then for all scoring functions \( s \),

\[
\forall \alpha \in (0, 1), \quad MV^*(\alpha) \overset{\text{def}}{=} MV_f(\alpha) \leq MV_s(\alpha)
\]

The lower is \( MV_s \) the better is \( s \)
Learning hyperparameters

Consider $\hat{\theta} = \mathcal{A}(S_n, \theta)$

Choose $\theta^*$ minimizing area under $\text{MV}_{\hat{\theta}}$

As $\text{MV}_{\hat{\theta}}$ depends on $P$, use empirical MV curve estimated on a test set

$$\hat{\text{MV}}_S : \alpha \in [0, 1) \mapsto \lambda_S(\hat{\alpha}_s^{-1}(\alpha))$$

where $\hat{\alpha}_s(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x, s(x) \geq t\}}(X_i).$
Algorithm

Choose $\theta^\dagger$ minimizing area under Mass Volume curve $AMV_{\hat{s}_0}$.
For each random split we may obtain a different $\theta^\dagger$

→ to reduce the variance of the estimator consider $B$ random splits of the data set

- for each random split $b$ we get $\theta^\dagger_b$ and $\hat{s}_{\theta^\dagger_b}^b$
- final scoring function

$$\hat{S} = \frac{1}{B} \sum_{b=1}^{B} \hat{s}_{\theta^\dagger_b}^b$$
Toy example

Minimum Volume set estimation with One-Class SVM and $B = 50$ ($\alpha = 0.95$)
Given an anomaly detection algorithm \( \mathcal{A} \), compare

- our approach \( \rightarrow s_{\text{tuned}} \)
- a priori fixed hyperparameters \( \rightarrow s_{\text{fixed}} \)

Performance criterion: relative gain

\[
G_{\mathcal{A}}(s_{\text{tuned}}, s_{\text{fixed}}) = \frac{\text{AMV}_{I}(s_{\text{fixed}}) - \text{AMV}_{I}(s_{\text{tuned}})}{\text{AMV}_{I}(s_{\text{fixed}})}
\]

where \( \text{AMV}_{I} \) is area under MV curve over interval \( I = [0.9, 0.99] \), computed on left out data.

If \( G_{\mathcal{A}}(s_{\text{tuned}}, s_{\text{fixed}}) > 0 \) then \( s_{\text{tuned}} \) better than \( s_{\text{fixed}} \)
Results

For $s_{\text{tuned}}$ we consider 50 random splits (80/20)

KLPE (Sricharan & Hero, 2011) — aKLPE: Average KLPE (Qian & Saligrama, 2012) — OCSVM: One-Class SVM (Schölkopf et al., 2001) — iForest: Isolation Forest (Liu et al., 2008) — KS: Kernel Smoothing
Consistency of $\hat{MV}_S$

We used $\hat{MV}_S$ as an estimate of $MV_S$ where

$$\forall \alpha \in [0, 1), \quad \hat{MV}_S(\alpha) = \lambda \{ x, s(x) \geq \hat{\alpha}_S^{-1}(\alpha) \}$$

with $\hat{\alpha}_S^{-1}$ the generalized inverse of $\hat{\alpha}_S$.

2 questions

- Consistency of $\hat{MV}_S$ as $n \to \infty$?
- How to build confidence regions?
Consistency

Let \( s \) be a scoring function and \( \varepsilon \in (0, 1] \).

- Assumptions on the random variable \( s(X) \) and its distribution \( F_s \)
- \( \lambda_s \) is \( C^2 \)

**Theorem (Clémenton and Thomas, 2017)**

(i) **Consistency.** With probability one,

\[
\sup_{\alpha \in [0, 1-\varepsilon]} \left| \hat{MV}_s(\alpha) - MV_s(\alpha) \right| \xrightarrow{n \to +\infty} 0
\]

(ii) **Functional CLT.** There exists a sequence of Brownian bridges \( \{B_n(\alpha)\}_{\alpha \in [0,1]} \) such that, almost-surely, uniformly over \([0, 1 - \varepsilon]\), as \( n \to \infty \),

\[
\sqrt{n} \left( \hat{MV}_s(\alpha) - MV_s(\alpha) \right) = \frac{\lambda'_s(\alpha_s^{-1}(\alpha))}{f_s(\alpha_s^{-1}(\alpha))} B_n(\alpha) + O(n^{-1/2} \log n)
\]
Confidence bands using smoothed bootstrap

$\hat{MV}_f$

90% confidence band

Up to log $n$ factors (Clémenton and Thomas, 2017),

- **Functional CLT**: rate in $O(n^{-1/2})$ but requires knowledge of $f_s$
- **Naive (non-smoothed) bootstrap**: rate in $O(n^{-1/4})$
- **Smoothed bootstrap**: rate in $O(n^{-2/5})$
Anomaly detection in extreme regions

Goal of multivariate Extreme Value Theory: model the tail of the distribution $F$ of a multivariate random variable

$$X = (X^{(1)}, \ldots, X^{(d)})$$

Motivation for unsupervised anomaly detection

- Anomalies are likely to be located in extreme regions, i.e. regions far from the mean $\mathbb{E}[X]$
- Lack of data in these regions makes it difficult to distinguish between large normal instances and anomalies

Relying on multivariate Extreme Value Theory we suggest an algorithm to detect anomalies in extreme regions.
Multivariate Extreme Value Theory

Common approach

- Standardization to unit Pareto: \( T(X) = V \in \mathbb{R}^d \) where

\[
V(j) = \frac{1}{1 - F_j(X(j))}, \quad \forall j \in \{1, \ldots, d\},
\]

\( F_j, 1 \leq j \leq d, \) being the margins. Note that \( X \) and \( V \) share the same dependence structure/copula.
Multivariate Extreme Value Theory

- \((r(v), \varphi(v)) = (\|v\|_\infty, v/\|v\|_\infty)\) polar coordinates
- \(S_{d-1}\) positive orthant of the unit hypercube

**Theorem (Resnick, 1987)**

With mild assumptions on the distribution \(F\), there exists a finite (angular) measure \(\Phi\) on \(S_{d-1}\) such that for all \(\Omega \subset S_{d-1}\),

\[
\Phi_t(\Omega) \overset{def}{=} t \cdot \mathbb{P}(r(V) > t, \varphi(V) \in \Omega) \xrightarrow{t \to \infty} \Phi(\Omega)
\]
Main idea

**Anomalies in extreme regions**: observations that deviate from the dependence structure of the tail

To find the most likely directions of the extreme observations we estimate a Minimum Volume set of the angular measure $\Phi$. 

Dependence

Independence
Minimum volume set estimation on the sphere

Solve empirical optimization problem (Scott and Nowak, 2006)

$$\min_{\Omega \in \mathcal{G}} \lambda_d(\Omega) \text{ subject to } \hat{\Phi}_{n,k}(\Omega) \geq \alpha - \psi_k(\delta)$$

where $\hat{\Phi}_{n,k}$ is estimated from $t \cdot \mathbb{P}(r(\mathbf{V}) > t, \varphi(\mathbf{V}) \in \Omega)$ with $t = n/k$:

$$\hat{\Phi}_{n,k}(\Omega) = \frac{1}{k} \sum_{i=1}^{n} \mathbb{1}\{r(\mathbf{V}_i) \geq \frac{n}{k}, \varphi(\mathbf{V}_i) \in \Omega\}$$
Theorem

$\Omega^*_\alpha$ being the true Minimum Volume set,

$$R(\Omega) = (\lambda_d(\Omega) - \lambda_d(\Omega^*_\alpha))_+ + (\alpha - \Phi_{n/k}(\Omega))_+$$

Theorem (Thomas, Clémençon, Gramfort, Sabourin, 2017)

Assume that

- the margins $F_j$ are known
- the class $\mathcal{G}$ is of finite VC dimension
- common assumptions for existence and uniqueness of MV sets are fulfilled by $\Phi_{n/k}$

Then there exists a constant $C > 0$ such that

$$\mathbb{E}[R(\hat{\Omega}_\alpha)] \leq \left( \inf_{\Omega \in \mathcal{G}_\alpha} \lambda_d(\Omega) - \lambda_d(\Omega^*_\alpha) \right) + C \sqrt{\frac{\log k}{k}}$$

where $\mathcal{G}_\alpha = \{\Omega \in \mathcal{G}, \Phi_{n/k}(\Omega) \geq \alpha\}$. 
Numerical experiments

Global scoring function

\[ \hat{s}(r(v), \varphi(v)) = \frac{1}{r(v)^2} \cdot \hat{s}_{\varphi}(\varphi(v)). \]

<table>
<thead>
<tr>
<th>Data set</th>
<th>OCSVM</th>
<th>Isolation Forest</th>
<th>Score ( \hat{s} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>shuttle</td>
<td>0.981</td>
<td>0.963</td>
<td>0.987</td>
</tr>
<tr>
<td>SF</td>
<td>0.478</td>
<td>0.251</td>
<td>0.660</td>
</tr>
<tr>
<td>http</td>
<td>0.997</td>
<td>0.662</td>
<td>0.964</td>
</tr>
<tr>
<td>ann</td>
<td>0.372</td>
<td>0.610</td>
<td>0.518</td>
</tr>
<tr>
<td>forestcover</td>
<td>0.540</td>
<td>0.516</td>
<td>0.646</td>
</tr>
</tbody>
</table>

ROC-AUC are computed on test sets made of normal and abnormal instances and restricted to the extreme region \( (k = \sqrt{n}) \).
References

- Anomaly detection in extreme regions via empirical MV sets on the sphere. A. Thomas, S. Clémençon, A. Gramfort, A. Sabourin. AISTATS 2017.

Code for hyperparameter tuning available on github https://github.com/albertctthomas/anomaly_tuning
Thank you