Extremes (2020) 23:501–520
https://doi.org/10.1007/s10687-020-00393-0

Extreme value theory for anomaly detection – the GPD classifier

Edoardo Vignotto[1] · **Sebastian Engelke**[1]

Received: 22 July 2019 / Revised: 10 August 2020 / Accepted: 12 August 2020 /
Published online: 24 September 2020
© The Author(s) 2020

Abstract
Classification tasks usually assume that all possible classes are present during the training phase. This is restrictive if the algorithm is used over a long time and possibly encounters samples from unknown new classes. It is therefore fundamental to develop algorithms able to distinguish between normal and abnormal test data. In the last few years, extreme value theory has become an important tool in multivariate statistics and machine learning. The recently introduced extreme value machine, a classifier motivated by extreme value theory, addresses this problem and achieves competitive performance in specific cases. We show that this algorithm has some theoretical and practical drawbacks and can fail even if the recognition task is fairly simple. To overcome these limitations, we propose two new algorithms for anomaly detection relying on approximations from extreme value theory that are more robust in such cases. We exploit the intuition that test points that are extremely far from the training classes are more likely to be abnormal objects. We derive asymptotic results motivated by univariate extreme value theory that make this intuition precise. We show the effectiveness of our classifiers in simulations and on real data sets.

Keywords Clustering · Novelty detection · Machine learning · Open set classification · Statistical methods

1 Introduction

Modern classifiers achieve human or super-human performance in a variety of tasks (Christopher 2016), including speech (Graves et al. 2013) and image recognition (He et al. 2016), but they are typically not able to discriminate between normal and abnormal classes and may give high confidence predictions for unrecognizable objects

✉ Edoardo Vignotto
edoardo.vignotto@unige.ch

[1] Research Center for Statistics, University of Geneva, Geneva, Switzerland

(Nguyen et al. 2015). Here and throughout, we call a class normal if we have examples of it during the training phase, whereas we call a class abnormal if we have no examples of it during the training phase. The ability to distinguish between these two cases is important if there is the possibility that new classes arise in the future or if there have not been any examples of some classes in the training set due to their rarity.

The task of distinguishing between normal and abnormal test data points is called anomaly detection (Chandola et al. 2009). We underline that in this context standard hyper parameter optimization procedures such as cross-validation are usually not available, since in the training set there are only normal objects. For this reason, an algorithm designed for anomaly detection should involve as few hyper parameters as possible.

The extreme value machine (EVM) introduced in Rudd et al. (2018) is an algorithm that uses extreme value theory to attack these problems. In the last few years, extreme value theory has become an important tool in multivariate statistics and machine learning. This is due to the fact that the extreme features, rather than the average ones, are the most important for discriminating between different objects (Scheirer 2017). The EVM strongly relies on the distances between the different classes in the training set. In particular, as we show in Section 8 using a simulated data set, this is the reason why the EVM fails to recognize abnormal points that arise relatively close to one class in the training set, even if the recognition task is fairly simple. For this reason, we propose an alternative approach that uses extreme value theory overcomes this problem. We call our method the GPD classifier (GPDC), according to the generalized Pareto distribution approximation from extreme value theory that it relies on. Moreover, we present also a second more heuristic algorithm, namely the GEV classifier (GEVC), based on the generalized extreme value distribution. We underline that, in the following, we consider only the Euclidean distance for simplicity, but the obtained results are more general and can be applied to any distance.

This paper is organized as follows. In Section 2 we summarize the previous work in this context. Section 3 states some fundamental results from extreme value theory that will be useful in the definition of the GEVC and GPDC. In Section 4 we describe the general framework of anomaly detection. In Section 5 we describe the EVM and explain the main limitations of this approach. In Sections 6 and 7 we describe the GPDC and GEVC and we theoretically justify their main properties. Finally, in Section 8, we evaluate and compare the performance of these anomaly detection algorithms on both simulated and real data.

2 Related work

Anomaly detection (Pimentel et al. 2014) is a well established field that shares similar though not identical goals with outlier detection (Walfish 2006). While the latter is focused on discovering which examples of the training data are not in agreement with the process that has generated the bulk of them, in anomaly detection we suppose to have observations from a number of normal classes and for a new sample

we would like to determine whether it belongs to one of those existing classes or a new one. In this sense, anomaly detection is more naturally related to classification. Machine learning techniques have been applied successfully to both tasks, improving considerably their performance (Abe et al. 2006; Shon and Moon 2007; Désir et al. 2013). Many anomaly detection methods first estimate the distribution of the normal data and then mark as abnormal the new points associated with low density regions (Bishop 1994). We follow a similar approach for the two techniques that we propose. A relevant work in this direction is Roberts (1999) who fit a Gaussian mixture model to the training data and use extreme value theory to decide if a new point is normal or abnormal. Our methods will not rely on parametric assumptions for the multivariate density, which makes them more widely applicable.

Extreme value theory has been recognized in the last years as a powerful tool to increase the performance of anomaly detection (Geng et al. 2018) and, more generally, machine learning techniques (Jalalzai et al. 2018) with particular interest to computer vision (Scheirer 2017). The main reason for this success is the fact that the tails of the distribution of distances between training observations can be modeled effectively using the asymptotic theory provided by extreme value theory (Scheirer et al. 2011). One example of this can be found in Fragoso et al. (2013) where the well-known RANSAC algorithm is improved using extreme value theory. Moreover, extreme value theory has been applied to extend the use of classical anomaly detection algorithms to extreme regions (Goix et al. 2016; Thomas et al. 2017) and to detect anomalies in time series (Siffer et al. 2017).

A feature that could be important for anomaly detection algorithms is the ability to be easily updated with the arrival of new training data, namely incremental learning. A simple approach in this direction is to use the k-nearest neighbor classifier (James et al. 2013), a route that is to some extent also followed by the GPDC and the GEVC. Another method that has the desired requirement to be naturally adapted to new data is the nearest class mean classifier (Mensink et al. 2012), which represents each class as a prototype vector that is the mean of all the examples for that class seen so far. More complex methods are based on machine learning approaches such as support vector machines (Ruping 2001), neural networks (Rebuffi et al. 2017) or random forests (Saffari et al. 2009) and are in general computationally more demanding.

Finally, we note that the formulation of GPDC is partially related to the estimation of the right end point of a univariate distribution. As we will argue in the following, in the GPDC this problem takes a simpler form than in the general problem of upper end point estimation in extreme value statistics (Hall 1982).

3 Extreme value theory

In this section we briefly recall the main results from univariate extreme value theory and introduce the notation that will be used throughout the paper.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables from the distribution function F, and denote by $M_n = \max(X_1, \ldots, X_n)$ the maximum of the first n samples.

The Fisher–Tippett–Gnedenko theorem (cf., Coles et al. 2001) states that if there exist sequences $a_n \in \mathbb{R}$, $b_n > 0$ such that

$$P\left(\frac{M_n - a_n}{b_n} \leq z\right) \to G(z), \qquad n \to \infty, \tag{1}$$

then if G is a non-degenerate distribution function it belongs to the class of generalized extreme value (GEV) distributions with

$$G(z) = \exp\left[-\left\{1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right\}^{-1/\xi}\right], \quad z \in \mathbb{R} : 1 + \xi\left(\frac{z-\mu}{\sigma}\right) > 0,$$

where $\xi \in \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma > 0$ are the shape, location and scale parameters, respectively. If the convergence in Eq. 1 holds, we say that X is in the max-domain of attraction of a GEV distribution with shape parameter ξ. A similar result can be formulated for threshold exceedances. In fact, if the convergence in Eq. 1 holds, then, for a threshold $u \in \mathbb{R}$ that tends to the upper endpoint of the distribution F of X, the distribution function of $X - u$, conditional on $X > u$, is approximately

$$H(y) = 1 - \left(1 + \frac{\xi y}{\bar{\sigma}}\right)^{-1/\xi}, \quad y > 0 : 1 + \frac{\xi y}{\bar{\sigma}} > 0, \tag{2}$$

and the distribution H is called the generalized Pareto distribution (GPD). The parameters of the two extreme value distributions are closely related: if M_n follows approximately a GEV distribution G with parameters ξ, μ_n and σ_n, then the GPD H has the same shape ξ and $\bar{\sigma} = \sigma_n + \xi(u - \mu_n)$.

If we are interested in computing the distribution of the maximum of a large number of independent copies of a given random variable, the first result suggests to fit a GEV distribution to various such maxima taken over blocks of the same lengths. On the other hand, if we are interested to model the right tail of a distribution, then the second result suggests to fit a GPD to all values above a high threshold of a sequence of independent replications of a random variable with this distribution. The two approaches are asymptotically equivalent and the parameters in the two distributions are deterministically linked. The following lemma will be useful to describe our algorithms.

Lemma 1 (Embrechts et al. 2013) *Suppose that the distribution function F of X has an upper endpoint denoted by b^*. Then it is in the max-domain of attraction of a GEV with $\xi < 0$ if and only if*

$$1 - F(b^* - 1/x) = \ell(x)x^{-1/\xi},$$

where ℓ is a slowly varying function at infinity. This is also equivalent to $(b^* - X)^{-1}$ having a regularly varying tail with index $-\xi$. In this case, we have the representation $b^* = u - \bar{\sigma}/\xi$ in terms of the parameters of the generalized Pareto distribution.

4 General setting

In this section we formalize the task of anomaly detection in a way suitable for all the algorithms that we describe in the following.

Denote the training data by $x_i \in \mathbb{R}^p$, each of them labeled as a class $y_i \in \{C_1, \ldots, C_J\}$, $i = 1, \ldots, n$. Here and throughout, $p \in \mathbb{N}$ is the dimension of the predictor space and $J \in \mathbb{N}$ the number of different classes in the training set. In total, then, we have J normal classes and we assume that each class is described by a continuous density function f_{C_j} defined on $\mathbb{R}^p$, $j = 1, \ldots, J$, where the probability that a point in class C_j falls in the set $A \subset \mathbb{R}^p$ is

$$\int_{x \in A} f_{C_j}(x)\, dx.$$

The process that has generated the training data set can be described as a mixture of these density functions, with unconditional density

$$f(x) = \sum_{j=1}^{J} w_j f_{C_j}(x), \quad x \in \mathbb{R}^p,$$

for weights $w_j \in [0, 1]$ with $\sum_{j=1}^{J} w_j = 1$. The value of the function f evaluated at some point is thus large if that point has high chance to be normal. Note that, in the following, we approximate directly f and then we do not have to estimate the weights w_j.

Further, suppose that we have a new unlabeled test point $x_0 \in \mathbb{R}^p$ that we would like to mark as normal or abnormal. More precisely, we define the point x_0 as normal if it was sampled from the training distribution with density f, and as abnormal if it was sampled from another distribution with unknown density f_0. The goal in anomaly detection is to decide if x_0 comes from the distribution with density f of the training data or not. Thus, we need to perform the hypothesis test

$$H_0\text{: } x_0 \text{ is normal,}$$
$$H_1\text{: } x_0 \text{ is abnormal.}$$

Note that most of the algorithms for anomaly detection avoid formal assumptions on the distribution with density f_0 and mark a test point x_0 as abnormal if the training density $f(x_0)$ is low. All the methods that we describe in the following are of this type.

5 The extreme value machine

In this section first briefly describe the extreme value machine introduced in Rudd et al. (2018) and then underline its limitations and incorrect assumptions.

5.1 Algorithm description

The idea of the EVM is to approximate the distribution of the margin distance of each point in each class using extreme value theory. A new point is then classified as normal if it is inside the margin of some point in the training set with high probability.

Rudd et al. (2018) introduce this concept of margin distance of a training point x_i as half of the minimum distance between x_i and all the points belonging to a different class in the training data set. More formally, the margin distance of x_i is

$$M^{(i)} = \min_{j:y_j \neq y_i} D_j^{(i)} = \min_{j:y_j \neq y_i} \frac{||x_i - x_j||}{2}.$$

The idea is to model the lower tail of the distribution of $M^{(i)}$ by using the $D_j^{(i)}$ as the input data. An equivalent problem, to which we can apply extreme value theory, is to retrieve the upper tail of the distribution of $\bar{M}^{(i)} = \max_{j:y_j \neq y_i} -D_j^{(i)}$. To this end, the authors propose to use the Fisher–Tippett–Gnedenko theorem (cf., Coles et al. 2001) to fit to the k largest observed $-D_j^{(i)}$ for each point x_i a GEV distribution

$$W^{(i)}(z) = \begin{cases} \exp\left\{-\left(-\frac{z}{\sigma_i}\right)^{\alpha_i}\right\}, & \text{if } z < 0, \\ 1, & \text{if } z \geq 0, \end{cases} \tag{3}$$

assuming $b^* = 0$ since $\bar{M}^{(i)}$ is bounded above by zero as a negated distance. We discuss issues with the use of this approximation in the next section.

Estimating the parameters α_i and σ_i for each training observation x_i separately yields estimates $\widehat{W}^{(i)}$ of the distribution functions $W^{(i)}$. Those estimates are used to label a new point x_0 as normal if it is likely enough to fall in at least one of the margins of the n training data, that is, if

$$\max_{i=1,\ldots,n} \widehat{W}^{(i)}\left(-||x_0 - x_i||\right) \geq \delta, \tag{4}$$

and it is labeled as abnormal otherwise. Here, δ is a probability threshold that is chosen by a heuristic formula in Rudd et al. (2018).

5.2 Limitations of the EVM

The EVM redefines the anomaly detection problem as a probabilistic framework within extreme value theory, but it has some drawbacks that we describe here and that will be the starting point to construct new methods that improve upon it.

First, assuming in Eq. 3 that the upper endpoint is $b^* = 0$ implies that all classes in the training set share the same support and that it is impossible that two classes are perfectly separated. Indeed, if two classes are perfectly separated, then the distance between two points sampled from them is bounded from above by a strictly positive constant and therefore $b^* < 0$. This assumption is rather restrictive and it is hardly verifiable on observed data. Second, estimating the upper tail of $\bar{M}^{(i)}$ based on the k largest observed $-D_j^{(i)}$ for each x_i should rely on the GPD instead of the GEV approximation since this amounts to a threshold exceedance approach with a GPD as limiting distribution.

A further limitation of the EVM is related to the choice of the threshold δ in Eq. 4. This threshold controls the anomaly detection error and a bigger δ implies a higher probability to label a new object as abnormal. In a statistical setting this type of thresholds are chosen by fixing the type I error, in this case, the probability to classify a normal object as abnormal. Unfortunately, for the EVM algorithm the authors do not propose any procedure to choose δ controlling the type I error and thus it is not clear how to properly tune it.

The most important drawback of the EVM is that it gives a non-justified premium to normal classes far from the others. In fact, if there are classes relatively close to each other and one class far from the others, then all the points of the latter class will have a large margin distance. For this reason, if we observe a new point from an abnormal class located closer to this far class than the other normal classes, we always classify the new point as belonging to the far away class. The main issue here is that we cannot use the distances between the normal classes to do anomaly detection because it does not convey any information regarding the abnormal ones. Knowing that class A is very far from class B, does not necessarily imply that a new object x_0 is normal only because the distance between it and the closest point in class A is much smaller than the distances between classes A and B. It might simply be that the new class is closer to class A than is class B. The EVM is implicitly assuming that the distances between the normal classes reflect also the behavior of the abnormal ones, a strong assumption that is again hard to verify. We will illustrate this problem in a simulation study in Section 8.

6 The GPD classifier

In this section we propose an algorithm for anomaly detection that uses the principles of extreme value theory and that does not rely on the distances between the different classes in the training set. This algorithm, which we call the GPDC, considers the distances between the new point that we want to mark as normal or abnormal, and the points composing the training data set. For simplicity, and since we are primarily interested in anomaly detection, all the normal classes are seen as one big class. If there is high confidence that the lower endpoint of the distribution of these distances is zero, then the process that has generated the training data can also have generated the new point, and we thus mark the new point as normal; otherwise we mark it as abnormal.

6.1 Extreme value theory and anomaly detection

In this setting the training data $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, can be thought as being sampled from only one class, the class of the normal data points, with density f. Given a new point x_0 without label, we are interested in deciding whether it is from a normal class and has been generated by f, or whether it belongs to an unseen class and has been generated by an abnormal density f_0.

We denote with D_i, $i = 1, \ldots, n$, the distances $||x_i - x_0||$ between x_0 and the points in the training set. In the following, we are interested in the lower tail of

the distribution D of the distance between x_0 and a generic point sampled from f. Intuitively, if the lower tail of D behaves similarly as the corresponding quantity for training samples, then x_0 has a high probability to be a normal point. More mathematically, let $B_\delta(x_0)$ denote a ball of radius δ centered in x_0. Then

$$P\{B_\delta(x_0)\} = P(D < \delta) = P(-D > -\delta),$$

where P is the probability measure that has density f. Moreover, the upper end point of $-D$ is equal to zero if and only if there is no $\delta > 0$ such that $P\{B_\delta(x_0)\} = 0$. Only in this case there is a positive probability that x_0 comes from the process that has generated the training data, i.e., that it is a normal point. Under weak assumptions, we can then specify the exact tail behavior of $-D$ at zero.

Proposition 1 *Assume that x_0 is in the interior of the support $\mathrm{supp}(f)$ of the normal classes and that f is continuous at x_0. Then the distribution of $-D$ is in the max-domain of attraction of a GEV distribution with shape parameter $\xi = -1/p$, where $p \in \mathbb{N}$ is the dimension of the predictor space.*

Proof We first note that the survival function of $-D$ behaves at zero as

$$\begin{aligned} P(-D > -\delta) &= P\{B_\delta(x_0)\} \\ &= \int_{B_\delta(x_0)} f(x)\,dx \\ &= \{f(x_0) + o(1)\}\, V\{B_\delta(x_0)\} \\ &= f(x_0)\frac{\pi^{p/2}}{\Gamma(p/2+1)}\delta^p + o(\delta^p), \end{aligned}$$

where $V\{B_\delta(x_0)\}$ is the volume of the ball $B_\delta(x_0)$ in $\mathbb{R}^p$, and the approximation holds as $\delta \to 0$. From this representation if follows directly that the distribution of $-D$ is in the max-domain of attraction of a GEV distribution with shape parameter $\xi = -1/p$. □

Note that f is, in general, a density on a multivariate space, whereas $P(-D > \cdot)$ is a simple univariate survival function and the above result describes its tail behavior. We can therefore model the right tail of $-D$ using extreme value statistics.

We further observe that the above proposition and Lemma 1 imply that $1/D$ has a regularly varying survival function

$$P(1/D > x) = \ell(x)x^{-p},$$

for some slowly varying function ℓ at infinity. The upper tail of $1/D$ is thus in the max-domain of attraction of a GEV distribution with shape parameter $\tilde{\xi} = 1/p > 0$. Therefore, a consistent estimator of $\tilde{\xi}$ is given by the Hill estimator (Hill 1975). The following theorem uses this fact to obtain a consistent estimator for ξ.

Theorem 1 *Assume that x_0 is in the interior of the support $\mathrm{supp}(f)$ of the normal classes and that f is continuous at x_0. Denote the distances between x_0 and the points in the training set with $D_1, \ldots, D_n$. Let further $R_{(n)} \geq R_{(n-1)} \cdots \geq R_{(1)}$ be*

the order statistics of $R_i = -D_i$. *For* $k = k(n) \to \infty$ *and* $k(n)/n \to 0$, *as* $n \to \infty$, *let*

$$\widehat{\xi}_n = \frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{R_{(n+1-i)}}{R_{(n-k)}} \right). \tag{5}$$

This estimator converges in probability

$$\widehat{\xi}_n \xrightarrow{P} -1/p,$$

where $p \in \mathbb{N}$ *is the dimension of the predictor space.*

Proof Rewriting the statistic $\widehat{\xi}_n$ slightly gives

$$\widehat{\xi}_n = -\frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{-1/R_{(n+1-i)}}{-1/R_{(n-k)}} \right),$$

and we recognize the (negated) classical Hill estimator applied to the sequence of independent copies $-1/R_1, \ldots, -1/R_n$ of $1/D$. The consistency of the Hill estimator (cf., Theorem 3.2.2 in De Haan and Ferreira (2007)) and the fact that $1/D$ has shape parameter $1/p$ yields the desired result. □

The above result motivates to use the statistic $\widehat{\xi}_n$ in Eq. 5 in an hypothesis test to decide whether the new point x_0 can come from a normal class, and the last part of the above theorem provides the theoretical background for this test. Indeed under the hypothesis that x_0 is in the interior of the support of f, the density of the normal classes, the statistic $\widehat{\xi}_n$ is approximately the negative reciprocal of the predictor space dimension. Moreover, it is possible to show that the estimator (5) is a special case of the classical Hill estimator (De Haan and Ferreira 2007) and hence it is asymptotically normal under a second order condition. Using this fact, we could even derive asymptotic confidence intervals for ξ based on it. The assumption on $k(n)$ is common in extreme value statistics to ensure the right trade-off between bias and variance of a tail estimator.

Under the alternative, namely that the new point x_0 is abnormal, we can encounter two situations. First, if the support of the new class is overlapping with the support of the normal classes $\text{supp}(f)$, then $-D$ might have upper endpoint zero if x_0 falls into the interior of $\text{supp}(f)$. Second, if the supports are not overlapping or $x_0 \notin \text{supp}(f)$, then the upper endpoint of $-D$ is strictly smaller than zero. In this second case, the statistic $\widehat{\xi}_n$ can be shown to converge to zero.

Theorem 2 *Let* r^* *be the upper endpoint of* $-D$ *and suppose that* $r^* < 0$. *Assume that* $k = k(n) \to \infty$ *and* $k(n)/n \to 0$, *as* $n \to \infty$. *Then the statistic in* Eq. 5 *converges almost surely to zero, that is,* $\widehat{\xi}_n \xrightarrow{a.s.} 0$.

Proof We note that

$$\begin{aligned} 0 \leq -\widehat{\xi}_n &= \frac{1}{k}\sum_{i=1}^{k}\left\{\log(-R_{(n-k)}) - \log(-R_{(n+1-i)})\right\} \\ &\leq \log(-R_{(n-k)}) - \log(-r^*). \end{aligned}$$

Since $-R_{(n-k)} \to -r^*$ almost surely as $n \to \infty$, this proves the assertion. □

We can use this fact to do a first test and mark a new point x_0 for which $p\widehat{\xi}_n$ is significantly larger than -1 as an abnormal point. Note that we use the quantity $p\widehat{\xi}_n$ instead of $\widehat{\xi}_n$ to stabilize the asymptotic variance and to avoid numerical instabilities as $-1/p \to 0$ as $p \to \infty$. In fact, under a second order condition and by asymptotic normality of the Hill estimator (cf., Theorem 3.2.5 in De Haan and Ferreira (2007)), the quantity $\sqrt{k}\, p\widehat{\xi}_n$ is approximately normal with mean -1 and variance 1. With this step we have excluded all the points that have no possibility to belong to one of the classes of the training set.

If, on the other hand, $p\widehat{\xi}_n$ is close to -1, we cannot reject the null hypothesis that x_0 is in the interior of $\mathrm{supp}(f)$. In this case, we can use the threshold $u = R_{(n-k)}$ to model the upper tail of $-D$, and since the upper endpoint of $-D$ is 0, Lemma 1 yields that the scale parameter of the approximating GPD satisfies $\bar{\sigma} = R_{(n-k)}\xi$. Using the estimator $\widehat{\xi}_n$ from Theorem 1 yields the tail approximation for $x > R_{(n-k)}$

$$P(-D > x) \approx \tfrac{k}{n}\left\{1 - \widehat{H}(x - R_{(n-k)})\right\} = \tfrac{k}{n}\left(x/R_{(n-k)}\right)^{-1/\widehat{\xi}_n}, \tag{6}$$

where $\widehat{H}$ is the GPD distribution with shape $\widehat{\xi}_n$ and scale $\bar{\sigma} = R_{(n-k)}\xi$. There is thus the possibility that x_0 was generated from one of the normal classes. Obviously, we have to exclude points that have positive, but very small probability of being normals. We therefore approximate the size of the ball around x_0 that contains a fixed amount $\gamma > 0$ of mass of f with the approximate distribution of $-D$ as given by Eq. 6. More precisely, we let $q_\gamma < 0$ be the $(1-\gamma)$-quantile of $-D$ for some small $\gamma < k/n$, which we can compute from Eq. 6 as

$$q_\gamma = R_{(n-k)} + \widehat{H}^{-1}(1 - n\gamma/k) = R_{(n-k)}(n\gamma/k)^{-\widehat{\xi}}, \tag{7}$$

such that $P\{B_{-q_\gamma}(x_0)\}$ is approximately equal to γ. Thus, the smaller $-q_\gamma$ the higher the training density $f(x_0)$ around x_0. If $-q_\gamma$ obtained for the new point x_0 is significantly larger than the corresponding quantity for a generic point in the training set, we mark x_0 as abnormal since the density $f(x_0)$ around x_0 is too small. Otherwise we mark it as normal. The probability γ should be chosen small enough to have a good approximation of the magnitude of the density around x_0, but at the same time large enough to obtain a reliable estimate of q_γ. We choose $\gamma = 1/n$ since this is sufficiently small and $R_{(n-k)} + \widehat{H}^{-1}(1 - 1/k)$ is usually a very good approximation of the true quantile.

Note that the quantile estimator in Eq. 7 is a special case of the estimator in Weissman (1978).

6.2 The GPDC algorithm

Using the results on the asymptotic behavior of the statistic $\widehat{\xi}_n$, we propose the following algorithm to classify a new point x_0.

1. Compute the negated distances $-D_1, \ldots, -D_n$ between x_0 and each point in the training set.
2. Estimate $\widehat{\xi}_n$ using only the biggest k negated distances $R_{(n)}, \ldots, R_{(n+1-k)}$.
3. If $p\widehat{\xi}_n$ is smaller than a given threshold $s > -1$, mark x_0 as possibly normal and go to the next point, otherwise mark it as abnormal and exit the algorithm.
4. Compute the $(1 - 1/n)$-quantile of $-D$ by $q_\gamma = R_{(n-k)} + \widehat{H}^{-1}(1 - 1/k)$ using the estimated GPD $\widehat{H}$ and $\gamma = 1/n$.
5. A large radius $-q_\gamma$ corresponds to a low density of f at x_0. Thus, if $-q_\gamma$ is bigger than a given threshold $t > 0$ mark x_0 as abnormal, otherwise mark it as normal.

Some of the steps require further explanation. The number of upper order statistics k in the second step that is used to obtain $\widehat{\xi}_n$ corresponds to the classical bias-variance trade-off of the Hill estimator. There are widely used diagnostics plots that help to determine the best k (Coles et al. 2001).

Note that step 3. marks x_0 as abnormal if it is not in the support of the training density f. Therefore, it can be seen as a preliminary hypothesis test with null hypothesis "x_0 is in the support of f" and alternative "x_0 is not in the support of f". If we do not reject this first null hypothesis, step 5. is needed to test the stronger null hypothesis "x_0 is normal" with alternative "x_0 is abnormal". In particular, we mark x_0 as abnormal if it lies in a region where the density $f(x_0)$ is low. Thus, steps 4. and 5. are especially necessary if the supports of f and f_0 are not clearly separated.

Step 3. and step 5. require the choice of the thresholds s and t to perform the hypothesis tests. It is common to fix the type I error to some probability α such as 5%, and then to compute the threshold that realizes this error. To this end we require the distribution of the test statistic under the null hypothesis. Instead of relying on asymptotic results, we propose to execute the algorithm using all the points in the training set as unknown one after the other in a jackknife fashion and to obtain in this way each time $\widehat{\xi}_n^{(i)}$ and $-q_\gamma^{(i)}$, $i = 1, \ldots, n$. We then jointly set the thresholds s and t to the $(1-\alpha/2)\%$ quantiles of $\widehat{\xi}_n^{(1)}, \ldots, \widehat{\xi}_n^{(n)}$ and $-q_\gamma^{(1)}, \ldots, -q_\gamma^{(n)}$, respectively, such that we mark at most $\alpha\%$ of the training data as abnormal following Bonferroni's correction for multiple testing (Shaffer 1995). Note that this requires to compute $O(nk \log n)$ distances, roughly the same as in the training phase of the EVM. We can do this procedure at the beginning of the life of the algorithm and repeat it once in a while if new training data arise.

We emphasize that, contrary to what is stated in the EVM paper (Rudd et al. 2018), one should be careful to choose hyper parameters based on cross-validation by randomly splitting the classes in the training set in normal and abnormal data points. This can be misleading since normal classes convey no information on the abnormal ones.

It is important to underline that this algorithm relies on the asymptotic approximation by a GPD and hence it is asymptotically exact as the number of training samples tends to infinity. Moreover, it is completely kernel free and its implementation is very efficient using fast nearest-neighbor searching (Arya et al. 1998). In fact, during the evaluation of a new point x_0 we have to compute only $O(k \log n)$ distances, whereas with the EVM we have to compute $O(n)$ distances. We stress also that the algorithm supports incremental learning since it does not require any training procedure once the thresholds for the hypothesis tests are fixed. If the interest is not only in anomaly detection our algorithm can be completed with an incremental classifier that performs standard classification only on the points marked as normals.

7 The GEV classifier

The GPDC in the previous section was based on the approximation of threshold exceedances by the GPD as limiting distribution. In this section we propose another algorithm based on a more heuristic approach to do anomaly detection. It is based on the approximation by the GEV distribution given the Fisher–Tippett–Gnedenko theorem (cf., Coles et al. 2001), and we thus refer to it as the GEVC.

For simplicity, as before, all normal classes are collapsed into one big class and we denote the training data with x_i, $i = 1, \dots, n$, and with f the probability density describing the distribution of the training data. For each $i = 1, \dots, n$, we compute the distance between the training point x_i and the nearest training point to it, i.e.,

$$D_i^{\min} = \min_{j \neq i} ||x_i - x_j||.$$

This phase, using fast neighbor search, requires to compute only $O(n \log n)$ distances. Note that the quantities

$$-D_i^{\min} = \max_{j \neq i} -||x_i - x_j||$$

are bounded above by zero. Even though these random variables are not exactly independent, their dependence seems weak enough to still apply methods from extreme value theory. Using this and Lemma 1 we can fit a GEV with negative shape parameter to $-D_1^{\min}, \dots, -D_n^{\min}$, and thus estimating the distribution of $-D^{\min}$, i.e., the distribution of the maximum of the negated distances between a normal point and all the remaining points in the training set. Denote this estimated distribution function by $\widehat{W}$. We note again that this approximation is motivated heuristically, since $-||x_i - x_j||$, conditionally on x_i, would be max-domain of attraction of a GEV with negative shape parameter. Since the data point x_i is random, this argument is formally no longer true, but in practice the GEV approximation with negative shape parameter seems still suitable.

When a new point x_0 arises and we want to mark it as normal or abnormal we consider the statistic

$$-d_0^{\min} = \max_{i=1,\dots,n} -||x_i - x_0||,$$

in order to perform the hypothesis test described in Section 4. Note that this operation requires to compute only $O(\log n)$ distances. Under the null hypothesis, $-d_0^{\min}$ is approximately a sample of the distribution $-D^{\min}$. Thus we expect the quantity

$$P(-D^{\min} < -d_0^{\min})$$

to be not too small, whereas under the alternative hypothesis this is not guaranteed. We can then fix a level $\alpha > 0$ for the type I error to determine if the new point x_0 is normal or abnormal. Given the estimated distribution $\widehat{W}$ that approximates $-D^{\min}$, we reject the null hypothesis if $\widehat{W}(-d_0^{\min}) < \alpha$. In this case, the smallest distance of the new point to a training sample is too large, and we therefore mark it as abnormal. Otherwise it is marked as normal. Note that, also with the GEVC, the threshold α permits to control directly the type I error. Moreover, the GEVC has no hyper parameters and it is fast to update when we add new data to the training set, since revising $-D_i^{\min}$ requires only to compute the minimum distance between x_i and the added points.

8 Application

In this section we compare the EVM introduced in Rudd et al. (2018) with our GPDC and the GEVC on simulated and real data. For completeness, when using real data, we compare the methods also with Isolation Forest (Liu et al. 2012) and One-Class SVM (Schölkopf et al. 2000), two state of the art methods for novelty detection. For this latter technique, we use a radial kernel and we select its hyper parameters using part of the test data as validation set, including both normal and abnormal objects, following what is proposed in the original paper, while for the former we follow the original implementation. Note that using part of the test set is an approach possible only in an experimental setting, since in real anomaly detection we do not have any access to the abnormal objects during the training phase.

8.1 Simulated data

We begin our experimental evaluation with a toy example that shows that the EVM may perform poorly when the distances between the different classes in the training set convey misleading information about the abnormal ones. The training set has $n = 600$ observations and it is composed of three classes, each one of them is sampled from a different bivariate normal distribution, i.e., the dimension of the predictor space is $p = 2$. The test data set has 800 observations and it is composed of examples from these three classes plus another abnormal class, sampled from another bivariate normal distribution. Both the train and the test data are shown in the left panel of Fig. 1. The abnormal objects are well separated from the normal ones.

We apply the EVM, the GPDC and the GEVC to this data set and, to solve the problem given by the fact that the probability threshold of the EVM is not interpretable as those of the GPDC and the GEVC, we choose to evaluate the performance

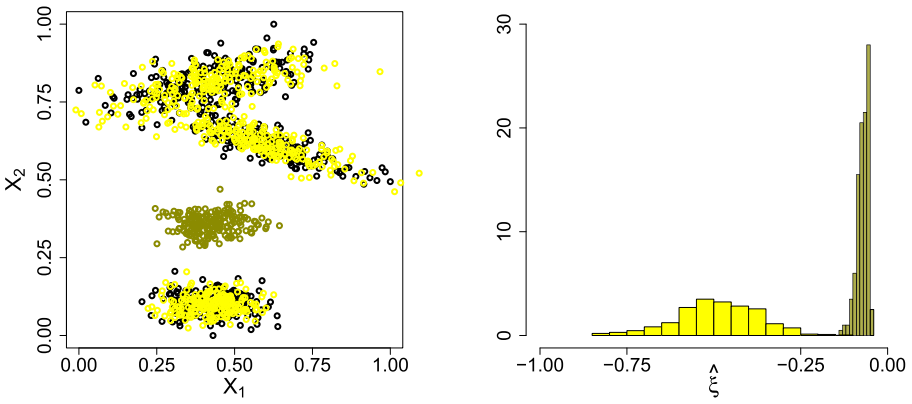

Fig. 1 Left panel: the simulated dataset for the toy example: training data (in black), the normal (bright yellow) and abnormal examples (dark yellow) from the test set. Right panel: the $\widehat{\xi}_n$ estimates for normal (bright yellow) and abnormal test data (dark yellow) for the simulated dataset

of the different algorithms using the obtained area under the ROC Curve (AUC) (Bradley 1997) that does not require to fix a threshold. For both the EVM and the GPDC we use $k = 20$, but the results are stable for different values of k. The EVM performs poorly in this rather simple scenario and has an AUC of 0.853. This is due to the fact that the abnormal class is relatively close to the normal class at the bottom compared to the other normal classes, even if it is perfectly separated from it. Conversely, the GPDC and the GEVC do not suffer of this type of issues since they do not rely on the distances between the normal classes to infer about the abnormal ones, and their results in this toy example are close to perfect for both algorithms, with AUC of 0.997 and 0.999, respectively. To show the effectiveness of the GPDC to determine whether new examples are in the support of the training distribution, we report the estimated $\widehat{\xi}_n$ of the test in point 3. of the algorithm in Section 6.2 for both normal and abnormal new data. The right panel of Fig. 1 shows for this dataset that the $\widehat{\xi}_n$ estimates for normal data are, as expected by Proposition 1, close to $-1/2$. On the other hand, the estimates corresponding to abnormal data are much closer to 0, as it is suggested by Theorem 2. Step 3 of the GPDC algorithm thus already has a good power to filter out the abnormal classes for this data set. We note that, strictly speaking, the supports of all classes are overlapping, but for this finite number of data they are effectively separated so that the test works well.

Figure 2 shows the decision boundaries of the GEVC and the GPDC for $\alpha = 0.01$ and $\alpha = 0.1$. It can be seen that both algorithms produce highly flexible decision boundaries that are capable to follow the shape of the training data. In some sense, one can see these boundaries as level sets of the training density, and a new point is considered as abnormal if it lies outside of these level sets. He and Einmahl (2017) also considered extrapolation of level sets into low density regions but using the more restrictive assumption of multivariate regular variation.

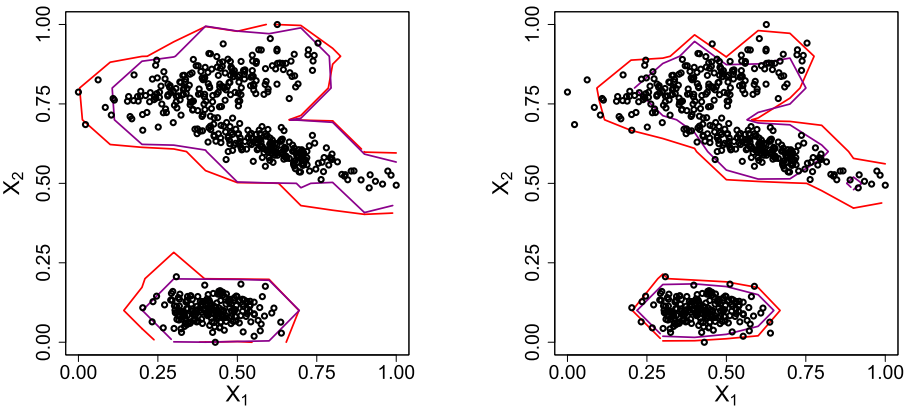

Fig. 2 Left panel: decision boundaries for the GPDC (in red) and the GEVC (in magenta) with $\alpha = 0.01$. Right panel: decision boundaries for the GPDC (in red) and the GEVC (in magenta) with $\alpha = 0.1$

8.2 OLETTER protocol

In order to evaluate the anomaly detection performance of the GPDC and the GEVC for real data we compare the two techniques with the EVM, One-Class SVM and Isolation Forest using the OLETTER protocol proposed in Bendale and Boult (2015). This protocol is based on the LETTER data set (Frey and Slate 1991) that contains a total of 20000 observations of 26 different classes corresponding to handwritten letters. The predictor space is composed of $p = 16$ features that have been extracted from the handwritten letters. The training set has 15000 observations. The protocol consists in randomly selecting 15 classes that are considered as normal during training and adding abnormal classes by incrementally including subsets of the remaining 11 classes during testing, varying in this way the amount of openness, i.e., the proportion of abnormal objects, during the test phase. This process is then repeated 20 times, in a cross-validation fashion. We evaluate the performance of the different algorithms using the obtained F-measure (Huang and Ling 2005) as done in the original paper (Bendale and Boult 2015). The F-measure is an evaluation metric that combines recall and precision according to the following formula

$$F\text{-measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where

$$\text{precision} = \frac{\text{number of correctly identified abnormal data points}}{\text{number of data points identified as abnormal}}$$

and

$$\text{recall} = \frac{\text{number of correctly identified abnormal data points}}{\text{number of abnormal data points}}.$$

For all the four algorithms we set dynamically the probability threshold using the heuristic rule proposed in Rudd et al. (2018) that accounts for the amount of openness at each step of the protocol. We set the hyper parameter $k = 75$ for the EVM as in

original paper, and $k = 22$ for the GPDC, roughly corresponding to use the 0.25% of the biggest negated distances.

The results for all the methods are reported in Fig. 3. It can be seen that all the three methods based on EVT perform better than Isolation Forest and One-Class SVM, even if they have fewer hyper parameters than the latter methods. Moreover, the hyper parameters of One-Class SVM were tuned using also abnormal examples and that is not realistic during a real use of an anomaly detection classifier. The GPDC is the best method in this case, while the GEVC is competitive with the other methods even if it has no hyper parameters.

8.3 Diagnostics of thyroid disease

We consider an application of our algorithms for anomaly detection to diagnostics of hypothyroidism, a type of thyroid disease. We analyse the thyroid dataset (Quinlan et al. 1986; Schiffmann et al. 1992), available at the UCI machine learning repository (Dua and Graff 2017). The original dataset contains raw clinical measurements from sick patients affected by hypothyroidism and healthy ones. These raw measurements

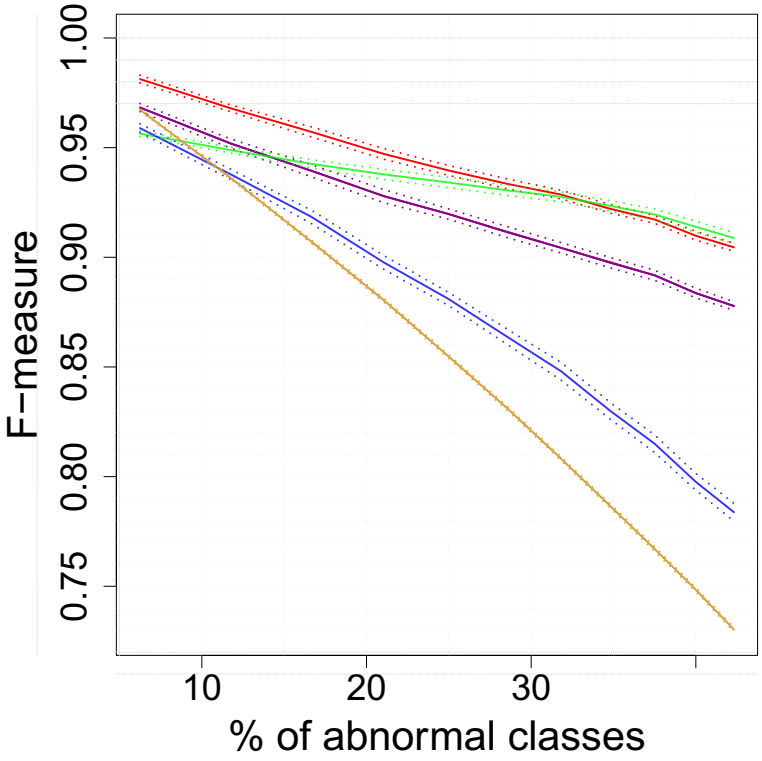

Fig. 3 Results for the OLETTER protocol for the EVM (green line), the GPDC (red line), the GEVC (magenta line), One-Class SVM (blue line) and Isolation Forest (golden line) with one standard deviation confidence intervals (dotted lines)

were pre-processed in order to easily apply directly neural networks and other common machine learning techniques, resulting in 21 features for each patient. In the final dataset there are 250 sick subjects and 6666 healthy subjects. We consider sick subjects as abnormal objects and healthy subjects as normal objects. We use all the 250 sick patients together with 250 randomly selected healthy patients to compose the test set. All the remaining healthy patients compose the training set.

Since in this case the normal objects belong to only one class, the EVM cannot be used with this dataset. For this reason, we compare the GEVC and the GPDC only with One-Class SVM and Isolation Forest. Also here, we use the AUC as evaluation measure. For the GPDC we train it with five different values of k, corresponding to using the most extreme 0.25%, 1%, 2.5%, 5% and 10% distances, respectively. These are common choices in extreme value applications and finally we consider the value of k that gives the best ROC curve. For One-Class SVM with radial kernel we consider different combinations of its hyper parameters and we retain the ones that gives the best results.

The ROC curves obtained for all the methods are shown in the left panel of Fig. 4. The performance of the GPDC, GEVC and One-Class SVM are comparable, with an AUC of 0.931, 0.897, 0.912 respectively, while Isolation Forest performs considerably worst with an AUC of 0.760. We underline again this result since it is particularly favorable to both the GPDC and the GEVC that are reaching the same performance as a state of the art kernel based method like One-Class SVM and better performance than Isolation Forest. These new methods are kernel free and the GPDC has only one hyper parameter that can be chosen based on EVT, whereas the GEVC has no hyper parameters at all.

The right panel of Fig. 4 shows the performance of the GPDC as a function of the number k of most extreme distances used. It can be seen that the GPDC shows a good performance for a wide range of thresholds and the best results are achieved for fairly small k as suggested by EVT.

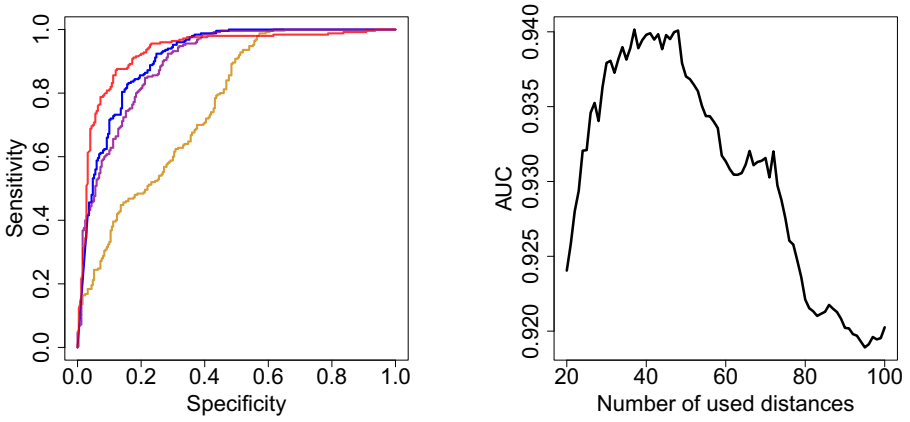

Fig. 4 Left panel: ROC curves obtained on the thyroid disease dataset for the GPDC (red line), the GEVC (magenta line), One-Class SVM (blue line) and Isolation Forest (golden line). Right panel: AUC as a function of the number k of most extreme used distances for the GPDC

9 Conclusion

We present two new kernel free algorithms that perform anomaly detection using extreme value theory. These algorithms, called the GPDC and the GEVC, are fast to update with the arrival of new data and they are easy to adapt to an incremental framework. Moreover, they do not use the distances between the classes in the training set to infer about the abnormals and are thus able to overcome certain restrictions of previously proposed methods. Their performances are therefore good even when the abnormal test data are close to a normal class relatively to the other normals. To show this fact and the effectiveness of the new GPDC and the GEVC, we compare them to the EVM (Rudd et al. 2018), another kernel free technique that uses EVT for anomaly detection, and the more classical kernel based One-Class SVM (Schölkopf et al. 2000) and tree based Isolation Forest (Liu et al. 2012), two state of the art techniques for novelty detection. The results of our methods on a simulated toy example and on real data sets are competitive. A major strength is that they perform well in very different situations. This good performance in general tasks is probably due to the fact that our methods rely on well-established statistical methods and asymptotically motivated approximations from univariate extreme value theory that apply under very mild conditions. We also underline that both the GPDC and the GEVC are computationally faster than the EVM during the evaluation phase.

The GPDC and the GEVC might be further improved by suitable representations of the data, for example using convolutional neural networks to extract features when working with images and we plan to perform more detailed evaluation of their performance in this direction on standard datasets such as ImageNet (Deng et al. 2009) or CIFAR-100 (Krizhevsky and Hinton 2009). Furthermore, to be fully incremental, they need to be capable to store only a subset of the training data. This may be achieved developing a suitable sampling technique that reduces the dimensionality of the training dataset without affecting the asymptotic correctness of the algorithms.

The result in Theorem 1 on the shape parameter of sample distances can be of independent interest. In particular, it will be studied how this can be used for a general method for extrapolation of level sets into low density regions, similarly as in Cai et al. (2011) and Einmahl et al. (2015) and He and Einmahl (2017).

Acknowledgments Edoardo Vignotto acknowledges funding from the Swiss National Science Foundation (Doc.Mobility Grant 188229). We gratefully acknowledge helpful comments by two anonymous referees and the editorial board. Sebastian Engelke was supported by the Swiss National Science Foundation; the paper was completed while he was a visitor at the Department of Statistical Sciences, University of Toronto.

Funding Open access funding provided by University of Geneva.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommonshorg/licenses/by/4.0/.

References

Abe, N., Zadrozny, B., Langford, J.: Outlier detection by active learning. In: International Conference on Knowledge Discovery and Data Mining. ACM (2006)

Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y.: An optimal algorithm for approximate nearest neighbor searching fixed dimensions. Journal of the ACM **45**(6) (1998)

Bendale, A., Boult, T.: Towards open world recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)

Bishop, C.M.: Novelty detection and neural network validation. IEE Proceedings-Vision, Image and Signal Processing **141**(4) (1994)

Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7) (1997)

Cai, J., Einmahl, J., De Haan, L., et al.: Estimation of extreme risk regions under multivariate regular variation. The Annals of Statistics **39**(3) (2011)

Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Computing Surveys (CSUR) **41**(3) (2009)

Christopher, M.B.: Pattern Recognition and Machine Learning. Springer, New York (2016)

Coles, S., Bawa, J., Trenner, L., Dorazio, P.: An Introduction to Statistical Modeling of Extreme Values. Springer, Berlin (2001)

De Haan, L., Ferreira, A.: Extreme Value Theory: an Introduction. Springer Science & Business Media, Berlin (2007)

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (2009)

Désir, C., Bernard, S., Petitjean, C., Heutte, L.: One class random forests. Pattern Recognition **46**(12) (2013)

Dua, D., Graff, C.: UCI machine learning repository. http://archive.ics.uci.edu/ml (2017)

Einmahl, J., Li, J., Liu, R., et al.: Bridging centrality and extremity: refining empirical data depth using extreme value statistics. The Annals of Statistics **43**(6) (2015)

Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events: for Insurance and Finance, vol. 33. Springer Science & Business Media, Berlin (2013)

Fragoso, V., Sen, P., Rodriguez, S., Turk, M.: EVSAC: accelerating hypotheses generation by modeling matching scores with extreme value theory. In: IEEE International Conference on Computer Vision (2013)

Frey, P.W., Slate, D.J.: Letter recognition using holland-style adaptive classifiers. Machine Learning **6**(2) (1991)

Geng, C., Huang, S., Chen, S.: Recent advances in open set recognition: a survey. Preprint arXiv:1811.08581 (2018)

Goix, N., Sabourin, A., Clemencon, S.: Sparse representation of multivariate extremes with applications to anomaly ranking. In: AISTATS (2016)

Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2013)

Hall, P.: On estimating the endpoint of a distribution. The Annals of Statistics **10**(2) (1982)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)

He, Y., Einmahl, J.: Estimation of extreme depth-based quantile regions. J. R. Stat. Soc. Ser. B. Stat. Methodol. **79** (2017)

Hill, B.M.: A simple general approach to inference about the tail of a distribution. The Annals of Statistics, pp 1163–1174 (1975)

Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering **17**(3) (2005)

Jalalzai, H., Clémençon, S., Sabourin, A.: On binary classification in extreme regions. In: Advances in Neural Information Processing Systems (2018)

James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning. Springer, Berlin (2013)

Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)

Liu, F.T., Ting, K.M., Zhou, Z.: Isolation-based anomaly detection. ACM Transactions on Knowledge Discovery from Data (TKDD) **6**(1) (2012)
Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Metric learning for large scale image classification: generalizing to new classes at near-zero cost. In: European Conference on Computer Vision. Springer, Berlin (2012)
Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: IEEE Conference on Computer Vision and Pattern Recognition (2015)
Pimentel, M.A.F., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Process. **99** (2014)
Quinlan, J.R., Compton, P.J., Horn, K.A., Lazarus, L.: Inductive knowledge acquisition: a case study. In: Proceedings of the second Australian Conference on the Applications of Expert Systems (1986)
Rebuffi, S., Kolesnikov, A., Lampert, C.H.: icaRL: incremental classifier and representation learning. In: Conference on Computer Vision and Pattern Recognition (2017)
Roberts, S.J.: Novelty detection using extreme value statistics. IEE Proceedings-Vision, Image and Signal Processing **146**(3) (1999)
Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3) (2018)
Ruping, S.: Incremental learning with support vector machines. In: IEEE International Conference on Data Mining (2001)
Saffari, A., Leistner, C., Santner, J., Godec, M., Bischof, H.: On-line random forests. In: IEEE International Conference on Computer Vision Workshops (2009)
Scheirer, W.J.: Extreme value theory-based methods for visual recognition. Synthesis Lectures on Computer Vision **7**(1) (2017)
Scheirer, W.J., Rocha, A., Micheals, R.J., Boult, T.E.: Meta-recognition: the theory and practice of recognition score analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(8) (2011)
Schiffmann, W., Joost, M., Werner, R.: Synthesis and performance analysis of multilayer neural network architectures. Technical report, University of Koblenz (1992)
Schölkopf, B., Williamson, R.C., Smola, A.J., Shawe-Taylor, J., Platt, J.C.: Support vector method for novelty detection. In: Advances in Neural Information Processing Systems (2000)
Shaffer, J.P.: Multiple hypothesis testing. Annual Review of Psychology **46**(1) (1995)
Shon, T., Moon, J.: A hybrid machine learning approach to network anomaly detection. Information Sciences **177**(18) (2007)
Siffer, A., Fouque, P., Termier, A., Largouet, C.: Anomaly detection in streams with extreme value theory. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017)
Thomas, A., Clemencon, S., Gramfort, A., Sabourin, A.: Anomaly detection in extreme regions via empirical MV-sets on the sphere. In: AISTATS (2017)
Walfish, S.: A review of statistical outlier methods. Pharmaceutical Technology **30**(11) (2006)
Weissman, I.: Estimation of parameters and large quantiles based on the k largest observations. J. Amer. Statist. Assoc. **73** (1978)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.