Conformal k-NN Anomaly Detector for Univariate Data Streams

Vladislav Ishimtsev VLADISLAV.ISHIMTSEV@SKOLKOVOTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Alexander Bernstein A.BERNSTEIN@SKOLTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Evgeny Burnaev E.BURNAEV@SKOLTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Ivan Nazarov IVAN.NAZAROV@SKOLKOVOTECH.RU
Skolkovo Institute of Science and Technology, Skolkovo, Moscow Region, Russia
Institute for Information Transmission Problems, Moscow, Russia

Editors: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Anomalies in time-series data give essential and often actionable information in many applications. In this paper we consider a model-free anomaly detection method for univariate time-series which adapts to non-stationarity in the data stream and provides probabilistic abnormality scores based on the conformal prediction paradigm. Despite its simplicity the method performs on par with complex prediction-based models on the Numenta Anomaly Detection benchmark and the Yahoo! S5 dataset.

Keywords: Conformal prediction, nonconformity, anomaly detection, time-series, nearest neighbours

1. Introduction

Anomaly detection in time-series data has important applications in many practical fields (Kejariwal, 2015), such as monitoring of aircraft's cooling systems in aerospace industry (Alestra et al., 2014), detection of unusual symptoms in healthcare, monitoring of software-intensive systems (Artemov and Burnaev, 2016), of suspicious trading activity by regulators or high frequency dynamic portfolio management in finance, etc.

General anomaly detection methods can be broadly categorized in five families, (Pimentel et al., 2014), each approaching the problem from a different angle: probabilistic, distance-based, prediction-based, domain-based, and information-theoretic techniques. The common feature of all families is the reliance on a negative definition of abnormality: "abnormal" is something which is not "normal", i.e. a substantial deviation from a typical set of patterns.

Prediction-based anomaly detection techniques rely on an internal regression model of the data: for each test example the discrepancy between the observed and the prediction, i.e. the reconstruction error, is used to decide its abnormality. For example, neural networks are used in this manner in (Augusteijn and Folkert, 2002) and (Hawkins et al., 2002; Williams et al., 2002), whereas in

(Chandola et al., 2009) the predictions are based on a comprehensive description of the variability of the input data. Other reconstruction methods include dimensionality reduction (Jolliffe, 2014), linear and kernel Principal Component Analysis (Dutta et al., 2007; Shyu et al., 2003; Hoffmann, 2007; Schölkopf et al., 1998).

Anomaly detection in time-series analysis is complicated by high noise and the fact that the assumptions of classical change point models are usually violated by either non-stationarity or quasi-periodicity of the time-series (Artemov et al., 2015; Artemov and Burnaev, 2016), or long-range dependence (Artemov and Burnaev, 2015a). Classical methods require strong pre- and post-change point distributional assumptions, when in reality change-points might exhibit clustering, or be starkly contrasting in nature between one another. Thus, the usual approach of detecting anomalies against a fixed model, e.g. the classical models (Burnaev, 2009; Burnaev et al., 2009), is unsubstantiated. This has compelled practitioners to consider specialized methods for anomaly model selection (Burnaev et al., 2015b), construction of ensembles of anomaly detectors (Artemov and Burnaev, 2015b), and explicit rebalancing of the normal and abnormal classes (Burnaev et al., 2015a), among others.

Time-series anomaly detection techniques include, among others, spatiotemporal self organising maps (Barreto and Aguayo, 2009), kurtosis-optimising projections of a VARMA model used as features for outlier detection algorithm based on the CUSUM (Galeano et al., 2006), Multidimensional Probability Evolution method to identify regions of the state space frequently visited during normal behaviour (Lee and Roberts, 2008), tracking the empirical outlier fraction of the one-class SVM on sliding data slices (Gardner et al., 2006), or applying one-class SVM to centred time-series embedded into a phase space by a sliding window (Ma and Perkins, 2003). The main drawback of these approaches is that they use explicit data models, which require parameter estimation and model selection.

Distance-based anomaly detection methods perform a task similar to that of estimating the pdf of data and do not require prior model assumptions. They rely on a metric, usually Euclidean of Mahalanobis, to quantify the degree of dissimilarity between examples and to derive either a distance-based or a local density score in order to assess abnormality. Such methods posit that the normal observations are well embedded within their metric neighbourhood, whereas outliers are not.

Despite being model-free, distance-based methods do not provide a natural probabilistic measure, which conveys detector's degree of confidence in abnormality of an observation. Indeed, there do exist distance-based methods, for example LoOP, (Kriegel et al., 2009), which output this kind of score, but typically they rely on quite limiting distributional assumptions. Such assumptions can potentially be avoided by using conformal prediction methods, (Shafer and Vovk, 2008). For instance, conformal prediction allows efficient construction of non-parametric confidence intervals (Burnaev and Nazarov, 2016).

This paper outlines an anomaly detection method in univariate time-series, which attempts to adapt to non-stationarity by computing "deferred" scores and uses conformal prediction to construct a non-parametric probabilty measure, which efficiently quantifies the degree of confidence in abnormality of new observations. We also provide technical details on boosting the performance of the final anomaly detector, e.g. signal pruning. The extensive comparison on Yahoo! S5 and Numenta benchmark datasets revealed that the proposed method performs on par with complex prediction-based detectors. The proposed method is among the top 3 winning solutions of the 2016 Numenta Anomaly Detection Competition, see (Numenta, 2016).

In section 2 we review general non-parametric techniques for assigning confidence scores to anomaly detectors. In sec. 3 we propose a conformal detector for univariate time-series based on k-NN (k Nearest Neighbours) and time-delay embedding, which attempts to tackle quasi-periodicity and non-stationarity issues. In section 4 we provide details on the comparison methodology and the Numenta Anomaly Detection benchmark, and in section 5 we compare the performance of the proposed method.

2. Conformal Anomaly Detection

Conformal Anomaly Detection (CAD), (Laxhammar, 2014), is a distribution-free procedure, which assigns a probability-like confidence measure to predictions of an arbitrary anomaly detection method. CAD uses the scoring output of the detector $A(X_{:t},\mathbf{x}_{t+1})$ as a measure of non-conformity (Non-Conformity Measure, NCM), which quantifies how much different a test object $\mathbf{x}_{t+1} \in \mathcal{X}$ is with respect to the *reference* sample $X_{:t} = (\mathbf{x}_s)_{s=1}^{t} \in \mathcal{X}$. Typical examples of NCMs are prediction error magnitude for a regression model, reconstruction error for dimensionality reduction methods, average distance to the k nearest neighbours, etc. The NCM may have intrinsic randomness independent of the data, (Vovk, 2013). For a sequence of observations $\mathbf{x}_t \in \mathcal{X}$, $t = 1,2,\ldots$, at each $t \geq 1$ CAD computes the scores

$$\alpha_s^t = A(X_{:t}^{-s},\mathbf{x}_s)\,,\, s = 1,\ldots,t\,, \tag{1}$$

where $X_{:t}^{-s}$ is the sample $X_{:t}$ without the s-th observation. The confidence that $\mathbf{x}_t$ is anomalous relative to the *reference* sample $X_{:(t-1)}$ is one minus the empirical p-value of its non-conformity score α_t^t in (1):

$$p(\mathbf{x}_t,X_{:(t-1)},A) = \frac{1}{t}\left|\{s = 1,\ldots,t \,:\, \alpha_s^t \geq \alpha_t^t\}\right|\,. \tag{CPv}$$

Basically, the more abnormal $\mathbf{x}_t$ is the lower its p-value is, since anomalies, in general, poorly conform to the previously observed reference sample $X_{:(t-1)}$.

In (Shafer and Vovk, 2008) it was shown that online conformal prediction, and by extension CAD, offers conservative coverage guarantees in online learning setting. Indeed, when iid sequence $\mathbf{x}_t \sim D$ is fed into the conformal anomaly detector one observation at a time, then for any NCM A and all $t \geq 1$

$$\mathbb{P}_{X\sim D^t}\big(p(\mathbf{x}_t,X^{-t},A) < \varepsilon\big) \leq \varepsilon\,, X = (\mathbf{x}_s)_{s=1}^{t}\,. \tag{2}$$

Intuitively, (CPv) is the empirical CDF, obtained on a sample $(A(X^{-s},\mathbf{x}_s))_{s=1}^{t}$, evaluated at a random point $A(X^{-t},\mathbf{x}_t)$ with the sample X drawn from an exchangeable distribution D^t. This means that the distribution of the p-value itself is asymptotically uniform. The NCM, used in (CPv), affects the tightness of the guarantee (2) and the volume of computations.

At any $t \geq 1$ in (1) CAD requires t evaluations of A with different samples $X_{:t}^{-s}$, which is potentially computationally heavy. (Laxhammar and Falkman, 2015) proposed the Inductive Conformal Anomaly Detection (ICAD) which uses a fixed proper training sample of size n as the reference in the non-conformity scores. If the sequence $(\mathbf{x}_t)_{t\geq 1}$ is relabelled so that it starts at $1-n$ instead of 1, then for each $t \geq 1$ the ICAD uses the following setup:

$$\underbrace{\mathbf{x}_{-n+1},\ldots,\mathbf{x}_0}_{\tilde{X}\text{ proper training}},\overbrace{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_{t-1}}^{\text{calibration}},\underbrace{\mathbf{x}_t}_{\text{test}},\ldots\,.$$

The conformal p-value of a test observation $\mathbf{x}_t$ is computed using (CPv) on the modified scores:

$$\alpha^t_s = A\left(\tilde{X}, \mathbf{x}_s\right), s = 1, \ldots, t, \tilde{X} = (\mathbf{x}_{-n+1}, \ldots, \mathbf{x}_0). \tag{3}$$

The ICAD is identical to CAD over the sequence $(\mathbf{x}_t)_{t \geq n+1}$ (relabelled to start at 1) with a non-conformity measure $\bar{A}$, which always *ignores* the supplied *reference* sample and uses the proper training sample $\tilde{X}$ instead. Therefore the ICAD has similar coverage guarantee as (2) with scores given by (3).

By trading the deterministic guarantee (2) for a PAC guarantee it is possible to make the ICAD use a fixed-size calibration set. The resulting "sliding" ICAD fixes the size of the calibration sample to m and forces it to move along the sequence $(\mathbf{x}_t)_{t \geq 1}$, i.e.

$$\underbrace{\mathbf{x}_{-n+1}, \ldots, \mathbf{x}_0}_{\tilde{X}\ \text{training}}, \ldots, \overbrace{\mathbf{x}_{t-m}, \ldots, \mathbf{x}_{t-1}}^{\text{calibration}}, \underbrace{\mathbf{x}_t}_{\text{test}}, \ldots.$$

The conformal p-value uses a subsample of the non-conformity scores (3):

$$p(\mathbf{x}_t, X_{:(t-1)}, A) = \frac{1}{m+1}\left|\{i = 0, \ldots, m : \alpha^t_{t-i} \geq \alpha^t_t\}\right|. \tag{CPv$_m$}$$

The guarantee for ICAD is a corollary to proposition (2) in (Vovk, 2013). In fact, the exchangeability of $(\mathbf{x}_t)_{t \geq 1}$ further implies a similar PAC-type validity result for the sliding ICAD, which states that for any $\delta, \varepsilon \in (0,1)$ for any fixed proper training set $\tilde{X}$ and data distribution D it is true that

$$\mathbb{P}_{\mathbf{x} \sim D}\left(p(\mathbf{x}, X, \bar{A}) < \varepsilon\right) \leq \varepsilon + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}, \tag{4}$$

with probability at least $1 - \delta$ over draws of $X \sim D^m$, and $\bar{A}$ is the NCM $\mathbf{x} \mapsto A(\tilde{X}, \mathbf{x})$, which uses $\tilde{X}$ as the *reference* sample.

3. Anomaly Detection in Univariate Time Series

In this section we outline the building blocks of the proposed model-free detection method which produces conformal confidence scores for its predictions. The conformal scores are computed using an adaptation of the ICAD to the case of potentially non-stationary and quasi-periodic time-series.

Consider a univariate time-series $X = (x_t)_{t \geq 1} \in \mathbb{R}$. The first step of the proposed procedure is to embed X into an l-dimensional space, via a sliding historical window:

$$\ldots, x_{t-l-1}, \mathrlap{\overbrace{\phantom{x_{t-l}, x_{t-l+1}, \ldots, x_{t-1}}}^{\mathbf{x}_{t-1}}}x_{t-l}, \underbrace{x_{t-l+1}, \ldots, x_{t-1}, x_t}_{\mathbf{x}_t}, x_{t+1}, \ldots. \tag{T-D}$$

In other words, $\mathbf{x}_t \in \mathbb{R}^l$ is l most recent observations of x_s, $s = t - l + 1, \ldots, t$. This embedding requires a "burn-in" period of l observations to accumulate at least one full window, unless padding is used.

This embedding of X permits the use of multivariate distance-based anomaly detection techniques. Distance-based anomaly detection uses a distance d on the input space $\mathcal{X}$ to quantify the degree of dissimilarity between observations. Such methods posit that the normal observations are generally closer to their neighbours, as opposed to outlying examples which typically lie farther. If the space $\mathcal{X}$ is $\mathbb{R}^{d\times 1}$ then, the most commonly used distance is the Mahalanobis metric, which takes into account the general shape of the sample and correlations of the data. In the following the distance, induced by the sample $\mathcal{S} = (\mathbf{x}_i)_{i=1}^n$, is $d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})'\hat{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})}$, where $\hat{\Sigma}$ is an estimate of the covariance matrix on $\mathcal{S}$.

The k-NN anomaly detector assigns the abnormality score to some observation $\mathbf{x} \in \mathcal{X}$ based on the neighbourhood proximity measured by the average distance to the k nearest neighbours:

$$\mathrm{NN}(\mathbf{x};k,\mathcal{S}) = \frac{1}{|N_k(\mathbf{x})|}\sum_{\mathbf{y}\in N_k(\mathbf{x})} d(\mathbf{x},\mathbf{y}), \tag{5}$$

where $N_k(\mathbf{x})$ are the k nearest neighbours of $\mathbf{x}$ within $\mathcal{S}$ excluding itself. The detector labels as anomalous any observation with the score exceeding some calibrated threshold. The main drawbacks are high sensitivity to k and poor interpretability of the score $\mathrm{NN}(\mathbf{x};k)$, due to missing natural data-independent scale. Various modifications of this detector are discussed in (Ramaswamy et al., 2000; Angiulli and Pizzuti, 2002; Bay and Schwabacher, 2003; Hautamaki et al., 2004) and (Zhang and Wang, 2006).

Alternatively, it is also possible to use density-based detection methods. For example the schemes proposed in (Breunig et al., 2000; Kriegel et al., 2009) are based on k-NN, but introduce the concept of *local data density*, a score that is inversely related to a distance-based characteristic of a point within its local neighbourhood. Similarly to the k-NN detector, these methods lack a natural scale for the abnormality score. Modifications of this algorithm are discussed in (Jin et al., 2006) and (Papadimitriou et al., 2003).

The combination of the embedding (T-D) and the scoring function (5) produces a non-conformity measure A for conformal procedures in sec. 2. The most suitable procedure is the sliding ICAD, since CAD and the online ICAD are heavier in terms of runtime complexity (tab. 3). However, the sliding ICAD uses a fixed proper training sample for *reference*, which may not reflect potential non-stationarity. Therefore we propose a modification called the Lazy Drifting Conformal Detector (LDCD) which adapts to normal regime non-stationarity, such as quasi-periodic or seasonal patterns. The LDCD procedure is conceptually similar to the sliding ICAD, and thus is expected to provide similar validity guarantees at least in the true iid case. The main challenge is to assess the effects of the calibration scores within the same window being computed on different windows of the training stream.

For the observed sequence $(\mathbf{x}_t)_{t\geq 1}$, the LDCD maintains two fixed-size separate samples at each moment $t \geq n+m$: the training set $\mathcal{T}_t = (\mathbf{x}_{t-N+i})_{i=0}^{n-1}$ of size n ($N = m+n$) and the calibration **queue** $\mathcal{A}_t$ of size m. The sample $\mathcal{T}_t$ is used as the *reference* sample for conformal scoring as in (1). The calibration **queue** $\mathcal{A}_t$ keeps m most recent non-conformity scores given by $\alpha_s = A(\mathcal{T}_s, \mathbf{x}_s)$ for $s = t-m,\ldots,t-1$. At each $t \geq n+m$ the samples $\mathcal{A}_t$ and $\mathcal{T}_t$ look as follows:

$$\begin{array}{rl}
\text{data:} & \ldots,\overbrace{\mathbf{x}_{t-m-n},\ldots,\mathbf{x}_{t-m-1}}^{\mathcal{T}_t\ \text{training}},\quad \mathbf{x}_{t-m},\ldots,\mathbf{x}_{t-1},\quad \overset{\text{test}}{\mathbf{x}_t},\ldots \\
\text{scores:} & \ldots,\alpha_{t-m-n},\ldots,\alpha_{t-m-1},\quad \underbrace{\alpha_{t-m},\ldots,\alpha_{t-1}}_{\mathcal{A}_t\ \text{calibration}},\quad \overset{\text{test}}{\alpha_t},\ldots
\end{array}$$

The procedure uses the current test observation $\mathbf{x}_t$ to compute the non-conformity score α_t used to obtain the p-value similarly to (CPv$_m$), but with respect to scores in the calibration queue $\mathcal{A}_t$. At the end of step t the calibration queue is updated by pushing α_t into $\mathcal{A}_t$ and evicting α_{t-m}.

The final conformal k-NN anomaly detector is defined by the following procedure:

1. the time-series $(x_t)_{t\geq 1}$ is embedded into $\mathbb{R}^l$ using (T-D) to get the sequence $(\mathbf{x}_{t+l-1})_{t\geq 1}$;
2. the LDCD uses k-NN average distance (5) for scoring $(\mathbf{x}_t)_{t\geq 1}$.

The proper training sample $\mathcal{T}_t$ for $t = n+m+1$ is initialized to the first n observations of $\mathbf{x}_t$, and the calibration queue $\mathcal{A}_t$ and is populated with the scores $\alpha_{n+s} = \mathrm{NN}(\mathbf{x}_{n+s}; k, \mathcal{T}_{n+m+1})$ for $s = 1, \ldots, m$.

4. Anomaly Detection Benchmark

The Numenta Anomaly Benchmark (NAB), (Lavin and Ahmad, 2015), is a corpus of datasets and a rigorous performance scoring methodology for evaluating algorithms for online anomaly detection. The goal of NAB is to provide a controlled and repeatable environment for testing anomaly detectors on data streams. The scoring methodology permits only automatic online adjustment of hyperparameters to each dataset in the corpus during testing. In this study we supplement the dataset corpus with additional data (sec. 4.1), but employ the default NAB scoring methodology (sec. 4.2).

4.1. Datasets

The NAB corpus contains 58 real-world and artificial time-series with 1000-22000 observations per series. The real data ranges from network traffic and CPU utilization in cloud services to sensors on industrial machines and social media activity. The dataset is labelled manually and collaboratively according to strict and detailed guidelines established by Numenta. Examples of time-series are provided in fig. 1.

We supplement the NAB corpus with Yahoo! S5 dataset, (S5), which was collected to benchmark detectors on various kinds of anomalies including outliers and change-points. The corpus contains 367 tagged real and synthetic time-series, divided into 4 subsets. The first group contains real production metrics of various Yahoo! services, and the other 3 – synthetic time-series with varying trend, noise and seasonality, which include either only outliers, or both outliers and change-points. We keep all univariate time-series from first two groups for benchmarking. Statistics of the datasets in each corpus are given in tab. 1.

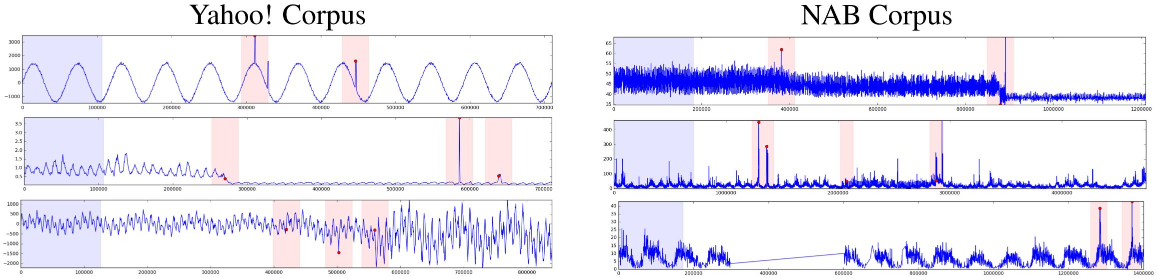

Figure 1: Examples of time-series data from Yahoo! and NAB corpora. The red shaded regions represent the anomaly windows centered at anomalies. The blue region marks the data which the benchmark offers for initial parameter estimation and hyperparameter tuning.

Corpus	Type	datasets	Observations			Total	Anomalies			Total
			Min	Mean	Max		Min	Mean	Max	
Yahoo!	Synthetic	33	1421	1594	1680	52591	1	4.03	8	133
	Real	67	741	1415	1461	94866	0	2.13	5	143
	Total	100	741	1475	1680	147457	0	2.76	8	276
NAB	Synthetic	11	4032	4032	4032	44352	0	0.55	1	6
	Real	47	1127	6834	22695	321206	0	2.43	5	114
	Total	58	1127	6302	22695	365558	0	2.07	5	120

Table 1: Description of the NAB and Yahoo! S5 corpora.

4.2. Performance scoring

Typical metrics, such as precision and recall, are poorly suited for anomaly detection, since they do not incorporate time. The Numenta benchmark proposes a scoring methodology, which favours timely true detections, softly penalizes tardy detections, and harshly punishes false alarms. The scheme uses anomaly windows around each event to categorize detections into true and false positives, and employs sigmoid function to assign weights depending on the relative time of the detection. Penalty for missed anomalies and rewards for timely detections is schematically shown in fig. 2.

The crucial feature of scoring is that all false positives decrease the overall score, whereas only the earliest true positive detection within each window results in a positive contribution. The number of false negatives is the number of anomaly windows in the time-series, with no true positive detections. True negatives are not used in scoring.

The relative costs of true positives (TP), false positives (FP) and false negatives (FN) vary between applications. In NAB this domain specificity is captured by the *application profile*, which multiplicatively adjusts the score contributions of TP, FP, and FN detections. NAB includes three prototypical application profiles, tab. 2. The "Standard" application profile mimics symmetric costs of misdetections, while the "low FP" and "low FN" profiles penalize either overly optimistic or conservative detectors, respectively. For the anomaly window of size $\approx 10\%$ of the span of the time-series, the standard profile assigns relative weights so that random detections made 10% of the time get on average a zero final score, (Lavin and Ahmad, 2015).

If X is the time-series with labelled anomalies, then the NAB score for a given detector and application profile is computed as follows. Each detection is matched to the anomaly window with the nearest right end after it. If τ is the relative position of a detection with respect to the right end

Metric	A_{TP}	A_{FP}	A_{TN}	A_{FN}
Standard	1.0	-0.11	1.0	-1.0
LowFP	1.0	-0.22	1.0	-1.0
LowFN	1.0	-0.11	1.0	-2.0

Table 2: The detection rewards of the default application profiles in the benchmark.

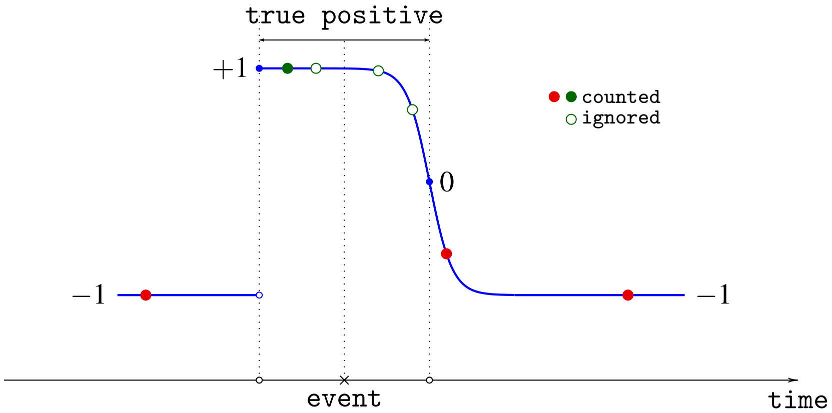

Figure 2: Score weighting in NAB: **all** detections outside the window are **false alarms**, whereas only **the earliest** detection within the window is a **true positive**, and later detections are ignored.

of the anomaly window of width W, then the score of this detection is

$$\sigma(\tau) = \begin{cases} A_{FP}, & \text{if } \tau < -W\,; \\ (A_{TP} - A_{FP})\bigl(1 + e^{5\tau}\bigr)^{-1} + A_{FP}, & \text{otherwise}\ . \end{cases}$$

The overall performance of the detector over X under profile A is the sum of the weighted rewards from individual detections and the impact of missing windows. It is given by

$$S^A_{\mathtt{det}}(X) = \sum_{d \in D_{\mathtt{det}}(X)} \sigma(\tau_d) + A_{FN} f_{\mathtt{det}}\,,$$

where $D_{\mathtt{det}}(X)$ is the set of all alarms fired by the detector on the stream X, τ_d is the relative position of a detection $d \in D_{\mathtt{det}}(X)$, and $f_{\mathtt{det}}$ is the number of anomaly windows which cover no detections at all. The raw benchmark score $S^A_{\mathtt{det}}$ of the detector is the sum of scores on each dataset in the benchmark corpus: $\sum_X S^A_{\mathtt{det}}(X)$.

The final NAB score takes into account the detector's responsiveness to anomalies and outputs a normalized score, (Lavin and Ahmad, 2015), computed by

$$\mathtt{NAB\_score}^A_{\mathtt{det}} = 100 \frac{S^A_{\mathtt{det}} - S^A_{\mathtt{null}}}{S^A_{\mathtt{perfect}} - S^A_{\mathtt{null}}}\,,$$

where $S_{\mathtt{perfect}}$ and $S_{\mathtt{null}}$ are the scores, respectively, for the detector, which generates true positives only, and the one which outputs no alarms at all. The range of the final score for any default profile is $(-\infty, 100]$, since the worst detector is the one which fires only false positive alarms.

5. Benchmark Results

In this section we analyse the runtime complexity of the proposed method (sec. 3) and conduct a comparative study on the anomaly benchmark dataset (sec. 4).

Tab. 3 gives the worst case runtime complexity for the conformal procedures in terms of the worst case complexity of the NCM $A(X_{:t}, \mathbf{x})$, denoted by $c_A(t)$. The CAD procedure is highly

method	Prediction on series $(\mathbf{x}_s)_{s=1-n}^{T}$	
	Scores	Pv
LDCD	$T c_A(n)$	$T m$
ICAD (sliding)	$T c_A(n)$	$T m$
ICAD (online)	$T c_A(n)$	$T \log T$
CAD	$\sum_{t=1}^{T} (t+n) c_A(t+n-1)$	$nT + \frac{1}{2} T(T+1)$

Table 3: Worst case runtime complexity of conformal procedures on $(\mathbf{x}_s)_{s=1-n}^{T}$, n is the length of the train sample.

computationally complex: for each $\mathbf{x}_t$ computing (1) requires a leave-one-out-like run of A over the sample of size $t+n$ and a linear search through new non-conformity scores. In the online ICAD it is possible to maintain a sorted array of non-conformity scores and thus compute each p-value via the binary search and update the scores in one evaluation of A on $\mathbf{x}_t$ and the *reference* train sample. In the sliding ICAD and the LDCD updating the calibration queue requires one run of A as well, but computing the p-value takes one full pass through m scores. The key question therefore is how severe the reliability penalty in (4) is, how well each procedure performs under non-stationarity or quasi-periodicity.

In sec. 4 we described a benchmark for testing detector performance based on real-life datasets and scoring technique, which mimics the actual costs of false negatives and false alarms. Almost all datasets in the Numenta Benchmark and Yahoo! S5 corpora exhibit signs of quasi-periodicity or non-stationarity. We use this benchmark to objectively measure the performance of the conformal k-NN detector, proposed in sec. 3.

The benchmark testing instruments provide each detector with the duration of the "probationary" period, which is 15% of the total length of the currently used time-series. Additionaly, the benchmark automatically calibrates each detector by optimizing the alarm decision threshold. We use the benchmark suggested thresholds and the probationary period duration as the size n of the sliding historical window for training and the size of the calibration queue m.

To measure the effect of conformal p-values on the performance we also test a basic k-NN detector with a heuristic rule to assign confidence. Similarly to sliding train and calibration samples in the proposed LDCD k-NN, the baseline k-NN detector uses the train sample $\mathcal{T}_t$ as in sec. 3, to compute the score of the t-th observation with (5):

$$\alpha_t = \mathrm{NN}(\mathbf{x}_t; k, \mathcal{T}_t)\,.$$

Then the score is dynamically normalized to a value within the $[0,1]$ range with a heuristic (DynR):

$$\mathtt{Pv}_t = \frac{\max_{i=0}^{m} \alpha_{t-i} - \alpha_t}{\max_{i=0}^{m} \alpha_{t-i} - \min_{i=0}^{m} \alpha_{t-i}}\,. \tag{DynR}$$

The conformal k-NN detector using the LDCD procedure performs the same historical sliding along the time-series, but its p-value is computed with (CPv$_m$) (sec. 3):

$$\mathtt{Pv}_t = \frac{1}{m+1} \Bigl| \{ i = 0, \ldots, m \,:\, \alpha_{t-i} \geq \alpha_t \} \Bigr|\,. \tag{LDCD}$$

Corpus	p-value	LowFN	LowFP	Standard
Numenta	DynR	-9.6	-185.7	-54.9
	LDCD	4.3	-143.8	-34.0
	DynR w. pruning	63.0	36.2	54.9
	LDCD w. pruning	64.1	42.6	56.8
Yahoo!	DynR	50.0	0.3	36.1
	LDCD	50.1	0.4	36.1
	DynR w. pruning	68.2	56.4	63.8
	LDCD w. pruning	68.8	56.9	64.3

Table 4: NAB scores of the k-NN detector $(27, 19)$ on the Numenta and Yahoo! S5 corpora.

The value $p_t = 1 - \mathtt{Pv}_t$ is the conformal abnormality score returned by each detector for the observation x_t.

We report the experiment results on two settings of k and l hyperparameters: $(27, 19)$ and $(1, 1)$ for the number of neighbours k and the (T-D) embedding dimension l respectively. The seemingly arbitrary setting $(27, 19)$ achieved the top-3 performance in the Numenta Anomaly Detection challenge, (Numenta, 2016). These hyperparameter values were tuned via grid search over the accumulated performance on the combined corpus of ≈ 400 time series, which makes the chosen parameters unlikely to overfit the data.

Preliminary experimental results have revealed that the LDCD k-NN detector has adequate anomaly coverage, but has high false positive rate. In order to decrease the number of false alarms, we have employed the following ad hoc pruning strategy in both detectors:

- output $p_t = 1 - \mathtt{Pv}_t$ for the observation x_t, and if p_t exceeds 99.5% fix the output at 50% for the next $\frac{n}{5}$ observations.

The results for k-NN detector with 27 neighbours and 19-dimensional embedding (T-D) are provided in table 4.

The key observation is that indeed the k-NN detector with the LDCD confidence scores performs better than the baseline DynR detector. At the same time the abnormality score produced by the dynamic range heuristic are not probabilistic in nature, whereas the conformal confidence scores of the k-NN with the LDCD are. The rationale behind this is that conformal scores take into account the full distribution of the calibration set, whereas the DynR, besides being simple scaling, addresses only the extreme values of the scores.

Tab. 5 shows the final scores for the k-NN detector with 1 neighbour and no embedding ($l = 1$). The table illustrates that the conformal LDCD procedure works well even without alarm thinning. Heuristically, this can be explained by observing that LDCD procedure on the k-NN with 1-D embeddings in fact a sliding-window prototype-based distribution support estimate. Furthermore, the produced p-values (LDCD) are closely related to the probability of an extreme observation relative to the current estimate of the support.

Tables 6 and 7 show the benchmark performance scores for detectors, which were competing in the Numenta challenge, (Numenta, 2016).

Corpus	*p*-value	LowFN	LowFP	Standard
Numenta	DynR	-167.0	-658.4	-291.0
	LDCD	62.3	34.8	53.8
	DynR w. pruning	52.2	4.2	39.0
	LDCD w. pruning	62.7	30.7	53.5
Yahoo!	DynR	30.8	-20.7	16.9
	LDCD	47.7	21.5	37.6
	DynR w. pruning	50.6	35.2	44.8
	LDCD w. pruning	53.8	36.2	46.9

Table 5: NAB scores of the *k*-NN detector $(1,1)$ on the Numenta and Yahoo! S5 corpora.

Detector	LowFN	LowFP	Standard
27-NN $l = 19$ LDCD w. pruning	68.8	56.9	64.3
1-NN $l = 1$ LDCD w. pruning	53.8	36.2	46.9
relativeEntropy	52.5	40.7	48.0
Numenta	44.4	37.5	41.0
Numenta™	42.5	36.6	39.4
bayesChangePt	43.6	17.6	35.7
windowedGaussian	40.7	25.8	31.1
skyline	28.9	18.0	23.6
Random ($p_t \sim \mathcal{U}[0,1]$)	47.2	1.2	29.9

Table 6: The performance of various detectors on the Yahoo! S5 dataset.

6. Conclusion

In this paper we proposed a conformal *k*-NN anomaly detector for univariate time series, which uses sliding historical windows both to embed the time series into a higher dimensional space for *k*-NN and to keep the most relevant observations to explicitly address potential quasi-periodicity. The proposed detector was tested using a stringent benchmarking procedure (Lavin and Ahmad, 2015), which mimics the real costs of timely signals, tardy alarms and misdetections. Furthermore we supplemented the benchmark dataset corpus with Yahoo! S5 anomaly dataset to cover more use-cases. The results obtained in sec. 5 demonstrate that the conformal *k*-NN has adequate anomaly coverage rate and low false negative score. The cases, when the conformal LDCD scores required the use of a signal pruning step, were also the cases when the baseline *k*-NN detector was over-sensitive. Nevertheless, in all cases, conformal abnormality confidence scores improved the benchmark scores.

Numenta held a detector competition in 2016 in which the prototype of the proposed procedure, (Burnaev and Ishimtsev, 2016), took the third place, (Numenta, 2016), competing against much more complex methods based on cortical memory, neural networks, etc. The favourable results on the NAB corpus (sec. 5) suggest that the theoretical foundations of the LDCD procedure,

Detector	LowFN	LowFP	Standard
27-NN $l = 19$ LDCD w. pruning	64.1	42.6	56.8
1-NN $l = 1$ LDCD w. pruning	62.7	30.7	53.5
Numenta	74.3	63.1	70.1
Numenta™	69.2	56.7	64.6
relativeEntropy	58.8	47.6	54.6
windowedGaussian	47.4	20.9	39.6
skyline	44.5	27.1	35.7
bayesChangePt	32.3	3.2	17.7
Random ($p_t \sim \mathcal{U}[0,1]$)	25.9	5.8	16.8

Table 7: The performance of various detectors on the Numenta dataset.

specifically the assumptions required for the proper validity guarantee, should be subject of further research. Besides the validity guarantees, the effects of the violations of the iid assumption should be investigated as well, especially since the embedded time-series vectors overlap.

Acknowledgments

The research, presented in Section 5 of this paper, was supported by the RFBR grants 16-01-00576 A and 16-29-09649 ofi_m; the research, presented in other sections, was also supported by the Russian Science Foundation grant (project 14-50-00150).

References

Stephane Alestra, Cristophe Bordry, Cristophe Brand, Evgeny Burnaev, Pavel Erofeev, Artem Papanov, and Cassiano Silveira-Freixo. Application of rare event anticipation techniques to aircraft health management. In *Advanced Materials Research*, volume 1016, pages 413–417. Trans Tech Publ, 2014.

Fabrizio Angiulli and Clara Pizzuti. Fast outlier detection in high dimensional spaces. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 15–27. Springer, 2002.

A. V. Artemov and Evgeny Burnaev. Optimal estimation of a signal, observed in a fractional gaussian noise. *Theory Probab. Appl.*, 60(1):126–134, 2015a.

Alexey Artemov and Evgeny Burnaev. Ensembles of detectors for online detection of transient changes. In *Eighth International Conference on Machine Vision*, pages 98751Z–98751Z. International Society for Optics and Photonics, 2015b.

Alexey Artemov and Evgeny Burnaev. Detecting performance degradation of software-intensive systems in the presence of trends and long-range dependence. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 29–36, Dec 2016. doi: 10.1109/ICDMW.2016.0013.

Alexey Artemov, Evgeny Burnaev, and Andrey Lokot. Nonparametric decomposition of quasi-periodic time series for change-point detection. In *Eighth International Conference on Machine Vision*, pages 987520–987520. International Society for Optics and Photonics, 2015.

M. F. Augusteijn and B. A. Folkert. Neural network classification and novelty detection. *International Journal of Remote Sensing*, 23(14):2891–2902, 2002. doi: 10.1080/01431160110055804.

Guilherme Barreto and Leonardo Aguayo. *Time Series Clustering for Anomaly Detection Using Competitive Neural Networks*, pages 28–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-02397-2. doi: 10.1007/978-3-642-02397-2_4. URL http://dx.doi.org/10.1007/978-3-642-02397-2_4.

Stephen D Bay and Mark Schwabacher. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38. ACM, 2003.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

E. Burnaev and V. Ishimtsev. Conformalized density- and distance-based anomaly detection in time-series data. *ArXiv e-prints*, August 2016.

E. Burnaev, P. Erofeev, and A. Papanov. Influence of resampling on accuracy of imbalanced classification. In *Eighth International Conference on Machine Vision*, pages 987521–987521. International Society for Optics and Photonics, 2015a.

E. Burnaev, P. Erofeev, and D. Smolyakov. Model selection for anomaly detection. In *Eighth International Conference on Machine Vision*, pages 987525–987525. International Society for Optics and Photonics, 2015b.

Evgeny Burnaev. Disorder problem for poisson process in generalized bayesian setting. *Theory of Probability & Its Applications*, 53(3):500–518, 2009.

Evgeny Burnaev and Ivan Nazarov. Conformalized kernel ridge regression. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 45–52, Dec 2016. doi: 10.1109/ICMLA.2016.0017.

Evgeny Burnaev, E. A. Feinberg, and A. N. Shiryaev. On asymptotic optimality of the second order in the minimax quickest detection problem of drift change for brownian motion. *Theory of Probability & Its Applications*, 53(3):519–536, 2009.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

Haimonti Dutta, Chris Giannella, Kirk D Borne, and Hillol Kargupta. Distributed top-k outlier detection from astronomy catalogs using the demac system. In *SDM*, pages 473–478. SIAM, 2007.

Pedro Galeano, Daniel Peña, and Ruey S. Tsay. Outlier detection in multivariate time series by projection pursuit. *Journal of the American Statistical Association*, 101(474):654–669, 2006. doi: 10.1198/016214505000001131. URL http://dx.doi.org/10.1198/016214505000001131.

A.B. Gardner, A.M. Krieger, G. Vachtsevanos, and B. Litt. One-class novelty detection for seizure analysis from intracranial eeg. *Journal of Machine Learning Research*, 7:1025–1044, 2006. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-33745235087&partnerID=40&md5=799e774d42e936de78fcc66a032a4498. cited By 140.

Ville Hautamaki, Ismo Karkkainen, and Pasi Franti. Outlier detection using k-nearest neighbour graph. In *Proc. of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ICPR '04, pages 430–433. IEEE Computer Society, 2004. ISBN 0-7695-2128-2. doi: 10.1109/ICPR.2004.671. URL http://dx.doi.org/10.1109/ICPR.2004.671.

Simon Hawkins, Hongxing He, Graham J. Williams, and Rohan A. Baxter. Outlier detection using replicator neural networks. In *Data Warehousing and Knowledge Discovery: 4th International Conference, DaWaK 2002 Aix-en-Provence, France, September 4–6, 2002 Proceedings*, DaWaK 2000, pages 170–180. Springer Berlin Heidelberg, 2002. ISBN 978-3-540-46145-6. doi: 10.1007/3-540-46145-0_17.

Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recogn.*, 40(3):863–874, March 2007. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.07.009.

Wen Jin, Anthony KH Tung, Jiawei Han, and Wei Wang. Ranking outliers using symmetric neighborhood relationship. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 577–593. Springer, 2006.

Ian Jolliffe. *Principal Component Analysis*. John Wiley & Sons, Ltd, 2014. ISBN 9781118445112. doi: 10.1002/9781118445112.stat06472.

Arun Kejariwal. Introducing practical and robust anomaly detection in a time series. *Twitter Engineering Blog. Web*, 15, 2015.

Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. Loop: local outlier probabilities. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1649–1652. ACM, 2009.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pages 38–44. IEEE, 2015.

Rikard Laxhammar. *Conformal anomaly detection*. PhD thesis, Ph. D. dissertation, University of Skövde, Skövde, Sweden, 2014.[Online]. Available: http://www. diva-portal. org/smash/get/-diva2: 690997/FULLTEXT02, 2014.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74 (1-2):67–94, 2015.

H. j. Lee and S. J. Roberts. On-line novelty detection using the kalman filter and extreme value theory. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, Dec 2008. doi: 10.1109/ICPR.2008.4761918.

J. Ma and S. Perkins. Time-series novelty detection using one-class support vector machines. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 3, pages 1741–1745 vol.3, July 2003. doi: 10.1109/IJCNN.2003.1223670.

Numenta. The numenta anomaly benchmark competition for real-time anomaly detection. http://numenta.com/blog/2016/08/10/numenta-anomaly-benchmark-nab-competition-2016-winners/, 2016. Accessed: 2017-03-06.

Spiros Papadimitriou, Hiroyuki Kitagawa, Phillip B Gibbons, and Christos Faloutsos. Loci: Fast outlier detection using the local correlation integral. In *Data Engineering, 2003. Proceedings. 19th International Conference on*, pages 315–326. IEEE, 2003.

Marco A.F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215 – 249, 2014. ISSN 0165-1684. doi: http://dx.doi.org/10.1016/j.sigpro.2013.12.026. URL http://www.sciencedirect.com/science/article/pii/S016516841300515X.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *ACM SIGMOD Record*, volume 29, pages 427–438. ACM, 2000.

Yahoo! S5. Yahoo! webscope s5: A labeled anomaly detection dataset, version 1.0. http://webscope.sandbox.yahoo.com. Accessed: 06/27/2016.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5):1299–1319, July 1998. ISSN 0899-7667. doi: 10.1162/089976698300017467.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. Technical report, DTIC Document, 2003.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92(2): 349–376, 2013. ISSN 1573-0565. doi: 10.1007/s10994-013-5355-6. URL http://dx.doi.org/10.1007/s10994-013-5355-6.

G. Williams, R. Baxter, He Hongxing, S. Hawkins, and Gu Lifang. A comparative study of rnn for outlier detection in data mining. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 709–712, 2002. doi: 10.1109/ICDM.2002.1184035.

Ji Zhang and Hai Wang. Detecting outlying subspaces for high-dimensional data: The new task, algorithms, and performance. *Knowl. Inf. Syst.*, 10(3):333–355, October 2006. ISSN 0219-1377. doi: 10.1007/s10115-006-0020-z. URL http://dx.doi.org/10.1007/s10115-006-0020-z.