Anomaly Detection for Symbolic Time Series Representations of Reduced Dimensionality

Konstantinos Bountrogiannis[1,2], George Tzagkarakis[1], and Panagiotis Tsakalides[1,2]
[1] *Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece*
[2] *Department of Computer Science, University of Crete, Heraklion, Greece*
E-mails: kbountro@ics.forth.gr, gtzag@ics.forth.gr, tsakalid@ics.forth.gr

***Abstract*—The systematic collection of data has become an intrinsic process of all aspects in modern life. From industrial to healthcare machines and wearable sensors, an unprecedented amount of data is becoming available for mining and information retrieval. In particular, anomaly detection plays a key role in a wide range of applications, and has been studied extensively. However, many anomaly detection methods are unsuitable in practical scenarios, where streaming data of large volume arrive in nearly real-time at devices with limited resources. Dimensionality reduction has been excessively used to enable efficient processing for numerous high-level tasks. In this paper, we propose a computationally efficient, yet highly accurate, framework for anomaly detection of streaming data in lower-dimensional spaces, utilizing a modification of the symbolic aggregate approximation for dimensionality reduction and a statistical hypothesis testing based on the Kullback-Leibler divergence.**

***Index Terms*—Online anomaly detection, kernel density estimator, symbolic representations, mode-bounding Lloyd-Max quantizer**

I. Introduction

Anomaly detection has a prominent role in monitoring and predicting critical processes and phenomena. It has been extensively applied in various distinct application scenarios employing both non-streaming and streaming data [1], such as network intrusion detection, fraud detection, detection of data abnormalities or instrumentation errors in the medical domain, novelty detection in textual data, etc. Focusing on the case of streaming data arriving in (near) real time, necessitates the design of fast anomaly detection algorithms, whereas anomaly detection in edge processing applications imposes additional computational constraints due to the limited power and memory resources available on-board small sensing devices. To address these issues, this work proposes an unsupervised, non-parametric method, characterized by low power and memory demands, for anomaly detection in unidimensional data, whereas guidelines for a generalization to higher-dimensional data are given in the last section.

More specifically, fast processing of the received streaming data is enabled by first applying a dimensionality reduction step. For this, we rely on the framework of symbolic aggregate approximation (SAX) [2]. SAX is a well-established method that transforms a given time series into a lower-dimensional symbolic sequence, and is widely used in a variety of data mining applications (ref. [3]–[7]). At the core of our method is a modification of the conventional SAX, in order to construct more accurately the symbolic representation by better adapting to the underlying data-generating process. The design of our SAX-based anomaly detection method is further motivated by the fact that symbolic representations can be coupled efficiently with the Kullback-Leibler Goodness-of-Fit process (ref. [8]) in order to track the time-evolving distribution of the generated symbols. The efficiency of our proposed method is evaluated by employing the Numenta Anomaly Benchmark (NAB) [9]. NAB consists of a highly comparative scoring system and provides a wide variety of real-world labeled datasets from diverse sources.

Other techniques which incorporate dimensionality reduction for efficient anomaly detection, either run in a supervised fashion, or make assumptions for the data statistics. For instance, the method proposed in [10] is supervised, whilst the method in [11] assumes Gaussian distribution of the data, which is very often inaccurate.

The rest of the paper is organized as follows: Section II introduces the core components of our method and discusses its differences with prior related works. Section III describes in detail our proposed anomaly detection method for symbolic time series representations, whilst Section IV evaluates its performance, whilst also investigating the relation between the anomaly detection accuracy and the degree of dimensionality reduction. Finally, Section V summarizes the main outcomes of this work and gives directions for further extensions.

II. Background and Related Works

In this section, we briefly introduce the building blocks of our method, namely, (i) the SAX framework, a modified version of which is used for dimensionality reduction by transforming a time series into a sequence of symbols in a data-adaptive fashion, and (ii) the Kullback-Leibler Goodness-of-Fit criterion, which is used to define our anomaly detection rule by tracking the time-varying distribution of the symbolic sequence.

A. Symbolic Aggregate Approximation

Let $U = (u_1, u_2, \ldots, u_N)$ be a discrete time series of N data samples, where u_i is the ith sample. The first step of SAX implements a piecewise aggregate approximation (PAA), which transforms the given time series U into a vector

This work is funded by the Interreg V-A Greece-Cyprus 2014-2020 programme, co-financed by the European Union (ERDF) and National Funds of Greece and Cyprus, under the project SmartWater2020.

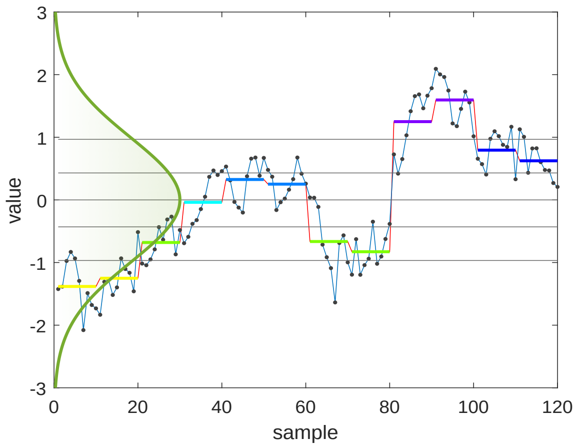

Fig. 1: SAX transformation. Each M-sized segment is averaged and assigned a codeword (color-coded) according to the α-quantiles of the standard Gaussian distribution. Here, $M = 12$ and $\alpha = 6$.

$Y = (y_1, \dots, y_M)$, with $M < N$. For this, U is divided into M segments of equal size and the average value is calculated for each segment. The ratio M/N determines the degree of dimensionality reduction.

In the second step, Y is transformed into a symbolic sequence S, by mapping the averages into a predefined set of symbols. More precisely, the original time series U is typically Z-normalized to mean zero and standard deviation equal to one, whereas it is assumed to follow a standard Gaussian distribution. Under this assumption, the M averages in Y are quantized within α equiprobable intervals under the standard Gaussian probability density function curve. Each quantization interval is assigned a codeword from an alphabet A with cardinality $|A| = \alpha$. The result is a symbolic sequence of length M. Fig. 1 illustrates the SAX process. Notably, the SAX transformation is fast and can be executed in real time. Moreover, processing the lower-dimensional symbolic sequence is much more efficient than processing the raw data.

B. Kullback-Leibler Goodness-of-Fit

The online anomaly detection step of our method is motivated by a hypothesis testing approach proposed by [8], which employs a goodness-of-fit test based on the Kullback-Leibler divergence. Hereafter, we denote the method introduced in [8] by “KL GoF”.

Specifically, for two discrete random variables X and Y defined in the same probability space $\mathcal{X}$, the Kullback-Leibler divergence of Y from X is defined as follows,

$$D(X\|Y) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{P_X(x)}{P_Y(x)} \ . \tag{1}$$

Let Q and $\hat{Q}$ be two discrete random variables. If the probability mass function of $\hat{Q}$ is the empirical mass function estimated from N samples of Q, the following result holds,

$$2N \cdot D(\hat{Q}\|Q) \longrightarrow \chi^2 \quad \text{uniformly,} \tag{2}$$

where χ^2 is the chi-squared distribution with $|\mathcal{X}| - 1$ degrees of freedom, with $|\mathcal{X}|$ being the cardinality of the sample space $\mathcal{X}$. The convergence rate is controlled by the quantile function $F^{-1}_{\chi^2}(\gamma)$. Specifically, the empirical random variable $\hat{Q}$ is considered to be asymptotically dissimilar from Q, if the following condition holds,

$$2N \cdot D(\hat{Q}\|Q) > F^{-1}_{\chi^2}(\gamma) \ ,$$

where γ is typically set equal to 0.05 or 0.01, as noted in [8].

The “KL GoF” method first performs a uniform quantization of the time series samples, with quantization intervals of equal size. Then, the empirical distribution of the quantized samples is estimated in sliding windows of length N. A window is flagged as “anomalous” if the associated empirical distribution is not close, according to the above threshold-based rule, to any of the distributions in the past windows.

III. Proposed Online Anomaly Detection Method

In this section, the proposed online anomaly detection method is analyzed. Specifically, a SAX-based dimensionality reduction step is performed first, followed by a statistical criterion based on the KL GoF to classify a window as anomalous or not.

A. Data-driven SAX-based Symbolic Representation

Although SAX yields a high representation accuracy for data following Gaussian statistics, however, its performance may degrade dramatically in more generic cases. Indeed, in practical scenarios, where the underlying probability distribution of a time series deviates significantly from a Gaussian, the accuracy of a SAX-based low-dimensional symbolic representation diminishes. To address this limitation, an alternative quantization method is employed, as described below.

1) Data-adaptive SAX quantization: An initial probability density function (pdf) is estimated for the data source from the first PAA segments of the associated time series. In a real streaming data scenario, a set of historical data from the same source can be employed. The pdf estimation is performed by means of a kernel density estimator (KDE) [12]. The KDE depends on two parameters, namely, the kernel function and the smoothness parameter. In our implementation, we employ the Epanechnikov kernel [13] and the Silverman’s rule [14, Sec. 3] for setting the smoothness parameter. Having estimated the pdf, a set of optimal quantization intervals is derived by applying the Lloyd-Max algorithm [15].

An illustration of these two steps (KDE and Lloyd-Max) is shown in the top plot of Fig. 2. Our modified SAX-based step yields symbolic representations of increased accuracy by better adapting to the pdf of the streaming data, without relying on any prior assumption for the statistics of the data source.

Nevertheless, it is important to highlight a problem, which is inherent to the way Lloyd-Max forms the quantization intervals and clusters the PAA segments in the distinct intervals. Specifically, the calculated intervals often split the true clusters (i.e., the intervals around the modes) of the source’s pdf (see top plot in Fig. 2). This is undesirable, since, although data

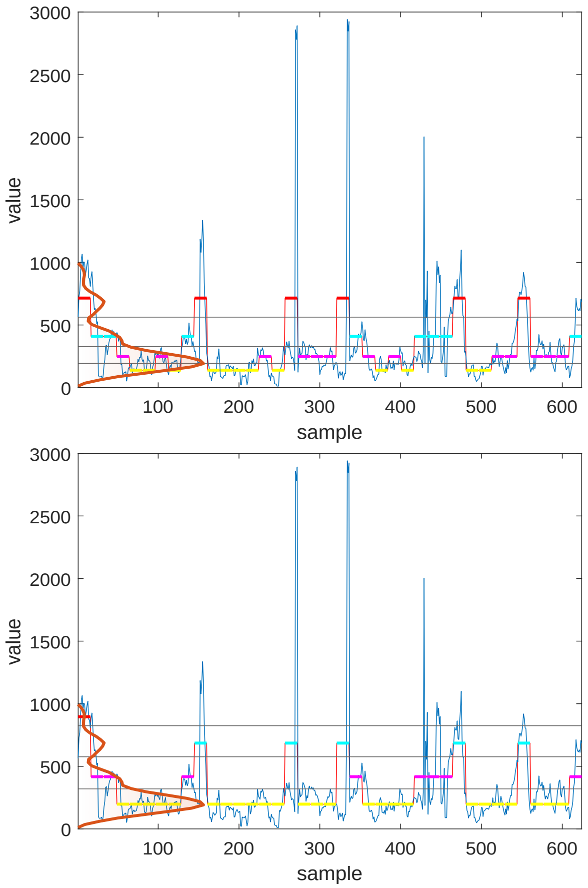

Fig. 2: Data-adaptive SAX via KDE and Lloyd-Max. The estimated pdf is drawn on the left of each plot. (a) Top: conventional Lloyd-Max, (b) Bottom: proposed mode-bounding Lloyd-Max. Note that the dominant mode is splitted in two intervals by the conventional Lloyd-Max, while a more accurate bounding is achieved by the mode-bounding Lloyd-Max.

Algorithm 1 Mode-bounding Lloyd-Max

```
Inputs: $\alpha, k$
Compute $B = k \cdot \alpha$ quantization intervals with bounds $\mathbf{M} = [m_1, \ldots, m_{B+1}]$ via Lloyd-Max.
while $|\mathbf{M}| > \alpha$ do
    remove $m_{j^*}$ from $\mathbf{M}$, with $j^* = \arg\min_j (m_j - m_{j-1})$
end while
```

falling in the same mode are assumed to be similar, however, splitting a mode may yield a misinterpretation as of the data belonging to distinct subclusters being significantly different. To overcome this drawback, a modification of Lloyd-Max quantizer is proposed below.

2) Mode-bounding Lloyd-Max: The proposed method is a simple modification of the conventional Lloyd-Max quantizer, aiming at better detecting the modes of a probability density function. Specifically, let α be a predefined number of quantization intervals. The mode-bounding Lloyd-Max quantizer first estimates a number of quantization intervals $k \cdot \alpha$, $k \in \mathbb{N}$. Then, a merging of the smallest intervals with their neighbours is carried out iteratively until the α largest intervals are left. This process is summarized in Algorithm 1, and illustrated in the bottom plot of Fig. 2.

The idea behind the proposed quantizer is that a finer quantization will fine-slice the peaks of the modes in the pdf, as the density around the peaks is high. Subsequently, merging the smallest intervals implicitly merges the intervals around the peaks, leaving the boundaries of the modes intact.

B. *SAX-KL Anomaly Detector*

The proposed anomaly detection method is applied in a lower-dimensional space by incorporating the modified SAX and Lloyd-Max algorithms described above, in conjunction with the KL GoF test overviewed in Sec. II-B.

In particular, working in a sliding window fashion, given the alphabet size α and the dimensionality reduction ratio M/N, the current window of length N is first transformed into a symbolic sequence S of length M. The transformation is carried out by the data-driven SAX-based method described in Sec. III-A.

Having generated the symbolic sequence of length M for the current window, the frequency distribution of the α alphabet symbols is calculated next for the M-sized sequence. Then, the goodness-of-fit test, described in Sec. II-B, is applied to classify the window as anomalous or not. Here, the cardinality of the sample space of the symbols in S, which is required for the definition of the chi-squared distribution (2), is equal to the alphabet size α.

Note that both α and M/N determine the degree of compressibility achieved by the symbolic sequence, and hence the computational and memory savings of the overall anomaly detection system. However, α and M/N affect the detector's efficiency in a different way. In the optimal case, the alphabet size should match the number of "states" in the given time series. For instance, a binary source with additive noise can be represented efficiently with a binary alphabet. Likewise, a CPU activity log may be represented adequately with an alphabet size equal to the expected number of activity states. On the other hand, the dimensionality reduction ratio should preserve the raw data patterns.

We emphasize again that our online anomaly detection method is distribution-free, by not relying on any prior assumption on the underlying data distribution. Furthermore, it does not require access to past data, but only to the probability distributions of the past (symbolic) windows. Memory-wise, this is more efficient, since a window of length N is represented by only α numbers, i.e., the probabilities of its symbols. Also, under high memory constraints, only the latest probability distributions can be saved in memory and utilized by the KL GoF test. Hereafter, our proposed anomaly detection method is denoted by "SAX-KL".

IV. Experimental Evaluation

In this section, the anomaly detection accuracy of our method is evaluated and compared against the results reported by NAB (ref. Section I). Specifically, the performance of anomaly detection methods is evaluated in terms of the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) rates. Standard performance metrics include the following information retrieval measures,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \ , \tag{3}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \ , \tag{4}$$

$$\text{F-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \ . \tag{5}$$

Precision quantifies the correctness of detected anomalies, whereas Recall measures the success in detecting them. F-score is the harmonic mean of Precision and Recall, which provides an overall measurement of the performance.

The above performance metrics are suitable for anomaly detection over batches of data or subsequences over a time window of predefined length (e.g. packets in communication networks, daily stock market prices, traffic in rush hours, etc.). On the other hand, the accuracy of anomaly detection in streaming data, which do not form predefined batches, cannot be evaluated directly with the above metrics. To alleviate this issue, the authors in [9] propose a benchmark algorithm (NAB) to tackle these limitations. The outcome is a scoring system tailored to streaming anomaly detectors, which contains a total of 58 synthetic and real-world streaming datasets with labeled anomalies. Moreover, the algorithm calculates three different scores by weighing FPs and FNs differently: (i) favoring fewer FPs, (ii) favoring fewer FNs, and (iii) a "standardized" score, that balances both FPs and FNs. The scores' values vary between 0 and 100 (the higher the better).

A core concept of NAB is the definition of anomalous windows, i.e., windows centered on anomaly points, with which a true positive is scored according to how early or late within the window it is located (the earlier the better). Naturally, the points that are classified as anomalous within an anomalous window are jointly accounted for as a single true positive. The length of the windows is set heuristically and separately for each dataset. This concept leads to the idea of splitting the streaming data, following the same heuristics as NAB does, into equal-sized windows, either anomalous or anomaly-free, into which the detected anomalies are merged. Adopting this approach allows us to exploit the commonly used performance metrics defined by (3)-(5).

In the following, the performance of our method is evaluated for varying dimensionality reduction ratios, and by employing both the NAB scores and the metrics in (3)-(5) (which are calculated according to the methodology described above). Doing so, we provide a complete quantification of the detector's performance. The results are averaged over 100 Monte Carlo iterations, although the divergence across the iterations was not significant. Regarding NAB, we also compare the scores achieved by our method with some of the currently best performing anomaly detection algorithms, whose scores are obtained directly from the online repository[1] maintained by the authors of [9].

Regarding KL GoF and SAX-KL, the following parameters setting is used for all datasets: window length $N = 50$, alphabet size $\alpha = 8$, and intervals multiplier for the mode-bounding Lloyd-Max (ref. Alg. 1) $k = 4$ (fixed for all the experiments hereafter). The statistic threshold is $\gamma = 10^{-4}$ in the case of no dimensionality reduction (i.e., when $M/N = 1.0$), whilst a larger $\gamma = 5 \cdot 10^{-3}$ is used when $M/N < 1.0$. We emphasize that the aforementioned online repository lists the values of the KL GoF under the name "Relative Entropy", and with scores significantly lower than those reported herein. The reason is that the anomaly detector they employed is executed for $\alpha = 5$ and $\gamma = 10^{-2}$, whereas we found out that our setting achieves higher scores for both methods.

As it can be seen in Table I, when no dimensionality reduction is applied, our anomaly detection method improves the performance of the KL GoF, whilst it clearly competes most of the currently best performing detectors. More importantly, the proposed method uniquely enables anomaly detection in a lower-dimensional space, due to the SAX-based dimensionality reduction method, which is explored next.

The second experiment investigates the effect of the degree of dimensionality reduction on the performance of our method. As described above, our proposed SAX-KL method enables dimensionality reduction of the original time series via the SAX-based step. Table II presents the anomaly detection performance of our method by varying the dimensionality reduction ratio $M/N \in [1/80, 1/1]$ (from large to low dimensionality reduction), in terms of the NAB scores, as well as the Precision and Recall. Furthermore, Fig. 3 shows the respective average F-score as a function of M/N.

Notably, according to the NAB scores in Table II, the performance of our method in the lower-dimensional space is still as good as the best performing detectors (Table I), even for large dimensionality reduction. An interesting observation is that the performance of the detector does not decrease monotonically with the dimensionality reduction. A more thorough study of the effect of the dimensionality reduction ratio on the detector's performance is left as a future work.

Notice also that the running time of SAX-KL and KL GoF method is exactly the same when $M/N = 1.0$, since the training phase of the KDE step is carried out only once during initialization, and thus can be disregarded in the subsequent application of the method on the streaming data. Practically, during the initial application of the detector, a better approximation of the true distribution function can be computed as samples come in and re-initialize the quantization intervals. Nevertheless, KDE's convergence speed is fast for smooth distributions (see for example [16]).

Overall, the method requires small memory and computationally benefits from the dimensionality reduction. As an example, we ran the top 5 detectors from Table I on the real-world dataset "machine_temperature_system_failure" from NAB's collection, which contains 22695 samples, on an Intel i7-6700@3.8GHz. The running time was 123.76 seconds for "Numenta HTM", 24.48 seconds for "CAD OSE", 93.05

[1] https://github.com/numenta/NAB/#scoreboard

TABLE I: NAB scores

Detector	Standard	Low FP	Low FN
Numenta HTM	70.1	63.1	74.3
CAD OSE	69.9	67.0	73.2
SAX-KL	67.1	61.6	71.0
KL GoF	63.2	59.0	66.4
earthgecko Skyline	58.2	46.2	63.9
KNN CAD	58.0	43.4	64.8
Random Cut Forest	51.7	38.4	59.7
Twitter ADVec v1.0.0	47.1	33.6	53.5

The proposed SAX-KL method has been set with $M/N = 1.0$ (i.e., no dimensionality reduction). The parameters of SAX-KL and KL GoF are the same and optimized for NAB's datasets: $N = 50$, $\alpha = 8$, $\gamma = 10^{-4}$.

TABLE II: Performance of the proposed SAX-KL method vs. dimensionality reduction ratio (M/N).

	NAB Scores				
M/N	Standard	Low FP	Low FN	Precision	Recall
1/1	67.1	71.0	61.6	0.3995	0.7955
1/2	63.3	67.5	57.3	0.3623	0.7696
1/4	65.0	69.1	59.3	0.3973	0.7832
1/8	59.7	63.7	54.7	0.3932	0.7237
1/16	60.4	64.8	54.7	0.3524	0.7462
1/32	56.8	61.1	51.8	0.3330	0.7042
1/64	53.3	58.1	47.9	0.2975	0.6888
1/80	51.2	55.8	45.2	0.2520	0.6648

seconds for "earthgecko Skyline" and in our MATLAB implementation of "KL GoF" it was 3.67 seconds. For the proposed "SAX-KL", including the KDE and Lloyd-Max steps, the running times versus M/N are: 5.11 seconds for $M/N = 1/1$, 2.5 seconds for $M/N = 1/2$, 0.65 seconds for $M/N = 1/8$.

V. Discussion and Future Work

This work proposes a new anomaly detection method for symbolic representations of time series, specifically designed for streaming data. At the core of the method is a data-driven SAX-based method, coupled with a modified version of KL GoF [8], which adapts directly to the underlying data distribution. To this end, a KDE-based estimator is combined with a Lloyd-Max quantizer, which clusters the data according to their probability density function. The proposed method achieves similar performance, or even it outperforms, the best performing methods available for streaming data. Most importantly, this is also the case even for large dimensionality reduction ratios (i.e., highly compressed data).

The application of the proposed method is currently limited to the uni-dimensional case. As a further extension, we are interested in generalizing our method for multi-dimensional data, exploiting the intra-dimensional correlation. To this end, a SAX-based transformation can be performed independently for each dimension. Since the data is discretized via the symbolic sequence, the computation of the Kullback-Leibler divergence is still efficient and hence the method can be employed for processing streaming data with low resources. Another extension concerns the design of a mapping from the multi-dimensional data points to the uni-dimensional space via a Hilbert curve [17], due to its strong locality-preserving property, before their subsequent processing.

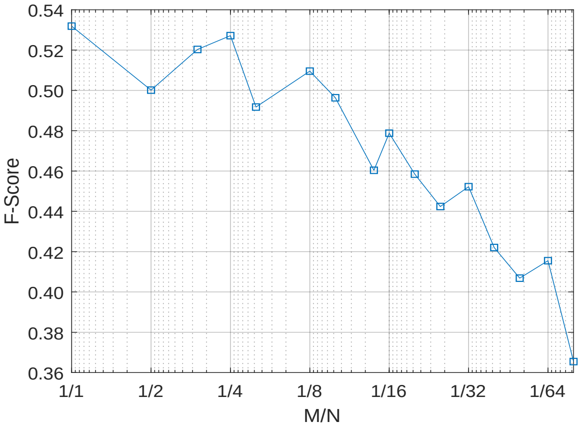

Fig. 3: Average F-score vs. dimensionality reduction ratio.

Lastly, a detector with multiple fidelity settings, determined by selectable dimensionality reduction ratios from a predefined set, is left as a future work. To this end, optimal dimensionality reduction ratios need to be determined, as the detector does not degrade monotonically with the dimensionality reduction. A thorough study of the relation between the dimensionality ratio and the detector's performance, probably with respect to the nature of the data source, is a step towards this direction.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, 07 2009.

[2] J. Lin et al., "Experiencing SAX: A novel symbolic representation of time series," *Data Min. Knowl. Discov.*, vol. 15, pp. 107–144, 08 2007.

[3] P. Ordonez et al., "Using modified multivariate bag-of-words models to classify physiological data," in *2011 IEEE 11th Intl' Conf. on Data Min. Worksh.*, Dec 2011, pp. 534–539.

[4] Y. Wang et al., "Clustering of electricity consumption behavior dynamics toward big data applications," *IEEE Trans. Smart Grid*, vol. 7, no. 5, 2016.

[5] C. Miller, Z. Nagy, and A. Schlueter, "Automated daily pattern filtering of measured building performance data," *Aut. in Constr.*, vol. 49, 2015.

[6] S. Aghabozorgi and Y. W. Teh, "Stock market co-movement assessment using a three-phase clustering method," *Exper. Sys. Appl.*, vol. 41, no. 4, Part 1, pp. 1301 – 1314, 2014.

[7] J. Shieh and E. Keogh, "iSAX: Indexing and mining terabyte sized time series," *Proc. ACM SIGKDD Intl' Conf. Knowl. Disc. Data Min.*, 2008.

[8] C. Wang et al., "Statistical techniques for online anomaly detection in data centers," in *Proc. 12th IFIP/IEEE Intl' Symp. on Integr. Netw. Manag., IM 2011*, 05 2011, pp. 385–392.

[9] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms – the numenta anomaly benchmark," in *2015 IEEE 14th Intl' Conf. on Mach. Lear. and Appl. (ICMLA)*, Dec 2015, pp. 38–44.

[10] Z. Li et al., "Dimensionality reduction for anomaly detection in electrocardiography: A manifold approach," in *2012 Ninth Intl' Conf. on Wear. and Implant. Body Sensor Netw.*, May 2012, pp. 161–165.

[11] A. Juvonen et al., "Online anomaly detection using dimensionality reduction techniques for http log analysis," *Comp. Netw.*, vol. 91, 2015.

[12] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, 09 1962.

[13] V. Epanechnikov, "Non-parametric estimation of a multivariate probability density," *Theory Probab. Appl.*, vol. 14, no. 1, pp. 153–158, 1969.

[14] B. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 1986.

[15] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inform. Theory*, vol. 6, no. 1, pp. 7–12, March 1960.

[16] P. Rigolett and R. Vert, "Optimal rates for plug-in estimators of density level sets," *Bernoulli*, vol. 15, no. 4, pp. 1154–1178, 2009.

[17] B. Moon et al., "Analysis of the clustering properties of Hilbert space-filling curve," *IEEE Trans. on Knowl. and Data Eng.*, vol. 13, 2001.