

Chapter 5

Learning and Adaptation to Detect Changes and Anomalies in High-Dimensional Data



Diego Carrera

Abstract The problem of monitoring a datastream and detecting whether the data generating process changes from normal to novel and possibly anomalous conditions has relevant applications in many real scenarios, such as health monitoring and quality inspection of industrial processes. A general approach often adopted in the literature is to learn a model to describe normal data and detect as anomalous those data that do not conform to the learned model. However, several challenges have to be addressed to make this approach effective in real world scenarios, where acquired data are often characterized by high dimension and feature complex structures (such as signals and images). We address this problem from two perspectives corresponding to different modeling assumptions on the data-generating process. At first, we model data as realization of random vectors, as it is customary in the statistical literature. In this settings we focus on the change detection problem, where the goal is to detect whether the datastream permanently departs from normal conditions. We theoretically prove the intrinsic difficulty of this problem when the data dimension increases and propose a novel non-parametric and multivariate change-detection algorithm. In the second part, we focus on data having complex structure and we adopt dictionaries yielding sparse representations to model normal data. We propose novel algorithms to detect anomalies in such datastreams and to adapt the learned model when the process generating normal data changes.

5.1 Introduction

The general problem we address here is the monitoring of datastreams to detect in the data-generating process. This problem has to be faced in several applications, since the change could indicate an issue that has to be promptly alarmed and solved. In particular, we consider two meaningful examples. Figure 5.1a shows a Scanning Electron Microscope (SEM) image acquired by an inspection system that monitors

D. Carrera (✉)
Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy
e-mail: diego.carrera@polimi.it

© The Author(s) 2020
B. Pernici (ed.), *Special Topics in Information Technology*, PoliMI SpringerBriefs,
https://doi.org/10.1007/978-3-030-32094-2_5

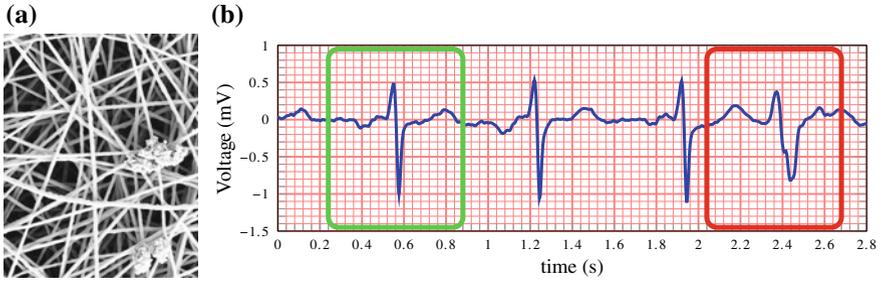


Fig. 5.1 **a** A detail of a SEM image acquired by the considered inspection system depicting nanofibrous material. The two small *beads*, i.e., fiber clots, on the right reduce the quality of the produced material. **b** Example of few seconds of ECG signal containing 4 normal heartbeats (one of which highlighted in green) and an anomalous one (highlighted in red) (Color figure online)

the quality of nanofibrous materials produced by a prototype machine. In normal conditions, the produced material is composed of tiny filaments whose diameter ranges up to 100 nanometers. However, several issues might affect the production process and introduce small defects among the fibers, such as the clots in Fig. 5.1a. These defects have to be promptly detected to improve the overall production quality.

The second scenario we consider is the online and long-term monitoring of ECG signals using wearable devices. This is a very relevant problem as it would ease the transitioning from hospital to home/mobile health monitoring. In this case the data we analyze are the heartbeats. As shown in Fig. 5.1b, normal heartbeats feature a specific morphology, while the shape of anomalous heartbeats, that might be due to potentially dangerous arrhythmias, is characterized by a large variability. Since the morphology of normal heartbeats depends on the user and the position of the device [16], the anomaly-detection algorithm has to be configured every time the user places the device.

Monitoring this kind of datastream raises three main challenges: at first data are characterized by complex structure and high dimension and there is no analytical model able to describe them. Therefore, it is necessary to learn models directly from data. However, only normal data can be used during learning, since acquiring anomalous data can be difficult if not impossible (e.g., in case of ECG monitoring acquiring arrhythmias might be dangerous for the user). Secondly, we have to careful design indicators and rules to assess whether incoming data fit or not the learned model. Finally, we have to face the domain adaptation problem, since normal condition might changes during time and the learned model might not be able to describe incoming normal data, thus it has to be adapted accordingly. For example, in ECG monitoring the model is learned over a training set of normal heartbeats acquired at low heart rate, but the morphology of normal heartbeats changes when the heart rate increases, see Fig. 5.2b.

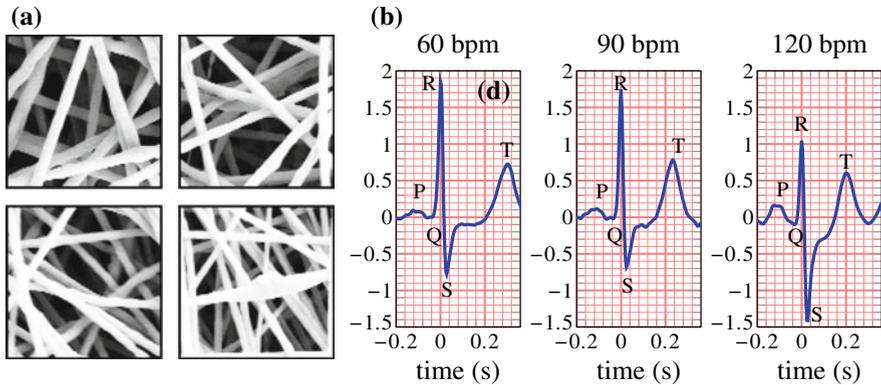


Fig. 5.2 **a** Details of SEM images acquired at different scales. The content of these images is different, although they are perceptually similar. **b** Examples of heartbeats acquired at different heart rate. We report the name of the waveforms of the ECG [13]

We investigate these challenges following two directions, and in particular, we adopt two different modeling assumptions on data-generating process. At first, we assume that data can be described by a smooth probability density function, as customary in the statistics literature. In these settings we focus on the change-detection problem, namely the problem of detecting permanent changes in the monitored datastreams. We investigate the intrinsic difficulty of this problem, in particular when the data dimension increases. Then, we propose QuantTree, a novel change detection algorithms based on histograms that enables the non-parametric monitoring of multivariate datastreams. We theoretically prove that any statistic compute over histograms defined by QuantTree does not depend on the data-generating process.

In the second part we focus on data having a complex structure, and address the anomaly detection problem. We propose a very general anomaly-detection algorithm based on a dictionary yielding sparse representation learned from normal data. As such, it is not able to provide sparse representation to anomalous data, and we exploit this property to design low dimensional indicators to assess whether new data conform or not to the learned dictionary. Moreover, we propose two domain adaptation algorithms to make our anomaly detector effective in the considered application scenarios.

The chapter is structure as follows: Sect. 5.2 presents the most relevant related literature and Sect. 5.3 formally state the problem we address. Section 5.4 focuses on the contribution of the first part of the thesis, where we model data as random vectors, while Sect. 5.5 is dedicated to the second part of the thesis, where we consider data having complex structures. Finally, Sect. 5.6 presents the conclusions and the future works.

5.2 Related Works

The first algorithms [22, 24] addressing the change-detection problem were proposed in the statistical process control literature [15] and consider only univariate datastreams. These algorithms are well studied and several properties have been proved due to their simplicity. Their main drawback is that they require the knowledge of the data generating distributions. Non-parametric methods, namely those that can operate when this distribution is unknown, typically employ statistics that are based on natural order of the real numbers, such as Kolmogorov–Smirnov [25] and Mann–Whitney [14]. Extending these methods to operate on multivariate datastreams is far from being trivial, since no natural order is well defined on \mathbb{R}^d . The general approach to monitor multivariate datastreams is to learn a model that approximates ϕ_0 and monitor a univariate statistic based on the learned model. One of the most popular non-parametric approximation of ϕ_0 is given by Kernel Density Estimation [17], that however becomes intractable when the data dimension increases.

All these methods assume that data can be described by a smooth probability density function (pdf). However, complex data such as signal and images live close to a low-dimensional manifold embedded in a higher dimensional space [3], and do not admit a smooth pdf. In these cases, it is necessary to learn meaningful representations to data to perform any task, from classification to anomaly detection. Here, we consider dictionary yielding sparse representations that have been originally proposed to address image processing problems such as denoising [1], but they were also employed in supervised tasks [19], in particular classification [20, 23]. The adaptation of dictionaries yielding sparse representations to different domain were investigated in particular to address the image classification problem [21, 27]. In this scenario training images are acquired under different conditions than the test ones, e.g. different lightning and view angles, and therefore live in a different domain.

5.3 Problem Formulation

Let us consider a datastream $\{\mathbf{s}_t\}_{t=1,\dots}$, where $\mathbf{s}_t \in \mathbb{R}^d$ is drawn from a process \mathcal{P}_N in normal condition. We are interested in detect whether $\mathbf{s}_t \sim \mathcal{P}_A$, i.e., \mathbf{s}_t is drawn from an alternative process \mathcal{P}_A representing the anomalous conditions. Both \mathcal{P}_N and \mathcal{P}_A are unknown, but we assume that a training set of normal data is available to approximate \mathcal{P}_N . In what follows we describe in details the specific problems we consider in the thesis.

Change Detection. At first, we model \mathbf{s}_t as a realization of a continuous random vector, namely we assume that \mathcal{P}_N and \mathcal{P}_A admit smooth probability density functions ϕ_0 and ϕ_1 , respectively. This assumption is not too strict, as it usually met after a feature extraction process.

We address the problem of detecting abrupt change in the data-generating process. More precisely, our goal is to detect if there is a change point $\tau \in \mathbb{N}$ in the

datastreams such that $\mathbf{s}_t \sim \phi_0$ for $t \leq \tau$ and $\mathbf{s}_t \sim \phi_1$ for $t > \tau$. For simplicity, we analyze the datastream in batches $W = \{\mathbf{s}_1, \dots, \mathbf{s}_v\}$ of v samples and detect changes by performing the following hypothesis test:

$$H_0 : W \sim \phi_0, \quad H_1 : W \approx \phi_0 \quad (5.1)$$

The null hypothesis is rejected whether $\mathcal{T}(W) > \gamma$, where \mathcal{T} is a statistic typically defined upon the model that approximate the density ϕ_0 , and γ is defined to guarantee a desired probability of false positive rate α , i.e. $P_{\phi_0}(\mathcal{T}(W) > \gamma) \leq \alpha$.

Anomaly Detection. The anomaly-detection problem is strictly related to change detection. The main difference is that in anomaly detection we analyze each \mathbf{s}_t independently to whether is draw from \mathcal{P}_N or \mathcal{P}_A , without taking into account the temporal correlation (for this reason we will omit the subscript t). In this settings we consider data having complex structure, such as heartbeats or patches, i.e., small region extracted from an image. We adopt dictionaries yielding sparse representation to describe $\mathbf{s} \sim \mathcal{P}_N$, namely we assume that $\mathbf{s} \approx D\mathbf{x}$, where $D \in \mathbb{R}^{d \times n}$ is a matrix called *dictionary*, that has to be learned from normal data and the coefficient vector $\mathbf{x} \in \mathbb{R}^n$ is *sparse*, namely it has few of nonzero components. To detect anomalies, we have to define a decision rule, i.e., a function $\mathcal{T} : \mathbb{R}^d \rightarrow \mathbb{R}$ and a threshold γ such that

$$\mathbf{s} \text{ is anomalous} \iff \mathcal{T}(\mathbf{s}) > \gamma, \quad (5.2)$$

where $\mathcal{T}(\mathbf{s})$ is defined using the sparse representation \mathbf{x} of \mathbf{s} .

Domain Adaptation. The process generating normal data \mathcal{P}_N may change over time. Therefore, a dictionary D learned on training data (i.e., in the source domain), might not be able to describe normal data during test (i.e., in the target domain). To avoid degradation in the anomaly-detection performance, D has to be adapted has soon as \mathcal{P}_N changes. For example, in case of ECG monitoring the morphology of normal heartbeats changes when the heart rate increases, while in case of SEM images, the magnification level of the microscope may change, and this modify the qualitative content of the patches, as shown in Fig. 5.2.

5.4 Data as Random Vectors

In this section we consider data modeled as random vectors. At first we investigate the detectability loss phenomenon, showing that the change-detection performance are heavily affected by the data dimension. Then we propose a novel change-detection algorithm that employs histograms to describe normal data.

5.4.1 Detectability Loss

As described in the Sect. 5.3, we assume that both \mathcal{P}_N and \mathcal{P}_A admit a smooth pdf $\phi_0, \phi_1: \mathbb{R}^d \rightarrow \mathbb{R}$. For simplicity, we also assume that ϕ_1 that can be expressed as $\phi_1(\mathbf{s}) = \phi_0(Q\mathbf{s} + \mathbf{v})$, where $\mathbf{v} \in \mathbb{R}^d$ and $Q \in \mathbb{R}^{d \times d}$ is an orthogonal matrix. This is quite a general model as it includes changes in the mean as well as in the correlations of components of \mathbf{s}_t . To detect changes, we consider the popular approach that monitor the loglikelihood w.r.t. the distribution generating normal data ϕ_0 :

$$\mathcal{L}(\mathbf{s}_t) = \log(\phi_0(\mathbf{s}_t)). \quad (5.3)$$

In practice, we reduce the multivariate datastream $\{\mathbf{s}_t\}$ to a univariate one $\{\mathcal{L}(\mathbf{s}_t)\}$. Since ϕ_0 is unknown, we should preliminary estimate $\hat{\phi}_0$ from data, and use it in (5.3) in place of ϕ_0 . However, in what follows we will consider ϕ_0 since it make easier to investigate how the data dimension affect the change-detection performance.

We now introduce two measures that we will use in our analysis. The *change magnitude* assesses how much ϕ_1 differs from ϕ_0 and is defined as $s\text{KL}(\phi_0, \phi_1)$, namely the symmetric Kullback–Leibler divergence between ϕ_0 and ϕ_1 [12]. In practice, large values of $s\text{KL}(\phi_0, \phi_1)$ makes the change very apparent, as proved in the Stein’s Lemma [12]. The *change detectability* assesses how the change is perceivable by monitoring the datastream $\{\mathcal{L}(\mathbf{s}_t)\}$ and is defined as the signal-to-noise ratio of the change $\phi_0 \rightarrow \phi_1$:

$$\text{SNR}(\phi_0, \phi_1) := \frac{\left(E_{s \sim \phi_0} [\mathcal{L}(\mathbf{s})] - E_{s \sim \phi_1} [\mathcal{L}(\mathbf{s})] \right)^2}{\text{var}_{s \sim \phi_0} [\mathcal{L}(\mathbf{s})] + \text{var}_{s \sim \phi_1} [\mathcal{L}(\mathbf{s})]}, \quad (5.4)$$

where E and var denote the expected value and the variance, respectively.

The following theorem proves *detectability loss* on Gaussian datastreams (the proof is reported in [2]).

Theorem 1 *Let $\phi_0 = \mathcal{N}(\mu_0, \Sigma_0)$ be a d -dimensional Gaussian pdf and $\phi_1 = \phi_0(Q\mathbf{s} + \mathbf{v})$, where $Q \in \mathbb{R}^{d \times d}$ is orthogonal and $\mathbf{v} \in \mathbb{R}^d$. Then, it holds*

$$\text{SNR}(\phi_0, \phi_1) \leq \frac{C}{d} \quad (5.5)$$

where the constant C depends only on $s\text{KL}(\phi_0, \phi_1)$.

The main consequences of Theorem 1 is that the change detectability decreases when the data dimension increases, as long as the change magnitude is kept fixed. Remarkably, this results is independent on how the changes affected the datastream (i.e., it is independent on Q and \mathbf{v}), but only on the change magnitude. Moreover, it does not depend on estimation error, since in (5.4) we have considered the true and unknown distribution ϕ_0 . However, the detectability loss becomes more severe when the loglikelihood is computed w.r.t. the estimated $\hat{\phi}_0$, as we showed in [2].

Finally, we remark that the detectability loss holds also for more general distributions ϕ_0 , such as Gaussian Mixture, and on real data. In that case the problem cannot be treated analytically, but we empirically show similar results using Controlling Change Magnitude (CCM) [6], a framework to inject changes of a given magnitude in real world datastreams.

5.4.2 QuantTree

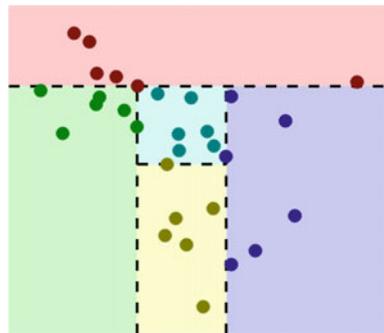
In this section we present QuantTree, a novel change detection algorithm that computes a histogram to approximate ϕ_0 . A histogram h is defined as $h = \{(B_k, \hat{\pi}_k)\}_{k=1, \dots, K}$, where the bins $\{B_k\}$ identify a partition of the data space, while the $\{\hat{\pi}_k\}$ are the probabilities associated to the bins. In practice, we estimate h to approximate ϕ_0 and π_k is an estimate of the probability of $s \sim \phi_0$ to fall inside B_k . As described in Sect. 5.3, we monitor the datastream in a batch-wise manner and we consider statistics $\mathcal{T}_h(W)$ defined over the histogram h , namely statistics that depend only on the number y_k of samples of W that fall in the bin B_k , for $k = 1, \dots, K$. Examples of such statistics are the Total Variation distance and the Pearson's statistic.

The proposed QuantTree algorithm takes as input target probability values $\{\pi_k\}$ and generates a partitioning $\{B_k\}$ such that the corresponding probability $\{\hat{\pi}_k\}$ are close to $\{\pi_k\}$. QuantTree is an iterative splitting scheme that generates a new bin at each iteration k . The bin is defined by splitting along a component chosen at random among the d available. The splitting point is selected to contain exactly $\text{round}(\pi_k N)$ samples of the training set, where S is the number of training samples. An example of partitioning computed by QuantTree is shown in Fig. 5.3.

The main property of QuantTree is that its peculiar splitting scheme makes the distribution of any statistics \mathcal{T}_h independent on the data-generating distribution ϕ_0 . This property is formally stated in the following theorem, that we proved in [4].

Theorem 2 *Let $\mathcal{T}_h(\cdot)$ be defined over an histogram h computed by QuantTree. When $W \sim \phi_0$, the distribution of $\mathcal{T}_h(W)$ depends only on v , N and $\{\pi_k\}_k$.*

Fig. 5.3 A partitioning with $K = 5$ bins computed by QuantTree over a training set of $N = 30$ samples. We set $\pi_k = 1/K$ to yield a uniform density histograms, thus all the bins contain 6 samples each



In practice, the distribution of any statistic \mathcal{T}_h depends only on the cardinalities of the training set and the window W and on the target probabilities $\{\pi_k\}$. The main consequence of Theorem 2 is that the threshold γ that guarantees a given false positive rate α does not depend on ϕ_0 . Therefore, we can precompute γ through by estimating the distribution of \mathcal{T}_h over synthetically generated samples through Montecarlo simulations. To the best of our knowledge, QuantTree is one of the first algorithms that performs non-parametric monitoring of multivariate datastreams.

We compare the histograms computed by QuantTree with other partitioning in the literature in [4] through experiments on Gaussian datastreams and real world datasets. Histograms computed by QuantTree yield a larger power and are the only ones that allows to properly control the false positive rate. We remark that also QuantTree suffers of the detectability loss, confirming the generality of our results on detectability loss.

5.5 Data Featuring Complex Structures

In this section we employ dictionaries yielding sparse representations to model data having complex structures. At first we present our general anomaly detection algorithm, then we introduce our two domain-adaptation solutions, specifically designed for the application scenarios described in Sect. 5.1.

5.5.1 Sparsity-Based Anomaly Detection

Our modeling assumption is that normal data $\mathbf{s} \sim \mathcal{P}_N$ can be well approximated as $\mathbf{s} \approx D\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$ is sparse, namely it has only few nonzero components and the dictionary $D \in \mathbb{R}^{d \times n}$ approximates the process \mathcal{P}_N . The sparse representation \mathbf{x} is computed by solving the sparse coding problem:

$$\mathbf{x} = \arg \min_{\tilde{\mathbf{x}} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{s} - D\tilde{\mathbf{x}}\|_2^2 + \lambda \|\tilde{\mathbf{x}}\|_1, \quad (5.6)$$

where the ℓ^1 norm $\|\tilde{\mathbf{x}}\|_1$ is used to enforce sparsity in $\tilde{\mathbf{x}}$, and the parameter $\lambda \in \mathbb{R}$ controls the tradeoff between the ℓ^1 norm and the reconstruction error $\|\mathbf{s} - D\tilde{\mathbf{x}}\|_2^2$. The dictionary D is typically unknown, and we have to learn it from a training set of normal data by solving the dictionary learning problem:

$$D, X = \arg \min_{\tilde{D} \in \mathbb{R}^{d \times n}, \tilde{X} \in \mathbb{R}^{n \times m}} \frac{1}{2} \|S_0 - \tilde{D}\tilde{X}\|_2^2 + \lambda \|X\|_1, \quad (5.7)$$

where $S_0 \in \mathbb{R}^{d \times m}$ is the training set that collects normal data column-wise.

To assess whether \mathbf{s} is generated or not from \mathcal{P}_N , we define a bivariate indicator vector $\mathbf{f} \in \mathbb{R}^2$ collecting the reconstruction error and the sparsity of the representation:

$$\mathbf{f}(\mathbf{s}) = \begin{bmatrix} \|\mathbf{s} - D\mathbf{x}\|_2 \\ \|\mathbf{x}\|_1 \end{bmatrix}. \quad (5.8)$$

In fact, we expect that when \mathbf{s} is anomalous it deviates from normal data either in the sparsity of the representation or in the reconstruction error. This means that $\mathbf{f}(\mathbf{s})$ would be an outlier w.r.t. the distribution ϕ_0 of \mathbf{f} computed over normal data. Therefore, we detect anomalies as in (5.2) by setting $\mathcal{T} = -\log(\phi_0)$, where ϕ_0 is estimate from a training set of normal data (different from the set S_0 used in dictionary learning) using Kernel Density Estimation [5].

We evaluate our anomaly-detection algorithm on a dataset containing 45 SEM images. In this case $\mathbf{s} \in \mathbb{R}^{15 \times 15}$ is a small squared patch extracted from the image. We analyze each patch independently to determine if it is normal or anomalous. Since each pixel of the image is contained in more than one patch, we obtain several decisions of each pixel. To detect defects at pixel level, we aggregate all the decisions through majority voting. An example of the obtained detections is shown in Fig. 5.4: our algorithm is able to localize all the defects by keeping the false positive rate small. More details on our solution and experiments are reported in [8].

In case of ECG monitoring, the data to be analyzed are the heartbeats, that are extracted from the ECG signal using traditional algorithms. Since our goal is to perform ECG monitoring directly on the wearable device, that has limited computational capabilities, we adopt a different sparse coding procedure, that is based on the ℓ^0 “norm” and it is performed by means of greedy algorithms [11]. In particular, we proposed a novel variant of the OMP algorithm [9], that is specifically designed for dictionary $D \in \mathbb{R}^{d \times n}$ where $n < d$, that is settings we adopt in ECG monitoring. Dictionary learning, that is required every time the device is positioned as the shape of normal heartbeats depends both of the users and device position, is performed on a host device, since the computational resources of wearable devices are not sufficient.

5.5.2 Multiscale Anomaly-Detection

In the quality inspection through SEM images, we have to face the domain adaptation problem since the magnification level of the microscope may change during monitoring. Therefore, we improve our anomaly-detection algorithm to make it scale-invariant. The three key ingredients we use are: (i) a multiscale dictionary that is able to describe patch extracted from image at different resolution, (ii) a multiscale sparse representation that captures the structure in the patch at different scales and (iii) a trivariate indicator vector, that is more powerful than the one in (5.8).

We build the multiscale dictionary as $D = [D_1 | \dots | D_L]$. Each subdictionary D_j is learned by solving problem (5.7) over patched extracted from synthetically rescaled

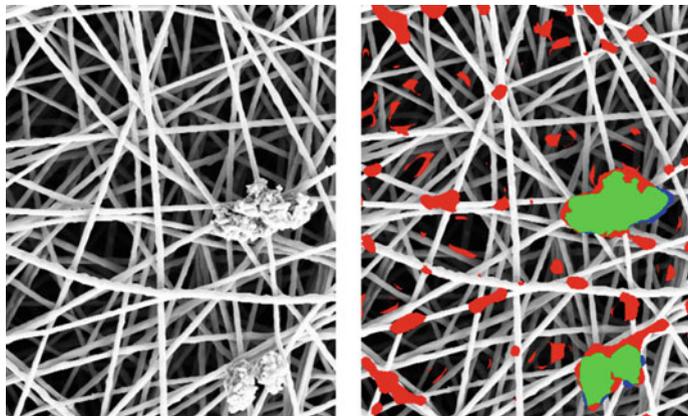


Fig. 5.4 Detail of defect detection at pixel level. Green, blue and red pixel identify the true positive, false negative and false positive, respectively. The threshold γ in (5.2) was set to yield a false positive rate equal to 0.05 (Color figure online)

version of training images. Therefore, we can learn a multiscale dictionary even if the training images are acquired at a fixed scale. To compute the multiscale representation \mathbf{x} we solve the following sparse coding problem:

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{s} - D\mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1 + \xi \sum_{j=1}^L \|\mathbf{x}_j\|_2, \quad (5.9)$$

where each \mathbf{x}_j refers to the subdictionary D_i and \mathbf{x} collects all the \mathbf{x}_j . The group sparsity term $\sum_{j=1}^L \|\mathbf{x}_j\|_2$ ensures that each patch is reconstructed using coefficients of \mathbf{x} that refers only to few scales. Finally, we define a trivariate indicator vector \mathbf{f} that includes the reconstruction error and the sparsity as in (5.8) and also the group sparsity term $\sum_j \|\mathbf{x}_j\|_2$. We detect anomalies using the same approach described in Sect. 5.5.1. Our experiments [7] show that employing a multiscale approach in all phases of the algorithm (dictionary learning, sparse coding and anomaly indicators) achieves good anomaly detection performance even when training and test images are acquired at different scales.

5.5.3 Dictionary Adaptation for ECG Monitoring

To make our anomaly-detection algorithm effective in long-term ECG monitoring, we have to adapt the user-specific dictionary. In fact, to keep the training procedure safe for the user, this dictionary is learned from heartbeats acquired in resting conditions, and it would be not able to describe normal heartbeats acquired during daily activities,

when the heart rate increases and normal heartbeats get transformed, see Fig. 5.2b. These transformations are similar for every user: the T and P waves approach the QRS complex and the support of the heartbeats narrows down. Therefore, we adopt user-independent transformations $F_{r,r_0} \mathbb{R}^{d_r \times d_{r_0}}$ to adapt the user-specific dictionary $D_{u,r_0} \in \mathbb{R}^{d_{r_0} \times n}$:

$$D_{u,r} = F_{r,r_0} D_{u,r_0}, \quad (5.10)$$

where r_0 and r denotes the heart rates in resting condition and during daily activities, respectively, and u indicates the user. The transformation F_{r,r_0} depends only on the heart rates and is learned from collections $S_{u,r}$ of training sets of normal heartbeats of several users acquired at different heart rates extracted from publicly available datasets of ECG signals. The learning has to be performed only once by solving the following optimization problem:

$$F_{r,r_0} = \arg \min_{F, \{X_u\}_u} \frac{1}{2} \sum_{u=1}^L \|S_{u,r} - F D_{u,r_0} X_u\|_2^2 + \mu \sum_{u=1}^L \|X_u\|_1 + \frac{\lambda}{2} \|W \odot F\|_2^2 + \xi \|W \odot F\|_1,$$

where the first term ensures that the transformed dictionaries $F D_{u,r_0}$ provide good approximation to the heartbeats of the user u . Moreover, we adopt three regularization terms: the first one is based on ℓ^1 norm and enforces sparsity in the representations of heartbeats at heart rate r for each user u . The other two terms represent a weighted elastic net penalization over F to improve the stability of the optimization problem, and the weighting matrix W introduces regularities in F_{r,r_0} .

The learned F_{r,r_0} (for several values of r and r_0) are then hard-coded in the wearable device to perform online monitoring [18]. We evaluate our solution on ECG signals acquired using the Bio2Bit Dongle [18]. Our experiments [10] show that the reconstruction error is kept small when the heart rate increases, implying that our transformations effectively adapt the dictionaries to operate at higher heart rate. Moreover, our solution is able to correctly identify anomalous heartbeats even at very large heart rate, when the morphology of normal heartbeats undergoes severe transformations.

5.6 Conclusions

We have addressed the general problem of monitoring a datastream to detect whether the data generating process departs from normal conditions. Several challenges have to be addressed in practical applications where data are high dimensional and feature complex structures. We address these challenges from two different perspectives by making different modeling assumption on the data generating process.

At first we assume that data can be described by a smooth probability density function and address the change detection problem. We prove the detectability loss phenomenon, that relates the change-detection performance and the data dimension, and we propose QuantTree, that is one of the first algorithms that enables non-

parametric monitoring of multivariate datastream. In the second part, we employ dictionary yielding sparse representation to model data having featuring complex structures and propose a novel anomaly-detection algorithm. To make this algorithm effective in practical applications, we propose two domain-adaptation solutions that turn to be very effective in long-term ECG monitoring and in a quality inspection of an industrial process.

Future works include the extension of QuantTree to design a truly multivariate change-detection algorithm, and in particular to control the average run length instead of the false positive rate. Another relevant directions is the design of models that provide good representation for detection tasks. In fact, our dictionaries are learned to provide good reconstruction to the data, and not to perform anomaly detection. Very recent works such as [26] show very promising results using deep learning, but a general methodology is still not available.

References

1. Aharon M, Elad M, Bruckstein A (2006) K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans Signal Process* 54(11):4311–4322
2. Alippi C, Boracchi G, Carrera D, Roveri M (2016) Change detection in multivariate datastreams: likelihood and detectability loss. In: *Proceedings of the international joint conference on artificial intelligence (IJCAI)*, vol 2, pp 1368–1374
3. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 35(8):1798–1828
4. Boracchi G, Carrera D, Cervellera C, Maccio D (2018) Quantree: histograms for change detection in multivariate data streams. In: *Proceedings of the international conference on machine learning (ICML)*, pp 638–647
5. Botev ZI, Grotowski JF, Kroese DP et al (2010) Kernel density estimation via diffusion. *Ann Stat* 38(5):2916–2957
6. Carrera D, Boracchi G (2018) Generating high-dimensional datastreams for change detection. *Big Data Res* 11:11–21
7. Carrera D, Boracchi G, Foi A, Wohlberg B (2016) Scale-invariant anomaly detection with multiscale group-sparse models. In: *Proceedings of the IEEE international conference on image processing (ICIP)*, pp 3892–3896
8. Carrera D, Manganini F, Boracchi G, Lanzarone E (2017) Defect detection in sem images of nanofibrous materials. *IEEE Trans Ind Inform* 13(2):551–561
9. Carrera D, Rossi B, Fragneto P, Boracchi G (2017) Domain adaptation for online eeg monitoring. In: *Proceedings of the IEEE international conference on data mining (ICDM)*, pp 775–780
10. Carrera D, Rossi B, Fragneto P, Boracchi G (2019) Online anomaly detection for long-term eeg monitoring using wearable devices. *Pattern Recognit* 88:482–492
11. Carrera D, Rossi B, Zambon D, Fragneto P, Boracchi G (2016) Ecg monitoring in wearable devices by sparse models. In: *Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML-PKDD)*, pp 145–160
12. Cover TM, Thomas JA (2012) *Elements of information theory*. Wiley, Hoboken
13. Felker GM, Mann DL (2014) *Heart failure: a companion to Braunwald's heart disease*. Elsevier Health Sciences
14. Hawkins DM, Deng Q (2010) A nonparametric change-point control chart. *J Qual Technol* 42(2):165–173
15. Hawkins DM, Qiu P, Chang WK (2003) The changepoint model for statistical process control. *J Qual Technol* 35(4):355–366

16. Hoekema R, Uijen GJ, Van Oosterom A (2001) Geometrical aspects of the interindividual variability of multilead ecg recordings. *IEEE Trans Biomed Eng* 48(5):551–559
17. Kreml G (2011) The algorithm apt to classify in concurrence of latency and drift. In: *Proceedings of the intelligent data analysis (IDA)*, pp 222–233
18. Longoni M, Carrera D, Rossi B, Fragneto P, Pessione M, Boracchi G (2018) A wearable device for online and long-term ecg monitoring. In: *Proceedings of the international conference on artificial intelligence (IJCAI)*, pp 5838–5840
19. Mairal J, Bach F, Ponce J (2012) Task-driven dictionary learning. *IEEE Trans Pattern Anal Mach Intell* 34(4):791–804
20. Mairal J, Ponce J, Sapiro G, Zisserman A, Bach F (2009) Supervised dictionary learning. In: *Advances in neural information processing systems (NIPS)*, pp 1033–1040
21. Ni J, Qiu Q, Chellappa R (2013) Subspace interpolation via dictionary learning for unsupervised domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 692–699
22. Page ES (1954) Continuous inspection schemes. *Biometrika* 41(1/2):100–115
23. Ramirez I, Sprechmann P, Sapiro G (2010) Classification and clustering via dictionary learning with structured incoherence and shared features. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 3501–3508
24. Roberts S (1959) Control chart tests based on geometric moving averages. *Technometrics* 1(3):239–250
25. Ross GJ, Adams NM (2012) Two nonparametric control charts for detecting arbitrary distribution changes. *J Qual Technol* 44(2):102
26. Ruff L, Goernitz N, Deecke L, Siddiqui SA, Vandermeulen R, Binder A, Müller E, Kloft M (2018) Deep one-class classification. In: *Proceedings of the international conference on machine learning (ICML)*, pp 4390–4399
27. Shekhar S, Patel VM, Nguyen HV, Chellappa R (2013) Generalized domain-adaptive dictionaries. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 361–368

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), 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 license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license 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.

