Signal Processing 99 (2014) 215–249

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Review

A review of novelty detection

Marco A.F. Pimentel *, David A. Clifton, Lei Clifton, Lionel Tarassenko

Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford OX3 7DQ, UK

ARTICLE INFO

Article history:
Received 17 October 2012
Received in revised form
16 December 2013
Accepted 23 December 2013
Available online 2 January 2014

Keywords:
Novelty detection
One-class classification
Machine learning

ABSTRACT

Novelty detection is the task of classifying test data that differ in some respect from the data that are available during training. This may be seen as "one-class classification", in which a model is constructed to describe "normal" training data. The novelty detection approach is typically used when the quantity of available "abnormal" data is insufficient to construct explicit models for non-normal classes. Application includes inference in datasets from critical systems, where the quantity of available normal data is very large, such that "normality" may be accurately modelled. In this review we aim to provide an updated and structured investigation of novelty detection research papers that have appeared in the machine learning literature during the last decade.

Contents

* Correspondence author.
E-mail address: marco.pimentel@eng.ox.ac.uk (M.A.F. Pimentel).

0165-1684/$ - see front matter © 2014 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.sigpro.2013.12.026

1. Introduction

Novelty detection can be defined as the task of recognising that test data differ in some respect from the data that are available during training. Its practical importance and challenging nature have led to many approaches being proposed. These methods are typically applied to datasets in which a very large number of examples of the "normal" condition (also known as positive examples) is available and where there are insufficient data to describe "abnormalities" (also known as negative examples).

Novelty detection has gained much research attention in application domains involving large datasets acquired from critical systems. These include the detection of mass-like structures in mammograms [1] and other medical diagnostic problems [2,3], faults and failure detection in complex industrial systems [4], structural damage [5], intrusions in electronic security systems, such as credit card or mobile phone fraud detection [6,7], video surveillance [8,9], mobile robotics [10,11], sensor networks [12], astronomy catalogues [13,14] and text mining [15]. The complexity of modern high-integrity systems is such that only a limited understanding of the relationships between the various system components can be obtained. An inevitable consequence of this is the existence of a large number of possible "abnormal" modes, some of which may not be known *a priori*, which makes conventional multi-class classification schemes unsuitable for these applications. A solution to this problem is offered by novelty detection, in which a description of normality is learnt by constructing a model with numerous examples representing positive instances (i.e., data indicative of normal system behaviour). Previously unseen patterns are then tested by comparing them with the model of normality, often resulting in some form of novelty score. The score, which may or may not be probabilistic, is typically compared to a decision threshold, and the test data are then deemed to be "abnormal" if the threshold is exceeded.

This survey aims to provide an updated and structured overview of recent studies and approaches to novelty detection that have appeared in the machine learning and signal processing literature. The complexity and main application domains of each method are also discussed. This review is motivated in Section 1.2, in which we examine previous reviews of the literature, concluding that a new review is necessary in light of recent research results.

1.1. Novelty detection as one-class classification

Conventional pattern recognition typically focuses on the classification of two or more classes. General multi-class classification problems are often decomposed into multiple two-class classification problems, where the two-class problem is considered the basic classification task [16,17]. In a two-class classification problem we are given a set of training examples $\boldsymbol{X} = \{(\mathbf{x}_i, \omega_i) | \mathbf{x}_i \in \mathbb{R}^D,\ i = 1 \ldots N\}$, where each example consists of a D dimensional vector $\mathbf{x}_i$ and its label $\omega_i \in \{-1, 1\}$. From the labelled dataset, a function $h(\mathbf{x})$ is constructed such that for a given input vector $\mathbf{x}'$ an estimate of one of the two labels is obtained, $\omega = h(\mathbf{x}'|\boldsymbol{X})$:

$$h(\mathbf{x}'|\boldsymbol{X}) : \mathbb{R}^D \rightarrow [-1, 1]$$

The problem of novelty detection, however, is approached within the framework of one-class classification [18], in which *one* class (the specified normal, positive class) has to be distinguished from all other possibilities. It is usually assumed that the positive class is very well sampled, while the other class(es) is/are severely under-sampled. The scarcity of negative examples can be due to high measurement costs, or the low frequency at which abnormal events occur. For example, in a machine monitoring system, we require an alarm to be triggered whenever the machine exhibits "abnormal" behaviour. Measurements of the machine during its normal operational state are inexpensive and easy to obtain. Conversely, measurements of failure of the machine would require the destruction of similar machines in all possible ways. Therefore, it is difficult, if not impossible, to obtain a very well-sampled negative class [19]. This problem is often compounded for the analysis of critical systems such as human patients or jet engines, in which there is significant variability between individual entities, thereby limiting the use of "abnormal" data acquired from other examples [20,19].

In the novelty detection approach to classification, "normal" patterns $\boldsymbol{X}$ are available for training, while "abnormal" ones are relatively few. A model of normality $M(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the free parameters of the model, is inferred and used to assign novelty scores $z(\mathbf{x})$ to

previously unseen test data $\mathbf{x}$. Larger novelty scores $z(\mathbf{x})$ correspond to increased "abnormality" with respect to the model of normality. A novelty threshold $z(\mathbf{x}) = k$ is defined such that $\mathbf{x}$ is classified "normal" if $z(\mathbf{x}) \leq k$, or "abnormal" otherwise. Thus, $z(\mathbf{x}) = k$ defines a decision boundary. Different types of models M, methods for setting their parameters $\boldsymbol{\theta}$, and methods for determining novelty thresholds k have been proposed in the literature and will be considered in this review.

Two interchangeable synonyms of novelty detection [21,1] often used in the literature are *anomaly detection* and *outlier detection* [22]. The different terms originate from different domains of application to which one-class classification can be applied, and there is no universally accepted definition. Merriam-Webster [23] defines "novelty" to mean "new and not resembling something formerly known or used". Anomalies and outliers are two terms used most commonly in the context of anomaly detection; sometimes interchangeably [24]. Barnett and Lewis [25] define an *outlier* as a data point that "appears to be inconsistent with the remainder of that set of [training] data". However, it is also used to describe a small fraction of "normal" data which lie far way from the majority of "normal" data in the feature space [9]. Therefore, outlier detection aims to handle these "rogue" observations in a set of data, which can have a large effect on the analysis of the data. In other words, outliers are assumed to contaminate the dataset under consideration and the goal is to cope with their presence during the model-construction stage. A different goal is to learn a model of normality $M(\boldsymbol{\theta})$ from a set of data that is considered "normal", in which the assumption is that the data used to train the learning system constitute the basis to build a model of normality and the decision process on test data is based on the use of this model. Furthermore, the notion of normal data as expressed in anomaly detection is often not the same as that used in novelty detection. Anomalies are often taken to refer to irregularities or transient events in otherwise "normal" data. These transient events are typically noisy events, which give rise to artefacts that act as obstacles to data analysis, to be removed before analysis can be performed. From this definition, novel data are not necessarily anomalies; this distinction has also been drawn by recent reviews in anomaly detection [24]. Nevertheless, the term "anomaly detection" is typically used synonymously with "novelty detection", and because the solutions and methods used in novelty detection, anomaly detection, and outlier detection are often common, this review aims to consider all such detection schemes and variants.

1.2. Overview of reviews on novelty detection

This review is timely because there has not been a comprehensive review of novelty detection since the two papers by Markou and Singh [26,27] in this journal ten years ago. A number of surveys have been published since then [26–32,24], but none of these attempts to be as wide-ranging as we are in our review. We cover not only the topic of novelty detection but also the related topics of outlier, anomaly and, briefly, change-point detection, using a taxonomy which is appropriate for the state of the art in the research literature today.

Markou and Singh distinguished between two main categories of novelty detection techniques: statistical approaches [26] and neural network based approaches [27]. While appropriate in 2003, these classifications are now problematic, due to the convergence of statistics and machine learning. The former are mostly based on using the statistical properties of data to estimate whether a new test point comes from the same distribution as the training data or not, using either parametric or non-parametric techniques, while the latter come from a wide range of flexible non-linear regression and classification models, data reduction models, and non-linear dynamical models that have been extensively used for novelty detection [33,34]. Another review of the literature of novelty detection using machine learning techniques is provided by Marsland [28]. The latter offers brief descriptions of the related topics of statistical outlier detection and novelty detection in biological organisms. The author emphasises some fundamental issues of novelty detection, such as the lack of a definition of how different a novel biological stimulus can be before it is classified as "abnormal", and how often a stimulus must be observed before it is classified as "normal". This issue is also acknowledged by Modenesi and Braga [35], who describe novelty detection strategies applied to the domain of time-series modelling.

Hodge and Austin [29], Agyemang et al. [30], and Chandola et al. [36] provide comprehensive surveys of outlier detection methodologies developed in machine learning and statistical domains. Three fundamental approaches to the problem of outlier detection are addressed in [29]. In the first approach, outliers are determined with no prior knowledge of the data; this is a learning approach analogous to *unsupervised clustering*. The second approach is analogous to *supervised classification* and requires labelled data ("normal" or "abnormal"). In this latter type, both normality and abnormality are modelled explicitly. Lastly, the third approach models only normality. According to Hodge and Austin [29], this approach is analogous to a *semi-supervised recognition* approach, which they term novelty detection or novelty recognition. As with Markou and Singh [26,27], outlier detection methods are grouped into "statistical models" and "neural networks" in [29,30]. Additionally, the authors suggest another two categories: machine learning and hybrid methods. According to Hodge and Austin [29], most "statistical" and "neural network" approaches require cardinal or ordinal data to allow distances to be computed between data points. For this reason, the machine learning category was suggested to include multi-type vectors and symbolic attributes, such as rule-based systems and tree-structure based methods. Their "hybrid" category covers systems that incorporate algorithms from at least two of the other three categories. Again, research since 2004 makes the use of these categories problematical.

The most recent comprehensive survey of methods related to anomaly detection was compiled by Chandola et al. [24]. Their focus is on the detection of anomalies; i.e., "patterns in data that do not conform to expected behaviour" [24, p. 15:1]. This survey builds upon the three previous methods discussed in [29,30,36] by expanding

the discussion of each method considered and adding two more categories of anomaly detection techniques: information theoretic techniques, which analyse the *information content* of the dataset using information-theoretic measures such as entropy; and spectral techniques, which attempt to find an approximation of the data using a combination of attributes that capture the bulk of the variability in the data. The surveys [29,30,24] agree that approaches to anomaly detection can be supervised, unsupervised, or semi-supervised. More recently, Kittler et al. [37] addressed the problem of anomaly detection in machine perception (where the key objective is to instantiate models to explain observations), and introduced the concept of *domain anomaly*, which refers to the situation when none of the models characterising a domain are able to explain the data. The authors argued that the conventional notions of anomalies in data (such as being an outlier or distribution drift) alone cannot detect all anomalous events of interest in machine perception, and proposed a taxonomy of domain anomalies, which distinguishes between component, configuration, and joint component and configuration domain anomaly events.

Some novelty detection methods have been the topic of a number of other very brief overviews that have recently been published [31,38–41,32]. Other surveys have focused on novelty detection methods used in specific applications such as cyber-intrusion detection [6,42,7] and wireless sensor networks [12].

Only a few of the recent surveys attempt to provide a comprehensive review of the different methods used in different application domains. Since the review paper by Markou and Singh [26,27], we believe that there has been no rigorous review of all the major topics in novelty detection. In fact, many reviews recently published contain fewer than 30 references (e.g., the reviews [35,40,41]), and do not include significant papers from the literature. The most recent comprehensive survey of a related topic (anomaly detection) was published by Chandola et al. [24]. However, as discussed in the previous subsection, although they can be seen as related topics, there are some fundamental differences between anomaly detection and novelty detection. Also, Chandola et al. [24] do not attempt to review the novelty detection literature, which itself has attracted significant attention within the research community as shown by the increasing number of publications in this field in the last decade. In this review, we therefore aim to provide a comprehensive overview of novelty detection research, but also include anomaly detection, outlier detection, and related approaches. To the best of our knowledge, this is the first attempt (since 2003) to provide such a structured and detailed review.

1.3. Methods of novelty detection

Approaches to novelty detection include both Frequentist and Bayesian approaches, information theory, extreme value statistics, support vector methods, other kernel methods, and neural networks. In general, all of these methods build some model of a training set that is selected to contain no examples (or very few) of the important (i.e., novel) class. Novelty scores $z(\mathbf{x})$ are then assigned to data $\mathbf{x}$, and deviations from normality are detected according to a decision boundary that is usually referred to as the novelty threshold $z(\mathbf{x}) = k$.

Different metrics are used to evaluate the effectiveness and efficiency of novelty detection methods. The effectiveness of novelty detection techniques can be evaluated according to how many novel data points are correctly identified and also according to how many normal data are incorrectly classified as novel data. The latter is also known as the *false alarm rate*. Receiver operating characteristic (ROC) curves are usually used to represent the trade-off between the detection rate and the false alarm rate. Novelty detection techniques should aim to have a high detection rate while keeping the false alarm rate low. The efficiency of novelty detection approaches is evaluated according to computational cost, and both time and space complexity. Efficient novelty detection techniques should be scalable to large and high-dimensional data sets. In addition, depending on the specific novelty detection task, the amount of memory required to implement the technique is typically considered to be an important performance evaluation metric.

We classify novelty detection techniques according to the following five general categories: (i) probabilistic, (ii) distance-based, (iii) reconstruction-based, (iv) domain-based, and (v) information-theoretic techniques. Approach (i) uses probabilistic methods that often involve a density estimation of the "normal" class. These methods assume that low density areas in the training set indicate that these areas have a low probability of containing "normal" objects. Approach (ii) includes the concepts of nearest-neighbour and clustering analysis that have also been used in classification problems. The assumption here is that "normal" data are tightly clustered, while novel data occur far from their nearest neighbours. Approach (iii) involves training a regression model using the training set. When "abnormal" data are mapped using the trained model, the reconstruction error between the regression target and the actual observed value gives rise to a high novelty score. Neural networks, for example, can be used in this way and can offer many of the same advantages for novelty detection as they do for regular classification problems. Approach (iv) uses domain-based methods to characterise the training data. These methods typically try to describe a domain containing "normal" data by defining a boundary around the "normal" class such that it follows the distribution of the data, but does not explicitly provide a distribution in high-density regions. Approach (v) computes the information content in the training data using information-theoretic measures, such as entropy or Kolmogorov complexity. The main concept here is that novel data significantly alter the information content in a dataset.

1.4. Organisation of the survey

The rest of the survey is organised as follows (see Fig. 1). We provide a state-of-the-art review of novelty detection research based on approaches from the different categories. Probabilistic novelty detection approaches are described in Section 2, distance-based novelty detection approaches are

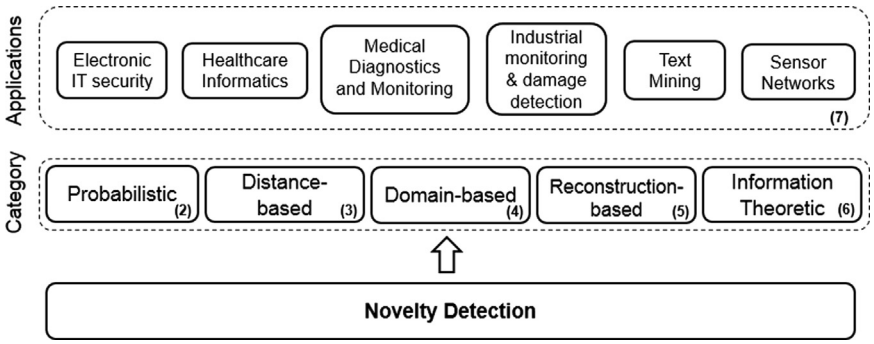

Fig. 1. Schematic representation of the organisation of the survey (the numbers within brackets correspond to the section where the topic is discussed).

presented in Section 3, reconstruction-based novelty detection approaches are described in Section 4. Sections 5 and 6 focus on domain-based and information-theoretic techniques, respectively. The application domains for all five categories of novelty detection methods discussed in this review are summarised in Section 7. In Section 8 we provide an overall conclusion for this review.

2. Probabilistic novelty detection

Probabilistic approaches to novelty detection are based on estimating the generative probability density function (pdf) of the data. The resultant distribution may then be thresholded to define the boundaries of normality in the data space and test whether a test sample comes from the same distribution or not. Training data are assumed to be generated from some underlying probability distribution D, which can be estimated using the training data. This estimate $\hat{D}$ usually represents a model of normality. A novelty threshold can then be set using $\hat{D}$ in some manner, such that it has a probabilistic interpretation.

The techniques proposed vary in terms of their complexity. The simplest statistical techniques for novelty detection can be based on statistical hypothesis tests, which are equivalent to *discordancy tests* in the statistical outlier detection literature [25]. These techniques determine whether a test sample was generated from the same distribution as the "normal" data or not, and are usually employed to detect outliers. Many of these statistical tests, such as the frequently used *Grubbs'* test [43], assume a Gaussian distribution for the training data and work only with univariate continuous data, although variants of these tests have been proposed to handle multivariate data sets; e.g., Aggarwal and Yu [44] recently proposed a variant of the Grubbs' test for multivariate data. The Grubbs' test computes the distance of the test data points from the estimated sample mean and declares any point with a distance above a certain threshold to be an outlier [43]. This requires a threshold parameter to determine the length of the tail that includes the outliers (and which is often associated with a distance of three standard deviations from the estimated mean). Another simple statistical scheme for outlier detection is based on the use of the *box-plot rule*. Solberg and Lahti [45] have applied this technique to eliminate outliers in medical laboratory reference data. Box-plots graphically depict groups of numerical data using five quantities: the smallest observation (sample minimum), lower quartile ($Q1$), median ($Q2$), upper quartile ($Q3$), and largest observation (sample maximum). The method used in [45] starts by transforming the original data so as to achieve a distribution that is close to a Gaussian distribution (by applying the Box-Cox transformation). Then, the lower and upper quartiles ($Q1$ and $Q3$, respectively) are estimated for the transformed data, and the interquartile range (IQR), which is defined by $IQR = Q3 - Q1$, is used to define two detection limits: $Q1 - (1.5 \times IQR)$ and $Q3 + (1.5 \times IQR)$. All values located outside the two limits are identified as outliers. Although experiments have shown that the algorithm has potential for outlier detection, they also suggest that the normalisation of distributions achieved by use of the transformation functions is not sufficient to allow the algorithm to work as expected.

Many other sophisticated statistical tests have been used to detect anomalies and outliers, as discussed in [25]. A description of these statistical tests is beyond the scope of this review. Instead, we will concentrate on more advanced statistical modelling methods that are used for novelty detection involving complex, multivariate data distributions.

The estimation of some underlying data density D from multivariate training data is a well-established field [46,47]. Broadly, these techniques fall into *parametric* and *non-parametric* approaches, in which the former impose a restrictive model on the data, which results in a large bias when the model does not fit the data, while the latter set up a very flexible model by making fewer assumptions. The model grows in size to accommodate the complexity of the data, but this requires a large sample size for a reliable fit of all free parameters. Opinion in the literature is divided as to whether various techniques should be classified as parametric or non-parametric. For the purposes of providing a probabilistic estimate $\hat{D}$, Gaussian mixture models (GMMs) and kernel density estimators have proven popular. GMMs are typically classified as a parametric technique [26,24,41], because of the assumption that the data are generated from a weighted mixture of Gaussian distributions. Kernel density estimators are typically classified as a non-parametric technique [33,26,24] as they are closely related to histogram methods, one of the earliest forms of non-parametric density estimation. Parametric and non-parametric approaches are discussed in the next two sub-sections (see Table 1).

Table 1
Examples of novelty detection methods using both parametric and non-parametric probabilistic approaches.

Probabilistic approach	Section	References
Parametric	2.1	
Mixture models	2.1.1	Filev and Tseng [48,49], Flexer et al. [50], Ilonen et al. [51], Larsen [52], Paalanen et al. [53], Pontoppidan and Larsen [54], Song et al. [55] and Zorriassatine et al. [56]
Extreme value theory	2.1.1	Clifton et al. [57–61], Hazan et al. [62], Hugueny et al. [63], Roberts [64,65], Sohn et al. [66] and Sundaram et al. [67]
State-space models	2.1.2	Gwadera et al. [68,69], Ihler et al. [70], Janakiram et al. [71], Lee and Roberts [72], McSharry et al. [73,74], Ntalampiras et al. [75], Pinto et al. [76], Qiau et al. [77], Quinn et al. [2,78], Siaterlis and Maglaris [79], Williams et al. [80], Wong et al. [81,82] Yeung and Ding [83] and Zhang et al. [84]
Non-parametric	2.2	
Kernel density estimators	2.2.1	Angelov [85], Bengio et al. [86,87], Erdogmus et al. [88], Kapoor et al. [89], Kemmler et al. [90], Kim and Lee [91], Ramezani et al. [92], Subramaniam et al. [93], Tarassenko et al. [94,95], Vincent et al. [96] and Yeung and Chow [97]
Negative selection	2.2.2	Dasgupta and Majumbar [98], Esponda et al. [99], Gómez et al. [100], González and Dasgupta [101], Surace and Worden [5] and Taylor and Corne [102]

2.1. Parametric approaches

Parametric approaches assume that normal data are generated from an underlying parametric distribution with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\theta}$ is finite, and probability density function $p(\mathbf{x}, \boldsymbol{\theta})$, where $\mathbf{x}$ corresponds to an observation. The parameters $\boldsymbol{\theta}$ are estimated from the given training data. The most commonly used form of distribution for continuous variables is the Gaussian. The parameters of which are estimated from the given training data using *maximum likelihood estimates* (MLE), for which there is a closed-form analytical solution for a Gaussian distribution. More complex forms of data distribution may be modelled by mixture models such as GMMs [34,103], or other mixtures of different types of distributions such as the gamma [104,105], the Poisson [106], the Student's t [107], and the Weibull [108] distributions. In the absence of prior information regarding the form of the underlying distribution of the data, the Gaussian distribution is often used because of its convenient analytical properties when determining the location of a novelty threshold (discussed later in this section).

2.1.1. Mixture models

GMMs estimate the probability density of the target class (here the normal class), given by a training set, typically using fewer kernels than the number of patterns in the training set [46]. The parameters of the model may be estimated using maximum likelihood methods (via optimisation algorithms such as conjugate gradients or expectation-maximisation, EM) or via Bayesian methods (such as *variational Bayes*) [34]. Mixture models, however, can suffer from the requirement of large numbers of training examples to estimate model parameters. A further limitation of parametric techniques is that the chosen functional form for the data distribution may not be a good model of the distribution that generates the data. However, GMMs have been used and explored in many studies for novelty detection, as described below.

One of the major issues in novelty detection is the selection of a suitable novelty threshold. Within a probabilistic approach, novelty scores can be defined using the unconditional probability distribution $z(\mathbf{x}) = p(\mathbf{x})$, and a typical approach to setting a novelty threshold k is to threshold this value; i.e., $p(\mathbf{x}) = k$. This method has been used for novelty detection in [25,1,109] among others. However, because $p(\mathbf{x})$ is a probability density function, a threshold on $p(\mathbf{x})$ has no direct probabilistic interpretation. Some authors [1,110] have interpreted the model output $p(\mathbf{x})$ probabilistically, by considering the cumulative probability P associated with $p(\mathbf{x})$; i.e., determining the probability mass obtained by numerically estimating the integral of $p(\mathbf{x})$ over the region $\mathbf{R}$ for which the value of $p(\mathbf{x})$ is above the novelty threshold k. For unimodal distributions, one can integrate from the mode of the probability density function to the probability contour defined by the novelty threshold $p(\mathbf{x}) = k$, which can be achieved in closed form for most regular distributions. For multi-modal distributions, however, this may need to be performed using Monte-Carlo techniques, as suggested by Nairac et al. [110]. An approximation in closed form for this was proposed by Larsen [52]. The sampling approach can then be used to set thresholds in relation to the actual probability of observing novel data.

Ilonen et al. [51] introduce a confidence measure for GMM pdfs that can be used in one-class classification problems in order to select a novelty threshold. In this case, confidence is used to estimate the reliability of a classification result where a class label is assigned to an unknown observation. If the confidence is low, it is probable that a wrong decision has been made. The method is based on a branch of functional theory dealing with *high-density regions* (HDR), also termed the minimal volume region, which was originally proposed by Hyndman [111]. To determine the confidence region Ilonen et al. [51] use an approximation to find a pdf value τ which is at the border of the confidence region. It is assumed that the gradient of the pdf is never zero in the neighbourhood of any point for which the pdf value is non-zero. The proposed measure is based on the pdf density quantile; specifically, τ is computed by rank-order statistics using the density quantile $F(\tau)$ and by generating data according to the pdf. In [53], the use of confidence information was demonstrated experimentally for face detection.

Another principled method of setting thresholds for novelty detection uses extreme value theory (EVT) [112].

EVT is a branch of statistics which deals with extreme deviations of a probability distribution; i.e., it considers extremely large or small values in the tails of the distribution that is assumed to generate the data. Existing work on the use of EVT for novelty detection is described in [59]. Multivariate extrema defined in the EVT literature [113] are those n-dimensional data $\mathbf{x}_n$ that are maxima or minima in one or more dimensions of n. Rather than considering extrema in each dimension independently (yielding the case of univariate distributions), extremes with regard to the multivariate model of normality are required. EVT was first used for novelty detection in multivariate data by Roberts [64,65], with models of normality represented by GMMs. According to the Fisher–Tippett theorem [114], upon which classical EVT is based, the distribution of the *maximum* of the training set must belong to one of three families of extreme value distributions: the Gumbel, Fréchet, or Weibull distributions. The method proposed by Roberts [64,65] is concerned with samples drawn from a distribution whose maxima distribution converges to the Gumbel distribution. Although the multi-modal distribution was represented by a mixture of Gaussian component distributions, the problem was reduced to a single-component problem by assuming that the component closest to the test data point (determined using the Mahalanobis distance) dominates the EVT statistics. In that case, the EVT probability for a test point is based on the Gumbel distribution corresponding to the closest component in the GMM. The contribution of the other components is assumed to be negligible. This method was applied to different biomedical datasets, such as EEG (electroencephalography) records, in order to detect regions of (abnormal or novel) epileptic activity, and MRI (Magnetic Resonance Imaging) data, in order to find the location of tumours in an image. However, Clifton et al. [59] demonstrate that this single-component approximation is unsuitable for novelty detection when the generative data distribution is multivariate or multimodal. In general, the assumption that only the closest component distribution to the test data point needs to be considered when determining the extreme value distribution is not valid. The main reason is that the effect of other components on the extreme value distribution may be significant due to the relative differences in variances between kernels. Clifton et al. [59] propose a numerical method of accurately determining the extreme value distribution of a multivariate, multimodal distribution, which is a transformation of the probability density contours of the generative distribution. This allows the extreme value distributions for mixture models of arbitrary complexity to be estimated by finding the MLE Gumbel distribution in the transformed space. A novelty threshold may then be set on the corresponding univariate cumulative density function in the transformed space, which describes where the most extreme samples generated from the normal distribution will lie.

Clifton et al. [59] have also proposed a closed-form, analytical solution for multivariate, unimodal models of normality. Classifying the extrema of unimodal distributions for novelty detection has been the focus of a large body of work in the field of EVT [115,57,58,63,66,67,116]. In [59], the use of an alternate definition of extrema and the derivation of closed-form solutions for the distribution function over pdf values allow accurate estimates of the extreme value distributions of multivariate Gaussian kernels to be obtained. The approach was applied to patient vital-sign data (comprising heart rate and breathing rate), to identify physiological deterioration in the patients being monitored.

EVT was also applied by Hazan et al. [62] in the context of vibration log-periodograms. They proposed the use of excess-value statistics instead of maximum statistics. Under mild conditions for the pdf of the training set, the probability of the excess value can be approximated by the Generalised Pareto distribution if the novelty threshold is sufficiently large. The fault detection algorithm begins by selecting a subset of the learning dataset, comprising N log-periodograms. For each frequency in the periodogram the maximum of the log-periodograms across the subset is computed, and a "mask" is obtained. Excesses over the mask in the remainder of the training set are identified and used for estimation of the parameters of the Generalised Pareto distribution. A detection threshold is then determined: any excess over the threshold is considered to be a fault. The algorithm was evaluated in a publicly available set of vibration data, and analysis of receiver operating characteristics (ROC) curves corresponding to the results of classification showed that bearing deterioration could be identified even when little wear was present.

Yamanishi et al. [117] provide a theoretical basis and demonstrate the effectiveness of "SmartSifter", an outlier detection algorithm based on statistical learning theory. This approach was first proposed in [118] for data mining, specifically in fraud detection, network intrusion detection, or network monitoring. It uses a probabilistic model which has a hierarchical structure. While the probability density over the domain of categorical data is represented by a histogram, a finite mixture is used to represent the probability density over the domain of continuous variables. When new data are presented as input, the algorithm employs an online learning scheme to update the model, using either a *sequentially discounting Laplace estimation* algorithm for learning the histogram or a *sequentially discounting expectation and maximising* algorithm for learning the finite mixture model. The two algorithms developed by the authors gradually "discount" the effect of past examples in the online process; i.e., recent examples take a greater weight in the update algorithm than older examples. A score is then assigned to each new vector based on the normal model, measuring how much the model has changed after learning the new vector, with a high score indicating significant model change, suggesting that the new vector is a statistical outlier. The authors claim that one of the advantages of their method is its ability to adapt to non-stationary sources of data. The method has been successfully tested using network intrusion and health insurance databases from Australia's Health Insurance Commission.

Disease outbreak detection has been proposed by detecting anomalies in the event logs of emergency hospital visits [119,120]. The authors propose a hierarchical Bayesian procedure for data that arise as multidimensional arrays

with each dimension corresponding to the levels of a categorical variable. Anomalies are detected by comparing information at the current time to historical data. The distribution of data (deviations at current time) is modelled with a two-component mixture of Gaussians, one for the normal data and another for the outliers. Using Bayesian inference, the probability of an observation being generated by the outlier model is estimated. The authors assume that the priors of an observation being normal or an outlier are known *a priori*. The algorithm uses EM to set the parameters of a mixture of models for the two classes, assuming that each data point is an anomaly with *a priori* probability λ, and normal with *a priori* probability $1-\lambda$.

Zorriassatine et al. [56] use GMMs for pattern recognition in condition monitoring of multivariate processes. The methodology for applying novelty detection to bivariate processes, which was described in previous work [121], was adapted to monitor the condition of a milling process by analysing signals for 10 process variates. Signals collected from the normal states of the machining process were used to create various models of the underlying pdf for the machining process, in which each model represents the healthy states of the cutter at given combinations of machining parameters, such as the depth of cut, spindle speed, and feed rate. The centre of each Gaussian was initialised using a k-means clustering algorithm, with all model parameters then determined using EM. The GMM with the lowest training error was used as the novelty detector, with a threshold set to be the minimum log likelihood of the training data (as in [21]). Previously unseen data were used to adjust the novelty threshold using a heuristic approach.

Pontoppidan and Larsen [54] describe a probabilistic change detection framework based on Independent Component Analysis (ICA), and compare it to a GMM-based approach. Bayesian ICA using mean field training [122] is used to train the ICA model. The noise is assumed to be Gaussian and independent of the sources, with a diagonal covariance matrix. This ICA-based method was successfully applied to the detection of changes in the condition of large diesel engines using acoustic emission signals.

Flexer et al. [50] apply novelty detection to retrieve songs from a library based on similarity to test songs and genre label information. Experiments considered 2522 songs from 22 genres. Two minutes of each song were evaluated and novel data detected using two algorithms. The first, named *ratio-reject*, uses a GMM in order to estimate the pdf of the training data. The second algorithm, named *Knn-reject*, defines a neighbourhood which is used to classify unseen data. Both methods were shown to perform equally well in terms of sensitivity, specificity, and accuracy within a genre-classification context.

Filev and Tseng [48,49] describe a real-time algorithm for modelling and predicting machine health status. It utilises the concepts of fuzzy k-nearest neighbour clustering and the GMM to model the data acquired from the machine as a collection of clusters that represent the dynamics of the machine's main operating modes. The Greedy EM clustering algorithm [123] is used to identify and initialise clusters corresponding to different operating modes. An unsupervised learning algorithm then continuously reads new feature data and recursively updates the structure and the parameters of the operating mode clusters. The algorithm was validated using a set of experimental vibration data collected in an accelerated testing facility.

Song et al. [55] propose a *conditional anomaly detection* (CAD) technique by assuming that attributes are already partitioned into *contextual* and *behavioural* attributes; that is, the context of the measurements is considered before identifying a data point as "anomalous". To detect anomalies, the CAD technique learns two models: the statistical behaviour of the monitored system, and the statistical behaviour of the environment. The probabilistic links between the models are also learnt, giving a combined model of likely data that co-occurs in the environment and the system. True anomalies can then be defined as being statistically unlikely events in the system parameters that occur when the environment is normal. Within the CAD literature, the parameters from the system under study are described as the *indicator* parameters, while those from the surrounding conditions are called the *environment* parameters. The technique used in [55] for learning the indicator and environment models is Gaussian mixture modelling.

Zhang et al. [124] propose a hierarchical probabilistic model for online document clustering. The generation of new clusters is modelled with a Dirichlet process mixture model, for which the base distribution can be treated as the prior of the general English model and the precision parameter is related to the generation rate of new clusters. The empirical Bayes method is used to estimate model hyperparameters based on a historical dataset. This probabilistic model was compared with existing approaches in the literature, such as logistic regression, and applied to the novelty detection task of *topic detection and tracking*. The objective of this task was to detect the earliest report for each news event as soon as the report arrives in a temporal sequence of new stories. Results showed that the performance of the proposed method is comparable to other methods in terms of topic-weighted measure (i.e., in terms of the *topic detection and tracking* evaluation measure in which the cost of the method is computed for every event, and then the average is taken).

Perner [125] outlines a new approach to novelty detection, which is based on a *case-based reasoning* (CBR) process. The author combines statistical and similarity inference methods. This view of CBR takes into account the properties of data such as uncertainty, and underlying concepts such as adaptation, storage, and learning. Known classes are described by statistical models. The performance of the models is improved by incremental updating based on newly available events, once a sufficiently large number of cases is collected. Information about cases is now implicitly represented by the model, and storage capacity is preserved. New events, not belonging to one of the known case classes, are classified as being novel events. These events are stored by a similarity-based registration procedure in a second case base. The similarity-based learning procedure ensures that similar cases are grouped into case classes, a representative for the case class is learnt, and generalisation over case classes can be performed. Novel events can be collected and grouped so that retrieval is

efficient. When a sufficiently large number of cases is available in the second case base, the case class is passed to the statistical learning unit to learn a new statistical model. The statistical learning strategy for updating a model and learning new models is based on the *minimum message length* learning principle.

Hempstalk et al. [126] introduce a technique that combines density estimation and class probability estimation for converting one-class classification problems into binary classification tasks. The initial phase of the method involves an examination of the training data for the normal class in order to determine its distribution (such as fitting a GMM to the data). This knowledge is subsequently used in the generation of an artificial dataset that represents an abnormal class. In the second phase, a standard binary classifier is trained based on the normal class and the generated abnormal class. Using Bayes′ rule, the authors demonstrate how the class density function can be combined with the class probability estimate to yield a description of the normal class. The authors conclude that the combination of the density function with a class probability estimate (i.e., a classification model) produces an improvement in accuracy beyond that which results from one-class classification with the density function alone, but this will depend on how well the artificial dataset represents the abnormal class.

Chen and Meng [127] propose a framework for patient-specific physiological monitoring, in which the ratio the densities of training and test data [128,129] are used to define a health status index. The density ratio is estimated by a linear model whose parameter values are found using a least-squares algorithm, without involving density estimation. The training and testing data were selected from the MIT-BIH Arrhythmia Database, which contains sequences of ECG (electrocardiogram) signals acquired from 10 patients. The authors claim that the method is advantageous because density ratio parameters are estimated without involving actual density estimation (a comprehensive review of density ratio estimation methods can be found in [130]).

Another strategy for novelty detection is to use time-series methods such as the well-known stochastic process, the *Autoregressive Integrated Moving Average* (ARIMA). This may be used to predict the next data point, and hence determine if it is artefactual or not; see, for example, [131]. In the latter, an online novelty detector, the *Automatic Dynamic Data Mapper* (ADDaM), is proposed, which is based on the construction of a pdf using Gaussian kernels. Two forms of pdf are possible: a static pdf for which the prior probability of each kernel is determined by the number of observations it represents, and a "temporal" pdf for which more recent observations have a higher prior probability. Testing against the current pdf assesses the novelty of the next test point. The performance of this method for artefact detection in heart rate data (from an automatic anaesthesia monitoring database) was compared with Kalman filtering, ARIMA, and moving average filters using ROC curves. The authors reported that both ADDaM-based methods outperformed all other methods tested. Their proposed method is similar to that proposed earlier by Roberts and Tarassenko [132], in which the authors describe an offline novelty detector similarly based on a time-varying GMM, which was tested using sleep EEG data. Their novelty detector was based on "growing" the mixture model over time using a process similar to reinforcement learning. During learning, training data were presented at random and were either added to regions that formed the basis of the Gaussian kernels in the model or used to estimate the parameters of new kernels. A variant of these regression-based techniques, which detects anomalies in multivariate time-series data generated by an *Autoregressive Moving Average* (ARMA) model, was proposed in [133]. Here, a multivariate time-series is transformed into a univariate time-series by linearly combining the components that are obtained using a *projection pursuit* technique, which maximises the kurtosis coefficient of the time-series data. Univariate statistical tests are then used for anomaly detection in each resulting projection. Similar approaches have been applied in ARIMA models [134], or have been used to detect anomalies by analysing the *Akaike Information Criterion* during model-fitting [135].

2.1.2. State-space models

State-space models are often used for novelty detection in time-series data. These models assume that there is some underlying hidden state that generates the observations, and that this hidden state evolves through time, possibly as a function of the inputs. The two most common state-space models used for novelty detection are the Hidden Markov Model (HMM) and the Kalman filter.

HMMs include temporal dependence through the use of a state-based representation updated at every time step. While the features are directly observable, the underlying system states are not, and hence they are called *unobservable*, *hidden*, or *latent* states. The transitions between the hidden states of the model are governed by a stochastic process [33,4]. The "emission" probabilities of the observable events are determined by a set of probability distributions associated with each state, which can be thresholded to perform novelty detection [83]. HMM parameters are typically learnt using a variant of the EM algorithm. Novelty detection with HMMs may also be achieved by constructing an "abnormal" state, a transition into which implies abnormal system behaviour [136].

Yeung and Ding [83] use HMMs for detection of intrusions based on shell command sequences within the network security domain. Two different types of behavioural models are presented: a dynamic modelling approach based on HMMs and a static modelling approach which is based on the frequency distributions of event occurrences. In the former, the parameters of an HMM for modelling normal system behaviour are estimated using an EM algorithm for mixture density estimation. The likelihood of an observation sequence, with respect to a given trained HMM, is computed using either the "forward" or "backward" algorithm. By comparing the likelihood of an observation sequence against a threshold (chosen to be the minimum likelihood among all training sequences), one can decide whether that sequence is abnormal or not. In the second approach, the probability distribution of normal behaviour of the system observed

over a certain period of time is modelled using a simple occurrence frequency distribution. The behaviour of the test system being monitored is modelled in the same way. An information-theoretic measure, *cross-entropy*, which is related to the *Kullback–Leibler measure*, is used to quantify separation between the two distributions (corresponding to "training" and test sequences). The Kullback–Leibler divergence is a statistical tool for estimating the difference in information content between two distributions. By checking whether the cross-entropy between the two distributions is larger than a certain threshold, chosen to be the maximum cross-entropy value computed between the entire training set and each time-series in the training set, one can decide whether the observed sequence should be considered an intrusion with respect to the model. Although the HMM is better suited for intrusion detection based on Unix system calls, the static modelling approach based on the information-theoretic technique outperformed the dynamic modelling approach across all experiments. Other similar intrusion detection methods based on HMMs have been proposed [77,84].

Ntalampiras et al. [75] explore HMMs for novelty detection applied to acoustic surveillance of abnormal situations, the goal being to help an authorised person take the appropriate action for preventing life or property loss. The HMM is a model commonly used in sound recognition, as it takes into account the temporal evolution of the audio signal. The framework was tested using a dataset that contains recordings from a smart-home environment, an open public space, and an office corridor.

A related state-based approach to novelty detection in time-series relies on Factorial Switching Kalman Filters [2]. A Kalman Filter can be seen as a generalisation of an autoregressive process, describing an observed process in terms of an evolving hidden state process. This may be generalised to the Switched Kalman Filter (SKF) [137], in which the evolving hidden process is dependent on a switching variable (which also evolves through time). The Factorial SKF (FSKF) is a dynamic extension of the SKF, in which a cross-product of multiple factors is used rather than a single variable. The FSKF allows the modelling of multiple time-series by assuming that a continuous, hidden state is responsible for data generation, the effects of which are observed through a noise process. An explicit abnormal mode of behaviour is included within the model which is used to identify departures from normality. This method was applied to the monitoring of premature infants in intensive care, and is described in [2,80]. The method was later extended in [78], which used a dataset of continuously observed physiological variables such as heart rate and blood pressure. Lee and Roberts [72] propose an online novelty detection framework using the Kalman filter and EVT. A multivariate Gaussian probability density over the target variables is obtained via a Kalman filter, with an auto-regression state model. This was used to model the dynamics of the state space and thereby to detect changes in the underlying system, as well as identify outliers in the observation sequence. EVT is then used to define a threshold and obtain a novelty measure on the univariate predictive distribution. Experiments were conducted on three univariate data sets: an artificial dataset, in addition to the Well-Log[1] and Dow Jones[2] datasets.

Also based on a dynamical model of time-series normal data, the Multidimensional Probability Evolution method [73,74] characterises normal data by using a non-linear state-space model; i.e., the pdf within a multidimensional state space is computed for each window of the time-varying signal. The regions of state space visited during normal behaviour are modelled, and departures from these, that can correspond to both linear and non-linear dynamical changes, are deemed abnormal. The performance of this technique was illustrated using a synthetic signal, in addition to electroencephalography (EEG) recordings to identify epileptic seizures.

A task related to that of time-series novelty detection is to determine whether a pattern discovered in the data is significant. Assuming an underlying statistical model for the data, one can estimate the expected number of occurrences of a particular pattern in the data. If the number of times a pattern actually occurs is significantly different from this expected value, then it could be indicative of unusual activity (and thus the pattern discovered may be regarded as being significant). Furthermore, since the statistics governing the data generation process are assumed to be known, it is possible to quantify the extent of deviation from the expected value that corresponds to a test pattern being classified as "significant". An application to the "frequent episode discovery problem" in temporal data mining is presented in [69]. It is shown that the number of sliding windows over the data in which a given episode occurs at least once converges to a Gaussian distribution with mean and variance that can be determined from the parameters of an underlying Bernoulli distribution (which are in turn estimated from some training data). For a pre-defined confidence level, upper and lower thresholds for the observed frequency of an episode can be determined, which can be used to decide whether an episode is over- or under-represented in the data. These ideas are extended in [138] to the case of determining significance for a set of frequent episodes, and in [68] to the case of a Markov model assumption for the data sequence.

Ihler et al. [70] consider the modelling of web click data. The proposed method is based on a time-varying Poisson process model that can account for anomalous events. The normal behaviour in a time-series is assumed to be generated by a non-stationary Poisson process while the outliers are assumed to be generated by a homogeneous Poisson process. The transition between normal and outlying behaviours is modelled using a Markov process. Markov Chain Monte Carlo (MCMC) is used to estimate the parameters of these processes. A test time-series is modelled using this process and the time points for which the outlying model is selected are considered as outliers.

[1] The Well-Log data set contains measurements of nuclear magnetic response while drilling a well.

[2] The Dow Jones data set contains daily stock market indexes (Industrial Average), that show how 30 large publicly owned companies based in the United States have traded during standard trading sessions on the stock market.

Both methods described above can be generalised to Dynamic Bayesian Networks (DBNs), which are more general state-space models [34]. DBNs generalise HMMs and Kalman filter models by representing the hidden and observed states in terms of state variables, which can have complex interdependencies. A DBN is a directed *probabilistic graphical model* of a stochastic process, which provides an easy way to specify these conditional independencies. They can also be seen as an extension of Bayesian networks to handle temporal models. A Bayesian network estimates the probabilistic relationship among variables of a dataset, in the form of a probabilistic graphical model. In addition to DBNs, Bayesian networks are sometimes termed *naïve Bayesian networks* or *Bayesian belief networks*. Janakiram et al. [71] propose a classification system based on a Bayesian belief network (BBN) to detect any missing or anomalous data in wireless sensor networks. Each node in the graph corresponds to a sensor, and models sensor measurements from neighbouring nodes in addition to its own. The authors then estimate the probability of each observed attribute using the BBN model. This model is suitable if some dependency exists between sensor variables and between nodes. Its accuracy depends on the number of neighbours that each node has. The technique requires offline training, and regeneration of a conditional probability table for each node if the network topology changes. BBNs are also used to incorporate prior probabilities into a novelty detection framework. Several variants based on naïve Bayes, which assumes independence between the variables, have been proposed for network intrusion detection [139,79] and for disease outbreak detection [81,82].

More recently, Pinto et al. [76] have proposed novelty threshold functions that operate on top of probabilistic graphical models instantiated dynamically from sensed semantic data in the context of room categorisation. By using thresholds on the distributions defined by the graph based solely on the conditional probability, as seen in [34], a novelty system can be implemented. However, it may not be suitable to perform novelty detection using graphs that are dynamically generated. Pinto et al. [76] show that the ratio between a conditional and unconditional probability is a suitable detector for implementing a threshold when samples are taken from dynamic distributions, under the assumption that the probability of a sample being generated by a (novel) unknown class is constant across all graph structures. This assumption may not be appropriate for some graph structures; e.g., a graph where there is only access to room-size information versus a graph where there is more information concerning the properties of the room available. The authors also show that correct estimation of unconditional probability plays an important role, and that unlabelled data can be used to construct an unconditional pdf that can then be used to optimise the novelty threshold. However, only synthetic data distributions were used to evaluate the effectiveness of the approach.

2.2. Non-parametric approaches

Non-parametric approaches do not assume that the *structure* of a model is fixed, i.e., the model grows in size as necessary to fit the data and accommodate the complexity of the data. The simplest non-parametric statistical technique is the use of histograms which graphically display tabulated frequencies. The algorithm typically defines a distance measure between a new test data point and the histogram-based model of normality to determine if it is an outlier or not [36]. For multivariate data, attribute-wise histograms are constructed and an overall novelty score for a test data point is obtained by aggregating the novelty scores from each attribute. This has been applied to network intrusion and web-based attack detection [117,140–142].

2.2.1. Kernel density estimators

A non-parametric approach to probability density estimation is the kernel density estimator [34]. In this approach, the probability density function is estimated using large numbers of kernels distributed over the data space. The estimate of the probability density at each location in data space relies on the data points that lie within a localised neighbourhood of the kernel. The kernel density estimator places a (typically Gaussian) kernel on each data point and then sums the local contributions from each kernel. This kernel density estimator is often termed the *Parzen windows* estimator [143]. This method has been used for novelty detection in applications such as network intrusion detection [97], oil flow data [21], and for mammographic image analysis [1]. In the Parzen Windows estimator, an isotropic Gaussian kernel is centred on each training point, with a single shared variance hyper-parameter. Training the Parzen density estimator consists of determining the variance of the kernels, which controls the smoothness of the overall distribution. The fixed width in each feature direction also means that the Parzen density estimator is sensitive to the scaling of the data. This problem is addressed in [21], in which the variance is determined using a nearest-neighbour method.

Vincent and Bengio [96] propose an approach to improve on this estimator, by using general covariance matrices for individual components set according to neighbourhoods local to each kernel. Not only are the localisation of the data point and its neighbours used but also their geometry, in order to try and infer the principal characteristics of the local shape of the manifold (where the density is concentrated), which can be summarised by the covariance matrix of the Gaussian. Bengio et al. [87] describe a non-local non-parametric density estimator which builds upon previously proposed GMMs with regularised covariance matrices to take into account the local shape of the manifold. The proposed approach builds upon the Manifold Parzen density estimator [96] that associates a regularised Gaussian with each training point, and upon previous work on non-local estimators of the tangent plane of a manifold [86]. The local covariance matrix characterising the density in the immediate neighbourhood of a data point is learnt as a function of that data point, with global parameters. Generalisation may then be possible in regions with little or no training data, unlike traditional, local, non-parametric models. The implicit assumption is that there is some kind of regularity in the shape of the density, such that learning about its shape in

one region could be informative of the shape in another region that is not adjacent. The proposed method was tested in three types of experiments involving artificial datasets and the USPS[3] dataset, which showed that the non-local estimator yielded improved density estimation and reduced classification errors when compared to local algorithms.

Erdogmus et al. [88] describe a multivariate density estimation method that uses Parzen windows to estimate marginal distributions from samples. The kernel size is optimised to minimise the Kullback–Leibler divergence of the true marginal distribution from the estimated marginal density. The estimated marginal densities are used to transform the random variables to be Gaussian-distributed, whereby joint statistics can be simply determined by sample covariance estimation. The proposed method was shown to be more data efficient than Parzen windows with a structured multidimensional kernel.

Subramaniam et al. [93] use kernel density estimators in a framework that computes an approximation to multidimensional data distributions in order to enable complex applications in resource-constrained sensor networks. The authors propose an online approximation of the data distribution by considering the values in a sliding window. The variance of the kernel for the values in the sliding window is computed using a histogram along the time axis. A network of nodes is considered, where the estimator updates are propagated around the network such that child nodes transmit updates to the parent nodes. Experiments showed that this method can achieve high precision for identifying outliers, but that it consumes a large amount of memory space and may not find all outliers.

Tarassenko et al. [94,95] propose an approach to patient monitoring based on novelty detection, in which a multivariate, multimodal model of the distribution of vital-sign data from "normal" high-risk patients is constructed using Parzen windows. Multivariate test data are then compared with this model to give a novelty score, and an alert is generated when the novelty score exceeds the novelty threshold. This system was used for monitoring patients in a clinical trial involving 336 patients [145], and it was able to provide early warning of adverse physiological events.

Ramezani et al. [92] consider the problem of novelty detection in video streams, and use a method derived from a kernel density estimator and an evolving clustering approach, "e-Clustering". In this method, the pdf of the colour intensity of the image frames is approximated by a Cauchy kernel. A recursive expression derived in [85,146] is then used to update this estimation online. The recursive density estimation clusters pixel colour intensities into "background" and "foreground" (i.e., pixels for which significant novelty is detected). The proposed approach gradually updates the background model, and it was found to be faster than the traditional kernel density estimate for background subtraction. The approach can also be extended to automatic object tracking when combined with Kalman filter or evolving Takagi-Sugeno fuzzy models [92,85].

More recently, one-class classification using Gaussian Processes (GPs) has been proposed [147,89–91], in which a point-wise approach to novelty detection is also taken, dividing the data space into regions with high support and low support depending on whether or not those regions are close to those occupied by normal training data, or not, respectively. It is often assumed that the desired mapping from inputs $\mathbf{x}$ to labels y can be modelled by $y=f(\mathbf{x})+\varepsilon$ where f is an unknown latent function and ε is a noise term. By choosing a proper GP prior, it is possible to derive useful membership scores for one-class classification. In [90], the authors use a mean of the prior with a smaller value than the positive class labels (e.g., $y=1$), such as a zero mean. This restricts the space of probable latent functions to functions with values gradually decreasing when the inputs are far from training points. Because the predictive probability is solely described by its first and second order moments, the authors also investigate the power of the predictive mean and variance as alternative membership scores: the mean decreases for inputs distant from the training data and can be directly utilised as a novelty detection measure, while the predictive variance increases which suggests that the negative variance value can serve as an alternative criterion for novelty detection. This latter concept is used in the context of clustering in [91]. Kemmler et al. [90] explore an heuristic measure: the predictive mean divided by the standard deviation, which was proposed in [89] as a combined measure for describing the uncertainty of the estimation.

A related approach involves a class of methods which are part of the well-established field of *changepoint* detection [148,149]. Here the problem setting is more specific, the aim being (typically) to detect whether the generative distribution of a sequence of observations has remained stable or has undergone some abrupt change. This may include not only detecting that a change has occurred but also, if it has occurred, estimating the time at which the change has occurred. Methods vary according to which restrictions are placed on the pre- and post-change distributions and the degree of knowledge of potential changes. Changepoint detection can also take place in a batch or online setting. The basic approach to the retrospective problem is to find a test statistic appropriate for testing the hypothesis that a change has occurred with respect to the hypothesis that no change has occurred. This statistic is usually based on a likelihood ratio, but other approaches exist [149,150]. The online setting has the goal of detecting a change as quickly as possible once it has occurred. The most basic approach to this is also based on likelihood ratios. The Cumulative Summation (CUSUM) algorithm is generally used in statistical process control to detect abrupt changes in the mean value of a stochastic process [151]. Non-parametric CUSUM algorithms sequentially accumulate data values that are higher than the mean value observed under normal conditions. An anomaly is detected by comparing this CUSUM value to a threshold, where the latter determines the sensitivity of the detector and the detection delay. This approach has been used in wireless sensor networks [152] and intrusion

[3] The USPS (United States Postal Service) dataset contains handwritten digit images, and comes from the UCI repository [144].

detection systems [151]. Bayesian approaches have also been proposed for both offline and online changepoint detection schemes [153]. The Bayesian framework incorporates prior knowledge about the distribution of the change time. The decision function is then based on the *a posteriori* probability of a change. Because it is often hard in practice to elicit specific information about the distribution of the changepoint, it is common to assume that the change time follows a geometric distribution, but finding the parameter of this distribution requires a preliminary estimation problem to be solved. There is a large body of literature on the topic of changepoint detection, and the interested reader is referred a book on this topic which has recently been published [154].

2.2.2. Negative selection

Negative selection approaches have been widely used for change detection and novelty detection [155,98,101,102]. The nature of the *negative selection algorithm* [155] is inspired by the properties of the immune system. The human immune system has the ability to detect *antigens*; i.e., anything which is not part of the human body (such as viruses, bacteria, etc.). The task of the immune system is to differentiate between antigens and the body itself, a process known as *self/non-self discrimination*, which is achieved when the antigen is "recognised" by a specific antibody called a T-cell receptor. These T-cell receptors are created by a random process of genetic rearrangements. Those cells that successfully bind with self-cells (normal cells) and thus incorrectly mark them for destruction, are destroyed in the thymus gland. Only those cells that fail to bind to self-cells are allowed to leave the thymus gland and become antibodies in the immune system. This process is called *negative selection*. In a similar way, novelty detection has the fundamental objective of distinguishing between *self* (which corresponds to the normal operation of the monitored system) and *non-self* (which corresponds to novel data).

The negative selection algorithm was first used by Forrest et al. [155] as a means of detecting unknown or illegal strings for virus detection in computer systems. A population of detectors is created to perform the function of T-cells, which are simple, fixed-length binary strings. A simple rule based on a distance measure is then used to compare bits in two such strings and decide whether a match has occurred. One major concern identified by the authors is the matching threshold, which has to be data specific for satisfactory system performance. Dasgupta and Majumdar [98] extended the approach for use with multi-dimensional data, where the dimensionality of the original data is first reduced to two dimensions using principal component analysis, and then binary encoded. The authors conclude that the encoding of self and non-self sets using binary strings results in a risk of destroying the semantic value of relationships between data items. Later, the authors propose a *real-valued negative selection* algorithm [101]. The main feature of the latter is that the self/non-self space corresponds to a subset of the original n-dimensional data space. A detector (antibody) is defined by a hypersphere; i.e., an n-dimensional vector that corresponds to the centre of the sphere and a scalar value that represents its radius. The matching rule is expressed by a membership function, which is a function of the detector-antigen Euclidean distance and the radius of the detector. The algorithm tries to evolve a set of points (antibodies or detectors) that covers the non-self space using an iterative process that updates the position of the detector. In order to detect if a detector matches a self point, the algorithm uses a nearest-neighbour approach to calculate a distance measure. A different approach is also taken in which a multi-layer perceptron trained with back-propagation is used for the detection problem. The system was tested on network traffic data sets. Gómez et al. [100] extended the negative characterisation approach to generate more flexible boundaries between self and non-self space using fuzzy rules. Taylor and Corne [102] demonstrate the feasibility of the above approaches in fault detection in refrigeration systems. Esponda et al. [99] describe a framework for outlier detection using this general approach.

Surace and Worden [5] apply the negative selection algorithm to more general feature sets, rather than the windowed time-series data used in previous studies. The authors consider Structural Health Monitoring (SHM) for situations where the normal condition of a structure may change due to time-varying environmental or operational conditions. Data were simulated for an offshore platform model with changing mass as a result of changing oil storage requirements. The method was also applied to the analysis of the structure of a transport aircraft, for which the effective mass decreases due to a reduction in fuel, simulated using a finite-element model. The negative selection algorithm proved capable of distinguishing various damage conditions in structures induced by time-varying oil storage and fuel use.

2.3. Method evaluation

Probabilistic approaches are mathematically well-grounded and can effectively identify novel data if an accurate estimate of the pdf is obtained. Also, after the model has been constructed, only a minimal amount of information is required to represent it, rather than requiring the storage of the entire set of data used for training. Probabilistic methods are also known to be "transparent" methods, i.e., their outputs can be analysed using standard numerical techniques. However, the performance of these approaches is limited when the size of the training set is very small, particularly in moderately high-dimensional spaces. As the dimensionality increases, the data points are spread through a larger volume in the data space. The problem encountered when applying density methods to sparsely populated training sets is that there is little control over the inherent variability introduced by the sparsity of the training data; i.e., the estimated quantiles can differ substantially from the true quantiles of the distribution. Different approaches have been proposed to overcome the problems associated with increasing dimensionality which both increase the processing time and distort the data distribution. In many real-life scenarios, no *a priori* knowledge of the data distributions is available, and so parametric approaches may be problematic if the

Table 2
Examples of novelty detection methods using distance-based approaches.

Distance-based approach	Section	References
Nearest neighbour	3.1	Angiulli and Pizzuti [156], Bay and Schwabacher [157], Boriah et al. [158], Breunig et al. [159], Chandola et al. [160], Chawla and Sun [161], Ghoting et al. [162,163], Hautamaki et al. [164], Jiang et al. [165], Kou et al. [166], Otey et al. [167], Palshikar [168], Pokrajac et al. [169], Wu and Jermaine [170] and Zhang and Wang [171]
Clustering	3.2	Barbará et al. [172,173], Budalakoti et al. [174], Clifton et al. [175,176], Filippone et al. [177], He et al. [178], Kim et al. [179], Srivastava and Zane-Ulman [180,181], Sun et al. [182], Syed et al. [183], Wang [184], Yang and Wang [185], Yong et al. [186,187], Yu et al. [188] and Zhang et al. [189]

data do not follow the assumed distribution. Thus, non-parametric techniques are appealing since they make fewer assumptions about the distribution characteristics. Kernel functions, for example, generally scale reasonably well for multivariate data and are not computationally expensive.

3. Distance-based novelty detection

Distance-based methods, including clustering or nearest-neighbour methods (Table 2), are another type of technique that can be used for performing a task equivalent to that of estimating the pdf of data. These methods rely on well-defined distance metrics to compute the distance (similarity measure) between two data points.

3.1. Nearest neighbour-based approaches

Nearest neighbour-based approaches are among the most commonly used methods for novelty detection. The k-nearest neighbour (k-NN) approach is based on the assumption that normal data points have close neighbours in the "normal" training set, while novel points are located far from those points [164]. A point is declared as an outlier if it is located far from its neighbours. Euclidean distance is a popular choice for univariate and multivariate continuous attributes, but other measures, such the Mahalanobis distance, can be used. For categorical attributes, a simple matching coefficient is often used, although other more complex measures have been proposed [158,160]. Several well-defined distance metrics to compute the distance (or similarity measure) between two data points can be used [33], which can broadly be divided into distance-based methods, such as the distance to the kth nearest neighbour [171], and local density-based methods in which the distance to the average of the k nearest neighbours is considered [164]. Many of these algorithms are unable to deal with high-dimensional data sets efficiently. A recent trend in high-dimensional outlier detection is to use the evolutionary search method where outliers are detected by searching for sparse subspaces.

The approach proposed in [156] considers a weighted sum of the distances from the k nearest neighbours to each data point, and classifies as outliers those points which have the largest weighted sums. The k nearest neighbours of each point are found by linearising the search space using a Hilbert space curve. This work is built upon previous techniques that prune the search space for nearest neighbours [190]. The latter partition the data space into a grid of hypercubes of fixed sizes. If a hypercube contains many data points, such points are likely to be normal. Conversely, if a test point lies in a hypercube that contains very few examples, the test point is likely to be an outlier. In [156], a high-dimensional data set is mapped onto the interval $[0,1]^n$ using Hilbert space-filling curves. Each successive mapping improves the estimate of the example's outlier score in the original high-dimensional space. In related works, [168] adapts the technique proposed in [190] to continuous sequences, and [166] incorporates spatial correlation between data. These analyses are related to the very well established application domain of Rough Sets, and indeed a formalisation of a similar approach within the framework of Rough Sets has been proposed in [165].

Zhang and Wang [171] describe the "HighDOD" method, the *High-Dimension Outlying Subspace Detection* method, for efficiently characterising the outlying subspaces of high-dimensional data spaces. The novelty score of a point is measured using the sum of distances between it and its k nearest neighbours, as in [156]. Two heuristic pruning strategies are proposed to perform fast pruning in the search, and an efficient dynamic method with a sample-based learning process is described. The dynamic subspace search method begins the search in those subspaces that have the highest *total saving factor*. The total saving factor of a subspace is defined to be the combined savings obtained by applying the two pruning strategies during the search process. As the search proceeds, the total saving factor of subspaces with different dimensions is updated and the set of subspaces with the highest values are selected for exploration in each subsequent step. The search process terminates when all the subspaces have been evaluated or pruned. Experiments were performed using both synthetic and real high-dimensional data sets, ranging from 8 to 160 dimensions.

Different approaches were taken in [157,163], which prune the search by ignoring data that cannot be considered to be outliers. Bay and Schwabacher [157] introduced the distance-based outlier detection algorithm called "ORCA". The authors showed that for sufficiently randomised data, a simple pruning step could result in the average complexity of the nearest neighbour search being approximately linear. After finding the nearest neighbours for a point, a threshold based on the score of the weakest outlier (i.e., the outlier that is closer to the point) found so far is set for any new data point. Therefore points that are close are discarded by the algorithm. To improve the performance of ORCA, Ghoting et al. [163] proposed the

Recursive Binning and Re-projection algorithm. In the first of two phases, a divisive hierarchical clustering algorithm is used to partition objects into bins or clusters. Objects in the same bin are reorganised according to a projection along their principal component. In the second phase, the strategy of ORCA [157] is employed on the clustered data. This pre-processing step allowed faster determination of the closest nearest neighbours compared with ORCA. A similar cluster-based pruning has been proposed in [191]. Wu and Jermaine [170] used a sampling algorithm to improve the efficiency of detection of the nearest neighbour-based technique. Rather than using the entire dataset, the nearest neighbour of every data point is computed only within a smaller sample of the dataset. This reduces the complexity of the proposed method according to the sample size chosen.

While most techniques discussed so far in this category have been proposed to handle continuous attributes, variants have been proposed to handle other data types. Wei et al. [192] propose a *hypergraph*-based technique called HOT, an efficient approach for detecting local outliers in categorical data. A hypergraph can be defined as a generalised graph, consisting of a set of vertices and hyper-edges. The authors use hyper-edges, which simply store "frequent itemsets" (commonly used terms in association rule mining) along with the data points (vertices) that contain these frequent itemsets. First, all the frequent itemsets are mined by using the Apriori algorithm [193]. Then, they are arranged in a hierarchy according to the containment relationship. The hierarchy is visited using a bottom-up strategy. Frequent itemsets I represent common attributes, while each attribute A not in I represents a potential exceptional attribute. For each itemset I, the histogram of frequencies associated with each attribute A not in I is stored, and used to compute the deviation of each value taken by A in the database. The objects assuming a value for the attribute A whose deviation is smaller than a defined threshold are returned as outliers. Two advantages of this method are that (i) it alleviates the problem of the curse of dimensionality in very large databases, and (ii) it uses the connectivity property of points to deal efficiently with missing values.

Otey et al. [167] and Ghoting et al. [162] propose a distance measure for data containing a mix of categorical and continuous attributes. The authors define links between two points by taking into account the dependencies between continuous and categorical attributes, where the distances for categorical and continuous attributes are considered separately. For categorical attributes, the distance between two points is defined to be the number of attributes which take the same values; two points are considered linked if they have at least one common attribute-value pair. The number of attribute-value pairs in common indicates the strength of the associated link between these two points. For continuous attributes, a covariance matrix is maintained to capture the dependencies between the continuous values. In a mixed attribute space, the dependence between the values with mixed continuous and categorical attributes is captured by incremental maintenance of the covariance matrix. Thus, a data point can be considered to be an outlier if the expected dependencies between categorical and continuous data are violated by it. We note that the construction of the covariance matrix implies an assumption that the data share the same distribution, which may not hold for real-life applications.

A density-based scheme for outlier detection has been proposed in [159], in which a *Local Outlier Factor* (LOF) is computed for each point. The LOF of a point is based on the ratios of the local density of the area around the point and the local densities of its neighbours. The size of the neighbourhood of a point is determined by the area containing a user-supplied minimum number of points. The LOF takes high values for outliers, because it quantifies how isolated the point is with regard to the density of its neighbourhood. Note that LOF ranks points by only considering the neighbourhood density of the points, and so it may miss potential outliers whose densities are close to those of their neighbours.

A similar technique called LOCI (*Local Correlation Integral*) is presented in [194]. LOCI addresses the difficulty of choosing values for the number of neighbours in the LOF technique by using data-driven methods. The local neighbourhood is defined such that each point has the same radius of neighbourhood, instead having a fixed number of neighbours. It uses the concept of a *multi-granularity deviation factor*, to measure the relative deviation of a point's local neighbourhood density from the average local neighbourhood density in its neighbourhood. A point can then be declared as an outlier by comparing its factor with a data-derived threshold value. The choice of an appropriate radius for the local neighbourhood becomes critical for high-dimensional datasets.

A variant of LOF was proposed in [195]. The *GridLOF* uses a simple grid-based technique to prune away some non-outliers and then only computes the LOF values for the remaining data. This avoids the computation of LOF for all points. Tang et al. [196] present a variation of the LOF that considers both the density of a test point in its neighbourhood and the degree that the point is connected to other points; it uses a *connectivity-based outlier factor* to identify outliers. This factor is calculated using the ratio of the average distance from the test point to its k-nearest neighbours and the average distance from its k-nearest neighbours to their own k-nearest neighbours. Points that have the largest factors are declared as outliers. This approach was found to be a more effective approach for outlier detection, especially for sparse data sets, where "non-outlier" patterns may have low densities.

Ren et al. [197] develop an efficient density-based outlier detection approach based on a relative density factor (RDF). This is another local density measurement for determining the degree of being an outlier by contrasting the density between a point and that of its neighbours. A *P-Trees* approach is used to efficiently prune some non-outliers, and the remaining subset of the data is then used to compute the RDF. Points with an RDF greater than a pre-defined threshold are considered outliers.

Yu et al. [198] propose an outlier detection approach for detecting "outliers" that may occur within the loci of "normal" data, rather than being far from the "normal" points in data space, for both categorical and numerical

data. Similarity between points is measured by a similarity graph, which is a weighted, connected, undirected graph. A weight for a pair of points specifies the similarity between them. A point can be considered to be an outlier if its similarity relationship with its neighbours is lower than the similarity relationships among its neighbours' neighbourhood. This use of similarity graphs overcomes the disadvantage of the traditional similarity measure (which assumes that outliers are far away from the "normal" points in data space) and can easily be applicable for categorical as well as numerical data. Several other variants of the LOF method have been proposed to handle different data types [199] and applied for detecting spatial outliers or anomalies in climate data [161,200], protein sequences [201], network intrusion [202,164], and video sensor data [169].

3.2. Clustering-based approaches

Clustering-based approaches to distance-based novelty detection include methods such as the *k*-means clustering. In this general type of methods, the "normal" class is characterised by a small number of prototype points in the data space. The minimum distance from a test point to the nearest prototype is often used to quantify abnormality. The methods use different approaches to obtain the prototype locations. The *k*-means clustering algorithm is perhaps the most popular method of clustering structured data due to its simplicity of implementation [180,181]. Kim et al. [179] use novelty detection to identify faulty wafers in semiconductor manufacturing. Among other methods, the authors employed the *k*-means clustering technique, GMMs, Parzen windows, and other non-probabilistic methods, such as one-class support vector machines and reconstruction methods based on principal component analysis. The *k*-means algorithm works by choosing *k* random initial cluster centres, computing the distances between these cluster centres and each point in the training set, and then identifying those points that are closest to each cluster centre. The corresponding cluster centres are moved to the centroid of those nearest points and the procedure is repeated. The algorithm converges when the cluster centres do not move from one iteration to the next. This procedure is often used as a pre-processing procedure [95].

Among the prototype-based clustering algorithms, we can identify many modifications of the *k*-means algorithm. Popular fuzzy-clustering algorithms are the fuzzy versions of the *k*-means algorithm with probabilistic and possibilistic descriptions of memberships: fuzzy *c*-means [203] and possibilistic *c*-means [204], respectively. The latter technique has recently been extended by Filippone et al. [177]. In the proposed extension, positive semidefinite kernels are used to map implicitly input patterns into a high-dimensional space, in which the mapped data are modelled by means of the possibilistic clustering algorithm. Wang [184] presents a hybrid approach that incorporates two kernel-based clustering methods (using the concepts of fuzzy and possibilistic *c*-means) for outlier identification and market segmentation.

Different clustering-based techniques have been proposed [188,178,205,182]. Yu et al. [188] propose an outlier detection approach based on a wavelet transform, which can be extended to detect outliers in datasets with different densities. This approach uses wavelets to transform the data and then find dense clusters in the transformed space. He et al. [178] present a new definition of a cluster-based local outlier, which takes into account both the size of a point's cluster and the distance between the point and its closest cluster. Each point is associated with a cluster-based local outlier factor, which is used to determine the likelihood of the point being an outlier. This approach partitions the data into clusters using a *squeezer* algorithm, which makes a single pass over the dataset and produces initial clustering results. The outlier factor is then computed for each point, and those points which have the largest factors are considered outliers. This approach is linearly scalable with respect to the number of data points and was found to work well with large datasets. Another technique that addressed computational efficiency was proposed by Sun et al. [182], in which an efficient indexing technique called *CD-trees* was used to partition data into clusters. Those points belonging to sparse clusters are declared anomalies.

Basu et al. [15] introduce a method for semisupervised clustering that employs Hidden Random Markov Fields (HMRFs) to use both labelled and unlabelled data in the clustering process. The method can be used with a number of distortion measures, including *Bregman divergences* (such as the Kullback–Leibler divergence) and directional measures. The authors propose an EM-based clustering algorithm, *HMRF-KMEANS* that incorporates supervision in the form of pairwise constraints at all stages of the algorithm: initialisation, cluster assignment, and parameter estimation. The HMRF method led to improved results when applied to realistic textual datasets, in comparison to unsupervised clustering methods. The algorithm discussed in [206] seeks to minimise class differences between nearby points and maximise class differences between distant points. It was applied to the classification of representations of handwritten digits and a synthetic control time-series. A similar approach was applied to network intrusion detection tasks in [207], combining factor analysis and the Mahalanobis distance metric.

Clustering data that are represented by a sequence of individual symbols was considered by Yang and Wang [185], who describe a clustering algorithm called *CLUSEQ* that produces a set of overlapped clusters, and which is able to adjust automatically the number of clusters and the boundary used to separate "normal" sequences from outliers. The similarity measure uses the conditional probability distribution derived from sequences, where a variation of the suffix tree (the probabilistic suffix tree) is used to estimate the distribution. The CLUSEQ algorithm takes a training set of sequences and a set of tree parameters as input from which it produces a set of clusters. An iterative process is used in which, for each iteration, a set of new clusters is generated from the set of unclustered sequences to augment the current set of clusters, which is followed by a sequential examination of every sequence to

evaluate its similarity to each cluster and to update its cluster membership. At the end of each iteration, a consolidation procedure is invoked to merge heavily overlapped clusters. Budalakoti et al. [174] propose a different outlier detection approach that efficiently clusters sequence data into groups and finds anomalous sequences that deviate from normal behaviour. A fast "normalised longest common subsequence" (nLCS) is used as the similarity measure for comparing sequences.

Syed et al. [183] explore *k*-NN and clustering-based novelty detection approaches to identify high-risk patients. For many clinical conditions, patients experiencing adverse outcomes comprise a small minority of the population. When evaluated with demographic and comorbidity data acquired from over 100,000 patients, the methods considered were able to identify patients at an elevated risk of mortality and morbidity following surgical procedures.

Barbará et al. [173] suggest the use of a bootstrapping technique that first separates normal data from outliers using frequent itemset mining. Data are windowed in time, and frequent itemsets then generated for each window. All itemsets which exist in more than one window are considered normal. Clusters were obtained using *COOLCAT*, a clustering tool for categorical data developed in [172]. This method was applied to intrusion detection tasks.

Ertöz et al. [208] explore a clustering algorithm based on a shared nearest-neighbour approach. This technique first finds the nearest neighbours of each point and then redefines the similarity between pairs of points in terms of how many nearest neighbours the two points share. Using this definition of similarity, the algorithm identifies "core points" and then builds clusters around the core points. The use of a shared nearest neighbour definition of similarity alleviates problems with varying densities and high dimensionality, and the use of core points handles problems with shape and size of the distribution. The number of clusters is automatically determined by the location and distribution of core points. Another novel aspect of the shared nearest neighbour clustering algorithm is that the resulting clusters do not contain all points, but contain only those points lying in regions of relatively uniform density. The authors apply this algorithm to the task of finding topics in collections of documents, for which it out-performed the *k*-means clustering method.

Zhang et al. [189] describe an unsupervised distance-based technique to identify global outliers in query-processing applications of sensor networks. The proposed technique reduces the communication overhead of sensor nodes using an aggregation tree. Each node in the tree transmits data to its parent after collecting all data sent from its children. The sink (also called the *base station*) uses the information from its children to identify the most likely outliers and "floods" these outliers for verification. If any node finds that it has two kinds of data which may modify the global result, it will send them to its parent in an appropriate time interval. This procedure is repeated until all nodes in the network agree on the results produced by the sink node. This technique considers only univariate data and the aggregation tree used may not be stable due to the dynamic changes of the network topology [12].

Clifton et al. [175,176] apply the *k*-means clustering algorithm to condition monitoring of aerospace gas-turbine engines. Novelty scores $z(\mathbf{x})$ are defined to be the number of standard deviations that a test point $\mathbf{x}$ lies from its closest cluster centre, relative to the distribution of all clusters.

Yong et al. [187,186] consider novelty detection in multiple-scene image sets. The framework starts with wildlife video frame image segmentation, followed by feature extraction and classification using the *k*-NN method. The labelled image blocks are then used to generate a co-occurrence matrix of object labels (called the *block label co-occurrence matrix*, BLCM), which represents the semantic context within the scene. Principal component analysis is used to perform dimensionality reduction, resulting in models for scene categories. The classification model is used to classify an image into scene classes; if it does not belong to any scene class, the image is classified as being "abnormal". Yong et al. [186] assume that in the BLCM feature space, for each scene type, there is a dense cluster related to normal images, while novel images are sparsely distributed around these clusters. During training, the centroid of the BLCM for each image group is calculated. The distances to the centroid for all points in the same scene are computed, and a threshold based on the mean and standard deviation of the distances is determined. Test images are then assessed by calculating their BLCM feature distance to the trained one-class centres: if it is smaller than a defined threshold for that class, the images are accepted by that one-class classifier, otherwise they are rejected from that class. If the test image is rejected by all classes, the image is classified as being novel. This multiple one-class classification with a distance thresholding algorithm was compared with a pdf-based one-class classifier [126] and a one-class support vector machine [209]. The proposed algorithm was shown to perform the most successfully at the task of detecting novel wildlife scenes.

Zhou et al. [210] present a method for distributed novelty detection on simulation mesh data. Large-scale simulation datasets are typically located on multiple computers and cannot be merged due to communication overhead and computational inefficiency. The proposed method consists of three steps. In the first step, local models from all distributed data sources are built using clustering-based methods. Novelty scores for test points are based on the distance to the nearest cluster centre. In the second step, all local outliers are collected from distributed sites and shared with each site, and all local models are rebuilt. Finally, in the third step, the local outliers' novelty scores are computed using the ensemble of the local models' results from the previous step. The ensemble methods consider both quality criteria of local models acting on local points and diversity criteria[4] of local models acting on all local outliers to detect novelty in a global view.

There are other variants of the methods described above. Spinosa et al. [211] describe an *online novelty and*

[4] Mutual information is used as a measure to indicate diversity.

drift detection algorithm that uses standard clustering methods to generate candidate clusters among examples that are not explained by the currently known concepts. Hassan et al. [212] propose a heuristic method based on a clustering approach in wireless sensor networks. Idé et al. [213] address the task of change-point analysis in correlated multi-sensor systems. Their approach is based on a neighbourhood preservation principle: if the system is working normally, the neighbourhood graph for each sensor is almost invariant with respect to fluctuations arising from experimental conditions. With this notion of a stochastic neighbourhood, the proposed method was able to compute novelty scores for each sensor. Onuma et al. [214] use clustering-based novelty detection for recommender systems. The authors apply a graph-based method to recommend items that are not in the user's current set of interests, but which lie in neighbouring areas of interest. This ensures novelty, and provides variety in the recommendations, which is seen favourably by users.

3.3. Method evaluation

Distance-based approaches do not require *a priori* knowledge of the data distribution and share some common assumptions with probabilistic approaches. Nearest neighbour-based techniques, however, rely on the existence of suitable distance metrics to establish the similarity between two data points, even in high-dimensional data spaces. Furthermore, most of them only identify novel data points globally and are not flexible enough to detect local novelty in data sets that have diverse densities and arbitrary shapes. Generally, in high-dimensional data sets it is computationally expensive to calculate the distance between data points and as a result these techniques lack scalability. Clustering-based approaches are capable of being used in incremental models, i.e., new data points can be fed into the system and tested to identify novelty. New techniques have been developed to optimise the novelty detection process and reduce the time complexity with respect to the size of data. However, these techniques suffer from having to choose an appropriate value of cluster width and are also susceptible to the curse of dimensionality.

Probabilistic and distance-based approaches rely on similar assumptions. They attempt to characterise the area of the data space occupied by normal data, with test data being assigned a novelty score based on some sort of distance metric. Nearest-neighbour and clustering-based techniques require a distance measure computation between a pair of data points. These techniques, when applied to novelty detection, assume that the distance measure can discriminate between novel and normal data points. Probabilistic techniques typically fit a probability model to the given data and determine whether or not a test data point comes from the same model by assuming that normal data points occur in the so called "high density regions" of the model. One of the main differences between these two types of approach is the computational complexity and scalability of the proposed techniques.

4. Reconstruction-based novelty detection

Reconstruction-based methods are often used in safety-critical applications for regression or classification purposes. They can autonomously model the underlying data, and when test data are presented to the system, the reconstruction error, defined to be the distance between the test vector and the output of the system, can be related to the novelty score. Neural networks and subspace-based methods can be trained in this way (Table 3).

4.1. Neural network-based approaches

Several types of neural networks have been proposed for novelty detection, a review of which can be found in [27]. In this section, we will concentrate on more recent methods that use neural networks.

Augusteijn and Folkert [215] investigate the ability of the back-propagation neural network architecture (a multi-layer perceptron, or MLP) to detect novel points. One novelty detection approach uses a threshold on the highest output value and declares a point to be novel if this value remains below the threshold; a second approach calculates the distance between the output and all target points and classifies the test point as novel if the minimum distance is found to exceed a predefined

Table 3
Examples of novelty detection methods using reconstruction-based approaches.

Reconstruction-based approach	Section	References
Neural networks	4.1	
Multi-layer perceptron		Augusteijn and Folkert [215] and Singh and Markou [216]
Hopfield networks		Crook et al. [217]
Autoassociative networks		Diaz and Hollmen [218], Hawkins et al. [219], Japkowicz [220], Manevitz and Yousef [221], Thompson et al. [222], Williams et al. [223]
Radial basis function		Bishop [21], Jakubek and Strasser [224], Li et al. [225] and Nairac et al. [110]
Self-organising networks		Albertini and de Mello [226], Barreto and Aguayo [227], Deng and Kasabov [228], García-Rodríguez et al. [229], Hristozov et al. [230], Kit et al. [231], Marsland et al. [232,233], Ramadas et al. [234] and Wu et al. [235]
Subspace methods	4.2	Chen and Malin [236,237], Günter et al. [238], Hoffmann [239], Lakhina et al. [240], Kassab and Alevandre [241], McBain and Timusk [242], Perera et al. [243], Ide and Kashima [244], Shyu et al. [245], Thottan and Ji [246], Toivola et al. [247] and Xiao et al. [248]

threshold. Both approaches were found to lead to a poor ability to identify novel data. The authors have also explored the applicability of the probabilistic neural network, which contains as many nodes as there are points in the training set, where the connections to these nodes are weighted with the feature values of the training data. Points belonging to the same category may first be clustered, and the cluster centres may then be used as initial connection weights. When presented with test data, each output unit, which can incorporate a prior probability and a cost of misclassification associated with the category, calculates a quasi-probability of the data belonging to that category. If the highest output value lies below a predefined threshold then the pattern can be assumed to belong to a class not represented by the network. This method showed superior performance as an overall classifier when compared to the MLP and was able to identify novel patterns.

Hawkins et al. [219] and Williams et al. [223] present an outlier detection approach for large multivariate datasets based on the construction of a *Replicator Neural Network* (RNN). The RNN is an MLP which has the same number of input and output neurons (that correspond to the features in the dataset), and three hidden layers. The aim of the RNN is to reproduce the input points at the output layer with the minimum reconstruction error, after compression through the hidden layers (which contain a smaller number of nodes than the input and output layers). If a small number of input points are not reconstructed well (they have large reconstruction errors), these points can be considered as outliers. An outlier factor based on the average reconstruction error is used as a novelty score. These techniques have been used in several investigations [218–223]. The auto-associative network described in [222], also termed an *autoenconder*, computes the bitwise difference between input and output to highlight novel components of the input. Diaz and Hollmen [218] also use an auto-associative neural network, where the residual mean-square error is used to quantify novelty. The method was used to detect outliers in vibration data for fraud detection in synchronous mechanical units.

Haggett et al. [249] present a dynamic predictive coding mechanism using a neural network model of circuits in the retina proposed in [250]. This latter model is a feed-forward network in which the connections may be modifiable synapses (weights). These modifiable weights are modulated according to an anti-Hebbian learning rule which causes the modifiable synapses to weaken when the activity at the pre-synaptic and post-synaptic neurons are correlated, and to strengthen when the activity is anti-correlated. Hosoya et al. [250] demonstrate the operation of this network with a number of artificially generated visual environments. This network is used as the basis for the novelty detector proposed in [249], which uses dynamic predictive coding. Three evolutionary algorithms, including a genetic algorithm and the *Neuro-evolution of Augmenting Topologies* (NEAT), are used to optimise the structure of the network to improve its performance using stimuli from a number of artificially generated visual environments. The authors have demonstrated that the optimised network evolved by NEAT outperforms other evolutionary algorithms and genetic algorithm approaches.

A novelty detection method applied to region-segmented outdoor scenes in video sequences is proposed in [216,9]. Their approach uses a feature-selection mechanism to encode image regions, and an MLP acting as a classifier. The MLP is used to reject any input not similar to the training data. A rejection filter is used to classify test data as either *known* or *unknown*. The known data points are classified by the neural network into one of the known classes on the basis of a "winner takes all" strategy. The rejected data points are collected in a "bin" for further processing. Post-processing of this filtered output has the goal of identifying clusters. Clusters that represent novel data are manually labelled and a new network is trained. This method reduces multi-class classification into a number of binary classifications; i.e., a classification problem with 10 classes is decomposed into a set of 10 binary problems. One neural network per class is trained, where data are labelled as either belonging or not belonging to the class. In this approach, the innovation is that *random rejects* (data that were labelled as not belonging to the class) are artificially generated. Although the performance of this framework was demonstrated using video data, the number of random rejects to be generated requires further investigation.

Generalised radial basis functions (RBF) neural networks have also been proposed in several different applications for novelty detection [21,110,225]. In this case, reverse connections from the output layer to the central layer are added, similar to a self-organising Bayesian classifier, which is capable of novelty detection. Each neuron in the central layer has an associated Normal distribution which is learned from the training data. Novel points have low likelihood with respect to these distributions and hence result in low values at each output node. In [110], a kernel is used to represent the distribution at each neuron, such that the distance of a test point from the nearest kernel centre can be determined and used to detect novelty. Jakubek and Strasser [224] propose a technique which uses neural networks with ellipsoid basis functions for fault detection. The advantage of these functions is that they can be fitted to the data with more accuracy than radially symmetric RBFs. The distribution of each cluster is represented by a kernel function. Results from experiments showed that the proposed network uses fewer basis functions than a RBF network of equivalent accuracy.

Crook et al. [217] applied *Hopfield Networks* to detect novelty in a mobile robot's environment. A Hopfield network uses binary threshold nodes with recurrent connections between them. A global energy function for the network with symmetric connections is defined and used to determine if a test point presented to the network is novel or not. The value of the energy function is lower for "normal" points and higher for novel points. It was demonstrated that the method can be used to learn a model of an environment during exploration by a robot and then detect novel features in subsequently encountered environments.

Kohonen maps, also called *Self-Organising Maps* (SOMs), can be used for novelty detection [251]. The SOM is a neural network with a grid-like architecture that is primarily used

as an unsupervised technique for identifying clusters in a dataset and which, in effect, moves the position of the nodes in the feature space to represent these identified clusters. When normal data are used to train a SOM, it creates a kernel-based representation of normality which can be used for novelty detection. The Euclidean distance between a test point and nodes in the SOM is evaluated, and used to determine novelty. SOMs have been used to detect network intrusions [252,234]. One important characteristic of this type of neural network is that they are topology-preserving; i.e., the network preserves neighbourhood relationships between the data by mapping neighbouring inputs onto neighbouring nodes in the map. The "Kohonen SOM" is a static SOM with a fixed structure: the grid size and the number of nodes have to be determined *a priori*. This results in a significant limitation on the final mapping as it is unlikely that the most appropriate structure is known beforehand. Several SOM variations, known as dynamic SOMs, and other neural network-based approaches, known as growing networks, have been introduced in the past to overcome these shortcomings. The latter include the "growing cell structures" model, the "incremental grid-growing" model, the "growing neural gas", the "growing SOM" and the "evolving SOM" [253–256,228,232].

Marsland et al. [232] propose a self-organising network that *Grows When Required* (GWR). In this method, each node is associated with a subset of the input space, and the network is initialised with a small number of nodes randomly located in the input space, and which are unconnected. At each training iteration, a new input is presented to the network, and existing nodes are either (i) moved or removed to better represent the distribution of the training data (the adaptation process), or (ii) new nodes and connections are added to the network (the growing process). A new node is added when the "activity" of the best matching node (the node that best matches the input) is not sufficiently high. The activity of nodes is calculated using the Euclidean distance between the weights for the node and the input. Novelty detection can be performed by describing how often a node has fired before (i.e., how often it has been the "winning" node when presented with input data). This network was applied to different tasks, including robot sonar scans, medical diagnosis, and machine fault detection. This method was also successfully applied in real-time automated visual inspection using mobile robots [10], in which colour statistics are used to encode visual features. In [233], the GWR network is combined with *habituation networks*. This method is based on learning to ignore previously encountered "normal" data, so that novel inputs are given more weight in the analysis, and become easier to detect. This is achieved using the GWR network with the addition of synapses connecting each node in the network to an output node. The algorithm was demonstrated with the task of mobile robot inspection of corridor environments, using inputs from sonar sensors and images from a camera mounted on the robot.

A different approach applied to data from a camera carried by a robot is taken by Kit et al. [231], who use growing neural gas [256] to detect changes. The model uses a growing neural gas network constructed using image data and any available spatial data. Growing neural gas is an unsupervised incremental clustering algorithm. Given some input distribution in the feature space, the method creates a network of nodes, where each node has a position in the feature space. It is an adaptive algorithm in the sense that if the input distribution slowly changes over time, the network is able to adapt by moving the nodes to cover the new distribution. The closest node to a test point is found, an error distance between the two can be determined, and the error compared to a threshold to determine novelty.

García-Rodríguez et al. [229] address the ability of self-organising neural network models to manage real-time applications, using a modified learning algorithm for a growing neural gas network. The modification proposed aims to satisfy real-time temporal constraints in the adaptation of the network. The proposed learning algorithm can add multiple neurons per iteration, the number of which is controlled dynamically. The authors concluded that the use of a large number of neurons made it difficult to obtain a representation of the distribution of training data with good accuracy in real-time.

Albertini and de Mello [226] propose a network that integrates features from SOM, GWR, and *adaptive resonance theory* networks. When an input is presented, the network searches through categories stored for a match. If no match is found, then the input is considered to be novel. A ligand-based virtual screening method based on using SOMs for novelty detection is described in [230]. Wu et al. [235] present a case study in which an online fault learning method based on SOM techniques was adopted for use with mechanical maintenance systems.

Barreto and Aguayo [227] evaluate the performance of different static and temporal SOM-based algorithms for identifying anomalous patterns in time series. The methodology consists of computing decision thresholds from the distribution of quantisation errors produced by normal training data, which are then used for classifying incoming data samples. Results from the experiments conducted show that temporal variants of the SOM are more suitable to deal with time-series data than static competitive neural networks.

4.2. Subspace-based approaches

A different type of reconstruction-based novelty detection (termed *spectral* methods in [24]) uses a combination of attributes to best describe the variability in the training data. These methods assume that data can be projected or embedded into a lower dimensional subspace in which "normal" data can be better distinguished from "abnormal" data. Principal Components Analysis (PCA) is a technique for performing an orthogonal basis transformation of the data into a lower-dimensional subspace. The number of features needed for effective data representation can thus be reduced. This technique can be used for novelty detection by constructing a model of the distribution of training data in the transformed space [257, ch. 10]. The first few principal components of a dataset correspond to vectors in the data space that account for most of the

variance of the data. The last few principal components can be used to find features that are not apparent with respect to the original variables. Dutta et al. [13] describe an outlier detection algorithm that uses approximate principal components. The last principal component enables identification of points which deviate significantly from the "correlation structure" of the data. Thus, a normal point that exhibits similar correlation structure to that of the training data will have a low value for such projections, and an outlier that deviates from the correlation structure will have a large value. This approach was applied to novelty detection in astronomy catalogues.

Shyu et al. [245] propose a novelty detection approach based on PCA, which can be seen as a robust estimator of the correlation matrix of normal data. Two functions of principal components to identify outliers are sequentially executed. The first function uses the major principal components to detect extreme points with large variances and covariances depending on the subset of original attributes. The second function, as in [13], uses the minor principal components to further identify the remainder of the outliers, which have different correlation structures from normal data. Experiments with network intrusion data indicated that the proposed scheme performed better than techniques based on clustering approaches, and that it can work in an unsupervised manner. Other authors have applied this PCA-based technique to network intrusion detection [240,246] and to the detection of anomalies in spacecraft components [258].

Hoffmann [239] explores *kernel PCA* in the context of the novelty detection task. Kernel PCA [259] extends the standard PCA to non-linear data distributions by mapping points into a higher-dimensional feature space before performing PCA. In this feature space, using a kernel, the originally linear operations of PCA are performed with a non-linear mapping (the "kernel trick"). This method has been applied to astronomical data for the prediction of stellar populations in space [14]. In [239], a principal-component subspace in an infinite-dimensional feature space describes the distribution of training data, and the reconstruction error of a test point with respect to this subspace is used as a measure of novelty. A strategy to improve the convergence of the kernel algorithm for iterative kernel PCA is described in [238]. Novelty detection with kernel PCA achieved better classification of novelty when applied to hand-written-digit and breast cancer databases, compared with a one-class support vector machine and a Parzen window density estimator. Kernel PCA is not robust to outliers in the "normal" training set due to the properties of the L2 norm used in the optimisation part of the training procedure [248]. Kwak [260] proposes a PCA method based on the L1 norm. Xiao et al. [248] extend this work to L1 norm-based kernel PCA. The proposed method was applied to novelty detection, and benefited from the robustness of the L1 norm to outliers.

Perera et al. [243] have developed a novelty detection method based on a recursive dynamic PCA approach for gas sensor array applications. Their method is based on a sliding window-based variance analysis algorithm. Under normal conditions, a certain variance distribution characterises sensor signals; however, in the presence of a new source of variance, the associated PCA decomposition changes. Sensor drift and other effects may be taken into account because the model is adaptive, and is updated in a recursive manner with minimal computational effort. The technique was applied to signals from oil vapour leakages in air compressors.

Further examples of reconstruction-based novelty detection with graph-based data have been proposed in recent years [261–263,244]. Ide and Kashima [244] approach the problem of Web-based systems as a weighted graph, using eigenvectors of adjacency matrices (that represent the activities of all of the services) from a time-series of graphs to detect novelty. At each time point, the principal component of the matrix is chosen as the *activity vector* for the given graph (which is represented as an adjacency matrix for a given time). The "normal" time-dependencies of the data are captured using the principal left-singular vector of the matrix which contains these time-series activity vectors as column vectors. For a new test graph, the angle between its activity vector (the principal component vector for the test adjacency matrix) and the principal-left singular vector obtained from previous graphs is computed and used to calculate a novelty score for the test graph. If the angle changes by more than some threshold, an abnormality is declared to be present. In a similar approach, Sun et al. [263] perform *Compact Matrix Decomposition* on the adjacency matrix of each graph on a sequence of graphs. An approximate version of the original matrix is constructed from the decompositions, and a time-series of approximation (or residual) errors between the original matrix and the approximation matrix is then determined and used to detect abnormalities.

Kassab and Alexandre [241] introduce an *Incremental data-driven Learning of Novelty Detector Filter* (ILoNDF) for one-class classification with application to high-dimensional noisy data. The novelty detection filter is implemented with a recurrent network of neuron-like adaptive elements. It consists of n fully connected neurons, where n is the dimensionality of the input vectors, such that all neurons are both input and output neurons. The weights associated with the feedback connections provide the variable internal state of the network, and are updated after presentation of an input vector using an anti-Hebbian learning rule. The network continuously integrates information relating to the distribution of the training data and their co-occurrence dependencies. Because it operates online without repeated training, the proposed method does not require extensive computational resources. Experiments conducted involving text categorisation tasks showed that ILoNDF tends to be more robust, is less affected by initial settings, and outperforms methods such as auto-associative neural networks and PCA-based models.

Chatzigiannakis et al. [264] present an approach that fuses data gathered from different nodes in a distributed wireless sensor network. An offline analysis step creates a model of normality; testing is then performed in real-time. PCA is used to identify test data that are considered "normal", where anomalies tend to result in large variations in the PCA residual. This procedure can be

computationally intense, and so a sliding window is used, with the principal components being re-estimated only when the deviation in one or more correlation coefficients (of all the monitored metrics) exceeds a threshold. The approach was demonstrated using meteorological data collected from a distributed set of sensor nodes.

Chen et al. [236,237] introduce the *community anomaly detection system* (CADS), an unsupervised algorithm for detecting computer access threats based on the access logs of collaborative environments. Collaborative system users tend to form community structures based on the subjects accessed (e.g., patients' records are typically viewed by healthcare providers). CADS is a "hybrid" method that consists of two schemes: it uses singular value decomposition to infer communities from relational networks of users, and k-NN to establish sets of nearest neighbours. The latter is used as a model to determine if users have deviated from the behaviour of existing communities. In [237], CADS is extended to MetaCADS to account for the semantics of subjects (e.g., patient diagnoses). The framework was empirically evaluated using three months of access logs from the electronic health record system of a large medical centre. When the number of "illicit" users is small, MetaCADS was shown to be the best-performing model of those considered, but as the number of illicit users grows, the original CADS algorithm was most effective.

Lämsä and Raiko [265] demonstrate how *non-linear factor analysis* (NFA), as a neural network-based method, can be used for separating structural changes from environmental and operational variation, and thereafter also for damage detection. Here, the relationships between the observations are described in terms of a few underlying but unobservable factors. The goal is to eliminate the adverse effects of these underlying factors from the observations resulting in new variables that can be used in damage detection. NFA is based on the assumption that the underlying factors are independent and normally distributed, which is typically not the case. A non-linear mapping from hidden factors to observations is modelled by a two-layer MLP. This method was applied to vibration data, and it was shown that damage detection via elimination of environmental and operational effects from damage features is feasible.

McBain and Timusk [242] propose a feature reduction technique for novelty detection. The method is similar to multiple discriminant analysis in that it attempts to find a subspace that maximises the difference between the average distance of the "normal" class and the average distance of the "abnormal" class. The effect of the reduced subspace on classification was shown to be better than that obtained from other dimensionality-reduction methods (such as PCA and kernel PCA), for machinery monitoring data.

Timusk et al. [266] describe a strategy for vibration-based online detection of faults in machinery. A selection of seven different types of novelty detection algorithms were implemented and compared, including a similar PCA-based approach to those described previously, probabilistic methods (such as a single Gaussian distribution density estimator), and clustering methods (such as k-NN and k-means). Results showed that the PCA-based approach was the best-performing fault detector.

In [247], three dimensionality-reduction approaches are assessed for novelty detection: the non-linear Curvilinear Component Analysis (CCA), classical PCA, and a computationally inexpensive *Random Projections* algorithm. As with Kohonen's SOM, the CCA aims to reproduce the topology of the original data in a projection subspace, but without fixing the configuration of the topology. It is an adaptive algorithm for non-linear dimensionality reduction, which minimises a cost function based on inter-point distances in both input and output space. Since the topology cannot be entirely reproduced in the projection subspace, which has a lower dimension than the original subspace, the local topology is favoured to the detriment of the global topology. Random Projections is a computationally inexpensive method of linear dimensionality reduction. It embeds a set of points from the original space in a randomly selected subspace whose dimensionality is logarithmic with respect to the dimension of the original space, such that pairwise distances between points before and after projection change only by a small factor. Four classification approaches were evaluated: three based on probabilistic models, such as the GMM, and one based on a nearest-neighbour method. Experiments showed that CCA was the best projection method for the novelty detectors assessed in this work. The authors mention that this agrees with expected results, since the CCA method should be more powerful due to its non-linear nature. Nevertheless, PCA was able to compete for datasets of lower dimensionality, when using a nearest-neighbour classifier.

4.3. Method evaluation

Reconstruction-based approaches belong to a very flexible class of methods that are trained to model the underlying data distribution without *a priori* assumptions on the properties of the data. Neural networks require the optimisation of a pre-defined number of parameters that define the structure of the model, and their performance may be very sensitive to these model parameters. They may therefore be very difficult to train in high-dimensional spaces. Moreover, networks that use constructive algorithms, in which the structure of the model is allowed to grow, suffer from the additional problem of having to select the most effective training method to enable the integration of new units into the existing model structure, and an appropriate stopping criterion (for when to stop adding new units). With subspace-based approaches, appropriate values must be selected for the parameters which control the mapping to a lower-dimensional space. It is difficult to determine which are the key attributes and it is computationally expensive to estimate the correlation matrix of normal patterns accurately.

5. Domain-based novelty detection

Domain-based methods require a boundary to be created based on the structure of the training dataset. These methods are typically insensitive to the specific sampling and density

of the target class, because they describe the target class boundary, or the *domain*, and not the class density. Class membership of unknown data is then determined by their location with respect to the boundary. As with two-class SVMs, novelty detection SVMs (most commonly termed "one-class SVMs" in the literature) determine the location of the novelty boundary using only those data that lie closest to it (in the transformed space); i.e., the support vectors. All other data from the training set (those that are not support vectors) are not considered when setting the novelty boundary. Hence, the distribution of data in the training set is not considered which is seen as "not solving a more general problem than is necessary" [209,267].

SVMs are a popular technique for forming decision boundaries that separate data into different classes. The original SVM is a network that is ideally suited for binary pattern classification of data that are linearly separable. The SVM uses a hyperplane that maximises the separating margin between two classes. The training points that lie near the boundary defining this separating margin are called *support vectors*. Since the introduction of the original idea, several modifications and improvements have been made. The *Robust Support Vector Machines* (RSVMs) algorithm [268,269] addresses the over-fitting problem introduced by noise in the training dataset. In this approach, an averaging technique (in the form of class centre) is incorporated into the standard SVM, which makes the decision surface smoother, controlling the amount of regularisation automatically. The number of support vectors of RSVMs is often fewer than that of standard SVMs, which leads to a faster execution speed. Experiments using system intrusion detection data showed that RSVMs can provide a reasonable ability to detect intrusions in the presence of noise. Nevertheless, it does not address the fundamental issue of the unbalanced nature between "normal" and "abnormal" training examples as typically exists for novelty detection applications.

Li et al. [270] propose an algorithm for online novelty detection with kernels in a *Reproducing Kernel Hilbert Space*. The technique differs from the original SVM formulation, in that training data are processed sequentially, and the Lagrangian dual equation is solved by maximising a quadratic equation in a given interval. The proposed approach has a simple update procedure with a much lower computational cost than with the equivalent procedure for conventional SVMs.

Diehl et al. [8] have implemented a real-time novelty detection mechanism for video surveillance. Their method is based on the extraction of monochromatic spatial features in image sequences to represent moving objects. A classifier based on a generalised SVM was trained offline with models of people and cars, and was later used to reject previously unseen moving objects, such as bicycles, vans, and trucks. The sequences of training images are processed to derive class labels for training the sequence classifier. Each training image sequence has a corresponding class label distribution which is simply the relative frequencies of each class label in the class label sequence. Using these class label distributions, a logistic linear classifier is constructed to partition the class label distribution space. Image sequences assigned to the same class are ranked based on class likelihood. This likelihood monotonically increases with increasing novelty in the sequence, which allows the user to focus on those examples that cause the greatest uncertainty of classification for the classifier.

SVMs have been used for novelty detection in two related approaches [271,209,267]. The idea of the *one-class SVM* approach proposed by Schölkopf et al. [209] is to define the novelty boundary in the feature space corresponding to a kernel, by separating the transformed training data from the origin in the feature space, with maximum margin. This approach requires fixing *a priori* the percentage of positive data allowed to fall outside the description of the "normal" class. This makes the one-class SVM more tolerant to outliers in the "normal" training data. However, setting this parameter strongly influences the performance of this approach (as discussed in [271]). Another approach, the *support vector data description* (SVDD) method, proposed by Tax and Duin [267], defines the novelty boundary as being the hypersphere with minimum volume that encloses all (or most) of the "normal" training data. The SVDD automatically optimises the model parameters by using artificially generated unlabelled data uniformly distributed in a hypersphere around the "normal" class. This causes the method to struggle with applications involving high-dimensional spaces. Novelty is assessed by determining if a test point lies within the hypersphere. In order to address the problem that the transformed data are not spherically distributed, Campbell and Bennett [272] use different kernels with linear programming optimisation methods, rather than the quadratic programming approaches typically used with SVMs. Other approaches to novelty detection have recently been proposed which are based on the use of either the SVDD or the one-class SVM (Table 4).

5.1. Support vector data description approaches

Some extensions to the SVDD approach have recently been proposed to improve the margins of the hyperspherically

Table 4
Examples of novelty detection methods using domain-based approaches.

Domain-based approach	Section	References
Support vector data description	5.1	Campbell and Bennett [272], Le et al. [273,274], Liu et al. [275,276], Peng and Xu [277], Tax and Duin [267], Wu and Ye [278] and Xiao et al. [279]
One-class support vector machine	5.2	Clifton et al. [280,281,3], Evangelista et al. [282], Gardner et al. [283], Hardoon et al. [284], Hayton et al. [285], Heller et al. [286], Lazarevic et al. [287], Lee and Cho [288], Ma and Perkins [289,290], Manevitz and Yousef [271], Rabaoui et al. [291], Schölkopf et al. [209] and Zhuang and Dai [292]

shaped novelty boundary. The first extension is proposed in [278], where the authors present a "small sphere and large margin" approach that surrounds the normal data with a hypersphere such that the margin from any outliers to the hypersphere is maximised. A further extension of the method is proposed by Le et al. [273], whose method aims to maximise (i) the margin between the surface of the hypersphere and abnormal data, and (ii) the margin between that surface and the normal data, while the volume of the hypersphere is minimised.

Xiao et al. [279] propose to use a number of hyperspheres to describe the normal data set. However, the method is heuristic and no demonstration that the multi-hypersphere approach can provide a better description of the data is provided. In [274], a more detailed multi-hypersphere approach to SVDD is described, in which a set of hyperspheres with different centres and radii is considered. The optimisation problem for this approach is solved by introducing slack variables and applying an iterative algorithm that consists of two alternative steps: one that calculates radii, centres of hyperspheres, and slack variables, and another that determines the hypersphere membership of each data point. Experimental results over 28 data sets showed that the multi-hypersphere SVDD performed better than the original SVDD in all cases. In [275], a *fast SVDD* method to improve the speed of the algorithm is proposed. The original SVDD centre is spanned by the images of support vectors in the feature space. Unlike traditional methods which try to compress a kernel expansion into one with fewer terms, the proposed fast SVDD directly finds the pre-image of a feature vector, and then uses a simple relationship between this feature vector and the SVDD centre to update the position of the centre. The decision function in this case contains only one kernel term, and thus the decision boundary of the fast SVDD is only spherical in the original space. Hence, the run-time complexity of the fast SVDD decision function is no longer linear in the support vectors, but is a constant, independent of the size of the training set. Results obtained from experiments using several real-world data sets (including a critical fabrication process for thin-film transistor liquid crystal display manufacturing [276]) are encouraging. Peng and Xu [277] also address the speed of the algorithm by proposing an *efficient SVDD*. The authors argue that in the fast SVDD, using a Gaussian kernel, the decision hypersphere being a sphere in the input space is not suitable for many real applications. The efficient SVDD first finds critical points using the kernel fuzzy *c*-means cluster technique [293] and then uses the images of these points to re-express the centre of the SVDD. The resulting decision function is linear in the number of clusters. Computational comparisons with one-class SVM, SVDD and fast SVDD in terms of prediction performance and learning time have shown that the proposed efficient SVDD achieves a faster test speed.

5.2. One-class support vector machine approaches

Gardner et al. [283] apply a one-class SVM to the detection of seizures in patients. The intracranial EEG time-series is mapped into corresponding sequences of novelty scores by classifying short-time, energy-based statistics computed from one-second windows of data. The model is trained using epochs of normal EEG. Epochs containing seizure activity exhibit changes in the distribution in feature space that increase the empirical outlier fraction, allowing seizure events to be detected.

Ma and Perkins [290] extend the one-class SVM approach for temporal sequences. The method unfolds time-series into a phase space using a time-delay embedding process, and projects all the vectors from a phase space to a subspace, thereby avoiding the bias created by extremely large or extremely small values. The one-class SVM is then applied to the projected data in the subspace. A different approach for online novelty detection in temporal sequences is presented by the same authors [289]. The latter algorithm is based on *support vector regression* (SVR), in which a linear regression function is constructed in the high-dimensional feature space of the kernel. Although it has good generalisation properties and can handle high-dimensional data efficiently, the SVR training algorithm requires retraining whenever a new sample is observed, which is not efficient for an online algorithm. An incremental SVR training algorithm is proposed to perform efficient updating of the model whenever a sample is added to or removed from the training set. To perform novelty detection, the authors define a matching function to determine how well a test sequence matches the normal model.

Roth [294,295] propose the *one-class kernel Fisher discriminant* classifier to overcome the "main conceptual shortcoming" of one-class SVM classifiers, which is that the expected fraction of outliers has to be specified in advance. The method relates kernelised one-class classification and Gaussian density estimation in the induced feature space. With respect to classification, the proposed model inherits the simple complexity control mechanism obtained by using regularisation techniques. The relation to Gaussian density estimation makes it possible to formalise the notion of novel objects by quantifying deviations from the Gaussian model. The parameters of the model are selected using a likelihood-based cross-validation procedure.

Hayton et al. [285] describe static and dynamic novelty detection methods for jet engine health monitoring. The one-class SVM for static novelty detection proposed in [209] is used to construct a model of normality to characterise the distribution of energy across the vibration spectra acquired from a three-shaft engine. A Kalman filter is then used as a linear dynamic model, with changes in test data identified using the normalised squared innovations of the Kalman filter. Both static and dynamic models were shown to perform well in detecting anomalous behaviour in jet engine vibration spectra.

Sotiris et al. [296] link SVM classifiers to Bayesian linear models to model the posterior class probability of test data. A PCA decomposition of the multivariate training data is performed, which defines a number of orthonormal subspaces, which can be used to estimate the joint class probability. A kernel density estimator is then computed for the projected data in two of the subspaces to estimate the likelihood of the "normal" class, from which the

"abnormal" class is estimated. An SVM classifier is constructed and a logistic distribution finally used to map SVM outputs onto posterior probabilities.

Clifton et al. [280,281] investigate the use of a one-class SVM using multivariate combustion data for the prediction of combustion instability. Wavelet analysis is used for feature extraction, from which detail coefficients are used as two-dimensional features. Novelty scores computed using the one-class SVM approach are obtained from each of the input time-series, and different classifier combination strategies are studied.

Heller et al. [286] consider an intrusion detection system which monitors accesses to the Microsoft Windows Registry using *Registry Anomaly Detection* (RAD). During normal computer activity, a certain set of registry keys are typically accessed by Windows programs. A one-class SVM is compared with a probabilistic algorithm which uses a Dirichlet-based hierarchical prior to smooth the distribution and account for the likelihoods of unobserved elements in sparse datasets by adjusting their probability mass based on the number of examples seen during training. The probabilistic algorithm was able to discriminate accurately between "normal" and "abnormal" examples. The authors suggest that a more effective selection of the SVM kernel should be used.

Lee and Cho [288] compare the performance of a one-class SVM with that of an auto-associative neural network, and results obtained from the analysis of six benchmark datasets show that the former performs consistently better than the latter. The one-class SVM has been used for novelty detection in: functional magnetic resonance imaging data [284]; audio recordings [291]; text data [292]; medical data to identify patient deterioration in vital signs [3]; and network intrusion detection [287].

Evangelista et al. [282] illustrate the impact of high dimensionality on kernel methods and, specifically, on the one-class SVM. It is shown that variance contributed by meaningless noisy variables confounds learning methods. The authors propose a framework to overcome this problem, which involves exploring subspaces of the data, training a separate model for each subspace, and then fusing the decision variables produced by the test data for each subspace using fuzzy logic aggregators. Experiments conducted using synthetic data sets showed that learning in the subspaces of high-dimensional data typically outperforms learning in the high-dimensional data space as a whole.

Munoz and Moguerza [297] apply the *one-class neighbour machine* to the problem of estimating high-density regions for the distribution of "normal" training data in the data space. This method is a block-based procedure that provides a binary decision function indicating whether or not a point is a member of a minimum volume set defined around the "normal" data. The algorithm replaces the task of estimating the density at each point by using a simpler measure that asymptotically preserves the order induced by the pdf. Numerical experiments showed that the proposed method performed consistently better than the one-class SVM in estimating minimum volume sets.

Li [298] performs novelty detection by constructing a closed decision surface around the "normal" training data through the derivation of surface normal vectors and identification of extreme data. Surface normal vectors are used to determine whether a point is extreme or not; i.e., a novel point is detected if it is located outside the region formed by the closed data surface. Experimental results demonstrated that the proposed method performs with high accuracy in detecting the novel class as well as identifying known classes. The performance of the proposed method, however, was not compared with that of other current approaches.

Sofman et al. [11] present an *anytime novelty detection* algorithm that deals with noisy and redundant high-dimensional feature spaces. A supervised dimensionality-reduction technique, multiple discriminant analysis, is used which causes the projected data to form clusters that are as compact as possible for within-class data, while being as far away as possible from cluster centres corresponding to other classes. The algorithm determines the influence of all previously seen novel data on a test point; if the accumulated influences exceed a novelty threshold, then the test point is identified as being novel, and is used for novelty prediction with subsequently observed test points. "Normal" data are not stored as they are assumed to have minimal impact on future novelty detection. This method was validated using data acquired by two mobile robots, and it was shown that the algorithm was able to identify all major unique objects (vegetation, barrels, fence, etc.) with a relatively small number of false positives.

5.3. Method evaluation

Domain-based approaches determine the location of the novelty boundary using only those data that lie closest to it and do not rely on the properties of the distribution of data in the training set. A drawback of these methods is the complexity associated with the computation of the kernel functions. Although some extensions have been proposed to overcome this problem, the choice of the appropriate kernel function may also be problematic. Additionally, it is not easy to select values for the parameters which control the size of the boundary region.

6. Information-theoretic novelty detection

Information theoretic methods compute the *information content* of a dataset using measures such as entropy, relative entropy, etc. These methods assume that novelty significantly alters the information content of the otherwise "normal" dataset. Typically, metrics are calculated using the whole dataset and then that subset of points whose elimination from the dataset induces the biggest difference in the metric is found. This subset is then assumed to consist of novel data.

He et al. [299,300] present a local-search heuristic approach, which involves entropy analysis, to identify outliers in categorical data. Entropy in information theory is a measure of the uncertainty associated with a random variable, and the authors use an entropy function to measure the degree of disorder of the remaining dataset after removal of high-entropy points. A point is considered to be an outlier if the entropy of the dataset decreases after its removal, compared with the entropy of the dataset

after removal of all previous outlier candidates. This procedure is repeated until k outliers are identified. Experimental results showed that this approach scales well with the size of the training sets.

Ando [301] considers the task of identifying clusters of "atypical" objects which strongly contrast from the bulk of the data in terms of their distribution, using an *information bottleneck* measure. The latter is derived from rate distortion theory and is used for unsupervised problems using two criteria associated with lossy data compression: *rate* and *distortion*. The former refers to the level of compression measured by the *mutual information*, while the latter refers to the disruption caused by compression. The algorithm was evaluated using a text classification task, and was shown to outperform a related method called *Bregman Bubble Clustering*, which is an extension of one-class clustering for multiple clusters.

Keogh et al. [302] propose parameter-free methods for data mining tasks (including clustering, anomaly detection and classification) based on compression theory. Subsequences of a given sequence of continuous observations are defined using a sliding window. Each sub-sequence is then compared with the entire sequence using a compression-based dissimilarity method, the *Kolmogorov complexity*, which is a measure of the computational resources needed to specify an object (i.e., it is a measure of randomness of strings based on their information content). This metric cannot be computed in the general case, and so the size of the compressed file that contains the string is used to approximate the measure. Empirical tests with time-series datasets showed that the approach is competitive with respect to others, such as the SVM, for anomaly detection tasks. Keogh et al. [303] propose a related technique (called HOT SAX) to solve the same problem for continuous time-series. Sub-sequences are extracted as before and then the Euclidean distance of each sub-sequence to its closest non-overlapping sub-sequences is calculated. This distance is then used as a novelty score for the sub-sequences. A similar approach is also applied to the domain of medical data in [304], in which the problem of finding the sequence that is least similar to all other sequences is addressed. The authors propose the use of the *Haar wavelet* transformation [305].

Gamon [306] investigates feature sets derived from a graph representation of sentences and sets of sentences using the information-theoretic metric of Kullback–Leibler divergence. The author shows that a highly connected graph produced using sentence-level term distances and pointwise mutual information can serve as a source to extract features for novelty detection.

Filippone and Sanguinetti [307] propose a new method to control the false-positive rate in novelty detection. Their method estimates the information content of the training set including and excluding the test point; the two resulting distributions are compared by computing the Kullback–Leibler divergence. The rationale for doing this is that this method explicitly takes into account the size of the training set in establishing a threshold for novelty. The method was shown to perform well in univariate and multivariate Gaussian cases, as well as in the mixture of Gaussians case, but it is not clear whether it can be extended to non-parametric methods. More recently, the same authors [308] have considered the online identification of event-based novelty in stationary linear autoregressive models with Gaussian noise. The authors use a perturbative approximation to the information-theoretic measure introduced previously in [307,309] for independent and identical distributed data. Again, they consider the Kullback–Leibler divergence between the estimates of the distributions of the stochastic term obtained before and after the test point arrives. This procedure yields a modified F-test, which more accurately incorporates the variability introduced by finite sample size effects.

Itti and Baldi [310] propose a Bayesian definition of surprise to capture subjective aspects of sensory information. "Surprise" measures how data affect an observer, in terms of differences between posterior and prior beliefs about the world, i.e., only data observations which substantially affect the observer's beliefs yield surprise. This difference or mismatch between expectations of the observer and the consequent perception of reality is calculated using the Kullback–Leibler divergence, which evaluates the statistics of visual attributes in specific scene types, or descriptions and layout of the scene.

6.1. Method evaluation

Information-theoretic approaches to novelty detection typically do not make any assumptions about the underlying distribution of the data. They require a measure that is sensitive enough to detect the effects of novel points in the dataset. The main drawback with this type of techniques is the selection of the information-theoretic measure. Typically, these measures can detect the presence of novel data points only if there is a significantly large number of novel data points present in the dataset. Thus, the performance of such techniques is very dependent on the choice of the information-theoretic measure. These techniques are also computationally expensive, although approximations have been proposed to deal with this problem. Finally, it may be difficult to associate a novelty score with a test point using an information theoretic-based method.

7. Application domains

Novelty detection has many practical real-life applications in different domains, and it is of crucial importance in applications that involve large datasets acquired from critical systems. These include the detection of faults in complex industrial systems, of structural damage, and of failure in electronic security systems. The main areas of application can be broadly categorised in six main domains, as shown in Table 5. Other domains of application not mentioned in the table include speech recognition, mobile robotics, astronomical data analysis and environmental monitoring. In the next sections we briefly discuss several applications of novelty detection, including the concept of novel data and the importance of novelty detection in each domain.

Table 5
Examples of novelty detection methods in the six main application domains covered in this review.

Application domain	Method (category)					References
	1	2	3	4	5	
Electronic IT security	✓	✓	✓	✓	✓	Helali [42], Heller et al. [286], Jyothsna et al. [7], Lakhina et al. [240], Peng et al. [151] and Yeung and Ding [83]
Healthcare informatics, medical diagnostics	✓	✓	✓	✓		Clifton et al. [3], Lin et al. [304], Quinn and Williams [2], Solberg and Lahit [45] and Tarassenko et al. [94,95]
Industrial monitoring and damage detection	✓	✓	✓	✓		Clifton et al. [175,176], Lämsä and Raiko [265], Surace and Worden [5] and Tarassenko et al. [4]
Image processing, video surveillance	✓	✓	✓	✓		Pokrajac et al. [169], Ramezani et al. [92], Singh and Markou [216,9] and Yong et al. [186,187]
Text mining	✓	✓	✓	✓	✓	Ando [301], Basu et al. [15], Ertöz et al. [208], Manevitz and Yousef [221], Zhang et al. [124] and Zhuang and Dai [292]
Sensor networks	✓	✓	✓			Chatzigiannakis et al. [264], Hassan et al. [212], Janakiram et al. [71], Phuong et al. [152], Subramaniam et al. [93] and Zhang et al. [12]

7.1. Electronic IT security

Novelty detection has generated much research in the field of electronic IT security systems, in which the goals include network intrusion detection and fraud detection. The former refers to the detection of malicious activity in a computer-related system. Frequent attacks on computer systems may result in systems being disabled, or even completely collapsing. The identification of such intrusions could help the discovery of malicious programs in the operating system and also detect unauthorised access to computer network systems. Intrusions may be associated with behaviours different from that of normal users.

Fraud detection involves the identification of criminal activities that occur in insurance claims, credit card purchases, mobile phone usage, and financial transactions, among others. The purchasing behaviour of people who steal credit cards (for example) is usually different from that of the owners of the cards. The identification of such changes in credit card use may prevent a subsequent period of fraud activity. The data used for constructing novelty detection models typically comprise several features, such as user ID, amount of money spent, time between consecutive card transactions, purchase records, and geographic location.

7.2. Healthcare informatics/medical diagnostics and monitoring

Healthcare informatics and medical diagnostics are an important domain of application of novelty detection approaches. Patient records with unusual symptoms or test results may indicate potential health problems for that patient. Ideally, the identification of unusual data should be able to discriminate between instrumentation or recording errors and clinically relevant changes in the condition of the patient, so that timely medical intervention may occur in the latter case. The data typically consist of records that have several types of features: patient age, weight, vital signs (such as heart rate), physiological signals (such as the electrocardiogram), blood test results, and medical image data.

7.3. Industrial monitoring and damage detection

Industrial assets deteriorate over time due to usage and normal wear. Such deterioration has to be identified early to prevent further escalation and losses, and for optimising machine performance while reducing maintenance and repair costs. High-value machinery is usually instrumented with sensors that are dedicated to monitoring their operation (including structural defects). The data collected by these sensors typically include temperature, pressure, and vibration amplitude.

7.4. Image processing/video surveillance

Novelty detection has been extensively applied to recognising novel objects in images and video streams. Specifically, extracting novel data from video streams is gaining attention because of the availability of large amounts of video data, and because of the lack of automated methods for extracting important details from such media. Detecting novel events in security or surveillance applications, in which there are particularly large streams of seemingly unimportant video data, is a very important task.

7.5. Text mining

The novelty detection problem applied to text data seeks an automatic means of detecting novel topics, new interesting events, or new stories in a given collection of documents or news articles. The data are typically very high-dimensional and very sparse, and comprise simple bag-of-word features and features derived from sophisticated linguistic representations.

7.6. Sensor networks

Sensor networks usually consist of a large number of small, low-cost sensor nodes distributed over a large area with one or more “sink” nodes gathering readings from sensor nodes. The sensor nodes are integrated with sensing, processing, and wireless communication capabilities.

Each node is equipped with a wireless radio transceiver, a small microcontroller, a power source, and multi-type sensors for recording temperature, pressure, sound and vibration. Novelty detection is applied to sensor networks to capture sensor faults and/or malicious attacks on the network.

8. Conclusion

The main goal of novelty detection is to construct classifiers when only one class is well-sampled and well-characterised by the training data. Novelty detection is an important learning paradigm and has drawn significant attention within the research community, as shown by the increasing number of publications in this field.

We have presented a review of the current state-of-the-art in novelty detection. We observe that a precise definition of novelty detection is difficult to achieve, nor is it possible to suggest what an "optimal" method of novelty detection would be. The variety of methods employed is a consequence of the wide variety of practical and theoretical considerations that arise from novelty detection in real-world datasets, such as the availability of training data, the type of data (including its dimension, continuity, and format), and application domain investigated. It is perhaps because of this great variety of considerations that there is no single universally applicable novelty detection algorithm.

There are several promising directions for further research in novelty detection, which are mainly associated with the open challenges that need to be tackled for the effective operation of the approach in growing domains of applicability. We observe that novelty detection algorithms can broadly be divided into five different categories, depending mainly on the assumptions made about the nature of the training data. Each category of methods discussed in this paper has its own strengths and weaknesses, and faces different challenges for complex datasets. Distance-based methods, which include nearest-neighbour and clustering-based approaches, require the definition of an appropriate distance measure for the given data, but are unable to deal with high-dimensional data efficiently, because distance measures in high dimensions are not able to differentiate between normal and abnormal data points. Also, such methods are typically heuristic and require manual selection of parameters, removing the possibility of using them for automatic construction of a model of normality (though approaches have been suggested for parameter selection using semi-automated techniques). Reconstruction-based methods are very flexible and typically address high-dimensionality problems, with no *a priori* assumptions about the properties of the data distribution. However, they require the optimisation of a pre-defined number of parameters that define the structure of the model, and may also be very sensitive to these model parameters. Furthermore, when constructive algorithms are used (in which the structure of the model is allowed to grow), two important problems arise: the selection of the most effective training method to enable the integration of new units into the existing structure, and a stopping criterion for when to stop adding new units.

Domain-based methods determine the location of the novelty boundary using only those data that lie closest to it, and do not make any assumption about data distribution. By focusing on the decision boundary, these methods are often influenced by outliers in the training set and they also depend on the choice of a suitable scaling for the features. In contrast, probabilistic methods make use of the distribution of the training data to determine the location of the novelty boundary. They are transparent methods, meaning that their outputs can be analysed using standard numerical techniques. However, the performance of such methods is limited when the size of the training set is very small. In general, the problem encountered when applying density methods to sparsely populated training sets, is that there is little control over the inherent variability introduced by the sparsity of the training data; i.e., the estimated quantiles can differ substantially from the true quantiles of the distribution. Information-theoretic methods require a measure that is sensitive enough to detect the effects of novel points in the dataset. Although they do not make any assumptions about the underlying distribution of the data, the performance of such methods is highly dependent on the choice of the information theoretic measure, and it may be difficult to associate a novelty score with a test point using an information theoretic-based method.

The computational complexity of these methods is also an important aspect. Generally, probabilistic, reconstruction-based, and domain-based methods have lengthy training phases, but with rapid testing. For many applications, this is not a problem as models can be trained offline, while testing is required to be in real time. On the other hand, distance-based and information theoretic-based methods, in general, are computationally expensive in the test phase, which may be an important limitation in real-world settings.

Acknowledgements

MAFP was supported by the RCUK Digital Economy Programme grant number EP/G036861/1 (Oxford Centre for Doctoral Training in Healthcare Innovation) and FCT – *Fundação* para a Ciência e Tecnologia under the grant SFRH/BD/79799/2011. DAC was supported by a Royal Academy of Engineering Research Fellowship, the Balliol Interdisciplinary Institute, the Maurice Lubbock Memorial Fund, and the Wellcome Trust and the EPSRC under grant number WT 088877/Z/09/Z. LC was supported by the NIHR Biomedical Research Centre Programme, Oxford.

References

[1] L. Tarassenko, P. Hayton, N. Cerneaz, M. Brady, Novelty detection for the identification of masses in mammograms, in: Proceedings of the 4th International Conference on Artificial Neural Networks, IET, 1995, pp. 442–447.

[2] J. Quinn, C. Williams, Known unknowns: novelty detection in condition monitoring, Pattern Recognit. Image Anal. 4477 (2007) 1–6.

[3] L. Clifton, D. Clifton, P. Watkinson, L. Tarassenko, Identification of patient deterioration in vital-sign data using one-class support vector machines, in: Proceedings of the Federated Conference on

Computer Science and Information Systems (FedCSIS), IEEE, 2011, pp. 125–131.
[4] L. Tarassenko, D. Clifton, P. Bannister, S. King, D. King, Novelty Detection, John Wiley & Sons, Ltd, 2009, pp. 1–22 (Chapter 35).
[5] C. Surace, K. Worden, Novelty detection in a changing environment: a negative selection approach, Mech. Syst. Signal Process. 24 (4) (2010) 1114–1128.
[6] A. Patcha, J. Park, An overview of anomaly detection techniques: existing solutions and latest technological trends, Comput. Netw. 51 (12) (2007) 3448–3470.
[7] V. Jyothsna, V.V.R. Prasad, K.M. Prasad, A review of anomaly based intrusion detection systems, Int. J. Comput. Appl. 28 (7) (2011) 26–35.
[8] C. Diehl, J. Hampshire, Real-time object classification and novelty detection for collaborative video surveillance, in: Proceedings of the International Joint Conference on Neural Networks, IJCNN'02, 2002, vol. 3, pp. 2620–2625.
[9] M. Markou, S. Singh, A neural network-based novelty detector for image sequence analysis, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1664–1677.
[10] H. Vieira Neto, U. Nehmzow, Real-time automated visual inspection using mobile robots, J. Intell. Robotic Syst. 49 (3) (2007) 293–307.
[11] B. Sofman, B. Neuman, A. Stentz, J. Bagnell, Anytime online novelty and change detection for mobile robots, J. Field Robot. 28 (4) (2011) 589–618.
[12] Y. Zhang, N. Meratnia, P. Havinga, Outlier detection techniques for wireless sensor networks: a survey, IEEE Commun. Surv. Tutor. 12 (2) (2010) 159–170.
[13] H. Dutta, C. Giannella, K. Borne, H. Kargupta, Distributed top-k outlier detection from astronomy catalogs using the DEMAC system, in: Proceedings of the 7th SIAM International Conference on Data Mining, IEEE, 2007.
[14] H. Escalante, A comparison of outlier detection algorithms for machine learning, in: Proceedings of the International Conference on Communications in Computing, Citeseer, 2005.
[15] S. Basu, M. Bilenko, R. Mooney, A probabilistic framework for semi-supervised clustering, in: Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2004, pp. 59–68.
[16] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.
[17] C. Sammut, G. Webb, Encyclopedia of Machine Learning. Springer, 2011. Springer reference.
[18] M. Moya, M. Koch, L. Hostetler, One-class classifier networks for target recognition applications, in: Proceedings of the World Congress on Neural Networks, International Neural Network Society, 1993, pp. 797–801.
[19] H. He, E. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.
[20] H.-J. Lee, S. Cho, The novelty detection approach for different degrees of class imbalance, in: I. King, J. Wang, L.-W. Chan, D. Wang (Eds.), Neural Information Processing, Lecture Notes in Computer Science, vol. 4233, Springer, Berlin/Heidelberg, 2006, pp. 21–30.
[21] C. Bishop, Novelty detection and neural network validation, in: Proceedings of the IEEE Conference on Vision, Image and Signal Processing, vol. 141, IET, 1994, pp. 217–222.
[22] G. Ritter, M. Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification, Pattern Recognit. Lett. 18 (6) (1997) 525–539.
[23] I. Merriam-Webster, Merriam-webster – an encyclopedia britannica company, May 2012. URL ⟨http://www.merriam-webster.com/dictionary/novel/⟩.
[24] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, ACM Comput. Surv. (CSUR) 41 (3) (2009) 1–58.
[25] V. Barnett, T. Lewis, Outliers in Statistical Data, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley and Sons, 1994.
[26] M. Markou, S. Singh, Novelty detection: a review – part 1: statistical approaches, Signal Process. 83 (12) (2003) 2481–2497.
[27] M. Markou, S. Singh, Novelty detection: a review – part 2: neural network based approaches, Signal Process. 83 (12) (2003) 2499–2521.
[28] S. Marsland, Novelty detection in learning systems, Neural Comput. Surv. 3 (2003) 157–195.
[29] V. Hodge, J. Austin, A survey of outlier detection methodologies, Artif. Intell. Rev. 22 (2) (2004) 85–126.
[30] M. Agyemang, K. Barker, R. Alhajj, A comprehensive survey of numeric and symbolic outlier mining techniques, Intell. Data Anal. 10 (6) (2006) 521–538.
[31] Z. Bakar, R. Mohemad, A. Ahmad, M. Deris, A comparative study for outlier detection techniques in data mining, in: Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems, IEEE, 2006, pp. 1–6.
[32] S. Khan, M. Madden, A survey of recent trends in one class classification, in: L. Coyle, J. Freyne (Eds.), Artificial Intelligence and Cognitive Science, Lecture Notes in Computer Science, vol. 6206, Springer, Berlin/Heidelberg, 2010, pp. 188–197.
[33] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, NY, USA, 2001.
[34] C. Bishop, Pattern Recognition and Machine Learning, vol. 4, Springer, New York, 2006.
[35] A. Modenesi, A. Braga, Analysis of time series novelty detection strategies for synthetic and real data, Neural Process. Lett. 30 (1) (2009) 1–17.
[36] V. Chandola, A. Banerjee, V. Kumar, Outlier Detection: A Survey, Technical Report 07-017, University of Minnesota, 2007.
[37] J. Kittler, W. Christmas, T. de Campos, D. Windridge, F. Yan, J. Illingworth, M. Osman, Domain anomaly detection in machine perception: a system architecture and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 99 (2013) 1.
[38] A.M. Bartkowiak, Anomaly, novelty, one-class classification: a comprehensive introduction, Int. J. Comput. Inf. Syst. Ind. Manage. Appl. 3 (2011) 61–71.
[39] Y. Gatsoulis, E. Kerr, J. Condell, N. Siddique, T. McGinnity, Novelty detection for cumulative learning, in: Proceedings of the Conference on Towards Autonomous Robotic Systems, 2010, pp. 62–67.
[40] E. Kerr, Y. Gatsoulis, N.H. Siddique, J.V. Condell, T.M. McGinnity, Brief overview of novelty detection methods for robotic cumulative learning, in: Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, 2010, pp. 171–180.
[41] D. Miljkovic, Review of novelty detection methods, in: Proceedings of the 33rd International Convention (MIPRO), IEEE, 2010, pp. 593–598.
[42] R. Helali, Data mining based network intrusion detection system: a survey, Novel Algoritm. Tech. Telecommun. Netw. (2010) 501–505.
[43] F.E. Grubbs, Procedures for detecting outlying observations in samples, Technometrics 11 (1) (1969) 1–21.
[44] C. Aggarwal, P. Yu, Outlier detection with uncertain data, in: Proceedings of the SIAM International Conference on Data Mining, 2008, pp. 483–493.
[45] H. Solberg, A. Lahti, Detection of outliers in reference distributions: performance of Horn′s algorithm, Clin. Chem. 51 (12) (2005) 2326–2332.
[46] C. Chow, On optimum recognition error and reject tradeoff, IEEE Trans. Inf. Theory 16 (1) (1970) 41–46.
[47] D.W. Scott, Frontmatter, John Wiley & Sons, Inc., 2008.
[48] D. Filev, F. Tseng, Real time novelty detection modeling for machine health prognostics, in: Proceedings of the Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), IEEE, 2006, pp. 529–534.
[49] D. Filev, F. Tseng, Novelty detection based machine health prognostics, in: International Symposium on Evolving Fuzzy Systems, 2006, pp. 193–199.
[50] A. Flexer, E. Pampalk, G. Widmer, Novelty detection based on spectral similarity of songs, in: Proceedings of 6th International Conference on Music Information Retrieval, 2005, pp. 260–263.
[51] J. Ilonen, P. Paalanen, J. Kamarainen, H. Kalviainen, Gaussian mixture pdf in one-class classification: computing and utilizing confidence values, in: Proceedings of the 18th International Conference on Pattern Recognition (ICPR), vol. 2, IEEE, 2006, pp. 577–580.
[52] J. Larsen, Distribution of the Density of a Gaussian Mixture, Technical Report, Informatics and Mathematical Modelling, DTU, 2003.
[53] P. Paalanen, J. Kamarainen, J. Ilonen, H. Kälviäinen, Feature representation and discrimination based on Gaussian mixture model probability densities – practices and algorithms, Pattern Recognit. 39 (7) (2006) 1346–1358.
[54] N. Pontoppidan, J. Larsen, Unsupervised condition change detection in large diesel engines, in: Proceedings of the IEEE 13th Workshop on Neural Networks for Signal Processing, NNSP′03, IEEE, 2003, pp. 565–574.
[55] X. Song, M. Wu, C. Jermaine, S. Ranka, Conditional anomaly detection, IEEE Trans. Knowl. Data Eng. 19 (5) (2007) 631–645.
[56] F. Zorriassatine, A. Al-Habaibeh, R. Parkin, M. Jackson, J. Coy, Novelty detection for practical pattern recognition in condition monitoring of multivariate processes: a case study, Int. J. Adv. Manuf. Technol. 25 (9) (2005) 954–963.
[57] D. Clifton, L. Clifton, P. Bannister, L. Tarassenko, Automated novelty detection in industrial systems, Adv. Comput. Intell. Ind. Syst. 116 (2008) 269–296.

[58] D. Clifton, S. Hugueny, L. Tarassenko, A comparison of approaches to multivariate extreme value theory for novelty detection, in: Proceedings of the IEEE/SP 15th Workshop on Statistical Signal Processing, IEEE, 2009, pp. 13–16.
[59] D. Clifton, S. Hugueny, L. Tarassenko, Novelty detection with multivariate extreme value statistics, J. Signal Process. Syst. 65 (3) (2011) 371–389.
[60] D. Clifton, S. Hugueny, L. Tarassenko, Pinning the tail on the distribution: a multivariate extension to the generalised Pareto distribution, in: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2011, pp. 1–6.
[61] D. Clifton, L. Clifton, S. Hugueny, D. Wong, L. Tarassenko, An extreme function theory for novelty detection, IEEE J. Sel. Top. Signal Process. 7 (1) (2013) 28–37.
[62] A. Hazan, J. Lacaille, K. Madani, Extreme value statistics for vibration spectra outlier detection, in: Proceedings of the 9th International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, 2012.
[63] S. Hugueny, D. Clifton, L. Tarassenko, Novelty detection with multivariate extreme value theory, part II: an analytical approach to unimodal estimation, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, IEEE, 2009, pp. 1–6.
[64] S. Roberts, Novelty detection using extreme value statistics, in: Proceedings of the IEEE Conference on Vision, Image and Signal Processing 146 (3) (1999) 124–129.
[65] S. Roberts, Extreme value statistics for novelty detection in biomedical data processing, in: Proceedings of the IEEE Conference on Science, Measurement and Technology, vol. 147, IET, 2000, pp. 363–367.
[66] H. Sohn, D.W. Allen, K. Worden, C.R. Farrar, Structural damage classification using extreme value statistics, J. Dyn. Syst. Meas. Control 127 (1) (2005) 125–132.
[67] S. Sundaram, D. Clifton, I. Strachan, L. Tarassenko, S. King, Aircraft engine health monitoring using density modelling and extreme value statistics, in: Proceedings of the 6th International Conference on Condition Monitoring and Machine Failure Prevention Technologies, 2009.
[68] R. Gwadera, M. Atallah, W. Szpankowski, Markov models for identification of significant episodes, in: Proceedings of 5th SIAM International Conference on Data Mining, 2005, pp. 404–414.
[69] R. Gwadera, M. Atallah, W. Szpankowski, Reliable detection of episodes in event sequences, Knowl. Inf. Syst. 7 (4) (2005) 415–437.
[70] A. Ihler, J. Hutchins, P. Smyth, Adaptive event detection with time-varying poisson processes, in: Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2006, pp. 207–216.
[71] D. Janakiram, V. Adi Mallikarjuna Reddy, A. Phani Kumar, Outlier detection in wireless sensor networks using Bayesian belief networks, in: Proceedings of the 1st International Conference on Communication System Software and Middleware (Comsware), IEEE, 2006, pp. 1–6.
[72] H.-J. Lee, S. Roberts, On-line novelty detection using the Kalman filter and extreme value theory, in: Proceedings of the 19th International Conference on Pattern Recognition (ICPR), 2008, pp. 1–4.
[73] P. McSharry, T. He, L. Smith, L. Tarassenko, Linear and non-linear methods for automatic seizure detection in scalp electroencephalogram recordings, Med. Biol. Eng. Comput. 40 (4) (2002) 447–461.
[74] P. McSharry, Detection of dynamical transitions in biomedical signals using nonlinear methods, in: Knowledge-Based Intelligent Information and Engineering Systems, Springer, 2004, pp. 483–490.
[75] S. Ntalampiras, I. Potamitis, N. Fakotakis, Probabilistic novelty detection for acoustic surveillance under real-world conditions, IEEE Trans. Multimed. 13 (4) (2011) 713–719.
[76] A. Pinto, A. Pronobis, L. Reis, Novelty detection using graphical models for semantic room classification, Prog. Artif. Intell. 7026 (2011) 326–339.
[77] Y. Qiao, X. Xin, Y. Bin, S. Ge, Anomaly intrusion detection method based on HMM, Electron. Lett. 38 (13) (2002) 663–664.
[78] J. Quinn, C. Williams, N. McIntosh, Factorial switching linear dynamical systems applied to physiological condition monitoring, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1537–1551.
[79] C. Siaterlis, B. Maglaris, Towards multisensor data fusion for dos detection, in: Proceedings of the ACM Symposium on Applied Computing, SAC '04, ACM, New York, NY, USA, 2004, pp. 439–446.
[80] C. Williams, J. Quinn, N. McIntosh, Factorial switching Kalman filters for condition monitoring in neonatal intensive care, Neural Inf. Process. (2006) 1513–1520.
[81] W. Wong, A. Moore, G. Cooper, M. Wagner, Rule-based anomaly pattern detection for detecting disease outbreaks, in: Proceedings of the National Conference on Artificial Intelligence, Menlo Park, CA; Cambridge, MA; London, AAAI Press; MIT Press; 1999, 2002, pp. 217–223.
[82] W. Wong, A. Moore, G. Cooper, M. Wagner, Bayesian network anomaly pattern detection for disease outbreaks, in: Proceedings of the 20th International Conference on Machine Learning, vol. 20, AAAI Press, 2003, pp. 808–815.
[83] D.-Y. Yeung, Y. Ding, Host-based intrusion detection using dynamic and static behavioral models, Pattern Recognit. 36 (1) (2003) 229–243.
[84] X. Zhang, P. Fan, Z. Zhu, A new anomaly detection method based on hierarchical HMM, in: Proceedings of the 4th International Conference on Parallel and Distributed Computing, Applications and Technologies, IEEE, 2003, pp. 249–252.
[85] P. Angelov, An approach for fuzzy rule-base adaptation using on-line clustering, Int. J. Approx. Reason. 35 (3) (2004) 275–289.
[86] Y. Bengio, M. Monperrus, Non-local manifold tangent learning, Adv. Neural Inf. Process. Syst. 17 (2005) 129–136.
[87] Y. Bengio, H. Larochelle, P. Vincent, Non-local manifold parzen windows, Adv. Neural Inf. Process. Syst. 18 (2006) 115–123.
[88] D. Erdogmus, R. Jenssen, Y. Rao, J. Principe, Multivariate density estimation with optimal marginal parzen density estimation and gaussianization, in: Proceedings of the 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, IEEE, 2004, pp. 73–82.
[89] A. Kapoor, K. Grauman, R. Urtasun, T. Darrell, Gaussian processes for object categorization, Int. J. Comput. Vis. 88 (2) (2010) 169–188.
[90] M. Kemmler, E. Rodner, J. Denzler, One-class classification with Gaussian processes, in: Asian Conference on Computer Vision (ACCV), vol. 6493, 2011, pp. 489–500.
[91] H. Kim, J. Lee, Pseudo-density estimation for clustering with Gaussian processes, Adv. Neural Netw. (ISNN) 3971 (2006) 1238–1243.
[92] R. Ramezani, P. Angelov, X. Zhou, A fast approach to novelty detection in video streams using recursive density estimation, in: Proceedings of the 4th International IEEE Conference Intelligent Systems, IS'08, IEEE, vol. 2, 2008, pp. 14–22.
[93] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, D. Gunopulos, Online outlier detection in sensor data using non-parametric models, in: Proceedings of the 32nd International Conference on Very Large Databases, VLDB Endowment, 2006, pp. 187–198.
[94] L. Tarassenko, A. Hann, A. Patterson, E. Braithwaite, K. Davidson, V. Barber, D. Young, Biosign™: multi-parameter monitoring for early warning of patient deterioration, in: Proceedings of the 3rd IEE International Seminar on Medical Applications of Signal Processing, IET, 2005, pp. 71–76.
[95] L. Tarassenko, A. Hann, D. Young, Integrated monitoring and analysis for early warning of patient deterioration, Br. J. Anaesth. 97 (1) (2006) 64–68.
[96] P. Vincent, Y. Bengio, Manifold parzen windows, Adv. Neural Inf. Process. Syst. 15 (2002) 825–832.
[97] D. Yeung, C. Chow, Parzen-window network intrusion detectors, in: Proceedings of the 16th International Conference on Pattern Recognition, vol. 4, IEEE, 2002, pp. 385–388.
[98] D. Dasgupta, N. Majumdar, Anomaly detection in multidimensional data using negative selection algorithm, in: Proceedings of the Congress on Evolutionary Computation (CEC), vol. 2, IEEE, 2002, pp. 1039–1044.
[99] F. Esponda, S. Forrest, P. Helman, A formal framework for positive and negative detection schemes, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 34 (1) (2004) 357–373.
[100] J. Gómez, F. González, D. Dasgupta, An immuno-fuzzy approach to anomaly detection, in: 12th IEEE International Conference on Fuzzy Systems (FUZZ '03), vol. 2, 2003, pp. 1219–1224.
[101] F. González, D. Dasgupta, Anomaly detection using real-valued negative selection, Genet. Program. Evolvable Mach. 4 (4) (2003) 383–403.
[102] D. Taylor, D. Corne, An investigation of the negative selection algorithm for fault detection in refrigeration systems, Artif. Immune Syst. 2787 (2003) 34–45.
[103] G.J. McLachlan, K.E. Basford, Mixture Models: Inference and Applications to Clustering, vol. 1, Marcel Dekker, New York, 1988.
[104] Y. Agusta, D. Dowe, Unsupervised learning of gamma mixture models using minimum message length, in: M.H. Hamza (Ed.),

Proceedings of the 3rd IASTED Conference on Artificial Intelligence and Applications, 2003, pp. 457–462.
[105] I. Mayrose, N. Friedman, T. Pupko, A gamma mixture model better accounts for among site rate heterogeneity, Bioinformatics 21 (2) (2005) 151–158.
[106] A. Carvalho, M. Tanner, Modelling nonlinear count time series with local mixtures of poisson autoregressions, Comput. Stat. Data Anal. 51 (11) (2007) 5266–5294.
[107] M. Svensén, C. Bishop, Robust Bayesian mixture modelling, Neurocomputing 64 (2005) 235–252.
[108] A. Stranjak, P. Dutta, M. Ebden, A. Rogers, P. Vytelingum, A multi-agent simulation system for prediction and scheduling of aero engine overhaul, in: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems: Industrial Track, International Foundation for Autonomous Agents and Multiagent Systems, 2008, pp. 81–88.
[109] L. Parra, G. Deco, S. Miesbach, Statistical independence and novelty detection with information preserving nonlinear maps, Neural Comput. 8 (2) (1996) 260–269.
[110] A. Nairac, T. Corbett-Clark, R. Ripley, N. Townsend, L. Tarassenko, Choosing an appropriate model for novelty detection, in: Proceedings of the 5th International Conference on Artificial Neural Networks, IET, 1997, pp. 117–122.
[111] R. Hyndman, Computing and graphing highest density regions, Am. Stat. 50 (2) (1996) 120–126.
[112] J. Pickands, Statistical inference using extreme order statistics, Ann. Stat. 3 (1) (1975) 119–131.
[113] P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling Extremal Events for Insurance and Finance, vol. 33, Springer Verlag, 1997.
[114] R. Fisher, L. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample, in: Proceedings of the Cambridge Philosophical Society, vol. 24, Cambridge University Press, 1928, pp. 180–190.
[115] D. Clifton, L. Tarassenko, N. McGrogan, D. King, S. King, P. Anuzis, Bayesian extreme value statistics for novelty detection in gas-turbine engines, in: Proceedings of the IEEE Aerospace Conference, IEEE, 2008, pp. 1–11.
[116] K. Worden, G. Manson, D. Allman, Experimental validation of a structural health monitoring methodology: part i. Novelty detection on a laboratory structure, J. Sound Vib. 259 (2) (2003) 323–343.
[117] K. Yamanishi, J. Takeuchi, G. Williams, P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Min. Knowl. Discov. 8 (3) (2004) 275–300.
[118] K. Yamanishi, J. Takeuchi, G. Williams, P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, in: Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2000, pp. 320–324.
[119] D. Agarwal, An empirical Bayes approach to detect anomalies in dynamic multidimensional arrays, in: Proceedings of the 5th IEEE International Conference on Data Mining, IEEE, 2005, pp. 26–33.
[120] D. Agarwal, Detecting anomalies in cross-classified streams: a Bayesian approach, Knowl. Inf. Syst. 11 (1) (2007) 29–44.
[121] F. Zorriassatine, J. Tannock, C. O'Brien, Using novelty detection to identify abnormalities caused by mean shifts in bivariate processes, Comput. Ind. Eng. 44 (3) (2003) 385–408.
[122] P. Højen-Sørensen, O. Winther, L. Hansen, Mean-field approaches to independent component analysis, Neural Comput. 14 (4) (2002) 889–918.
[123] J. Verbeek, N. Vlassis, B. Kröse, Efficient greedy learning of Gaussian mixture models, Neural Comput. 15 (2) (2003) 469–485.
[124] J. Zhang, Z. Ghahramani, Y. Yang, A probabilistic model for online document clustering with application to novelty detection, in: NIPS, 2005.
[125] P. Perner, Concepts for novelty detection and handling based on a case-based reasoning process scheme, Eng. Appl. Artif. Intell. 22 (1) (2009) 86–91.
[126] K. Hempstalk, E. Frank, I. Witten, One-class classification by combining density and class probability estimation, in: W. Daelemans, B. Goethals, K. Morik (Eds.), Machine Learning and Knowledge Discovery in Databases, Lecture Notes in Computer Science, vol. 5211, Springer, Berlin/Heidelberg, 2008, pp. 505–519.
[127] D. Chen, M. Meng, Health status detection for patients in physiological monitoring, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 4921–4924.
[128] T. Kanamori, S. Hido, M. Sugiyama, A least-squares approach to direct importance estimation, J. Mach. Learn. Res. 10 (2009) 1391–1445.
[129] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, T. Kanamori, Statistical outlier detection using direct density ratio estimation, Knowl. Inf. Syst. 26 (2) (2011) 309–336.
[130] M. Sugiyama, T. Suzuki, T. Kanamori, Density ratio estimation: a comprehensive review, RIMS Kokyuroku (2010) 10–31.
[131] S. Hoare, D. Asbridge, P. Beatty, On-line novelty detection for artefact identification in automatic anaesthesia record keeping, Med. Eng. Phys. 24 (10) (2002) 673–681.
[132] S. Roberts, L. Tarassenko, A probabilistic resource allocating network for novelty detection, Neural Comput. 6 (2) (1994) 270–284.
[133] P. Galeano, D. Peña, R. Tsay, Outlier detection in multivariate time series by projection pursuit, J. Am. Stat. Assoc. 101 (474) (2006) 654–669.
[134] D. Chen, X. Shao, B. Hu, Q. Su, Simultaneous wavelength selection and outlier detection in multivariate regression of near-infrared spectra, Anal. Sci. 21 (2) (2005) 161–166.
[135] K. Kadota, D. Tominaga, Y. Akiyama, K. Takahashi, Detecting outlying samples in microarray data: a critical assessment of the effect of outliers on sample classification, Chem-Bio Informat. 3 (1) (2003) 30–45.
[136] P. Smyth, Markov monitoring with unknown states, IEEE J. Sel. Areas Commun. 12 (9) (1994) 1600–1612.
[137] Z. Ghahramani, G. Hinton, Variational learning for switching state-space models, Neural Comput. 12 (4) (2000) 831–864.
[138] M. Atallah, W. Szpankowski, R. Gwadera, Detection of significant sets of episodes in event sequences, in: Proceedings of the 4th IEEE International Conference on Data Mining, ICDM'04, IEEE, 2004, pp. 3–10.
[139] A. Sebyala, T. Olukemi, L. Sacks, Active platform security through intrusion detection using naive Bayesian network for anomaly detection, in: London Communications Symposium, Citeseer, 2002.
[140] C. Kruegel, G. Vigna, Anomaly detection of web-based attacks, in: Proceedings of the 10th ACM Conference on Computer and Communications Security, ACM, 2003, pp. 251–261.
[141] C. Kruegel, D. Mutz, W. Robertson, F. Valeur, Bayesian event classification for intrusion detection, in: Proceedings of the 19th Annual Computer Security Applications Conference, IEEE, 2003, pp. 14–23.
[142] M. Mahoney, P. Chan, Learning nonstationary models of normal network traffic for detecting novel attacks, in: Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2002, pp. 376–385.
[143] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (3) (1962) 1065–1076.
[144] A. Frank, A. Asuncion, UCI machine learning repository, 2010.
[145] M. Hravnak, L. Edwards, A. Clontz, C. Valenta, M. DeVita, M. Pinsky, Defining the incidence of cardiorespiratory instability in patients in step-down units using an electronic integrated monitoring system, Arch. Internal Med. 168 (12) (2008) 1300–1308.
[146] P. Angelov, D. Filev, An approach to online identification of Takagi-Sugeno fuzzy models, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 34 (1) (2004) 484–498.
[147] R. Adams, I. Murray, D. MacKay, The Gaussian process density sampler, in: Advances in Neural Information Processing Systems (NIPS) 21, 2009, pp. 9–16.
[148] G. Lorden, Procedures for reacting to a change in distribution, Ann. Math. Stat. 42 (6) (1971) 1897–1908.
[149] M. Basseville, I.V. Nikiforov, Detection of Abrupt Changes: Theory and Application, vol. 104, Prentice Hall, Englewood Cliffs, 1993.
[150] J. Reeves, J. Chen, X.L. Wang, R. Lund, Q.Q. Lu, A review and comparison of changepoint detection techniques for climate data, J. Appl. Meteorol. Climatol. 46 (6) (2007) 900–915.
[151] T. Peng, C. Leckie, K. Ramamohanarao, Information sharing for distributed intrusion detection systems, J. Netw. Comput. Appl. 30 (3) (2007) 877–899.
[152] T. Van Phuong, L. Hung, S. Cho, Y. Lee, S. Lee, An anomaly detection algorithm for detecting attacks in wireless sensor networks, Intell. Secur. Informat. 3975 (2006) 735–736.
[153] A.G. Tartakovsky, G.V. Moustakides, State-of-the-art in Bayesian changepoint detection, Seq. Anal. 29 (2) (2010) 125–145.
[154] J. Chen, A.K. Gupta, Parametric Statistical Change Point Analysis: with Applications to Genetics, Birkhäuser Boston, Medicine, and Finance, Boston, 2012.
[155] S. Forrest, A. Perelson, L. Allen, R. Cherukuri, Self-nonself discrimination in a computer, in: Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy, IEEE, 1994, pp. 202–212.
[156] F. Angiulli, C. Pizzuti, Fast outlier detection in high dimensional spaces, in: Proceedings of the 6th European Conference on

Principles of Data Mining and Knowledge Discovery, PKDD '02, Springer-Verlag, London, UK, 2002, pp. 15–26.
[157] S. Bay, M. Schwabacher, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, in: Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2003, pp. 29–38.
[158] S. Boriah, V. Chandola, V. Kumar, Similarity measures for categorical data: a comparative evaluation, in: Proceedings of the 8th SIAM International Conference on Data Mining, 2008, pp. 243–254.
[159] M. Breunig, H. Kriegel, R. Ng, J. Sander, LOF: identifying density-based local outliers, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, vol. 29, ACM, 2000, pp. 93–104.
[160] V. Chandola, S. Boriah, V. Kumar, Understanding Categorical Similarity Measures for Outlier Detection, Technical Report 08-008, University of Minnesota, 2008.
[161] S. Chawla, P. Sun, SLOM: a new measure for local spatial outliers, Knowl. Inf. Syst. 9 (4) (2006) 412–429.
[162] A. Ghoting, M. Otey, S. Parthasarathy, Loaded: link-based outlier and anomaly detection in evolving data sets, in: Proceedings of the 4th IEEE International Conference on Data Mining, ICDM'04, IEEE, 2004, pp. 387–390.
[163] A. Ghoting, S. Parthasarathy, M. Otey, Fast mining of distance-based outliers in high-dimensional datasets, Data Min. Knowl. Discov. 16 (3) (2008) 349–364.
[164] V. Hautamaki, I. Karkkainen, P. Franti, Outlier detection using k-nearest neighbour graph, in: Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004, vol. 3, IEEE, 2004, pp. 430–433.
[165] F. Jiang, Y. Sui, C. Cao, Outlier detection using rough set theory, in: D. Slezak, J. Yao, J. Peters, W. Ziarko, X. Hu (Eds.), Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Lecture Notes in Computer Science, vol. 3642, Springer, Berlin Heidelberg, 2005, pp. 79–87.
[166] Y. Kou, C. Lu, D. Chen, Spatial weighted outlier detection, in: Proceedings of the SIAM Conference on Data Mining, 2006.
[167] M. Otey, A. Ghoting, S. Parthasarathy, Fast distributed outlier detection in mixed-attribute data sets, Data Min. Knowl. Discov. 12 (2) (2006) 203–228.
[168] G. Palshikar, Distance-based outliers in sequences, Distrib. Comput. Internet Technol. 3816 (2005) 547–552.
[169] D. Pokrajac, A. Lazarevic, L. Latecki, Incremental local outlier detection for data streams, in: Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2007, pp. 504–515.
[170] M. Wu, C. Jermaine, Outlier detection by sampling with accuracy guarantees, in: Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2006, pp. 767–772.
[171] J. Zhang, H. Wang, Detecting outlying subspaces for high-dimensional data: the new task, and performance, Knowl. Inf. Syst. 10 (3) (2006) 333–355.
[172] D. Barbará, Y. Li, J. Couto, COOLCAT: an entropy-based algorithm for categorical clustering, in: Proceedings of the 11th International Conference on Information and Knowledge Management, ACM, 2002, pp. 582–589.
[173] D. Barbará, Y. Li, J. Couto, J. Lin, S. Jajodia, Bootstrapping a data mining intrusion detection system, in: Proceedings of the ACM Symposium on Applied Computing, ACM, 2003, pp. 421–425.
[174] S. Budalakoti, A. Srivastava, R. Akella, E. Turkov, Anomaly Detection in Large Sets of High-Dimensional Symbol Sequences, Technical Report NASA TM-2006-214553, NASA Ames Research Center, 2006.
[175] D. Clifton, P. Bannister, L. Tarassenko, Learning shape for jet engine novelty detection, Adv. Neural Netw. (ISNN) 3973 (2006) 828–835.
[176] D. Clifton, P. Bannister, L. Tarassenko, A framework for novelty detection in jet engine vibration data, Key Eng. Mater. 347 (2007) 305–310.
[177] M. Filippone, F. Masulli, S. Rovetta, Applying the possibilistic c-means algorithm in kernel-induced spaces, IEEE Trans. Fuzzy Syst. 18 (3) (2010) 572–584.
[178] Z. He, X. Xu, S. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9) (2003) 1641–1650.
[179] D. Kim, P. Kang, S. Cho, H. Lee, S. Doh, Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing, Expert Syst. Appl. 39 (4) (2011) 4075–4083.
[180] A. Srivastava, B. Zane-Ulman, Discovering recurring anomalies in text reports regarding complex space systems, in: Proceedings of the IEEE Aerospace Conference, IEEE, 2005, pp. 3853–3862.
[181] A. Srivastava, Enabling the discovery of recurring anomalies in aerospace problem reports using high-dimensional clustering techniques, in: Proceedings of the IEEE Aerospace Conference, IEEE, 2006, pp. 1–17.
[182] H. Sun, Y. Bao, F. Zhao, G. Yu, D. Wang, CD-trees: an efficient index structure for outlier detection, Adv. Web-Age Inf. Manage. 3129 (2004) 600–609.
[183] Z. Syed, M. Saeed, I. Rubinfeld, Identifying high-risk patients without labeled training data: anomaly detection methodologies to predict adverse outcomes, in: AMIA Annual Symposium Proceedings, vol. 2010, American Medical Informatics Association, 2010, pp. 772–776.
[184] C.-H. Wang, Outlier identification and market segmentation using kernel-based clustering techniques, Exp. Syst. Appl. 36 (2) (2009) 3744–3750.
[185] J. Yang, W. Wang, CLUSEQ: efficient and effective sequence clustering, in: Proceedings of the 19th International Conference on Data Engineering, IEEE, 2003, pp. 101–112.
[186] S.-P. Yong, J.D. Deng, M.K. Purvis, Novelty detection in wildlife scenes through semantic context modelling, Pattern Recognit. 45 (9) (2012) 3439–3450.
[187] S. Yong, J. Deng, M. Purvis, Wildlife video key-frame extraction based on novelty detection in semantic context, Multimed. Tools Appl. 62 (2) (2013) 359–376.
[188] D. Yu, G. Sheikholeslami, A. Zhang, Findout: finding outliers in very large datasets, Knowl. Inf. Syst. 4 (4) (2002) 387–412.
[189] K. Zhang, S. Shi, H. Gao, J. Li, Unsupervised outlier detection in sensor networks using aggregation tree, Adv. Data Min. Appl. 4632 (2007) 158–169.
[190] E. Knorr, R. Ng, Algorithms for mining distance-based outliers in large datasets, in: Proceedings of the International Conference on Very Large Data Bases, Citeseer, 1998, pp. 392–403.
[191] Y. Tao, X. Xiao, S. Zhou, Mining distance-based outliers from large databases in any metric space, in: Proceedings of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2006, pp. 394–403.
[192] L. Wei, W. Qian, A. Zhou, W. Jin, J. Yu, Hot: hypergraph-based outlier test for categorical data, Adv. Knowl. Discov. Data Min. 2637 (2003) 562.
[193] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, vol. 1215, 1994, pp. 487–499.
[194] S. Papadimitriou, H. Kitagawa, P. Gibbons, C. Faloutsos, LOCI: fast outlier detection using the local correlation integral, in: Proceedings of the 19th International Conference on Data Engineering, IEEE, 2003, pp. 315–326.
[195] A. Chiu, A. Fu, Enhancements on local outlier detection, in: Proceedings of the 7th International Database Engineering and Applications Symposium, IEEE, 2003, pp. 298–307.
[196] J. Tang, Z. Chen, A. Fu, D. Cheung, Enhancing effectiveness of outlier detections for low density patterns, Adv. Knowl. Discov. Data Min. 2336 (2002) 535–548.
[197] D. Ren, B. Wang, W. Perrizo, RDF: a density-based outlier detection method using vertical data representation, in: Proceedings of the 4th IEEE International Conference on Data Mining, ICDM'04, IEEE, 2004, pp. 503–506.
[198] J. Yu, W. Qian, H. Lu, A. Zhou, Finding centric local outliers in categorical/numerical spaces, Knowl. Inf. Syst. 9 (3) (2006) 309–338.
[199] J. Tang, Z. Chen, A. Fu, D. Cheung, Capabilities of outlier detection in large datasets, framework and methodologies, Knowl. Inf. Syst. 11 (1) (2007) 45–84.
[200] P. Sun, S. Chawla, On local spatial outliers, in: Proceedings of the 4th IEEE International Conference on Data Mining, IEEE, 2004, pp. 209–216.
[201] P. Sun, S. Chawla, B. Arunasalam, Mining for outliers in sequential databases, in: Proceedings of the 6th SIAM International Conference on Data Mining, vol. 124, Society for Industrial Mathematics, 2006.
[202] P. Chan, M. Mahoney, M. Arshad, A Machine Learning Approach to Anomaly Detection, Technical Report, Department of Computer Science, Florida Institute Technology Melbourne, 2003.
[203] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Kluwer Academic Publishers, Norwell, MA, USA, 1981.
[204] R. Krishnapuram, J. Keller, A possibilistic approach to clustering, IEEE Trans. Fuzzy Syst. 1 (2) (1993) 98–110.
[205] A. Pires, C. Santos-Pereira, Using clustering and robust estimators to detect outliers in multivariate data, in: Proceedings of the International Conference on Robust Statistics, 2005.
[206] A. Vinueza, G. Grudic, Unsupervised Outlier Detection and Semi-Supervised Learning, Technical Report CU-CS-976-04, University of Colorado at Boulder, 2004.

[207] N. Wu, J. Zhang, Factor analysis based anomaly detection, in: Proceedings of the Information Assurance Workshop, IEEE Systems, Man and Cybernetics Society, IEEE, 2003, pp. 108–115.
[208] L. Ertöz, M. Steinbach, V. Kumar, Finding topics in collections of documents: a shared nearest neighbor approach, Clust. Inf. Retr. 11 (2003) 83–103.
[209] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, Adv. Neural Inf. Process. Syst. 12 (3) (2000) 582–588.
[210] J. Zhou, Y. Fu, C. Sun, Y. Fang, Unsupervised distributed novelty detection on scientific simulation data, J. Comput. Inf. Syst. 7 (5) (2011) 1533–1540.
[211] E. Spinosa, A. deLeon, F. de Carvalho, J. Gama, Novelty detection with application to data streams, Intell. Data Anal. 13 (3) (2009) 405–422.
[212] A.F. Hassan, H.M.O. Mokhtar, O. Hegazy, A heuristic approach for sensor network outlier detection, Int. J. Res. Rev. Wirel. Sensor Netw. (IJRRWSN) 1 (4) (2012) 66–72.
[213] T. Idé, S. Papadimitriou, M. Vlachos, Computing correlation anomaly scores using stochastic nearest neighbors, in: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM), IEEE, 2007, pp. 523–528.
[214] K. Onuma, H. Tong, C. Faloutsos, Tangent: a novel,'surprise me', recommendation algorithm, in: Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2009, pp. 657–666.
[215] M. Augusteijn, B. Folkert, Neural network classification and novelty detection, Int. J. Remote Sens. 23 (14) (2002) 2891–2902.
[216] S. Singh, M. Markou, An approach to novelty detection applied to the classification of image regions, IEEE Trans. Knowl. Data Eng. 16 (4) (2004) 396–407.
[217] P. Crook, S. Marsland, G. Hayes, U. Nehmzow, A tale of two filters-on-line novelty detection, in: Proceedings of the IEEE International Conference on Robotics and Automation, ICRA'02, vol. 4, IEEE, 2002, pp. 3894–3899.
[218] I. Diaz, J. Hollmen, Residual generation and visualization for understanding novel process conditions, in: Proceedings of the International Joint Conference on Neural Networks, IJCNN'02, vol. 3, IEEE, 2002, pp. 2070–2075.
[219] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier detection using replicator neural networks, Data Wareh. Know. Discov. 2454 (2002) 113–123.
[220] N. Japkowicz, Supervised versus unsupervised binary-learning by feedforward neural networks, Mach. Learn. 42 (1) (2001) 97–122.
[221] L. Manevitz, M. Yousef, One-class document classification via neural networks, Neurocomputing 70 (7) (2007) 1466–1481.
[222] B. Thompson, R. Marks, J. Choi, M. El-Sharkawi, M. Huang, C. Bunje, Implicit learning in autoencoder novelty assessment, in: Proceedings of the International Joint Conference on Neural Networks, IJCNN′02, vol. 3, IEEE, 2002, pp. 2878–2883.
[223] G. Williams, R. Baxter, H. He, S. Hawkins, L. Gu, A comparative study of RNN for outlier detection in data mining, in: Proceedings of the IEEE International Conference on Data Mining, IEEE, 2002, pp. 709–712.
[224] S. Jakubek, T. Strasser, Fault-diagnosis using neural networks with ellipsoidal basis functions, in: Proceedings of the American Control Conference, vol. 5, IEEE, 2002, pp. 3846–3851.
[225] Y. Li, M. Pont, N. Barrie Jones, Improving the performance of radial basis function classifiers in condition monitoring and fault diagnosis applications where unknown faults may occur, Pattern Recognit. Lett. 23 (5) (2002) 569–577.
[226] M.K. Albertini, R.F. de Mello, A self-organizing neural network for detecting novelties, in: Proceedings of the 2007 ACM Symposium on Applied Computing, SAC ′07, ACM, New York, NY, USA, 2007, pp. 462–466.
[227] G. Barreto, L. Aguayo, Time series clustering for anomaly detection using competitive neural networks, in: J. Príncipe, R. Miikkulainen (Eds.), Advances in Self-Organizing Maps, Lecture Notes in Computer Science, vol. 5629, Springer, Berlin Heidelberg, 2009, pp. 28–36.
[228] D. Deng, N. Kasabov, On-line pattern analysis by evolving self-organizing maps, Neurocomputing 51 (2003) 87–103.
[229] J. García-Rodríguez, A. Angelopoulou, J. García-Chamizo, A. Psarrou, S. Orts Escolano, V. Morell Giménez, Autonomous growing neural gas for applications with time constraint: optimal parameter estimation, Neural Netw. 32 (2012) 196–208.
[230] D. Hristozov, T. Oprea, J. Gasteiger, Ligand-based virtual screening by novelty detection with self-organizing maps, J. Chem. Inf. Model. 47 (6) (2007) 2044–2062.
[231] D. Kit, B. Sullivan, D. Ballard, Novelty detection using growing neural gas for visuo-spatial memory, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2011, pp. 1194–1200.
[232] S. Marsland, J. Shapiro, U. Nehmzow, A self-organising network that grows when required, Neural Netw. 15 (8–9) (2002) 1041–1058.
[233] S. Marsland, U. Nehmzow, J. Shapiro, On-line novelty detection for autonomous mobile robots, Robot. Auton. Syst. 51 (2) (2005) 191–206.
[234] M. Ramadas, S. Ostermann, B. Tjaden, Detecting anomalous network traffic with self-organizing maps, in: Recent Advances in Intrusion Detection, Springer, 2003, pp. 36–54.
[235] F. Wu, T. Wang, J. Lee, An online adaptive condition-based maintenance method for mechanical systems, Mech. Syst. Signal Process. 24 (8) (2010) 2985–2995.
[236] Y. Chen, B. Malin, Detection of anomalous insiders in collaborative environments via relational analysis of access logs, in: Proceedings of the 1st ACM Conference on Data and Application Security and Privacy, ACM, 2011, pp. 63–74.
[237] Y. Chen, S. Nyemba, B. Malin, Detecting anomalous insiders in collaborative information systems, IEEE Trans. Dependable Secur. Comput. 9 (3) (2012) 332–344.
[238] S. Günter, N. Schraudolph, S. Vishwanathan, Fast iterative kernel principal component analysis, J. Mach. Learn. Res. 8 (2007) 1893–1918.
[239] H. Hoffmann, Kernel PCA for novelty detection, Pattern Recognit. 40 (3) (2007) 863–874.
[240] A. Lakhina, M. Crovella, C. Diot, Mining anomalies using traffic feature distributions, ACM SIGCOMM Comput. Commun. Rev. 35 (4) (2005) 217–228.
[241] R. Kassab, F. Alexandre, Incremental data-driven learning of a novelty detection model for one-class classification with application to high-dimensional noisy data, Mach. Learn. 74 (2) (2009) 191–234.
[242] J. McBain, M. Timusk, Feature extraction for novelty detection as applied to fault detection in machinery, Pattern Recognit. Lett. 32 (7) (2011) 1054–1061.
[243] A. Perera, N. Papamichail, N. Bârsan, U. Weimar, S. Marco, On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions, IEEE Sens. J. 6 (3) (2006) 770–783.
[244] T. Ide, H. Kashima, Eigenspace-based anomaly detection in computer systems, in: Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2004, pp. 440–449.
[245] M. Shyu, S. Chen, K. Sarinnapakorn, L. Chang, A Novel Anomaly Detection Scheme Based on Principal Component Classifier, Technical Report, DTIC Document, 2003.
[246] M. Thottan, C. Ji, Anomaly detection in IP networks, IEEE Trans. Signal Process. 51 (8) (2003) 2191–2204.
[247] J. Toivola, M. Prada, J. Hollmén, Novelty detection in projected spaces for structural health monitoring, Adv. Intell. Data Anal. IX 6065 (2010) 208–219.
[248] Y. Xiao, H. Wang, W. Xu, J. Zhou, L1 norm based KPCA for novelty detection, Pattern Recognit. 46 (1) (2013) 389–396.
[249] S. Haggett, D. Chu, I. Marshall, Evolving a dynamic predictive coding mechanism for novelty detection, Knowl. Based Syst. 21 (3) (2008) 217–224.
[250] T. Hosoya, S. Baccus, M. Meister, Dynamic predictive coding by the retina, Nature 436 (7047) (2005) 71–77.
[251] T. Kohonen, The self-organizing map, Proc. IEEE 78 (9) (1990) 1464–1480.
[252] K. Labib, R. Vemuri, NSOM: a real-time network-based intrusion detection system using self-organizing maps, Netw. Secur. (2002) 1–6.
[253] D. Alahakoon, S. Halgamuge, B. Srinivasan, Dynamic self-organizing maps with controlled growth for knowledge discovery, IEEE Trans. Neural Netw. 11 (3) (2000) 601–614.
[254] J. Blackmore, R. Miikkulainen, Incremental grid growing: encoding high-dimensional structure into a two-dimensional feature map, in: Proceedings of the IEEE International Conference on Neural Networks, vol. 1, 1993, pp. 450–455.
[255] B. Fritzke, Growing cell structures – a self-organizing network for unsupervised and supervised learning, Neural Netw. 7 (9) (1994) 1441–1460.
[256] B. Fritzke, A growing neural gas network learns topologies, Adv. Neural Inf. Process. Syst. 7 (1995) 625–632.
[257] I. Jolliffe, MyiLibrary, Principal Component Analysis, vol. 2, Wiley Online Library, 2002.

[258] R. Fujimaki, T. Yairi, K. Machida, An approach to spacecraft anomaly detection problem using kernel feature space, in: Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining (SIGKDD), ACM, 2005, pp. 401–410.
[259] B. Schölkopf, A. Smola, K. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (5) (1998) 1299–1319.
[260] N. Kwak, Principal component analysis based on l1-norm maximization, IEEE Trans. Pattern Anal. Mach. Intell. 30 (9) (2008) 1672–1680.
[261] C. Noble, D. Cook, Graph-based anomaly detection, in: Proceedings of the 9th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2003, pp. 631–636.
[262] J. Sun, H. Qu, D. Chakrabarti, C. Faloutsos, Neighborhood formation and anomaly detection in bipartite graphs, in: Proceedings of the 5th IEEE International Conference on Data Mining, IEEE, 2005, pp. 418–425.
[263] J. Sun, Y. Xie, H. Zhang, C. Faloutsos, Less is more: compact matrix decomposition for large sparse graphs, in: Proceedings of the 7th SIAM International Conference in Data Mining, 2007.
[264] V. Chatzigiannakis, S. Papavassiliou, M. Grammatikou, B. Maglaris, Hierarchical anomaly detection in distributed large-scale sensor networks, in: Proceedings of the 11th IEEE Symposium on Computers and Communications, ISCC'06, IEEE, 2006, pp. 761–767.
[265] V. Lämsä, T. Raiko, Novelty detection by nonlinear factor analysis for structural health monitoring, in: Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2010, pp. 468–473.
[266] M. Timusk, M. Lipsett, C. Mechefske, Fault detection using transient machine signals, Mech. Syst. Signal Process. 22 (7) (2008) 1724–1749.
[267] D. Tax, R. Duin, Support vector domain description, Pattern Recognit. Lett. 20 (11) (1999) 1191–1199.
[268] Q. Song, W. Hu, W. Xie, Robust support vector machine with bullet hole image classification, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 32 (4) (2002) 440–448.
[269] W. Hu, Y. Liao, V. Vemuri, Robust anomaly detection using support vector machines, in: Proceedings of the International Conference on Machine Learning, 2003, pp. 282–289.
[270] G. Li, C. Wen, Z. Li, A new online learning with kernels method in novelty detection, in: Proceedings of the 37th Annual Conference on IEEE Industrial Electronics Society (IECON), IEEE, 2011, pp. 2311–2316.
[271] L. Manevitz, M. Yousef, One-class SVMs for document classification, J. Mach. Learn. Res. 2 (2002) 139–154.
[272] C. Campbell, K. Bennett, A linear programming approach to novelty detection, in: Proceedings of the Conference on Advances in Neural Information Processing Systems, vol. 13, The MIT Press, 2001, pp. 395–401.
[273] T. Le, D. Tran, W. Ma, D. Sharma, An optimal sphere and two large margins approach for novelty detection, in: Proceedings of the International Joint Conference on Neural Networks (IJCNN), IEEE, 2010, pp. 1–6.
[274] T. Le, D. Tran, W. Ma, D. Sharma, Multiple distribution data description learning algorithm for novelty detection, Adv. Knowl. Discov. Data Min. 6635 (2011) 246–257.
[275] Y.-H. Liu, Y.-C. Liu, Y.-J. Chen, Fast support vector data descriptions for novelty detection, IEEE Trans. Neural Netw. 21 (8) (2010) 1296–1313.
[276] Y.-H. Liu, Y.-C. Liu, Y.-Z. Chen, High-speed inline defect detection for TFT-LCD array process using a novel support vector data description, Exp. Syst. Appl. 38 (5) (2011) 6222–6231.
[277] X. Peng, D. Xu, Efficient support vector data descriptions for novelty detection, Neural Comput. Appl. 21 (8) (2012) 2023–2032.
[278] M. Wu, J. Ye, A small sphere and large margin approach for novelty detection using training data with outliers, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 2088–2092.
[279] Y. Xiao, B. Liu, L. Cao, X. Wu, C. Zhang, Z. Hao, F. Yang, J. Cao, Multi-sphere support vector data description for outliers detection on multi-distribution data, in: Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, 2009, pp. 82–87.
[280] L. Clifton, H. Yin, Y. Zhang, Support vector machine in novelty detection for multi-channel combustion data, Adv. Neural Netw. (ISNN) 3973 (2006) 836–843.
[281] L. Clifton, H. Yin, D. Clifton, Y. Zhang, Combined support vector novelty detection for multi-channel combustion data, in: Proceedings of the IEEE International Conference on Networking, Sensing and Control, IEEE, 2007, pp. 495–500.
[282] P.F. Evangelista, M.J. Embrechts, B.K. Szymanski, Taming the curse of dimensionality in kernels and novelty detection, in: Applied Soft Computing Technologies: The Challenge of Complexity, Springer Verlag, 2006, pp. 431–444.
[283] A. Gardner, A. Krieger, G. Vachtsevanos, B. Litt, One-class novelty detection for seizure analysis from intracranial EEG, J. Mach. Learn. Res. 7 (2006) 1025–1044.
[284] D.R. Hardoon, L.M. Manevitz, fMRI analysis via one-class machine learning techniques, in: Proceedings of the 19th International Joint Conference on Artificial intelligence, IJCAI'05, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005, pp. 1604–1605.
[285] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, L. Tarassenko, Static and dynamic novelty detection methods for jet engine health monitoring, Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. 365 (1851) (2007) 493–514.
[286] K. Heller, K. Svore, A. Keromytis, S. Stolfo, One class support vector machines for detecting anomalous windows registry accesses, in: Proceedings of the Workshop on Data Mining for Computer Security, 2003.
[287] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, J. Srivastava, A comparative study of anomaly detection schemes in network intrusion detection, in: Proceedings of the 3rd SIAM International Conference on Data Mining, vol. 3, SIAM, 2003, pp. 25–36.
[288] H. Lee, S. Cho, Application of LVQ to novelty detection using outlier training data, Pattern Recognit. Lett. 27 (13) (2006) 1572–1579.
[289] J. Ma, S. Perkins, Online novelty detection on temporal sequences, in: Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2003, pp. 613–618.
[290] J. Ma, S. Perkins, Time-series novelty detection using one-class support vector machines, in: Proceedings of the International Joint Conference on Neural Networks, vol. 3, IEEE, 2003, pp. 1741–1745.
[291] A. Rabaoui, H. Kadri, N. Ellouze, New approaches based on one-class SVMs for impulsive sounds recognition tasks, in: Proceedings of the IEEE Workshop on Machine Learning for Signal Processing, IEEE, 2008, pp. 285–290.
[292] L. Zhuang, H. Dai, Parameter optimization of kernel-based one-class classifier on imbalance learning, J. Comput. 1 (7) (2006) 32–40.
[293] Z. Wu, W. Xie, J. Yu, Fuzzy c-means clustering algorithm based on kernel method, in: Proceedings of the 5th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA), IEEE, 2003, pp. 49–54.
[294] V. Roth, Outlier detection with one-class kernel fisher discriminants, in: Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS), 2004.
[295] V. Roth, Kernel fisher discriminants for outlier detection, Neural Comput. 18 (4) (2006) 942–960.
[296] V. Sotiris, P. Tse, M. Pecht, Anomaly detection through a Bayesian support vector machine, IEEE Trans. Reliab. 59 (2) (2010) 277–286.
[297] A. Munoz, J. Moguerza, Estimation of high-density regions using one-class neighbor machines, IEEE Trans. Pattern Anal. Mach. Intell. 28 (3) (2006) 476–480.
[298] Y. Li, A surface representation approach for novelty detection, in: Proceedings of the International Conference on Information and Automation (ICIA), IEEE, 2008, pp. 1464–1468.
[299] Z. He, S. Deng, X. Xu, An optimization model for outlier detection in categorical data, Adv. Intell. Comput. 3644 (2005) 400–409.
[300] Z. He, S. Deng, X. Xu, J. Huang, A fast greedy algorithm for outlier mining, Adv. Knowl. Discov. Data Min. 3918 (2006) 567–576.
[301] S. Ando, Clustering needles in a haystack: an information theoretic analysis of minority and outlier detection, in: Proceedings of the 7th IEEE International Conference on Data Mining, ICDM'07, IEEE, 2007, pp. 13–22.
[302] E. Keogh, S. Lonardi, C. Ratanamahatana, Towards parameter-free data mining, in: Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), ACM, 2004, pp. 206–215.
[303] E. Keogh, J. Lin, S. Lee, H. Herle, Finding the most unusual time series subsequence: algorithms and applications, Knowl. Inf. Syst. 11 (1) (2007) 1–27.
[304] J. Lin, E. Keogh, A. Fu, H. Van Herle, Approximations to magic: finding unusual medical time series, in: Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, IEEE, 2005, pp. 329–334.
[305] A. Fu, O. Leung, E. Keogh, J. Lin, Finding time series discords based on haar transform, Adv. Data Min. Appl. 4093 (2006) 31–41.
[306] M. Gamon, Graph-based text representation for novelty detection, in: Proceedings of the First Workshop on Graph Based Methods

for Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2006, pp. 17–24.
[307] M. Filippone, G. Sanguinetti, Information theoretic novelty detection, Pattern Recognit. 43 (3) (2010) 805–814.
[308] M. Filippone, G. Sanguinetti, A perturbative approach to novelty detection in autoregressive models, IEEE Trans. Signal Process. 59 (3) (2011) 1027–1036.
[309] M. Filippone, G. Sanguinetti, Novelty Detection in Autoregressive Models Using Information Theoretic Measures, Technical Report CS-09-06, Department of Computer Science, University of Sheffield, 2009.
[310] L. Itti, P. Baldi, Bayesian surprise attracts human attention, Vis. Res. 49 (10) (2009) 1295–1306.