Change detection in streaming data analytics: 
A comparison of Bayesian online and martingale approaches

Bernadin Namoano*, Christos Emmanouilidis* , Cristobal Ruiz-Carcel*, and Andrew G Starr*

* Cranfield University, United Kingdom (email: {bernadin.namoano; christosem; c.ruizcarcel; a.starr}@cranfield.ac.uk)

Abstract: On line change detection is a key activity in streaming analytics, which aims to determine whether the current observation in a time series marks a change point in some important characteristic of the data, given the sequence of data observed so far. It can be a challenging task when monitoring complex systems, which are generating streaming data of significant volume and velocity. While applicable to diverse problem domains, it is highly relevant to monitoring high value and critical engineering assets. This paper presents an empirical evaluation of two algorithmic approaches for streaming data change detection. These are a modified martingale and a Bayesian online detection algorithm. Results obtained with both synthetic and real world data sets are presented and relevant advantages and limitations are discussed.

Keywords: streaming analytics; change detection; martingale; Bayesian online detection;

1. INTRODUCTION

Engineering asset resilience management critically depends on the ability to detect events that may prevent assets from safely and dependably delivering their intended function. Complex asset monitoring produces data often characterised by significant volume and velocity. With the greater integration of internet of things technologies in such monitoring, data generation creates demands for considerable transmission bandwidth and significant computing resources. While the value of the data asset itself is increasingly acknowledged, collected data are often left poorly exploited, failing to take appropriate advantage of their potential for enhancing asset management performance (Kubler et al., 2015). Part of the challenge lies with the difficulty in understanding when any observable change in the data corresponds to events of interest, which in turn must require intervention actions. Directing the attention on significant events remains a challenging problem in asset monitoring. Terms such as outlier detection (André et al., 2008), novelty detection (Markou and Singh, 2003), and anomaly detection (Chandola et al., 2009), are all employed in this context and are relevant to change detection when monitoring engineering assets (Worden et al., 2000). However, collected data are typically not linked with validated event cases, making it hard to apply any supervised type of learning from data, making unsupervised (Filev et al., 2010) or semi-supervised types of learning more applicable in practice (Kingma et al., 2014). Most employed algorithmic approaches still require calibration and adjustment, as the dynamic characteristics of streaming data greatly vary across domains.

This paper performs an empirical evaluation of two change detection approaches, namely a non-parametric modified martingale type (Ho, 2005) and a Bayesian approach (Adams and MacKay, 2007). These are applied to event detection problems on both synthetic and real world data to enable better insight into their performance. The paper is structured as follows. Section 2 places the present work in the context of the broader literature in the field. Section 3 presents a typical martingale change point detection algorithm, and introduces adaptations to address shortcomings of the original one. Section 4 describes the Bayesian online change detection algorithm. Results from the two approaches are presented and discussed in section 5. Section 6 is the conclusion.

2. RELATED WORK

Efficient change detection requires appropriate processing of streaming data to ensure that the delay time between a true change and its detection, as well as the rate of missed change events, are kept minimal (Ho, 2005). Depending on the availability of data annotated with labelled events, change detection methods can apply supervised, unsupervised, and semi-supervised learning and a range of optimisation methods.

In supervised learning-based methods, offline streaming data are labelled and then used to train models to perform change detection. These methods include classification or regression-based approaches, and it would be beyond the scope of this paper to mention all such applicable techniques. It is of interest though to highlight that major issues with this category of algorithms are their inability to detect unseen classes and their greediness in terms of data. Moreover, they require appropriate handling of unbalanced data, i.e. data with uneven numbers of pattern exemplars per class. For example, long time series with sparse events often present such learning challenges. In real world applications, the typical case is that there is a lack of annotated sample data which are representative enough of the range of possible circumstances and therefore a purely supervised approach is rarely applicable in practice.

Unsupervised learning based methods do not operate on labelled data and are therefore a more natural choice in practice. They typically seek to assign incoming data points into different clusters or simply to detect when an incoming
data point sufficiently deviates from a single or more clusters. A change is detected when consecutive data points are either assigned to different clusters or cannot be assigned to any cluster with a sufficient degree of confidence. The former may involve any type of clustering algorithms, whereas the latter typically rely on employing some type of distance metrics. Relevant methods include the cumulative sum (CUSUM) (Peach et al., 1995), martingale test (Vovk et al., 2003), minimax (Unnikrishnan et al., 2009), Bayesian inference based models (Adams and MacKay, 2007) (Ge et al., 2014) (Mohammad-Djafari and Féron, 2006), wavelets (Wang et al., 2018), as well as various forms of distance-based ones. Results from applying such methods are highly dependent on appropriately calibrating them to the problem at hand.

Semi-supervised learning, in which both supervised and unsupervised learning are combined for change point detection. Such methods are well suited for real world problems whereby sparsely labelled data are only available. Examples include feed-forward neural network (Zhang et al., 2017), used for offline training and the model built can be used online with non-negligible time delay. While they address the challenge of labelled data sparsity, they too rely on calibration. Change point detection problem can also be assimilated into an optimisation problem, wherein a cost function is minimised under some constraints (Truong et al., 2020). Many heuristics have also been developed to this end. Examples encompass sliding windows and bottom-up algorithm (SWAB) (Keogh et al., 2001), binSeg (Truong et al., 2020) in which global or local error cost is minimised. Appropriate choice of the cost function, sizing of the sliding window, and handling known challenges for the relevant optimisation approaches are among the typical challenges of such methods.

The choice of a change detection approach depends on the nature of the problem and the involved data characteristics. The motivation for the present work has been the need to develop streaming data change detection solutions for railway rolling stock asset monitoring. The aim is to evaluate current approaches (Namoano et al., 2019), study their shortcomings and strengths, and propose solutions and improvements to enhance change detection performance. Given this motivation, the paper focuses on fully online methods, i.e. methods that determine whether or not the current observation marks a change, given the data observed so far but not given the complete time series. The Bayesian online change detection (BOCD) (Adams and MacKay, 2007) and the martingale exchangeability test (Ho, 2005) are appropriate for such problems as they can handle univariate and multivariate data and can be adjusted to the problem at hand. For example, in BOCD, one can set the prior distribution of the underlying process or run a classical statistical test such as T-test to set hyperparameters. For the martingale methods, the computation of the so called p-values and the hypothesis test can be adapted to the problem. The two methods are presented next.

3. MARTINGALES AND EXCHANGEABILITY

3.1. Basic Martingale algorithm

In a martingale process, the conditional expectation of the upcoming data point is the same as the current one, given all observed values so far. The idea of the martingale concept of change detection is that by learning the statistical properties underlying the observed data, one can analyse any deviation by testing the data exchangeability. A time sequence of a vector of random variables $X = \{X_i \mid i = 1, \ldots, n \}$ is exchangeable if its joint distribution does not change by any alteration of the time sequence ordering of the observations $X_1, \ldots, X_n$. If the expected value of a sequence $M = \{M_i \mid i = 1, \ldots, n \}$ is a measurable function of $X$, then $M$ is a martingale with respect to $X$. Considering a sequence of data $S = \{S_1, S_2, \ldots, S_i, \ldots, S_{n-1}\}$ where $S_i$ is a multidimensional data point, a martingale test can be completed for each new point $S_n$ involving three types of parameters as follows (Ho, 2005):

1) **Strangeness or non-conformity measure.** Denoted as $\alpha_i$, it expresses the dissimilarity between a specific point and other data points.

2) **p-values:** the p-value $p_n$ for the new set $S \cup \{S_n\}$ defined as: $p_n = \prod_{i=1}^{n} p_{\alpha_i}$, where $\alpha_i$ is the strangeness, $\theta_i$ is randomly chosen from $[0, 1]$, $i=1...n-1$ and $\#\{i \mid \theta_i \leq \alpha_i\}$ is a counting function. A key property of the p-values is that as long a change does not occur, they are uniformly distributed in the interval $[0, 1]$.

3) **Computation of the martingale values $M^\varepsilon_n = \prod_{i=1}^{n} (\varepsilon \ p_{\alpha_i}^{-1})$, where $\varepsilon$ is chosen from $[0, 1]$ and controls the sensitivity of the change detection, while $p_i$ are the p-values. The martingale values can be computed without storing all previous values as $M^\varepsilon_n = \varepsilon p_{\alpha_n}^{-1} M^\varepsilon_{n-1}$. To remove the dependency on $\varepsilon$, a simple mixture of martingales can be used (Fedorova et al., 2012):

$$M_n = \int_0^1 M^\varepsilon_n \ d\varepsilon$$

When a change occurs, the p-values distribution becomes skewed, and the exchangeability condition is not met (Ho, 2005). The absence of change is represented by the hypothesis $H_0: 0 < M_n < \lambda$ where $\lambda$ is a defined threshold. The $H_0$ is rejected when change occurs, hence $M^\varepsilon_n \geq \lambda$. Testing the exchangeability online involves computing power martingale values using p-values (Vovk et al., 2003). The main idea is to construct martingales which attain large values when small p-values are generated. The exchangeability test was applied to detect changes in sequential streaming data (Ho, 2005). This can work on unlabelled data streams, with multiple martingale tests with multiple strangeness and thresholds being used to determine whether a change occurs. Martingale difference values can be used to test whether a concept change occurred.

3.2. Adaptations and modified martingale algorithm

Complex processes often exhibit such dynamic behaviour that selecting a constant threshold for the martingale test leads to higher false alarm rates. To reduce such effects, the martingale value can be reset at specified time intervals, when a change has not been detected (Balasubramanian et al., 2014). Furthermore, an adaptive threshold setting approach with Reproducing Kernel Hilbert Space (RKHS) projection (Wang et al., 2017) can be applied, but this does not estimate the value
of the initial threshold. Alternatively, different martingale tests with different thresholds and composite change detection criteria can be employed (Ho and Wechsler, 2007b). However, while this may improve performance, it also increases the risk of adding false alarms produced by individual martingale tests.

A modified martingale algorithm (Table 1) is introduced in this paper to address some of the challenges facing the original approach. The motivation for the modification is to address a common challenge for martingale approaches, related to the time delay between the true change point and the change point detected by the test. The potential significance of this often depends on the application context. In long run processes, the estimation of the strangeness becomes computationally expensive. This is due to the fact that when the frequency of changes is low or when there are no changes, the size of the buffer set of samples ($T$ in Table 1), used to compute the strangeness, can become excessively large. To improve the computational efficiency of estimating the strangeness, downsampling (Chawla, 2009) and windowing techniques can be used. Engineering processes often exhibit rapid changes in a limited number of time steps before going back to a steady-state. For example, a driver’s acceleration is a typical case. A driving journey may switch between coasting and cruising to modes that involve gear change, acceleration and deceleration with braking. In such cases, the growth of the martingale value is not fast enough to capture the change. Lowering the detection threshold will result in confusing true changes with noise, resulting in higher false detections.

Another difficulty is that the martingale value converges to zero for a streaming process with no changes. This often causes change points to be missed, as the growing struggle to reach the defined threshold. Moreover, when a change is detected, significant time delays may be observed. To reduce this effect, empirical tests carried out (Volkhonskiy et al., 2017) suggest that instead of basing the change determination on the original martingale $M_n$ values, martingale growth values $G_n$ can instead be utilised:

$$G_n = \max \{0, G_{n-1} + \log(M_n)\}$$

where $M_n$ is the original martingale. Depending on the context, the martingale values $G_n$ grows faster than the traditional martingale $M_n$ and can hence, exhibit a reduced delay in detecting a change. The present paper introduces a further change in the way a change is detected via a decision variable $B_n$, computed as follows:

$$B_n = \max(G_n, M_n) \text{ with } B_0 = 0.$$  

Hence, with this modified martingale the hypothesis $H_0$ is rejected when $B_n \geq \lambda_B$, where $\lambda_B$ a threshold, or accept it otherwise. This modification enables the algorithm to detect a change whether this is picked by growth in $G_n$ or in $M_n$. Such modifications aim at addressing some of the aforementioned challenges of the original algorithm.

4. BAYESIAN ONLINE CHANGE DETECTION

The BOCD approach applies online Bayesian reasoning by estimating for each current observation in a time series the probability to be a change point based on the data observed so far ($X_{1:t}$). The underlying assumption is that the generated data are independent and identically distributed (i.i.d) random variables and the change point segments are not overlapping. These are not assumptions that can be assured to hold in practice. However, it allows a simplification it the involved estimations which in many cases can still offer adequate results. The underlying idea of the algorithm is computing the posterior probabilities $P(r_t | X_{1:t})$ over the run lengths $r_t$.

$$P(r_t | X_{1:t}) = \sum \limits_{r_{t-1}} P(r_t, r_{t-1} | X_{1:t})$$

Another difficulty is that the martingale value converges to zero for a streaming process with no changes. This often causes change points to be missed, as the growing struggle to reach the defined threshold. Moreover, when a change is detected, significant time delays may be observed. To reduce this effect, empirical tests carried out (Volkhonskiy et al., 2017) suggest that instead of basing the change determination on the original martingale $M$ values, martingale growth values $G$ can instead be utilised:

$$G_n = \max \{0, G_{n-1} + \log(M_n)\}$$
The run-length is increased by one when the present data point is determined to belong to the same distribution with the previous data. However, it is reset to zero when a change occurs, indicating that the present measurement point belongs to a new distribution (Fig. 1-2). The run length drops to zero when a change occurs. The corresponding estimated posteriors are shown in Fig. 3. Further details of the algorithm can be found here in the literature (Adams and MacKay, 2007).

![Diagram of the message passing methods describing the posterior probabilities computed.](image)

In practice, one of the drawbacks of BOCD methods is the quadratic growth of the run-length table with the growth of the size of the time-series. Pruning techniques exist to overcome this issue but may affect the efficiency of the algorithm.

5. EXPERIMENTS AND RESULTS

The experiments carried out aim to compare the modified martingale and the Bayesian online method and seek to identify performance characteristics for each approach. The work is motivated by industrial requirements to apply change detection on real world datasets from railway rolling stock monitoring. However, as the involved datasets are not publicly available, a choice was made to apply also the evaluated methods on publicly available data as well. For better evaluation of both methods, the present investigation used different data sources with different characteristics of change points (regular change points and random change points) and different dimensions and numbers of records. Specifically, two categories of datasets are used in the experiments. The first is a synthetic publicly available benchmarking dataset. Three types of synthetic data streams with induced changes are simulated in this dataset. The first is generated using normally distributed clusters data generator called NDC. It consists of a dataset with 1000 attributes, including 100,000 data points. The second is the USPS three-digit handwritten data, with 256 attributes and 7,291 data points. The third synthetic dataset is a modified version of nursery binary data (UCI) consisting of 5 attributes and 12,960 data points. For each dataset, change points include the engine frequency, the charge air pressure, the second stage oil, the exhaust gases pressure and temperature, and the ambient temperature and its pressure. The dataset contains 295 change points. Each change represents changing driving modes (idle, acceleration, maximum speed and deceleration). The second dataset represents the 2016 soccer UEFA championship eurogame with 16 games. The change points include game events, such as start, end, goals, and substitutions. The details on the data processing as well as its transformation are available on (Goutte et al., 2019).

5.1. Algorithmic parameters setting

For the modified martingale, an incremental SVM is used for the computation of strangeness, while $\lambda_y$ was empirically set to 10, after experimentation. For the engines data as well as the eurogame datasets, the root mean square (RMS) method is used to compute the strangeness, while $\lambda_b$ was empirically set to 14. In both cases, the k-means (k=2) algorithm is employed to improve the change point location by separating the observations since the last change into two classes. The first class represents the statistical properties learned from the observations before the change and the second those after the change. For the martingale change detection, the algorithm of Table 1 is employed and when the adapted martingale value computed is greater or equal to the defined threshold, a change is detected. For the Bayesian online algorithm, the Gaussian distribution is used to update sufficient statistics referred to in (Adams and MacKay, 2007). Table 2 summarises the parameters used for both methods. The parameters were chosen via a nonlinear optimiser available in Matlab.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Martinagle parameters</th>
<th>BOCD parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDC</td>
<td>SVM, $\lambda_b = 10$</td>
<td>Gaussian, hazard=2000</td>
</tr>
<tr>
<td>USPS</td>
<td>SVM, $\lambda_b = 10$</td>
<td>Gaussian, hazard=1500</td>
</tr>
<tr>
<td>UCI</td>
<td>SVM, $\lambda_b = 10$</td>
<td>Gaussian, hazard=1500</td>
</tr>
<tr>
<td>Engines</td>
<td>RMS, $\lambda_b = 14$</td>
<td>Gaussian, hazard=100</td>
</tr>
<tr>
<td>Eurogame</td>
<td>RMS, $\lambda_b = 14$</td>
<td>Gaussian, hazard=100</td>
</tr>
</tbody>
</table>

5.2. Evaluation approach

The standard performance assessment employed in typical classification problems is insufficient for change detection. In real world problems it can be highly misleading to detect a change too early or too late. Therefore, typical performance metrics are supplemented by an additional indicator that assesses the timeless of the detection. Four performance indicators are therefore involved: recall, precision, the F1-score and the mean time delay (MTD). The precision gives the percentage of the correctness for a detection, meaning the likelihood of detecting true change. Recall measures the percentage of detected changes that are true changes. The F1-score is a metric of test accuracy. It measures the correctness of all identified cases. The MTD represents an average detection time lag, which is the duration between the true change and the one detected by the system.
5.3. Results and discussion

For each dataset, 50 runs have been performed, and then performance is averaged over all runs. These results are compared with the traditional martingale method (Ho, 2005). As seen in Table 3, the modified version of the martingale method has improved performance. The precision and the recall as well as the mean time delay have been improved in both synthetic and engines dataset.

Table 3: Original vs modified martingale performance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Precision</th>
<th>Recall</th>
<th>MTD (in s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDC data</td>
<td><strong>0.87</strong></td>
<td><strong>0.93</strong></td>
<td><strong>70</strong></td>
</tr>
<tr>
<td>Train engine data</td>
<td><strong>0.78</strong></td>
<td><strong>0.85</strong></td>
<td><strong>10</strong></td>
</tr>
</tbody>
</table>

Table 4: Results overall employed datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Modified martingale</th>
<th>BOCD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P(%)</td>
<td>R(%)</td>
</tr>
<tr>
<td>Synthetic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDC</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>USPS</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>UCI</td>
<td>94</td>
<td>92</td>
</tr>
<tr>
<td>Engines1</td>
<td>45</td>
<td>44</td>
</tr>
<tr>
<td>Engines2</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Engines3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>CROESP</td>
<td>62</td>
<td>50</td>
</tr>
<tr>
<td>ENGLIS</td>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>FRAALB</td>
<td>43</td>
<td>26</td>
</tr>
<tr>
<td>FRAIRL</td>
<td>29</td>
<td>6</td>
</tr>
<tr>
<td>FRAISL</td>
<td>36</td>
<td>25</td>
</tr>
<tr>
<td>GERFRA</td>
<td>53</td>
<td>57</td>
</tr>
<tr>
<td>GERTA</td>
<td>60</td>
<td>35</td>
</tr>
<tr>
<td>GERPOL</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>POLPOR</td>
<td>44</td>
<td>14</td>
</tr>
<tr>
<td>POLRAT</td>
<td>38</td>
<td>15</td>
</tr>
<tr>
<td>FRAFPR</td>
<td>59</td>
<td>25</td>
</tr>
<tr>
<td>FORWAL</td>
<td>43</td>
<td>35</td>
</tr>
<tr>
<td>RUSWAL</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SUFFRA</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>WALBEL</td>
<td>50</td>
<td>39</td>
</tr>
<tr>
<td>WALTNR</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 4 summarises the results of the experiments. The colour represents a visualisation of the performance for each metric. Green indicates better performance. A straightforward observation is that in terms of time delay, the BOCD MTD is consistently better than the martingale. On the synthetic data with regularly spaced change points, the martingale method highlights better F1-scores in comparison to the BOCD. The precision shows that the BOCD is less robust in term of noise.

On the real world dataset experiments, as there is no regularity at the timing of the change points, it is observed that the modified martingale method creates more false detections when fast paced changes occur in a short period (RUSWAL dataset result for example, in Table 4). Overall, BOCD outperformed the martingale, having lower mean time delay and better accuracy. Further work and more experiments are needed to replace the empirical choice for the calibration parameters with a systematic approach for their setting. Considering both types of datasets, the experiments show that when changes occur over long periods, the martingale is more accurate (F1-score) than the BOCD. However, for changes in short time windows, BOCD accuracy is higher than the martingale methods.

6. CONCLUSION

This paper proposes a modified version of the martingale method for change detection in multi dimensional streams as well as an empirical evaluation of both the modified martingale and Bayesian online change detection. The experiments show that the proposed modified martingale method is effective in real world as well as synthetic data. The comparison between the martingale method and BOCD shows that the martingale method achieved better performance regarding robustness to noise and detection accuracy when the observed time series is sparse in changes and the changes are of longer duration. BOCD is more accurate when changes of shorter duration occur, but generates high false alarm rates in the presence of noise. Future work will examine performance on real world datasets, using thresholds optimisation and mixing both methods through an ensemble learning process to enhance performance, taking into account the strangeness magnitude. Further research is needed to clarify the following:

**Strangeness measure:** For the employed data, unsuccessful attempts to use as strangeness measure distance-based metrics such as Euclidean and cosine have been made. It is therefore desirable to find a suitable or adaptive strangeness measure for the martingale methods so as to improve as well as the MTD.

**Modified martingale values:** Although empirical tests show better performance with the modified version, further research is needed for a sound justification regarding the conditions under which such improved performance is achieved.

**BOCD:** For this approach, there is a need for further research to improve its robustness and computational time complexity. This is needed to address the fact that when changes are sparse in time, the computational time complexity is quadratic.

ACKNOWLEDGEMENT

The authors wish to thank Unipart Rail, Instrumentel and the Engineering and Physical Sciences Research Council (EPSRC) for their financial and technical support.

REFERENCES


