# Comparing Data Distribution Using Fading Histograms

**Raquel Sebastião**[1] and **João Gama**[2] and **Teresa Mendonça**[3]

**Abstract.** The emergence of real temporal applications under non-stationary scenarios has drastically altered the ability to generate and gather information. Nowadays, under dynamic scenarios, potentially unbounded and massive amounts of information are generated at high-speed rate, known as data streams. Dealing with evolving data streams imposes the online monitoring of data in order to detect changes. The contribution of this paper is to present the advantage of using fading histograms to compare data distribution for change detection purposes. In an windowing scheme, data distributions provided by the fading histograms are compared using the Kullback-Leibler divergence. The experimental results support that the detection delay time is smaller when using fading histograms to represent data instead of standard histograms.

## 1 INTRODUCTION

Nowadays, a massive amount of information is gathered at high-speed rate, known as data streams. When dealing with data streams in dynamics environments, besides remembering discarded data, it is also necessary forgetting outdated data. To accomplish such assignments, this paper advances fading histograms, which weight data examples according to their age. Thus, while remembering the discarded data, fading histograms gradually forget old data.

Moreover, dynamic environments raise the need of online detecting changes and the delay between the occurrence of a change and its detection must be minimal. Widely used in the data stream context [4, 3, 1, 9, 8], windowing approaches for detecting changes in data consist of monitoring distributions over two different time-windows, performing tests to compare distributions and decide if there is a change. This paper proposes a windowing model for change detection, which evaluates, through the Kullback-Leibler divergence, the distance between data distributions provided by fading histograms.

This paper is organized as follows. It start introducing fading histograms and Section 3 presents a windowing model to compare data distributions for detecting changes. Next, in Section 4, the performance of the proposed model for detecting changes is evaluated on an artificial data set and is compared with the Page-Hinkley Test (PHT) when detecting distribution changes on a real data set. Finally, Section 5 presents conclusions on the proposed approach and advances directions for further research.

[1] Signal Processing Lab, IEETA, University of Aveiro, 3810-193 Aveiro, Portugal & LIAAD-INESC TEC, Campus da FEUP, Rua Dr. Roberto Frias, 378, 4200 - 465 Porto, Portugal, email: raquel.sebastiao@ua.pt

[2] LIAAD-INESC TEC & Fac. Economia da Universidade do Porto (FEP), Rua Dr. Roberto Frias, 4200-464 Porto, Portugal, email: jgama@fep.up.pt

[3] Dep. Matemática, Fac. Ciências da Universidade do Porto (FCUP), Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal & CIDMA, Dep. de Matemática, University of Aveiro, 3810-193 Aveiro, Portugal, email: tmendo@fc.up.pt

## 2 FADING HISTOGRAMS

A histogram is a synopsis structure that allows accurate approximations of the underlying data distribution and provides a graphical representation of a random variable. A standard histogram attributes the same importance to all observations. However, in dynamic scenarios, recent data is usually more important than old data. Therefore, outdated data can be gradually forgotten attributing different weights to data observations. Following an exponential forgetting, data distribution can be computed with fading histograms [10]. In this sense, data observations with high weight (the recent ones) contribute more to the fading histogram than observations with low weight (the old ones). With the recursive form, the fading histograms counts ($FHC$)can be constructed in the flow:

$$FHC_t = \alpha * FHC_{t-1}; FHC_t(k) = FHC_t(k) + 1; \quad (1)$$

where $k$ is the correspondent bin of the actual observation and $\alpha$ is an exponential fading factor, such that $0 \ll \alpha < 1$.

## 3 ADAPTIVE CUMULATIVE WINDOWS MODEL (ACWM)

A windows-based change detection method considers two time windows and the data distribution on both windows is monitored and compared to detect changes [1, 4, 8].

The ACWM for change detection evaluates the distance between data distributions (provided by fading histograms) through the Kullback-Leibler divergence (KLD) [5]. Considering two discrete distributions, from a reference window with probabilities $P_{RW}(i)$ and from a current sliding window with probabilities $P_{CW}(i)$, the KLD of $P_{RW}$ with respect to $P_{CW}$ is defined by:

$$KLD(P_{RW}||P_{CW}) = \sum_i P_{RW}(i) log \frac{P_{RW}(i)}{P_{CW}(i)}.$$

Due to the asymmetric property of the KLD, if the distributions are similar, the difference between $KLD(P_{RW}||P_{CW})$ and $KLD(P_{CW}||P_{RW})$ is small. The ACWM decision rule for detecting changes in based on the KLD asymmetry of KLD: $|KLD(P_{RW}||P_{CW}) - KLD(P_{CW}||P_{RW})| > \delta$, where $\delta$ is a user defined threshold. In the ACWM, the reference window (RW) has a fixed length and reflects the data distribution observed in the past. The current window (CW) is cumulative and it is updated sliding forward and receiving the most recent data. The evaluation step length is determined automatically depending on the data similarity: increases if the distance between distributions is small and decreases otherwise. More details on ACWM and the pseudo-code can be found in [8].

## 4 EXPERIMENTAL EVALUATION

This Section presents an evaluation of the advantage of using fading histograms to compare data distributions.

**Experiments on Artificial Data**

Different kinds of changes were simulated varying the mean and the standard deviation of normal distributions. Details on the artificial data design can be found in [8]. Data distributions, within the reference and the current windows, were computed using fading histograms with different values of fading factors: 1 (no forgetting at all), 0.9994, 0.9993, 0.999, 0.9985 and 0.997. The ACWM-fh is also suitable to detect changes on data streams with different amounts of noise and with different lengths of stationary phases [8].

Table 1 presents a summary of the detection delay time, showing that detection delay time decreases by decreasing the fading factor. The increase of false alarms when using a fading histogram with $\alpha = 0.997$ suggests that it is over reactive, therefore $\alpha \leq 0.997$ are not suitable for use in this data set. A Wilcoxon signed rank test was performed, at a significance level of 5%. Considering the very low p-values obtained, there is strong statistical evidence that the detection delay time of ACWM-fh is smaller than of ACWM.

**Table 1**: Detection delay time (average and standard deviation of 30 runs) of ACWM-fh. The number of runs, if any, where the ACWM-fh misses detection or signals a false alarm are in the form (Miss; False Alarm).

| Parameter changed | Mag. | Rate | Fading Factor 1 | 0.9985 | 0.997 |
|---|---|---|---|---|---|
| Mean | Abrupt | Low | 260 ± 57 | 233 ± 77 | 226 ± 70 (0;5) |
| | | Medium | 153 ± 24 (1;1) | 140 ± 32 (0;2) | 125 ± 36 (0;2) |
| | | Sudden | 19 ± 4 (0;1) | 16 ± 6 (0;1) | 13 ± 6 (0;1) |
| | Medium | Low | 410 ± 131 (0;1) | 365 ± 148 (0;2) | 311 ± 96 (0;5) |
| | | Medium | 242 ± 125 (0;1) | 205 ± 58 (0;3) | 186 ± 54 (0;5) |
| | | Sudden | 36 ± 22 (0;1) | 22 ± 9 (0;1) | 36 ± 110 (0;2) |
| | Smooth | Low | 516 ± 171 (7;2) | 448 ± 173 (2;2) | 369 ± 150 (0;9) |
| | | Medium | 371 ± 233 (5;0) | 289 ± 168 | 250 ± 151 (0;4) |
| | | Sudden | 233 ± 229 (1;0) | 138 ± 180 | 66 ± 76 (0;1) |
| STD | Abrupt | Low | 240 ± 34 | 204 ± 45 | 186 ± 52 |
| | | Medium | 168 ± 16 | 143 ± 25 | 138 ± 40 |
| | | Sudden | 71 ± 10 | 42 ± 19 | 23 ± 18 |
| | Medium | Low | 368 ± 87 | 294 ± 79 | 249 ± 92 |
| | | Medium | 213 ± 28 | 159 ± 35 | 140 ± 40 |
| | | Sudden | 65 ± 15 | 43 ± 13 | 39 ± 14 |
| | Smooth | Low | 517 ± 158 | 380 ± 145 | 316 ± 113 (0;2) |
| | | Medium | 362 ± 127 | 260 ± 85 | 204 ± 52 |
| | | Sudden | 162 ± 60 (1;0) | 87 ± 46 | 62 ± 40 |

**Experiments on Real Data**

This industrial data set was obtained within the scope of the work presented in [2], with the objective of designing different machine learning classification methods for predicting surface roughness in high-speed machining. Data was obtained by performing tests in a Kondia HS1000 machining center equipped with a Siemens 840D open-architecture CNC. These tests were done with different cutting parameters, using sensors for registry vibration and cutting forces. For change detection purposes, the measurements of the cutting speed on $X$ axes from 7 tests were joined sequentially in order to have only one data set with 6 changes with different magnitudes and sudden and low rates. The ability for detecting changes in data distribution of the ACWM-fh ($\alpha = 0.997$ and $\alpha = 1$) was compared with the Page-Hinkley Test (PHT)[7], which is a sequential analysis technique typically used for monitoring change detection in the average of a Gaussian signal [6].

Table 2 presents the results. The ACWM-fh is able to detect the 6 changes in the data with smaller detection delay time than when using histograms constructed over the entire data. Moreover, with both approaches for data representations, the model did not miss any change. Although data has different kinds of changes, both ACWM and ACWM-fh presented a performance which was highly resilient to false alarms. The same is not true for the PHT, which presented 18 false alarms in this experiment. Moreover, the average detection delay time obtained with the PHT is greater than when performing the ACWM-fh. Considering the detection delay, the false alarms and the miss detection rates, the ACWM-fh outperforms the ACWM and the PHT. Although ACWM-fh presents smaller detection delay times, the Wilcoxon signed rank test results do not support statistical evidence of that. However, it must be pointed out that this conclusion is based on a very small number of cases (6).

**Table 2**: Detection delay time on the industrial data set.

| | True Change | 4500 | 9000 | 21000 | 25500 | 37500 | 42000 | Average |
|---|---|---|---|---|---|---|---|---|
| Method | ACWM | 240 | 84 | 328 | 2410 | 284 | 128 | **3474** |
| | ACWM-fh | 170 | 14 | 258 | 499 | 348 | 192 | **1481** |
| | PHT | 19 | 27 | 706 | 450 | 889 | 50 | **2141** |

# 5 CONCLUSIONS AND FURTHER RESEARCH

This paper presents the advantage of using fading histograms to compare data distribution for change detection purposes. The experimental results show that when using fading histograms to represent data instead of standard histograms, the time to detect a change is significantly reduced. The proposed ACWM-fh but does not provide insights on the description of changes and further research must address change analysis. Moreover, fading histograms must be extended to a multidimensional perspective.

# ACKNOWLEDGEMENTS

This work is financed by the ERDF – European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281. The authors also acknowledge the support of the European Commission through the project MAESTRA (Grant number ICT-2013-612944).

# REFERENCES

[1] Albert Bifet and Ricard Gavaldá, 'Learning from time-changing data with adaptive windowing', in *In SIAM International Conference on Data Mining*, Berlin, Heidelberg, (2007).

[2] M. Correa, C. Bielza, and J. Pamies-Teixeira, 'Comparison of bayesian networks and artificial neural networks for quality detection in a machining process.', *Expert Syst. Appl.*, **36**(3), 7270–7279, (2009).

[3] Tamraparni Dasu, Shankar Krishnan, Suresh Venkatasubramanian, and Ke Yi, 'An information-theoretic approach to detecting changes in multi-dimensional data streams', in *In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications*, (2006).

[4] Daniel Kifer, Shai Ben-David, and Johannes Gehrke, 'Detecting change in data streams', in *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, VLDB '04, pp. 180–191. VLDB Endowment, (2004).

[5] S. Kullback and R. A. Leibler, 'On information and sufficiency', *Annals of Mathematical Statistics*, **22**, 49–86, (1951).

[6] H. Mouss, D. Mouss, N. Mouss, and L. Sefouhi, 'Test of page-hinckley, an approach for fault detection in an agro-alimentary production system', in *Control Conference, 2004. 5th Asian*, volume 2, pp. 815 –818, (2004).

[7] E. S. Page, 'Continuous inspection schemes', *Biometrika*, **41**(1-2), 100–115, (1954).

[8] Raquel Sebastião, 'Learning from Data Streams: Synopsis and Change Detection', *PhD Thesis*, University of Porto, Portugal, (2014).

[9] Raquel Sebastião, João Gama, and Teresa Mendonça, 'Learning from data streams: Synopsis and change detection.', in *STAIRS*, eds., Amedeo Cesta and Nikos Fakotakis, volume 179 of *Frontiers in Artificial Intelligence and Applications*, pp. 163–174. IOS Press, (2008).

[10] Raquel Sebastião, João Gama, and Teresa Mendonça, 'Constructing fading histograms from data streams', *Progress in Artificial Intelligence*, 1–14, (2014).