Efficient Change Detection for High Dimensional Data Streams

Spiros V. Georgakopoulos
Department of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
spirosgeorg@dib.uth.gr

Sotiris K. Tasoulis
Department of Applied Mathematics
Liverpool John Moores University
Liverpool, United Kingdom
S.Tasoulis@ljmu.ac.uk

Vassilis P. Plagianakos
Department of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece
vpp@dib.uth.gr

***Abstract*—The recent technological advancements in cloud computing and the access in increasing computational power has led in undertaking the data processing derived by mobile devices. In particular, when these data are high dimensional this is indispensable, since the mobile device has to balance its processing functionalities to additional services. However, developing efficient algorithms could allow various types of analysis to be performed locally, avoiding the necessity of a constantly connected device. In this work, we present a methodology that combines lightweight dimensionality reduction and change detection techniques. The experimental results justify its impressive performance and subsequently its usefulness in several tasks.**

***Index Terms*—High Dimensional Data, Data streams, Cumulative Sum, Incremental Principal Component Analysis.**

1. Introduction

The recent years, within the field of sensor networks, various wearable sensors are used to collect human body information. Furthermore, advances in Artificial Intelligent and Machine Learning allow data processing [1], [2] in an attempt to aid the medical treatment, social welfare, sports, etc. In many cases smartphone devices, having a variety of built-in sensors are used to collect these data. However, as the data dimensionality tends to grow, the limited memory and computational power of mobile devices such as the smartphones, Raspberry Pi or Unmanned Aerial Vehicle, is hindering the efficient data processing.

To deal with this problem, the wireless network capabilities of the devices are used and data are processed in remote servers or more recently in high computational power cloud infrastructure [3]. Nevertheless, this approach gives birth to a new series of problems [4], such as network connectivity, device energy consumption, etc.

In this work, we provide a methodology that fits on the low memory and computational capabilities of smartphones. To test our approach, we use the publicly available dataset "Human Activities and Postural Transitions" (HAPT) [5] which is a time series dataset characterized by high dimensionality. To this end, we employ an online dimensionality reduction technique to reduce the original space to an 1-dimensional space coupled with a lightweight statistical method for time series analysis. Our aim is to capture in real time a specific state in the signal every time it is appearing, using only the smartphone device.

The rest of the paper is structured as follows: In Section 2, we provide information regarding the dataset used. In Section 3, background material for dimensionality reduction and classification methods are provided. In Section 4, we present the proposed methodology and the experimental results. Finally, Section 5 contains concluding remarks and pointers for future work.

2. Dataset

As a case study to examine our methodology, we use a multivariant time series dataset [6], constructed using a series of basic human activities which are obtained using the sensor signals of a smartphone. To assemble the dataset, experiments were carried out within a group of 30 volunteers at the age bracket of 19-48 years. All the participants were wearing a smartphone (Samsung Galaxy S II) on their waist during the experiment execution. 3-axial linear acceleration and 3-axial angular velocity were captured at a constant rate of 50Hz using the built-in accelerometer and gyroscope.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of 561 features was obtained by calculating variables from the time and frequency domain.

3. Background Methods

In this section, we briefly review the basic tools used in the proposed methodology. In particular, we present the

Incremental Principal Component Analysis (IPCA) [7] for the dimensionality reduction task and the Cumulative Sum (CuSum) algorithm [8] for the online change detection.

3.1. Incremental Principal Components Analysis

The typical computational approach to PCA requires all the data input to be available in order to compute the eigenvalues and eigenvectors of the sample covariance matrix, and thus it belongs to the category of batch methods. This approach is not feasible when the data are incrementally derived from an on-line stream. Thus, an incremental method is required to estimate the principal components for observations arriving sequentially.

This can be achieved by updating the principal components for each arriving observation vector, while avoiding to estimate the covariance matrix as an intermediate result. Here, for that purpose we employ the Candid Covariance-free IPCA (CCIPCA) method [7], which is based on the works of Oja and Karhunen [9] and Sanger [10]. A short description of the method follows.

Let $d_1, d_2, \ldots$ be the sample vectors that are acquired sequentially at each time point and let u_1 be the first principal component. Each $d_n, n = 1, 2, \ldots$, is a a-dimensional vector, where each dimension corresponds to a sensor signal for the case at hand. Without loss of generality, we can assume that d_n has a zero mean, since the mean may be incrementally estimated and subtracted out. Then, the n-th step estimate u_1^n of u_1 is given by

$$u_1^n = \frac{n-1-l}{n} u_1^{n-1} + \frac{1+l}{n} d_n d_n^T \frac{u_1^{n-1}}{\|u_1^{n-1}\|},$$

where $(n-1)/n$ is the weight for the last estimate and $1/n$ is the weight of the new data, while the one dimensional projection y_n onto u_1^n is given by

$$y_n = u_1^n d_n.$$

The positive parameter l is called the amnesic parameter. With the presence of l, larger weight is given to new samples and the effect of old samples will gradually fade out. Finally, to begin the iteration, we set $u_1^0 = d_1$, the first direction of data spread. A mathematical proof of the convergence of CCIPCA can be found in [11].

3.2. Cumulative Sum

The Cumulative Sum (CuSum) is a change detection algorithm that can be used for off-line or on-line change detection. The CuSum has been firstly proposed in [8]. In this work, we utilize the on-line version of the algorithm.

To describe the functionality of the algorithm, we consider a sequence of independent random variables y_n, which correspond to a signal for various discrete time instances n, with a probability density $p_\theta(y)$, which depends only to the parameter θ. To capture the change of the signal at an unknown time instance t_0, the parameter θ have to change from the initial value θ_0 (θ is equal to θ_0) to θ_1 where $\theta_0 \neq \theta_1$. The approximation of these parameters that can be addressed using a training set of sample data.

More specifically, we assume the following hypotheses concerning the parameter θ:

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta = \theta_1 \end{aligned} \tag{1}$$

In the problem considered here, the signal samples are captured by the sensors of the smartphone in real time and the proper hypotheses must be computed. When the decision is continuously in favour of hypothesis H_0, there is absence of signal change, while a decision in favour of H_1 corresponds to signal samples that indicate a change.

In this paper, the following notation will be used. Let

$$S_n = \sum_{i=1}^{n} s_i, \text{ where } s_i = ln \frac{p_{\theta_1}(y_i)}{p_{\theta_0}(y_i)} \tag{2}$$

is the log-likelihood ratio for the observations from y_i to y_n, where n is the current time instant and refer to s_i as sufficient statistic. Considering the particular case where the distribution is Gaussian, with μ the mean value and σ the constance variance, when the changing parameter θ corresponds to μ. The probability density is denoted as

$$p_\theta(y) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp^{-(y-\theta)^2/2\sigma^2} \tag{3}$$

and the sufficient statistic s_i as

$$s_i = \frac{\theta_1 - \theta_0}{\sigma^2}(y_i - \frac{\theta_0 + \theta_1}{2}). \tag{4}$$

To detect the change, at each instant time, the following decision rule is utilized:

$$g_n = S_n - \mu_n \geq h, \text{ where } \mu_n = \min_{1 \geq j \geq n} S_j. \tag{5}$$

The stopping time, denoted as t_a, is

$$t_a = \min \{n : g_n \geq h\} \tag{6}$$

and can be rewritten as

$$t_a = \min \{n : S_n \geq \mu_n + h\}. \tag{7}$$

The above decision rule compares the cumulative sum S_n and the adaptive thresholding $\mu_n + h$, since the μ_n keeps memory of the past observations whereas, the parameter h must be specified. The parameter h is critical, affects the decision rule threshold and therefore the signal change correctness.

4. Proposed Method and Experimental Results

In this section, we describe the technical details of the proposed methodology and we examine its efficiency on processing the publicly available dataset "Human Activities and Postural Transitions" (HAPT) to detect the Lying state from the different walking states. There are several activities categorized in the dataset which are further discriminated as active or non active. Here we are interested in detecting

changes between these two different states. As a follow up on our previous work on fall detection [2], [12] we are particularly interested in significant changes of the body orientation. Detecting every single one of such changes in a time series that correspond to a time window could further help on measuring walking stability or danger of accident occurrence. In addition such measurements are of great use in athletic exercise applications that are extremely popular in recent wearable accessories.

4.1. Proposed Method

To begin with, we apply the Principal Components Analysis method on the train set by projecting it to the First Principal Component. The significance of this process is dual; firstly, we produce an initial eigenspace, which is used as an initialization for the Incremental Principal Components Analysis method and secondly we calculate the one dimensional mean values of the two classes that are later required for our hypotheses testing. The mean value of Lying state denote the signal change parameter θ_1, while the mean value of the other active state correspond to θ_0.

Subsequently, for the test dataset we employ Incremental Principal Components Analysis to sequentially reduce the dimensionality of each data sample from the 561-dimensional space to an 1-dimensional space. Then using CUSUM algorithm decision rule g_n we aim to detect the lying states in the dataset. Notice that there is no need for any intermediate calculation of the covariance matrix, while as the projected sample is 1-dimensional there is no need to employ modifications of the CUSUM algorithm or any other more complex change detection algorithms that are applicable in multivariate time series.

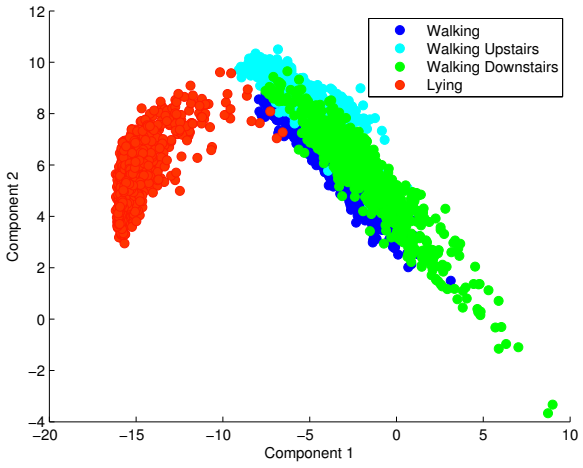

Figure 1. The training data transformed by PCA into 2-dimensional space.

4.2. Experimental Results

Firstly, we make an attempt to visually investigate the effect of the dimensionality reduction on the data structure.

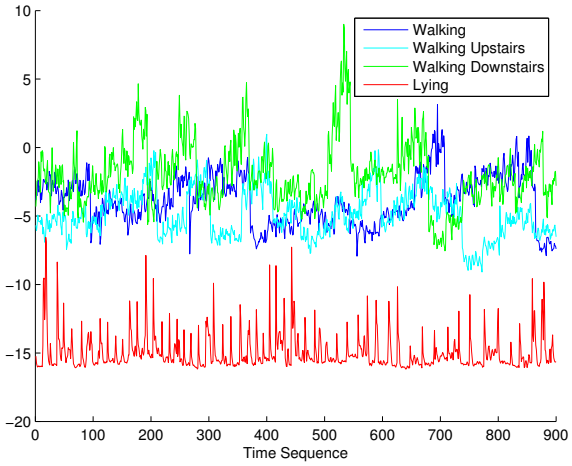

Figure 2. The training data transformed by PCA into 1-dimensional space. Different colors denote the four different states present in the dataset.

For this purpose we employ the 2-dimensional (see Figure 1) and 1-dimensional (see Figure 2) projection onto the first two and the first Principal Components, respectively. In Figure 2 the sample of each different state has been grouped together and presented in the same time frame. As shown in both cases the class separability is exposed, while the presence of possible outliers is also detected.

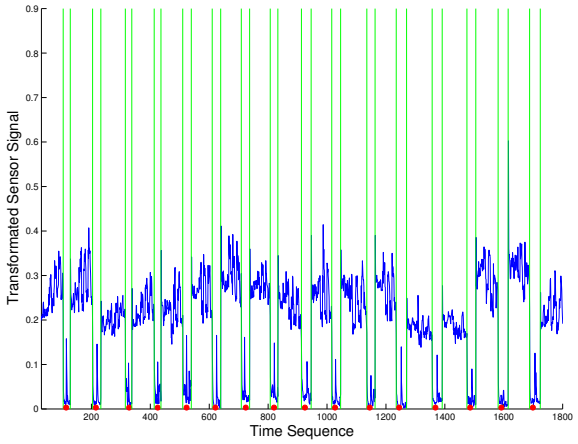

Figure 3. The test dataset transformed by IPCA into 1-dimensional space maintaining. The red dots represent the CUSUM algorithm triggering.

We apply the proposed methodology, using the default amnesic parameter $l = 2$ for the CCIPCA method [7] and the CUSUM parameter $h = 0.2$, while to further investigate the robustness of CUSUM we use different values of h parameter. The performance of the proposed methodology is illustrated in Figure 3, where the 1-dimension test data transformed using IPCA is depicted. Each two consecutive green vertical lines represent the Lying state in the time series while red dots represent the exact times that the algorithm

detected the change to the new state. As shown every Lying state is recognized every single time. We indicate that during this process, the CUSUM stops triggering and capture the next Lying without the need of any reset. In Figure 4, we present an example of the CUSUM decision function for the test sample, where we observe how the decision function is fast approaching zero values.

To examine the sensitivity of the proposed approach we perform a series of experiments using different values for the h parameter. In Table 1, we present the ratio of detected changes to the actual Lying states. We observe that the performance is only affected for values higher than 0.4.

TABLE 1. LYING STATE DETECTION USING DIFFERENT CUSUM PARAMETER h

	Parameter h					
	0.02	0.1	0.2	0.3	0.4	0.5
Lying state Detected	16/16	16/16	16/16	16/16	16/16	11/16

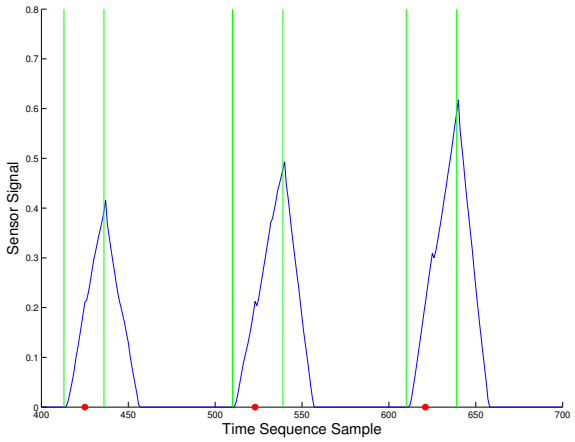

Figure 4. The CUSUM decision function of data test sample. The vertical lines represent the real Lying state and the red dots the CUSUM triggering.

5. Conclusion

In this work, we present a methodology for change detection in multidimensional mobile sensors data. We use the publicly available HAPT dataset, where all the changes from active states to the Lying state are accurately detected. We focus on computational efficiency by incorporating an online dimensionality reduction approach combined with a lightweight change detection algorithm. In our further work, we intend to release an open source implementation of this approach for low computational power devices such as smartphones, Raspberry Pi, Unmanned Aerial Vehicle, etc.

Acknowledgments

The authors would like to thank the European Union (European Social Fund ESF) and Greek national funds through the Operational Program Education and Lifelong Learning of the National Strategic Reference Framework (NSRF) Research Funding Program: Thalis: Interdisciplinary Research in Affective Computing for Biological Activity Recognition in Assistive Environments, for financially supporting this work.

References

[1] X. Lai, Q. Liu, X. Wei, W. Wang, G. Zhou, and G. Han, "A survey of body sensor networks," *Sensors*, vol. 13, no. 5, pp. 5406–5447, 2013.

[2] S. K. Tasoulis, C. N. Doukas, V. P. Plagianakos, and I. Maglogiannis, "Statistical data mining of streaming motion data for activity and fall recognition in assistive environments," *Neurocomputing*, vol. 107, pp. 87–96, 2013.

[3] M. Choi, "A platform-independent smartphone application development framework," in *Computer Science and Convergence*, ser. Lecture Notes in Electrical Engineering, J. J. (Jong Hyuk) Park, H.-C. Chao, M. S. Obaidat, and J. Kim, Eds. Springer Netherlands, 2012, vol. 114, pp. 787–794.

[4] S. Tarkoma, M. Siekkinen, E. Lagerspetz, and Y. Xiao, *Smartphone Energy Consumption: Modeling and Optimization*. Cambridge University Press, 2014.

[5] J. L. Reyes-Ortiz, A. Ghio, X. Parra, D. Anguita, J. Cabestany, and A. Català, "Human activity and motion disorder recognition: towards smarter interactive cognitive environments," in *21st European Symposium on Artificial Neural Networks, ESANN 2013, Bruges, Belgium, April 24-26, 2013*, 2013.

[6] J.-L. Reyes-Ortiz, L. Oneto, A. Sam, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, 2015.

[7] J. Weng, Y. Zhang, and W. shiuan Hwang, "Candid covariance-free incremental principal component analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1034–1040, 2003.

[8] E. S. Page, "Continuous Inspection Schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[9] E. Oja and J. Karhunen, "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix," *Journal of Mathematical Analysis and Applications*, vol. 106, pp. 69–84, 1985.

[10] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, no. 6, pp. 459–473, 1989.

[11] Y. Zhang and J. Weng, "Convergence analysis of complementary candid incremental principal component analysis," Comput. Sci. Eng., Michigan State Univ., East, Tech. Rep., 2001.

[12] S. V. Georgakopoulos, S. K. Tasoulis, I. Maglogiannis, and V. P. Plagianakos, "On-line fall detection via mobile accelerometer data," in *procceding of the 11th International Conference on Artificial Intelligence Applications and Innovations, (IFIP AIAI 2015)*, 2015, to appear.