# PCA Feature Extraction for Change Detection in Multidimensional Unlabelled Streaming Data

Ludmila I Kuncheva and William J Faithfull
*School of Computer Science, Bangor University*
*Bangor, LL57 1UT, United Kingdom*
*l.i.kuncheva@bangor.ac.uk, w.faithfull@bangor.ac.uk*

## Abstract

*While there is a lot of research on change detection based on the streaming classification error, finding changes in multidimensional unlabelled streaming data is still a challenge. Here we propose to apply principal component analysis (PCA) to the training data, and mine the stream of selected principal components for change in the distribution. A recently proposed semi-parametric log-likelihood change detector (SPLL) is applied to the raw and the PCA streams in an experiment involving 26 data sets and an artificially induced change. The results show that feature extraction prior to the change detection is beneficial across different data set types, and specifically for data with multiple balanced classes.*

## 1. Introduction

Adaptive classification in the presence of concept drift is one of the main challenges of modern machine learning and data mining [14].[1] The increasing interest in this field reflects the abundance of application areas, including engineering, finance, medicine and computing. Monitoring a single variable such as the classification error rate has been thoroughly studied [2, 6, 9–11]. However, in many applications, the class labels of the streaming data are not readily available, and thus the error rate cannot serve as a performance gage.

An indirect performance indicator would be a change in the distribution of the streaming data. There are at least two caveats related to this approach. First, the concept of change becomes context-dependent. For example, in comparing X-ray images, a hair-line discrepancy in a relevant segment of the image may be a sign of an important change. At the same time, if colour distribution is monitored, such a change will be left unregistered. Second, not all substantial changes of the distribution of the unlabelled data will manifest themselves as an increase of the error rate of the classifier. In some cases the same classifier may still be optimal for the new distributions. Hence there are two starting assumptions: (1) changes likely to affect adversely the performance of the classifier are detectable from the unlabelled data, and (2) changes of the distribution of the unlabelled data will be reasonably correlated with the classification error.

Our research hypothesis is that feature extraction is beneficial for change detection from multidimensional unlabelled streaming data. Section 2 gives the details of the proposed approach and the change detecting criterion. Section 3 contains the experiment, and Section 4, the conclusions.

## 2. Feature extraction for change detection

There is a rich body of literature on change detection including strategies for choosing, sampling, splitting, growing and shrinking a pair of sliding windows for optimal change detection [2, 6, 13]. Here we assume that the two windows of data, $W_1$ and $W_2$, are given. The first window contains the training data, and the second window is a sample from the streaming data.

Figure 1 shows the two major scenarios for change detection. When the labels of the data are available straight after classification, or even with some delay, the classification error can be monitored directly. When substantial increase is found, change is signalled. Most of the existing change detection methods and criteria are developed under this assumption.

Within the second scenario, labels are not available, and the question is whether the streaming data distribution matches the training one. The two scenarios share a

[1] See also `http://www.cs.waikato.ac.nz/~abifet/PAKDD2011/`.

978-4-9906441-1-6 ©2012 IAPR

distribution modelling block in the diagram. The modelling is sometimes implicit, and is included in the calculation of the change detection criterion. Compared to the multi-dimensional case, approximating distributions in the one-dimensional case can be much more accurate and useful. This explains the greater interest in the one-dimensional case. Methods such as Hidden Markov Models (HMM), Gaussian Mixture Modelling (GMM), Parzen windows, kernel-based approximation and martingales have been proposed for this task. The most common approach to the multidimensional case is clustering [5] followed by monitoring of the clusters' characteristics over time. Song et al. [12] propose a kernel estimation, and Dasu et al. [3] consider approximation via $kdq$-trees. A straightforward solution from statistics is to treat the two windows as two groups and apply the Hotelling's $t^2$ test to check whether the means of the two groups are the same [7]. The output of the data modelling block, which can also be labelled "criterion evaluation", is a value that is compared with a threshold to declare change or no change.

We propose the Feature extraction block, highlighted in the diagram. Our rationale for inserting an extra block in the unlabelled scenario is that distribution modelling of multidimensional raw data may be difficult. Intuitively, extracting features which are meant to *capture and represent the distribution in a lower dimensional space* may simplify this task.

We use a recently proposed semi-parametric log-likelihood criterion (SPLL) for change detection [8]. It comes as a special case of a log-likelihood framework and is modified to ensure computational simplicity. The SPLL statistic is calculated as follows. (1) Cluster the data in $W_1$ into $K$ clusters using the k-means algorithm ($K$ is a parameter of the algorithm; it was found that $K = 3$ works well). (2) Calculate the weighted intra-cluster covariance matrix $S$. (3) For each object in window $W_2$, calculate the Mahalanobis distance to each cluster centre using $S^{-1}$. Calculate the average of the minimum distances as

$$SPLL(W_1, W_2) = \frac{1}{M_2} \sum_{\mathbf{x} \in W_2} (\mathbf{x} - \mu_{i*})^T S^{-1} (\mathbf{x} - \mu_{i*}),$$

where $\mu_{i*}$ is the centre of the cluster closest to $\mathbf{x}$ and $M_2$ is the number of objects in window $W_2$. The criterion is derived as the upper limit of the negative log-likelihood of the sample in $W_2$ with respect to the approximated distribution from $W_1$. If $W_2$ comes from the same distribution as $W_1$, the squared Mahalanobis distances have a chi-square distribution with $n$ degrees of freedom (where $n$ is the dimensionality of the feature space). Leaving the problem of determining an optimal threshold aside, we are more interested in finding out whether the $SPLL$ statistic correlates with the classification accuracy. Since larger values of SPLL indicate change, large negative correlations are desirable. We expect the correlation between SPLL and the accuracy to be stronger for the PCA features compared to that for the raw data.

## 3. Experiment

The experiment was run on 26 data sets listed alphabetically in Table 1, with differing numbers of instances, features and classes. The sets were sourced from UCI [1] and a private collection. The experiment compared the correlation of the SPLL change statistic and classifier accuracy, with and without PCA. We used the SVM classifier from the MATLAB bioinformatics toolbox. PCA is beneficial if it results in a stronger negative correlation of the change statistic and classifier accuracy. Our initial hypothesis was that the more relevant PCs will lead to better change detection. A pilot experiment revealed that, in fact, the opposite is true. The PCs responsible for the last 10% of the variability of the data were more indicative of change and can be used for identifying outliers [4]. The following procedure was applied 30 times to each data set

(1) Take a stratified random sample of size $M$ as the window with the training data, $W_1$, and train an SVM classifier on it.
(2) Run PCA on $W_1$ and keep the components responsible for the last 10% of the variance of the data.
(3) Construct a random sample of 1000 instances from the remaining data as an i.i.d. testing data stream, and induce artificial concept drift between time moments $T_1$ and $T_2$ by setting $k$ features to zero.
(4) Run a sliding window $W_2$ of size $M$ along the testing stream. Calculate and store $SPLL(W_1, W_2)$ on the row data and on the PCA-transformed data. Calculate and store the classification accuracy of the SVM on $W_2$.
(5) Calculate the correlation coefficient between the classification accuracy and the two SPLL change statistics.

In this experiment we used $M = 50$, $T_1 = 300$, $T_2 = 450$, and $k$ was set to $1/6$ and $1/4$ of all features. The change that we applied reflects a possible real-life case where a group of sensors stop working due to a technical fault. The results are shown in Table 1. The correlation coefficient between the classification accuracy and SPLL calculated from the raw data is denoted by $\rho_{\text{raw}}$, and the one for the features extracted through PCA, by $\rho_{\text{PCA}}$. The coefficients where the PCA "wins" over the raw data detection are underlined. Using the 30

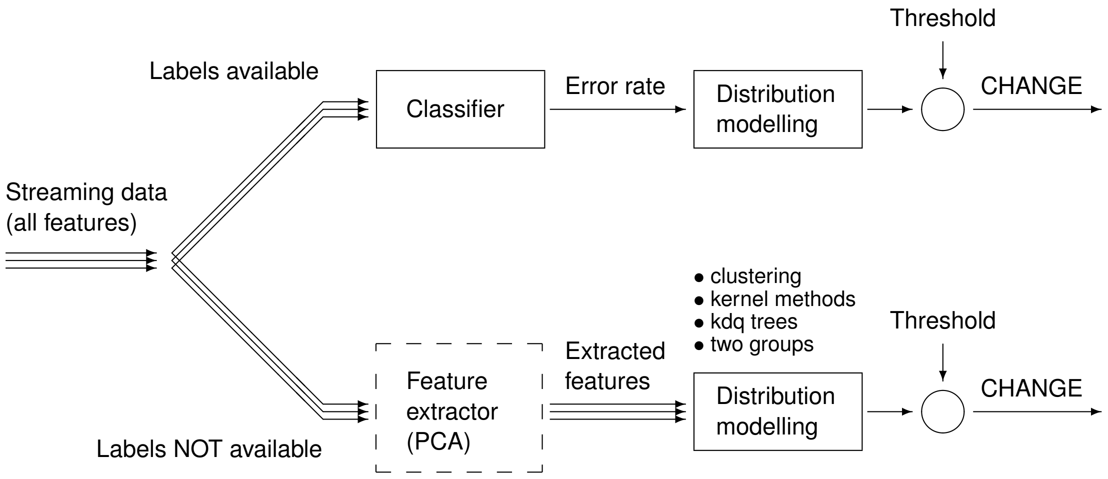

**Figure 1. Feature extraction for change detection**

replicas of the experiment, we carried out a paired two-tailed t-test for each data set. Statistically significant differences ($\alpha = 0.05$) are indicated in the table with • if PCA was better, and with ○ if the raw data detection was better. The results can be summarised as follows with respect to PCA versus raw data:

| | Wins$_{(\alpha=0.05)}$ | Better | Draw | Losses |
|---|---|---|---|---|
| 1/6th zeros | 12 | 18 | 13 | 1 |
| 1/4th zeros | 13 | 21 | 13 | 0 |

The results demonstrate that feature extraction through PCA leads to better change detection and therefore stronger correlation with the classification accuracy than using the raw unlabelled data. We carried out further analyses to establish which characteristics of the data sets may be related to the feature extraction success. Figure 2 shows a scatter plot where each point corresponds to a data set for the 1/6th null features experiment (the less favourable of the two). The x-axis is the prior probability of the largest class and the y-axis is the prior probability of the smallest class. The feasible space is within a triangle, as shown in the figure. The right edge corresponds to 2-class problems, and the number of classes increases from this edge towards the origin (0,0). The left edge of the triangle corresponds to equiprobable classes. This edge can be thought of as the edge of balanced problems. The balance disappears towards the bottom right corner. The pinnacle of the triangle corresponds to two equiprobable classes. The marker signifies which of the two approaches is better. Diamond marker means that PCA wins over the raw data, and a circle around it indicates that the difference is found to be statistically significant. The data sets where PCA was worse than raw data are shown with triangle markers. The only significant difference in that direction is indicated with a square around the triangle.

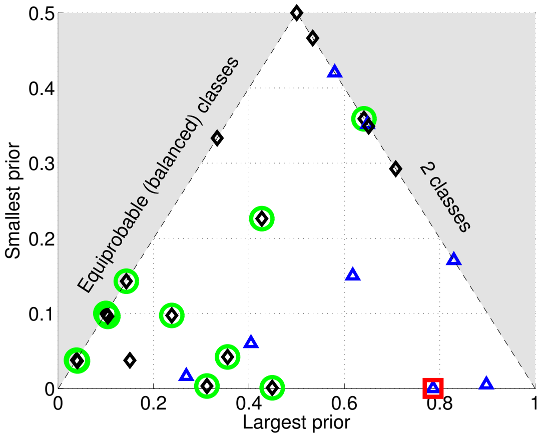

**Figure 2. Winning approach for 1/6th null features (diamond = PCA, triangle = raw data)**

The figure suggests that the PCA has a stable and consistent behaviour for multi-class, fairly balanced data sets (bottom left of the scatterplot). For imbalanced classes (bottom right) raw data gave better results.

## 4. Conclusions

We propose PCA feature extraction for change detection in unlabelled, multidimensional streaming data. We carried out an experiment where we simulated equipment fault through setting a proportion of all fea-

**Table 1. Results on streaming data classification with three induced changes.**

| | | | | | | Sixth null | | Quarter null | |
|---|---|---|---|---|---|---|---|---|---|
| Name | $n$ | $c$ | $N$ | $\frac{N_{\max}}{N_{\min}}$ | % PCA | $\rho_{\text{raw}}$ | $\rho_{\text{PCA}}$ | $\rho_{\text{raw}}$ | $\rho_{\text{PCA}}$ |
| breast | 9 | 2 | 277 | 2.42 | 62 | -2.6 | $\underline{-8.2}$– | 0.1 | $\underline{-22.7}$● |
| contraceptives | 9 | 3 | 1473 | 1.89 | 96 | -7.9 | $\underline{-22.3}$● | -4.0 | $\underline{-21.4}$● |
| glass | 9 | 6 | 214 | 8.44 | 72 | -37.1 | $\underline{-75.5}$● | -23.6 | $\underline{-89.9}$● |
| image | 19 | 7 | 2310 | 1.00 | 88 | -0.9 | $\underline{-55.1}$● | 0.8 | $\underline{-69.1}$● |
| ionosphere | 34 | 2 | 351 | 1.79 | 73 | -5.2 | $\underline{-37.6}$● | -17.7 | $\underline{-42.1}$● |
| iris | 4 | 3 | 150 | 1.00 | 100 | -89.7 | $\underline{-91.0}$– | -89.8 | -87.0– |
| isolet | 617 | 26 | 1559 | 1.02 | 96 | -27.8 | $\underline{-87.0}$● | -16.2 | $\underline{-89.1}$● |
| laryngeal3 | 16 | 3 | 353 | 4.11 | 94 | -17.9 | -8.3– | -9.8 | $\underline{-18.6}$– |
| letters | 16 | 26 | 20000 | 1.11 | 54 | -60.4 | $\underline{-75.6}$● | -64.5 | $\underline{-80.8}$● |
| liver | 6 | 2 | 345 | 1.38 | 73 | -16.8 | -12.5– | -39.4 | -34.5– |
| madelon | 500 | 2 | 2600 | 1.00 | 93 | -13.1 | $\underline{-15.4}$– | -13.4 | $\underline{-14.9}$– |
| magic_telescope | 10 | 2 | 19020 | 1.84 | 72 | -30.1 | -27.4– | -24.2 | -21.1– |
| multiple_features | 649 | 10 | 2000 | 1.00 | 100 | -4.5 | $\underline{-64.5}$● | -4.6 | $\underline{-76.0}$● |
| OCR_digits | 64 | 10 | 5620 | 1.03 | 78 | -0.9 | $\underline{-12.7}$● | -4.7 | $\underline{-5.9}$– |
| page_blocks | 10 | 5 | 5473 | 175.46 | 100 | -15.1 | -11.8– | -24.6 | -9.9– |
| pendigits | 16 | 10 | 10992 | 1.08 | 66 | -49.5 | $\underline{-87.8}$● | -33.8 | $\underline{-87.3}$● |
| pima | 8 | 2 | 768 | 1.87 | 91 | -22.8 | $\underline{-28.9}$– | -11.7 | $\underline{-20.1}$– |
| robot | 24 | 4 | 5456 | 6.72 | 50 | -34.8 | -34.0– | -54.1 | -53.4– |
| satimage | 36 | 6 | 6435 | 2.45 | 94 | -84.3 | $\underline{-91.2}$● | -87.7 | $\underline{-94.7}$● |
| scrapie | 14 | 2 | 3113 | 4.86 | 46 | -0.4 | 8.2– | -0.0 | $\underline{-4.8}$– |
| shuttle | 9 | 7 | 58000 | 4558.60 | 99 | -34.0 | -8.1○ | -22.2 | $\underline{-35.3}$– |
| sonar | 60 | 2 | 208 | 1.14 | 85 | -23.9 | $\underline{-25.9}$– | -43.1 | $\underline{-47.2}$● |
| soybean_large | 35 | 15 | 266 | 4.00 | 73 | -4.8 | $\underline{-12.0}$– | -4.4 | $\underline{-30.6}$● |
| voice_9 | 10 | 9 | 428 | 16.43 | 90 | -40.1 | -39.9– | -26.1 | $\underline{-34.4}$– |
| wine_quality | 11 | 7 | 4898 | 439.60 | 97 | -63.3 | $\underline{-67.4}$● | -48.9 | $\underline{-51.2}$– |
| yeast | 8 | 10 | 1484 | 92.60 | 43 | 3.2 | $\underline{-32.1}$● | -5.5 | $\underline{-39.4}$● |

tures to 0 for a period of time. We found that many data sets benefit significantly from using the PCA.

The main purpose of this paper was proof of concept. Other feature extraction methods, classifiers, detectors, type of changes, non-i.i.d. data streams, etc., should be examined to gain more insight into the potential of feature extraction for change detection in streaming data.

## References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] A. Bifet and R. Gavaldà. Learning from time-changing data with adaptive windowing. In *Proc. 7th SIAM Int Conf on Data Mining*, pages 443 – 448, Minneapolis, Minnesota, USA, 2007.

[3] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi. An information-theoretic approach to detecting changes in multi-dimensional data streams. In *Proc 38th Symposium on the Interface of Statistics, Computing Science, and Applications*, Pasadena, CA, 2006.

[4] B. S. Everitt. and G, Dunn. *Applied Multivariate Data Analysis*. Arnold, London, 2 edition, 2001.

[5] M. Gaber and P. Yu. Classification of changes in evolving data streams using online clustering result deviation. In *Proc 3rd Int Workshop on Knowledge Discovery in Data Streams*, Pittsburgh PA, USA, 2006.

[6] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In Proc 17th Brazilian Symposium on AI, volume LNCS 3171, pages 286–295. Springer Verlag, 2004.

[7] H. Hotelling. The generalization of Student's ratio. *Annals of Mathematical Statistics*, 2(3):360–378, 1931.

[8] L. I. Kuncheva. Change detection in streaming multivariate data using likelihood detectors. *IEEE Tran on Knowledge and Data Engineering*, (to appear), 2011.

[9] K. Nishida and K. Yamauchi. Detecting concept drift using statistical testing. In *Proc 10th Int Conf on Discovery Science*, pages 264–269, 2007.

[10] M. R. Reynolds Jr and Z. G. Stoumbos. The SPRT chart for monitoring a proportion. *IIE Transactions*, 30:545–561, 1998.

[11] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand. Exponentially weighted moving average charts for detecting concept drift. *Pattern Recognition Letters*, 33:191–198, 2012.

[12] X. Song, M. Wu, C. Jermaine, and S. Ranka. Statistical change detection for multi-dimensional data. In *Proc 13th ACM SIGKDD*, pages 667–676, California, USA, 2007.

[13] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23:69–101, 1996.

[14] I. Zliobaite, A. Bifet, G. Holmes, and B. Pfahringer. MOA concept drift active learning strategies for streaming data. *JMLR: Workshop and Conference Proceedings*, volume 17, pages 48–55, 2011.