PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data

Ludmila I. Kuncheva, *Member, IEEE*, and William J. Faithfull

***Abstract*— When classifiers are deployed in real-world applications, it is assumed that the distribution of the incoming data matches the distribution of the data used to train the classifier. This assumption is often incorrect, which necessitates some form of change detection or adaptive classification. While there has been a lot of work on change detection based on the classification error monitored over the course of the operation of the classifier, finding changes in multidimensional unlabeled data is still a challenge. Here, we propose to apply principal component analysis (PCA) for feature extraction prior to the change detection. Supported by a theoretical example, we argue that the components with the lowest variance should be retained as the extracted features because they are more likely to be affected by a change. We chose a recently proposed semiparametric log-likelihood change detection criterion that is sensitive to changes in both mean and variance of the multidimensional distribution. An experiment with 35 datasets and an illustration with a simple video segmentation demonstrate the advantage of using extracted features compared to raw data. Further analysis shows that feature extraction through PCA is beneficial, specifically for data with multiple balanced classes.**

***Index Terms*— Change detection, feature extraction, log-likelihood detector, pattern recognition.**

I. Introduction

ADAPTIVE classification in the presence of concept drift is one of the main challenges of modern machine learning and data mining [1], [2].[1] The increasing interest in this field reflects the variety of application areas, including engineering, finance, medicine, and computing. Monitoring a single variable such as the classification error rate has been thoroughly studied [3]–[10]. The most notable application is engineering, where control charts have been used for process quality control [4]. Classical examples of control charts are Shewhart's method, CUmulative SUM (CUSUM), and Wald's sequential probability ratio test (SPRT) [5], [11], [12]. One of the main assets of the univariate change detection methods is their statistical soundness. Advanced as they are, these methods cannot handle directly multidimensional data with concept drift.

In many applications, the class labels of the incoming data are not readily available, and thus the error rate cannot serve as a performance gauge. An indirect performance indicator would be a change in the distribution of the unlabeled multidimensional data. Typically, a change detector relies on comparing two distributions, one estimated from the old data, and one from the new data. In addition to defining the criterion, a strategy for finding the exact change point must be put in place. There is a wealth of literature on such strategies, for example, choosing, sampling, splitting, growing, and shrinking a pair of sliding windows [8], [9], [13]–[18]. In this paper, we propose a new approach to formulating a change detection criterion, which can be used with any such strategy.

There are at least three caveats in choosing or designing a criterion for change detection from multidimensional unlabeled data. First, change detection is an ill-posed problem, especially in high-dimensional spaces. The concept of change is highly context-dependent. How much of a difference, and in what feature space, constitutes a change? 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 color distribution is monitored, such a change will be left unregistered. The second caveat is that not all substantial changes of the distribution of the unlabeled 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. Fig. 1 shows three examples of substantial distribution changes that do not affect the error rate of the classifier built on the original data. Conversely, classification error may increase with an adverse change in the class labels, without any manifestation of this change in the distribution of the unlabeled data. An example scenario is change of user interest preferences on a volume of articles. Fig. 2 illustrates a label change which will corrupt the classifier but will not be picked up by a detector operating on the unlabeled data.

Finally, change detection depends on the window size. Small windows would be more sensitive to change compared to large windows.

To account for the uncertainties and lack of a clear-cut definition, we make the following starting assumptions: 1) changes that are likely to affect adversely the performance of the classifier are detectable from the unlabeled data; 2) changes of the distribution of the unlabeled data are reasonably correlated with the classification error; and 3) the window sizes for the old and the new distributions are specified.

Given the context-dependent nature of concept change, feature extraction can be beneficial for detecting changes. For example, extracting edge information from frames in a video stream can improve the detection of scene change [19]. A more general approach to change detection in

Manuscript received May 15, 2012; revised January 31, 2012; accepted February 12, 2013. Date of publication March 13, 2013; date of current version December 13, 2013.

The authors are with the School of Computer Science, Bangor University, Bangor LL57 1UT, U.K. (e-mail: l.i.kuncheva@bangor.ac.uk; w.faithfull@bangor.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2013.2248094

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

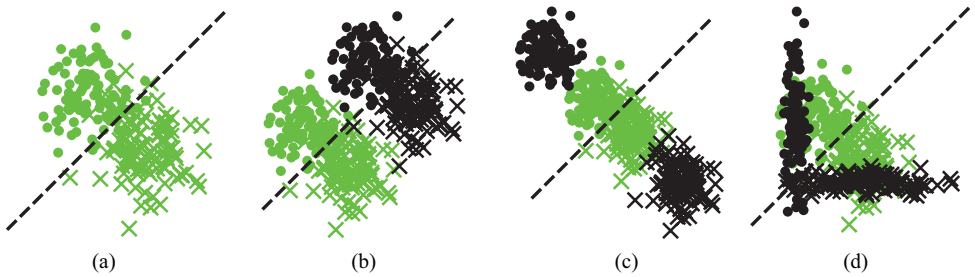

Fig. 1. Example of three changes (plotted in black) that lead to the same optimal classification boundary as the original data (dashed line). (a) Original. (b) Change 1. (c) Change 2. (d) Change 3.

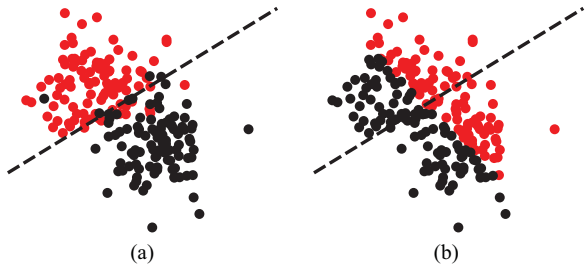

Fig. 2. Example of a change in classification accuracy with no change in the unlabeled pdf. (a) Before change. (b) After change.

multivariate time series is identifying and removing stationary subspaces [20].

In the absence of a bespoke heuristic, we propose that principal component analysis (PCA) can be used as a general method for feature extraction to improve change detection from multidimensional unlabeled incoming data. The theoretical grounds of our approach are detailed in Section II. Section III describes the criterion for the change detection. Section IV contains the experiment with 35 datasets, and Section V gives an illustration of change detection with feature extraction for a simple video segmentation task.

II. Feature Extraction for Change Detection

Fig. 3 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 signaled. 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 incoming data distribution matches the training one. The two scenarios share a distribution modeling block in the diagram. The modeling is sometimes implicit, and is included in the calculation of the change detection criterion. Compared to the multidimensional case, approximating distributions in the 1-D case can be much more accurate and useful. This explains the greater interest in the 1-D case. Methods such as hidden Markov models, Gaussian mixture modeling, Parzen windows, kernel-based approximation, and martingales have been proposed for this task. The most common approach to the multidimensional case is clustering [21] followed by monitoring of the clusters' characteristics over time. Nikovski and Jain [22] base their two detection methods on the average distance between all pairs of observations, one from the old window and one from the new window. Song *et al.* [23] propose a kernel estimation, and Dasu *et al.* [24] consider approximation via kdq-trees. A straightforward solution from statistics is to treat the two windows as two groups and apply Hotelling's t^2 test to check whether the means of the two groups are the same [25] or the multirank test for equal medians [26]. The output of the data modeling block, which can also be labeled "criterion evaluation," is a value that is compared with a threshold to declare change or no change.

A. Rationale

We propose to include a feature extraction block (highlighted in the diagram). Distribution modeling of multidimensional raw data is often difficult. Intuitively, extracting features that are meant to capture and represent the distribution in a lower dimensional space may simplify this task.

PCA is routinely used for preprocessing of multispectral remote sensing images for the purposes of change detection [27]. The concept of change, however, is different from the interpretation we use here. In remote sensing, change is understood as the process of identifying differences in the state of an object in space by observing it at different times, for example, a vegetable canopy.

If there is no knowledge of what the change may be, it is not clear whether the representation in a lower dimensional space will help. Our hypothesis is that, if the change is blind to the data distribution and class labels, the principal components with a smaller variance will be more indicative compared to the components with larger variance. This means that, contrary to standard practice, the components that should be retained and used for change detection are not the most important ones but the least important ones. Such blind change could be, for example, equipment failure, where signal is replaced by random noise, or signals bleeding into one another.

By leaving the most important principal components aside, we are not necessarily neglecting important classification information. PCA does not take into account class labels, and therefore less relevant components may still have high discriminatory value.

Therefore we propose to use the components of lowest variance for detecting a change between data windows W_1 and W_2.

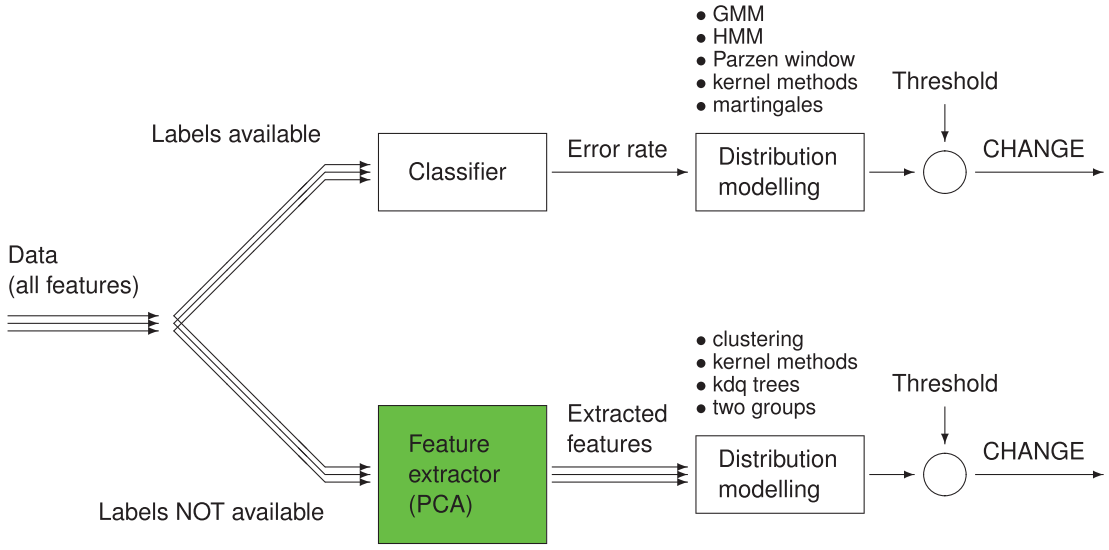

Fig. 3. Feature extraction for change detection.

B. Theoretical Example

PCA merely rotates the coordinate system in $\Re^n$ so that the axes are orientated along directions with progressively decreasing variance of the data. Consider a 2-D Gaussian dataset already rotated through PCA. Let x_1 and x_2 be the principal components. The mean of the distribution is $\mathbf{0} = [0, 0]^T$ and the covariance matrix is

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \tag{1}$$

where $\sigma_1 > \sigma_2 > 0$. Consider a change in the original space which leads to a new Gaussian distribution. We will examine the projection of the change on each of the PC axes to show that the second component (x_2) is more sensitive to blind changes than the first component (x_1).

We chose one of the most widely used distance measures between distributions, namely the Bhattacharyya distance. Let $p(y)$ and $q(y)$ be probability distributions of the random variable y. Assuming, without loss of generality, that p and q are continuous, the Bhattacharyya distance between the two distributions is

$$D_B(p, q) = -\ln \int \sqrt{p(y)q(y)}dy. \tag{2}$$

Denote by o_1 and o_2 the original distributions

$$\begin{aligned} o_1 &\equiv x_1 \sim N(0, \sigma_1^2) \\ o_2 &\equiv x_2 \sim N(0, \sigma_2^2) \end{aligned} \tag{3}$$

and by c_1 and c_2 the respective marginal distributions after the change.

The following propositions demonstrate that the second principal component is more sensitive than the first one to three standard types of changes: translation, rotation, and change in the variance. To show this, we prove that the Bhattacharyya distance between the old and the new distribution is always larger for the second component

$$D_B(o_1, c_1) < D_B(o_2, c_2). \tag{4}$$

Lemma: For univariate normal distributions, $p \equiv y \sim N(m_p, \sigma_p^2)$ and $q \equiv y \sim N(m_q, \sigma_q^2)$

$$D_B(p, q) = -\frac{1}{2}\ln\left(\frac{2\sigma_p\sigma_q}{\sigma_p^2 + \sigma_q^2}\right) + \frac{(m_p - m_q)^2}{4(\sigma_p^2 + \sigma_q^2)}. \tag{5}$$

Proof: The result is arrived at by substituting the expressions for the normal distributions in (2) followed by standard algebraic manipulations.

Proposition 1: Let the change be a translation of the mean of the original distribution to $\Delta = [\Delta_1, \Delta_2]^T$, where Δ is a random variable following a radially symmetric distribution centered at $(0, 0)$. Then the following holds:

$$\mathbb{E}_\Delta[\text{sign}\,\{D_B(o_1, c_1) - D_B(o_2, c_2)\}] < 0 \tag{6}$$

where $\mathbb{E}_\Delta$ is the expectation across Δ.

Proof: A translation will change the means but not the variances of the projected distributions c_1 and c_2. From the lemma

$$D_B(o_1, c_1) = -\frac{1}{2}\ln\left(\frac{2\sigma_1\sigma_1}{\sigma_1^2 + \sigma_1^2}\right) + \frac{(0 - \Delta_1)^2}{4(\sigma_1^2 + \sigma_1^2)} \tag{7}$$

$$= \frac{\Delta_1^2}{8\sigma_1^2}. \tag{8}$$

Similarly

$$D_B(o_2, c_2) = \frac{\Delta_2^2}{8\sigma_2^2}. \tag{9}$$

Form the difference

$$\begin{aligned} D_B(o_1, c_1) - D_B(o_2, c_2) &= \frac{\Delta_1^2}{8\sigma_1^2} - \frac{\Delta_2^2}{8\sigma_2^2} \\ &= \frac{1}{8}\left(\frac{\Delta_1}{\sigma_1} - \frac{\Delta_2}{\sigma_2}\right)\left(\frac{\Delta_1}{\sigma_1} + \frac{\Delta_2}{\sigma_2}\right). \end{aligned} \tag{10}$$

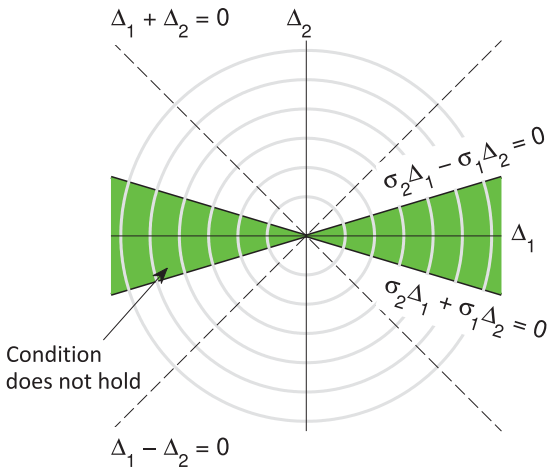

Fig. 4. Regions determined by inequalities (11) and (12). The second principal component is more sensitive than the first if the translation moves the mean of the data to any point in the unshaded area.

For this difference to be negative, one of the following must hold

$$\sigma_2\Delta_1 - \sigma_1\Delta_2 < 0 \text{ and } \sigma_2\Delta_1 + \sigma_1\Delta_2 > 0 \tag{11}$$

or

$$\sigma_2\Delta_1 - \sigma_1\Delta_2 > 0 \text{ and } \sigma_2\Delta_1 + \sigma_1\Delta_2 < 0. \tag{12}$$

The diagram in Fig. 4 illustrates the regions in the (Δ_1, Δ_2) space. The shaded region contains the translation points where the first principal component is more sensitive than the second component. Shown in the plot are the two bisecting diagonals where $\Delta_1 = \Delta_2$ and $\Delta_1 = -\Delta_2$. The shaded region will always occupy less than half of the space because the slopes of the bounding lines are $\sigma_2/\sigma_1 < 1$ and $-\sigma_2/\sigma_1 > -1$.

Let $p(\Delta)$ be a radially symmetrical distribution centered at the origin, which governs the translation coordinates Δ_1 and Δ_2. Denote the shaded region by R_+, the nonshaded region by R_-, and the borders by $R_=$. Then

$$\text{sign}\{D_B(o_1, c_1) - D_B(o_2, c_2)\} = \begin{cases} 1, & \text{if } \Delta \in R_+ \\ -1, & \text{if } \Delta \in R_- \\ 0, & \text{if } \Delta \in R_=. \end{cases} \tag{13}$$

Then

$$\mathbb{E}_\Delta[\text{sign}\{D_B(o_1, c_1) - D_B(o_2, c_2)\}] = \int \text{sign}\{D_B(o_1, c_1) - D_B(o_2, c_2)\} p(\Delta) d\Delta \tag{14}$$

$$= -\int_{R_-} p(\Delta) d\Delta + \int_{R_+} p(\Delta) d\Delta. \tag{15}$$

Because of the radial symmetry of p, the integrals will be proportional to the angles of the respective regions. The angle between the Δ_1 axis and line $\sigma_2\Delta_1 + \sigma_2\Delta_2 = 0$ is

$$\alpha = \arctan\frac{\sigma_2}{\sigma_1}. \tag{16}$$

Since $\frac{\sigma_2}{\sigma_1} < 1$, $\alpha < \frac{\pi}{4}$. Then

$$\mathbb{E}_\Delta[\text{sign}\{D_B(o_1, c_1) - D_B(o_2, c_2)\}] = \frac{4}{\pi}\arctan\frac{\sigma_2}{\sigma_1} - 1 < 0. \tag{17}$$

■

Proposition 2: Inequality (4) holds for a rotation transformation for any rotation angle θ.

Proof: The distribution after the change will be centered at $(0, 0)$ and rotated at θ. It will be a normal distribution with covariance matrix $\Sigma_c = R\Sigma R^T$, where R is the rotation matrix

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{18}$$

Then

$$\Sigma_c = R\Sigma R^T = \begin{bmatrix} \cos^2(\theta)\sigma_1^2 + \sin^2(\theta)\sigma_2^2, & \sin(\theta)\cos(\theta)(\sigma_1^2 - \sigma_2^2) \\ \sin(\theta)\cos(\theta)(\sigma_1^2 - \sigma_2^2), & \sin^2(\theta)\sigma_1^2 + \cos^2(\theta)\sigma_2^2 \end{bmatrix}. \tag{19}$$

The diagonal elements of Σ_c are the respective variances of the changed distributions c_1 and c_2. From the lemma, taking into account that the second term is 0

$$D_B(o_1, c_1) = -\frac{1}{2}\ln\left(\frac{2\sqrt{\cos^2(\theta) + \sin^2(\theta)\frac{\sigma_2^2}{\sigma_1^2}}}{1 + \cos^2(\theta) + \sin^2(\theta)\frac{\sigma_2^2}{\sigma_1^2}}\right). \tag{20}$$

Similarly

$$D_B(o_2, c_2) = -\frac{1}{2}\ln\left(\frac{2\sqrt{\cos^2(\theta) + \sin^2(\theta)\frac{\sigma_1^2}{\sigma_2^2}}}{1 + \cos^2(\theta) + \sin^2(\theta)\frac{\sigma_1^2}{\sigma_2^2}}\right). \tag{21}$$

Noticing that the expressions are the same, apart from the inversed ratio of the two original variances σ_1^2 and σ_2^2, we can form the difference $D_B(o_1, c_1) - D_B(o_2, c_2)$ and prove that it is always negative. Let $t = \sigma_2^2/\sigma_1^2$ $(0 < t < 1)$ and $a = \sin^2(\theta)$ $(\cos^2(\theta) = 1 - a)$. Then

$$D_B(o_1, c_1) - D_B(o_2, c_2) = -\frac{1}{2}\ln\left(\frac{2\sqrt{1 - a + at}}{2 - a + at}\right) + \frac{1}{2}\ln\left(\frac{2\sqrt{1 - a + \frac{a}{t}}}{2 - a + \frac{a}{t}}\right) \underbrace{-\frac{1}{2}\ln\left(\sqrt{\frac{1 - a + at}{1 - a + \frac{a}{t}}}\,\frac{(2 - a + \frac{a}{t})}{(2 - a + at)}\right)}_{A}. \tag{22}$$

For the difference to be negative, the argument of the logarithm, A, must be greater than 1. Manipulating the inequality $A > 1$ leads to

$$a(1 - a)\frac{(1 + t)(1 - t)^3}{t^2} > 0. \tag{23}$$

Since $0 < t < 1$, the inequality always holds; hence (4) holds, too, for any rotation angle θ. ■

Proposition 3: Inequality (4) holds for a transformation whereby the variances of both components change by the same amount.

Proof: Let a be a constant, $a > -\min\{\sigma_1, \sigma_2\}$, so that the variances of the two components after the change are respectively $(\sigma_1 + a)^2$ and $(\sigma_2 + a)^2$. From the lemma

$$D_B(o_1, c_1) = -\frac{1}{2}\ln\left(\frac{2\sigma_1(\sigma_1 + a)}{\sigma_1^2 + (\sigma_1 + a)^2}\right) \quad (24)$$

$$D_B(o_2, c_2) = -\frac{1}{2}\ln\left(\frac{2\sigma_2(\sigma_2 + a)}{\sigma_2^2 + (\sigma_2 + a)^2}\right). \quad (25)$$

To show that $D_B(o_1, c_1) < D_B(o_2, c_2)$, it is sufficient to show that the following function is monotonically decreasing with respect to its argument x

$$f(x) = -\frac{1}{2}\ln\left(\frac{2x(x_1 + a)}{x^2 + (x + a)^2}\right).$$

The first derivative of $f(x)$ is

$$\frac{\partial f}{\partial x} = -\frac{a^2(2x + a)}{2x(x + a)(x^2 + (a + x)^2)}.$$

By definition $x > 0$ and $a > -x$. Therefore the derivative is negative, which completes the proof. ∎

III. Choosing the Change Detection Criterion

Here we detail a recently proposed semi-parametric log-likelihood criterion (SPLL) for change detection [28], and argue our choice by comparing it with three criteria used in multidimensional change detection: Hotelling test, multirank [26] and Kulback–Leibler (K-L) distance [24].

A. Semi-Parametric Log-Likelihood Change Detector (SPLL)

SPLL comes as a special case of a log-likelihood framework, and is modified to ensure computational simplicity. Suppose that the data before the change come from a Gaussian mixture $p_1(\mathbf{x})$ with c components with the same covariance matrix. The parameters of the mixture are estimated from the first window of data W_1. The change detection criterion is derived using an upper bound of the log-likelihood of the data in the second window, W_2. The criterion is calculated as

$$\text{SPLL}(W_1, W_2) = \frac{1}{M_2}\sum_{\mathbf{x}\in W_2}(\mathbf{x} - \mu_{i*})^T \Sigma^{-1}(\mathbf{x} - \mu_{i*}) \quad (26)$$

where M_2 is the number of objects in W_2, and

$$i* = \arg\min_{i=1}^{c}\left\{(\mathbf{x} - \mu_i)^T \Sigma^{-1}(\mathbf{x} - \mu_i)\right\} \quad (27)$$

is the index of the component with the smallest squared Mahalanobis distance between $\mathbf{x}$ and its center. If the assumptions for p_1 are met, and if W_2 comes from p_1, the squared Mahalanobis distances have a chi-square distribution with n degrees of freedom (where n is the dimensionality of the feature space) [29]. The expected value is n and the standard deviation is $\sqrt{2n}$. If W_2 does not come from the same distribution, then the mean of the distances will deviate from n. Too large or too small a value will indicate a change.

Here we propose to fold the criterion to make it monotonic. This can be done by estimating p_1 from W_1 and assessing the fit of the data from W_2, and then swap the two windows and calculate the criterion again. Thus the final value of SPLL will be

$$\text{SPLL} = \max\{\text{SPLL}(W_1, W_2), \text{SPLL}(W_2, W_1)\}. \quad (28)$$

Given two data windows W_1 and W_2, the SPLL statistic is calculated as follows: 1) cluster the data in W_1 into K clusters using the c-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 squared Mahalanobis distance to each cluster center using S^{-1}, and calculate SPLL(W_1, W_2) as the average of the minimum distances; 4) swap windows W_1 and W_2 and follow the same steps to find SPLL(W_2, W_1); and 5) take forward the maximum of the two values as in (28).

In practice, the SPLL assumptions are rarely met, which makes it difficult to set up a threshold or determine a confidence interval. This difficulty is not uncommon for change detection criteria in general. Bootstrap Monte Carlo sampling and permutation tests have been suggested for estimating a suitable threshold [6], [23], [24]. Here we are interested in the raw values of the criteria and will leave the problem for selecting a threshold for future studies.

B. Comparison With Hotelling, Multirank, and K–L

We have found that SPLL statistic compares favorably for detecting changes with its main competitor, the Hotelling t^2 test [28]. The reason behind this finding is that a Gaussian mixture is usually a more reasonable model than the single Gaussian assumed for the Hotelling test. The Hotelling criterion will not be able to detect change in the variance of the data, while the SPLL criterion is equipped to do so. The same holds for the nonparametric version of this test based on multidimensional ranking. The Multirank test [26] compares the medians of the distributions in the two windows but again leaves aside changes in the variance.

To support our criterion choice, we include here a simulation example. One hundred points were sampled as window W_1 from a 5-D normal distribution with mean $\mathbf{0}$ and a diagonal covariance matrix S. The variances of the features were sampled from the positive half of the standard normal distribution. Denote this distribution by P_1. Window W_2 was sampled once from P_1 (with the same covariance matrix) and then according to three types of changes. Fig. 5 shows scatterplots of the two windows in the space of the first two features.

1) Translation. A new mean was sampled from $2y$, where $y \sim N(0, 1)$. W_2 was sampled anew from P_1 and the new mean was added [Fig. 5(b)].
2) Random linear transformation. A random matrix R of size 5×5 was generated, where each element was sampled from $N(0, 1)$. Window W_2 was sampled from P_1 and all objects were multiplied by R [Fig. 5(c)].
3) Change of variance. W_2 was sampled from a normal distribution with mean $\mathbf{0}$ and covariance matrix $S \times D$, where D is a diagonal matrix with diagonal elements sampled from $3|y|$, where $y \sim N(0, 1)$ [Fig. 5(d)].

The procedure of generating W_1 and four versions of W_2 was repeated 100 times. Four change detection criteria were

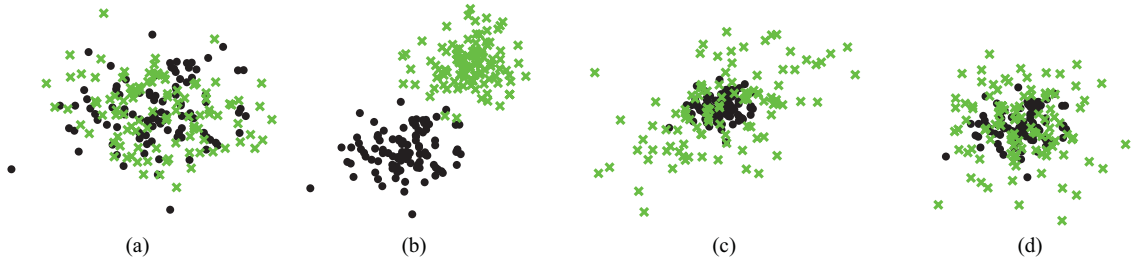

Fig. 5. Example of windows W_1 (black) and W_2 (green) for comparing the change detection criteria. (a) Same distribution. (b) Translation. (c) Random linear transformation. (d) Change of variance.

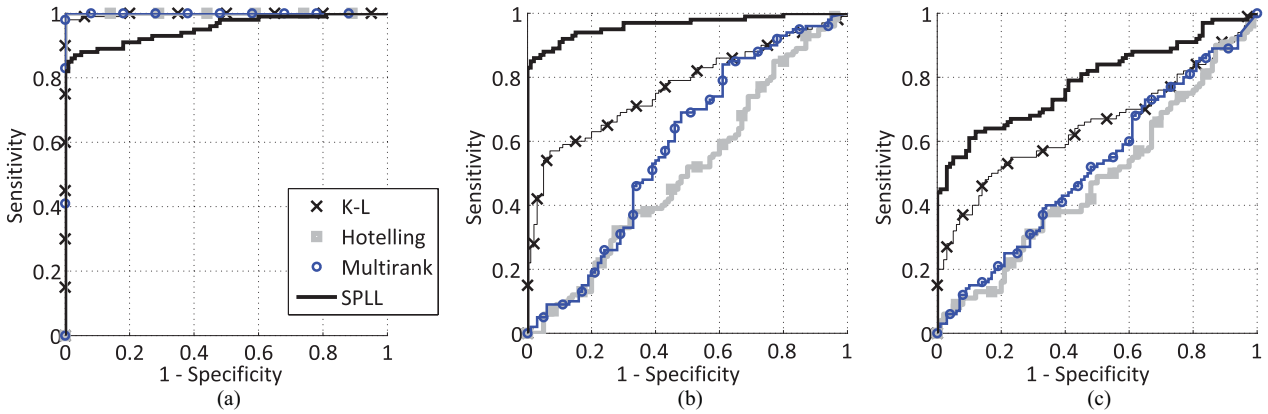

Fig. 6. ROC curves for the four criteria and the three types of change. (a) Translation. (b) Random linear transformation. (c) Change of variance.

calculated: K-L distance, the Hotelling's t^2, multirank [26], and SPLL. Both K-L and SPLL were used with three clusters. Note that no thresholds were applied, as we were evaluating the raw criteria values. The receiver operating characteristic (ROC) curves were constructed for each criterion and each change type. Fig. 6 shows the curves for the three changes. The graphs illustrate the behavior of the four criteria. While for the mean change the two bespoke criteria (Hotelling and Multirank) are superior to K-L and SPLL, the two latter changes favor SPLL. This is why we take SPLL for the experiment reported in the next section. We note that the choice of the criterion is not crucial for supporting our hypothesis that change detection will be aided by preserving the low-variance principal components.

IV. Experiment

A. Preliminaries

Our aim is to compare SPLL with and without PCA in order to demonstrate the benefit of feature extraction.

1) Acid Test: It is difficult to find an acid test for change detection in unlabeled multidimensional data. Here we chose two change heuristics which could be regarded as instances of equipment failure.

a) Shuffle values: A random integer k, $1 \leq k \leq n$, was generated to determine how many features out of n will be affected. k random features were chosen, and the values of each feature were randomly permuted within window W_2.

b) Shuffle features: Again, a random integer k, $1 \leq k \leq n$, was generated to determine how many features will be affected. k random features were chosen, and their columns were randomly permuted within window W_2.

The shuffle values change resembles a case where a group of sensors stop working as a result of a technical fault and produce random readings within the sensor ranges. The shuffle features change can be likened to bleeding of signals into one another. We previously experimented with setting a number of features to zero or infinity, but that seemed to be too easy a change to detect.

2) Change Detection is Context-Specific: We should also bear in mind that identifying changes is the first step in a process. The concept of change depends on what we will be using the result for. There could be, for example, a scenario where a change in the mean of the distribution is irrelevant, and only a change in the variance should be flagged. The magnitude of change is also context-dependent. How big a change should be accepted as worthy of triggering an alarm?

Therefore, here we do not offer a change detector as such. We investigate the ability of a criterion (SPLL and PCA+SPLL) to respond to changes. Setting up a threshold for this criterion is a separate problem. Such a threshold may be data-specific, and can be tuned to the desired level of false positives versus true positives.

3) Indirect Detection for Classification: In the context of classification, there may be a problem-specific threshold on the classification error that should not be exceeded. Any changes of the distributions of the classes that do not lead to increased error can be perceived as insignificant.

As we argued in Section I, not all changes in the unconditional pdf will lead to change in the classification error. Thus

a genuine change detected through the criterion may fail to correlate with the classification error. On the other hand, classification error may suffer with no change in the distribution of the unlabeled data. Even though such a correlation is an indirect quality measure, we include it here because of the importance of classification performance measure.

B. Experimental Protocol

The experiment was run on 35 datasets listed alphabetically in Table I, with differing numbers of instances, features, and classes. The sets were sourced from UCI [30] and a private collection. All datasets were standardized prior to the experiments.

1) Experiment 1: In the first experiment, we examined the difference between change detection on raw data and PCA data. For the PCA feature extraction, we varied the proportion of dismissed variance as $K = \{0\%$(keep all components), 50%, 80%, 85%, 90%, and 95%$\}$. The following procedure was applied 50 times to each dataset.

1) Take a stratified random sample of size M as window W_1.
2) Run PCA on W_1 and keep the components beyond the $K\%$ of dismissed variance. For example, consider $K = 90\%$ and a 4-D dataset, whose eigenvalues are $\{12, 8, 5, 2\}$. Taking the cumulative sum and dividing by the sum of the eigenvalues, the cumulative explained variance (in %) is $\{44, 74, 93, 100\}$. The first three components explain 93% of the variance in the data. We dismiss these components and keep only the last component which explains the remaining 7% of the variability of the data. Denote the PCA-transformed and clipped dataset as $W_{1,\mathrm{PCA}}$.
3) Repeat for $i = 1{:}100$.
   a) Take a random sample of M instances from the remaining data as the i.i.d. window W_2. Calculate SPLL for windows W_1 and W_2 as in (28) and store the criterion value in $b(i)$.
   b) Transform W_2 in the PC space using the eigenvectors of the retained components. Call this set $W_{2,PCA}$. Calculate SPLL for windows $W_{1,PCA}$ and $W_{2,PCA}$ as in (28) and store the result in $c(i)$.
   c) Apply a change (described above: value shuffle or feature shuffle) to W_2 to obtain a new set called W_2'. Calculate SPLL for windows W_1 and W_2' as in (28) and store the result in $b'(i)$.
   d) Transform W_2' in the PC space using the eigenvectors of the retained components. Call this set $W_{2,PCA}'$. Calculate SPLL for windows $W_{1,PCA}$ and $W_{2,PCA}'$ as in (28) and store the result in $c'(i)$.
4) Concatenate the values SPLL for the cases with and without a change, to obtain $B = [b, b']$ and $C = [c, c']$. Calculate the ROC curves from B and C and the areas under the curves (AUCs). If our hypothesis is correct, the AUC for B will be smaller than the AUC for C.

2) Experiment 2: The purpose of the second experiment was to find out how the SPLL change statistic correlates with the classification accuracy with and without PCA.[2] Larger values of SPLL signify a change in the distribution, which is likely to result in lower classification accuracy. Therefore we hypothesize that SPLL in the selected PCA space results in a stronger negative correlation compared to SPLL calculated from the raw data. The following procedure was applied 50 times to each dataset.

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.[3]
2) Run PCA on W_1 and keep the components beyond the $K = 95\%$ of explained variance. Denote the PCA-transformed and clipped dataset as $W_{1,PCA}$.
3) Repeat for $i = 1{:}100$.
   a) Take a random sample of M instances from the remaining data as the i.i.d. window W_2. Calculate the classification accuracy of the SVM trained on W_1, say $a(i)$. Calculate SPLL for windows W_1 and W_2 as in (28) and store the result in $b(i)$.
   b) Transform W_2 in the PC space using the eigenvectors of the retained components. Call this set $W_{2,PCA}$. Calculate SPLL for windows $W_{1,PCA}$ and $W_{2,PCA}$ as in (28) and store the result in $c(i)$.
   c) Apply a change (described above) to W_2 to obtain a new set called W_2'. Calculate the classification accuracy of the SVM trained on W_1 and store in $a'(i)$. Calculate SPLL for windows W_1 and W_2' as in (28) and store the result in $b'(i)$.
   d) Transform W_2' in the PC space using the eigenvectors of the retained components. Call this set $W_{2,PCA}'$. Calculate SPLL for windows $W_{1,PCA}$ and $W_{2,PCA}'$ as in (28) and store the result in $c'(i)$.
4) Concatenate the accuracies and the SPLL for the cases with and without a change, to obtain $A = [a, a']$, $B = [b, b']$ and $C = [c, c']$. Calculate and store the correlation between A and B, and A and C. If our hypothesis is correct, A (accuracy) and C (SPLL from PCA-transformed data) will have a stronger negative correlation than A and B (SPLL from raw data).

The window size M is a parameter of the algorithm; we used $M = 50$. By carrying out 50 runs of this procedure for each data set, 50 correlation coefficients are obtained.

C. Results

1) Experiment 1: Fig. 7 shows the mean difference AUC(PCA)−AUC(raw) across the 35 datasets as a function of the percentage of dismissed variance K. The differences are positive if the low-variance components are retained. Using the 35 datasets, we carried out a paired two-tailed t-test between AUC(raw) and AUC(PCA, K), for the six values of K. The test was applied only for values of K for which the

[2]We used the support vector machine (SVM) classifier from the MATLAB bioinformatics toolbox.

[3]For multiple classes, we applied SVM to all pairs of classes and labeled the data point to the class with the most votes.

TABLE I
RESULTS FROM THE EXPERIMENTS WITH TWO TYPES OF CHANGE

							Shuffle Values		Shuffle Features	
Name	N	n	c	$P_{\max}$	$P_{\min}$	#PCA	ρ_{raw}	ρ_{PCA}	ρ_{raw}	ρ_{PCA}
Breast	277	9	2	0.708	0.292	2.28	−0.2983	−0.3451•	−0.1696	−0.2841•
Contrac	1473	9	3	0.427	0.226	2.18	−0.2544	−0.3320•	−0.1844	−0.2983•
Contractions	98	27	2	0.500	0.500	16.38	−0.8262	−0.8169○	−0.6719	−0.6811•
Ecoli	336	7	8	0.426	0.006	2.94	−0.5667	−0.7546•	−0.6066	−0.6161–
German	1000	24	2	0.700	0.300	8.20	−0.1395	−0.3500•	−0.0918	−0.3330•
Glass	214	9	6	0.355	0.042	4.32	−0.4585	−0.6713•	−0.3134	−0.5876•
Image	2310	19	7	0.143	0.143	12.58	−0.6516	−0.8294•	−0.3206	−0.6878•
Intubation	302	17	2	0.500	0.500	6.00	−0.5045	−0.6702•	−0.3571	−0.6016•
Ionosphere	351	34	2	0.641	0.359	21.64	−0.6755	−0.7811•	−0.3253	−0.5368•
Laryngeal1	213	16	2	0.620	0.380	9.02	−0.6387	−0.6791•	−0.4225	−0.5262•
Laryngeal2	692	16	2	0.923	0.077	9.02	−0.4272	−0.5304•	−0.2845	−0.4525•
Laryngeal3	353	16	3	0.618	0.150	9.38	−0.5976	−0.6728•	−0.3683	−0.5140•
Lenses	24	4	3	0.625	0.167	1.00	0.2319	0.2586–	0.2524	0.1843•
Letters	20000	16	26	0.041	0.037	6.22	−0.7074	−0.8155•	−0.5456	−0.7715•
Liver	345	6	2	0.580	0.420	1.98	−0.3360	−0.3856•	−0.1154	−0.2779•
Lymph	148	18	4	0.453	0.014	5.48	−0.2127	−0.2466•	−0.0597	−0.2015•
Pendigits	10992	16	10	0.104	0.096	8.12	−0.9156	−0.9436•	−0.8133	−0.8996•
Phoneme	5404	5	2	0.707	0.293	1.02	−0.3219	−0.3285–	−0.1969	−0.1443○
Pima	768	8	2	0.651	0.349	2.02	−0.3230	−0.4637•	−0.0855	−0.2192•
Rds	85	17	2	0.529	0.471	6.06	−0.8013	−0.8302•	−0.6035	−0.6954•
Satimage	6435	36	6	0.238	0.097	31.98	−0.9285	−0.9012○	−0.5080	−0.6296•
Scrapie	3113	14	2	0.829	0.171	4.10	−0.0832	−0.0999–	−0.0438	−0.3151•
Shuttle	58000	9	7	0.786	0.000	6.94	0.0709	−0.4929•	0.2515	−0.4491•
Sonar	208	60	2	0.534	0.466	40.42	−0.6630	−0.7119•	−0.4413	−0.5570•
Soybean_large	266	35	15	0.150	0.038	17.64	−0.7492	−0.9187•	−0.5760	−0.8726•
Spam	4601	57	2	0.606	0.394	37.34	−0.0492	−0.1566•	−0.0074	−0.1130•
Spect_continuous	349	44	2	0.728	0.272	28.14	−0.3655	−0.4721•	0.0682	−0.2115•
Thyroid	215	5	3	0.698	0.140	1.98	−0.6682	−0.6517–	−0.4921	−0.6281•
Vehicle	846	18	4	0.258	0.235	12.94	−0.7721	−0.8396•	−0.4387	−0.7444•
Voice_3	238	10	3	0.706	0.076	4.20	−0.6433	−0.6895•	−0.4300	−0.5481•
Voice_9	428	10	9	0.269	0.016	4.00	−0.5985	−0.6552•	−0.4132	−0.5356•
Votes	232	16	2	0.534	0.466	6.06	−0.8193	−0.7874○	−0.6825	−0.6254○
Vowel	990	11	10	0.091	0.091	3.54	−0.7907	−0.8654•	−0.6813	−0.7560•
Wbc	569	30	2	0.627	0.373	22.84	−0.7728	−0.7707–	−0.1849	−0.4653•
Wine	178	13	3	0.399	0.270	4.98	−0.8970	−0.8933–	−0.7403	−0.8029•

Jarque–Bera hypothesis test indicated normality of the pairwise differences of the AUC. For the remaining values of K, we used the Wilcoxon signed rank test for zero median of the differences. The circled points correspond to statistically significant differences. Thresholds $K = 90\%$ and $K = 95\%$ lead to significantly better change detection than raw data. Interestingly, using all principal components ($K = 0\%$) leads to significantly worse AUC compared to detection from raw data. One possible explanation for this finding is that PCA fools the clustering algorithm so that the (anyway rough) approximation of the pdf as a mixture of Gaussians becomes inadequate.

The points where the AUC for the PCA data is significantly better than AUC with raw data are enclosed in circles. The points where PCA loses to raw data are enclosed in grey squares.

Fig. 8 shows a scatterplot of the 35 datasets in the space of AUC(raw) and AUC (PCA, $K = 95\%$) for the two types of changes. The reference diagonal for which the PCA extraction does not make any difference is also plotted. It can be seen that most points are above the diagonal, demonstrating the improved change detection capability of the PCA features.

2) Experiment 2: Table I shows the correlation coefficients averaged across 50 runs for each dataset. The correlation coefficient between the classification accuracy and SPLL calculated from the raw data is denoted by ρ_{raw}, and the one for the features extracted through PCA by ρ_{PCA}. Using the 50 replicas of the experiment, we carried out a paired two-tailed t-test for the datasets for which the Jarque–Bera hypothesis test indicated normality of the pairwise differences of the correlation coefficients. For the remaining datasets, we used the Wilcoxon signed rank test for zero median of the differences. Statistically significant differences ($\alpha = 0.05$) are marked in Table I with • if PCA was better, and with ○ if the raw data detection was better. Shown in the table are also the prevalences of the largest and the smallest classes in the data ($P_{\max}$ and $P_{\min}$) estimated from the whole dataset. The column labeled "# PCA" contains the percentage of retained principal components.

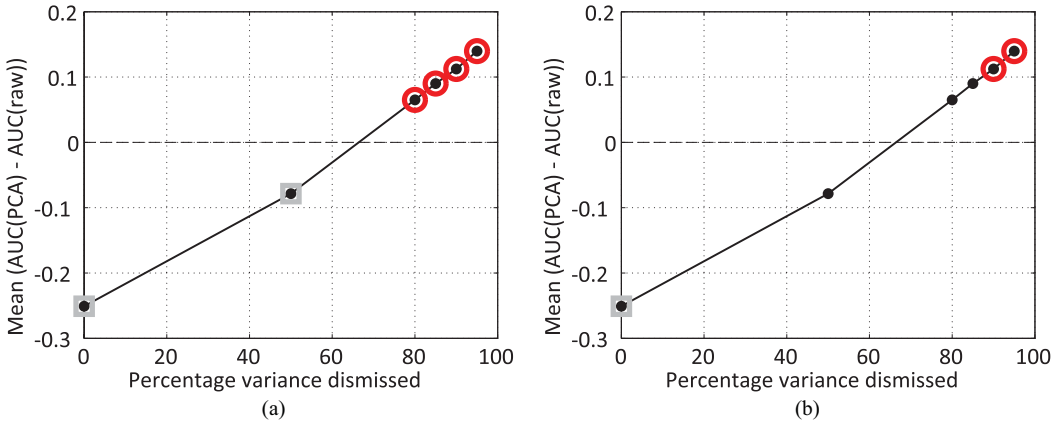

Fig. 7. Average difference AUC(PCA)−AUC(raw). (a) Change value shuffle. (b) Change feature shuffle.

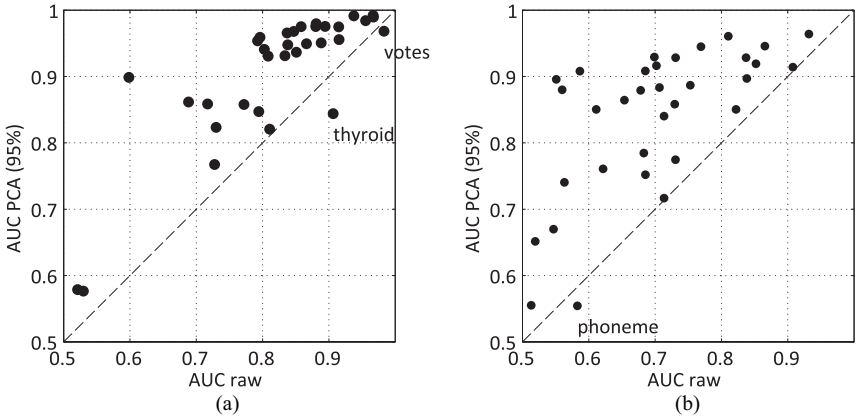

Fig. 8. Scatterplot of the 35 datasets in the space of AUC(raw) and AUC(PCA, $K = 95\%$). (a) Change value shuffle. (b) Change feature shuffle.

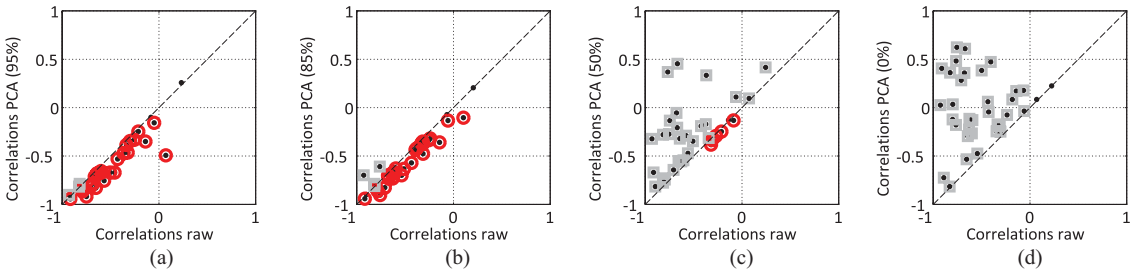

Fig. 9. Shuffle values. Scatterplot of the 35 datasets in the space space (ρ_{raw}, ρ_{PCA}). (a) $K = 95\%$. (b) $K = 85\%$. (c) $K = 50\%$. (d) $K = 0\%$.

Figs. 9 and 10 show scatterplots of the 35 datasets in the space (ρ_{raw}, ρ_{PCA}) for four values of K. The differences that were found to be statistically significant are marked with circles if favorable to PCA and with grey squares if favorable to the raw data.

The results demonstrate that feature extraction through PCA leads to markedly better change detection and therefore stronger correlation with the classification accuracy than using the raw unlabeled data.

D. Further Analyses

We carried out further analyses to establish which characteristics of the datasets may be related to the feature extraction success. Fig. 11 shows a scatter plot where each point corresponds to a dataset. 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 two-class problems, because the smallest and the largest priors sum up to 1. The number of classes increases from this edge toward the origin (0, 0). The left edge of the triangle corresponds to equiprobable classes. The largest prior on this edge is equal to the smallest prior, which means that all classes have the same prior probabilities. This edge can be thought of as the edge of balanced problems. The balance disappears toward the bottom right corner. The pinnacle of the triangle corresponds to two equiprobable classes. The size of the marker signifies the

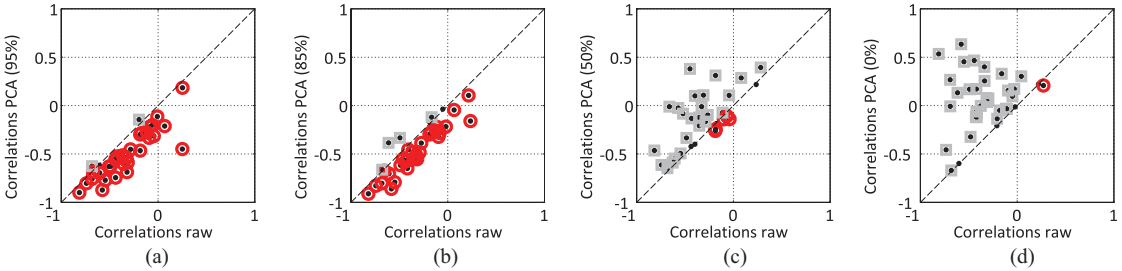

Fig. 10. Shuffle features. Scatterplot of the 35 datasets in the space space (ρ_{raw}, ρ_{PCA}). (a) $K = 95\%$. (b) $K = 85\%$. (c) $K = 50\%$. (d) $K = 0\%$.

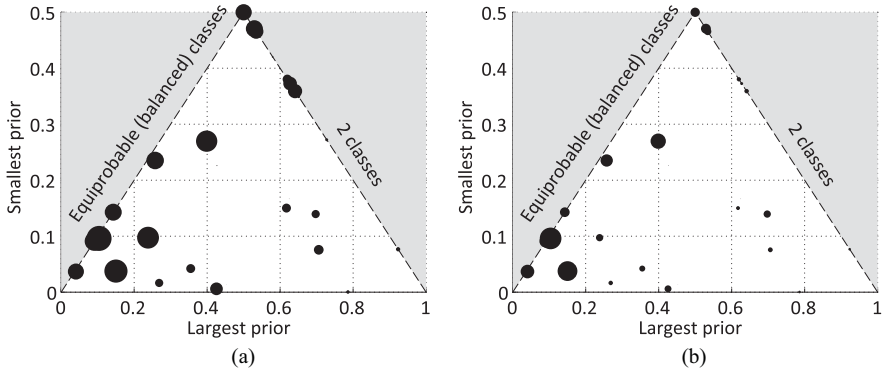

Fig. 11. Scatterplot of the 35 datasets in the space of the largest and smallest prior probabilities. The size of the marker signifies the strength of the correlation between SPLL with PCA and the classification accuracy. (a) Change value shuffle. (b) Change feature shuffle.

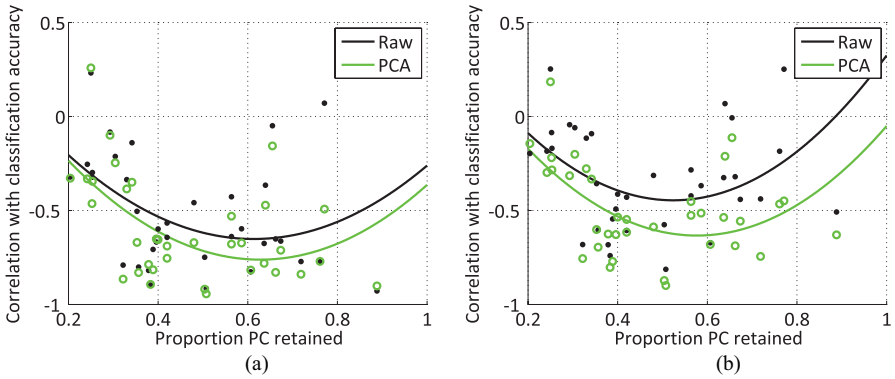

Fig. 12. Example of three changes (plotted in black) which lead to the same optimal classification boundary as the original data (dashed line). (a) Change Value shuffle. (b) Change Feature shuffle.

strength of the correlation between SPLL with PCA and the classification accuracy.

The figure suggests that the PCA has a stable and consistent behavior for multiclass fairly balanced datasets (bottom left of the scatterplot). For a smaller number of imbalanced classes (bottom right), the correlation ρ_{PCA} is not very strong. Our further analyses did not find interesting relationship patterns between the data characteristics and the correlations, except for the pronounced dip for both correlations ρ_{PCA} and ρ_{Raw} with respect to the number of retained principal components.

Fig. 12 shows the two correlations as functions of the proportion of retained principal components. The fit with the parabolas is not particularly tight but shows a tendency. For both heuristics, change detection is most related to the classification accuracy if about half of the principal components explain 95% of the variance; hence we retain the remaining half. As can be expected, the PCA curve lies beneath the curve for the raw data, demonstrating the advantage of feature extraction for change detection. The pattern, however, is similar for both correlation coefficients. It may be related to the type of changes and the way we induced them, but may also benefit from a data-related interpretation. Since we are interested in comparing feature extraction to raw data change detection, we relegate the further analysis of this pattern to future studies.

V. Simple Video Segmentation

We applied the change detection with and without PCA to a simple video segmentation problem. A short video clip of an office environment was produced, with small movements of the chairs and the posture of one of the assistants in

(a) (b) (c)

Fig. 13. Frames from the three parts of the video being segmented. (a) Beginning. (b) Middle. (c) End.

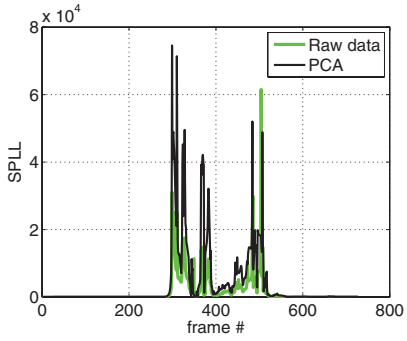

Fig. 14. SPLL criteria values for the video frames.

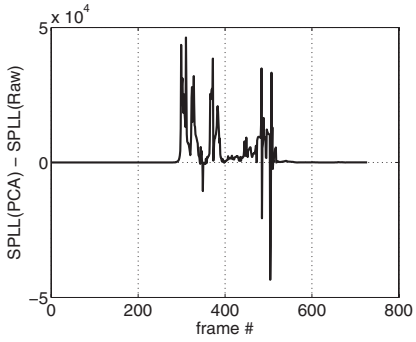

Fig. 15. Difference between the two SPLL criteria.

the office. The change was introduced in the middle part of the video by blocking the camera with the palm of a hand. The hand was made into a fist and opened again before removing it from view. Sample frames from the beginning, middle, and end part of the video are shown in Fig. 13.

For the purposes of showcasing the feature extraction, we were only interested in the admittedly easy detection of the change in the middle. The features that formed the online multidimensional stream were the red, green, and blue averages of each frame. We set W_1 to be the sequence of the first 50 frames, and took a sliding window of 25 frames as W_2. The PCA was applied to W_1 only. Fig. 14 plots the SPLL value with and without PCA across the frame sequence. Both criteria identify correctly the middle part with the change, but the values obtained through PCA are much larger. Fig. 15 depicts the difference between SPLL with PCA and without PCA. Again, the results favor the feature extraction approach to change detection.

VI. Conclusion

The lack of a rigorous methodology for feature extraction for the purposes of change detection in multidimensional unlabeled data has been noted in the literature. This paper offered a step in this direction. We argued that, after applying PCA, the components with the smaller variance should be kept because they are likely to be more sensitive to a general change.

This paper was concerned only with comparing two given windows of data. There are many more issues to be taken into account, such as non i.i.d data, window sizes, policy for signaling a change, establishing thresholds for the criteria involved, etc. Many of these issues need to be explored together with the feature extraction scenario proposed here. For example, it is interesting to find answers to the following questions.

1) How sensitive is PCA-based change detection to the sizes of windows W_1 and W_2?
2) Is there a "middle part" of principal components which are both relatively important and relatively sensitive to change?
3) How will the correlation coefficients behave for different classifier models?
4) How efficient will feature extraction be in detecting changes for very high-dimensional data such as functional magnetic resonance imaging (fMRI)? The exceptionally high feature redundancy in this case may require a different approach in terms of which components are retained.
5) How much computational complexity is added by the feature extraction step?

References

[1] I. Zliobaite, A. Bifet, G. Holmes, and B. Pfahringer, "MOA concept drift active learning strategies for streaming data," in *Proc. 2nd Workshop Appl. Pattern Anal.*, vol. 17. 2011, pp. 48–55.
[2] Q. Zhenzheng, W. Tao, Z. Zipeng, and G. Yuhai, "Study on the classification of data streams with concept drift," in *Proc. 8th Int. Conf. Fuzzy Syst. Knowl. Discovery*, vol. 3. Jul. 2011, pp. 1673–1677.
[3] G. J. Ross, N. M. Adams, D. K. Tasoulis, and D. J. Hand, "Exponentially weighted moving average charts for detecting concept drift," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 191–198, Jan. 2012.
[4] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[5] M. R. Reynolds and Z. G. Stoumbos, "The SPRT chart for monitoring a proportion," *Inst. Ind. Eng. Trans.*, vol. 30, no. 6, pp. 545–561, 1998.
[6] D. Kifer, S. Ben-David, and J. Gehrke, "Detecting change in data streams," in *Proc. 30th Int. Conf. Very Large Data Bases*, vol. 30. 2004, pp. 180–191.
[7] S.-S. Ho, "A martingale framework for concept change detection in time-varying data streams," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 321–327.
[8] A. Bifet and R. Gavaldà, "Learning from time-changing data with adaptive windowing," in *Proc. 7th SIAM Int. Conf. Data Mining*, 2007, pp. 443–448.
[9] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Proc. 17th Brazilian Symp. Artif. Intell.*, 2004, pp. 286–295.
[10] K. Nishida and K. Yamauchi, "Detecting concept drift using statistical testing," in *Proc. 10th Int. Conf. Discovery Sci.*, 2007, pp. 264–269.
[11] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, nos. 1–2, pp. 100–114, 1954.
[12] M. R. Reynolds and Z. G. Stoumbos, "A general approach to modeling CUSUM charts for a proportion," *Inst. Ind. Eng. Trans.*, vol. 32, no. 6, pp. 515–535, Jun. 2000.
[13] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," *Mach. Learn.*, vol. 23, no. 1, pp. 69–101, Apr. 1996.
[14] I. Koychev and R. Lothian, "Tracking drifting concepts by time window optimisation," in *Proc. 25th SGAI Int. Conf. Innovat. Tech. Appl. Artif. Intell.*, 2005, pp. 46–59.

[15] M. Baena-García, J. Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavaldà, and R. Morales-Bueno, "Early drift detection method," in *Proc. 4th Int. Workshop Knowl. Discovery Data Streams*, 2006, pp. 77–86.
[16] R. Klinkenberg and I. Renz, "Adaptive information filtering: Learning in the presence of concept drifts," in *Proc. Workshop Notes ICML/AAAI Workshop Learn. Text Categorizat.*, 1998, pp. 33–40.
[17] M. Lazarescu and S. Venkatesh, "Using selective memory to track concept drift effectively," in *Proc. Int. Conf. Intell. Syst. Control*, vol. 388. Jun. 2003, pp. 14–19.
[18] C. Alippi, G. Boracchi, and M. Roveri, "A hierarchical, nonparametric, sequential change-detection test," in *Proc. Int. Joint Conf. Neural Netw.*, 2011, pp. 2889–2896.
[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2000, pp. 556–562.
[20] D. A. J. Blythe, P. von Bunau, F. C. Meinecke, and K.-R. Muller, "Feature extraction for change-point detection using stationary subspace analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 4, pp. 631–643, Apr. 2012.
[21] M. Gaber and P. Yu, "Classification of changes in evolving data streams using online clustering result deviation," in *Proc. 3rd Int. Workshop Knowl. Discovery Data Streams*, 2006, pp. 1–12.
[22] D. D. Nikovski and A. Jain, "Fast adaptive algorithms for abrupt change detection," *Mach. Learn.*, vol. 79, no. 3, pp. 283–306, 2009.
[23] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Statistical change detection for multi-dimensional data," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 667–676.
[24] T. Dasu, S. Krishnan, S. Venkatasubramanian, and K. Yi, "An information-theoretic approach to detecting changes in multi-dimensional data streams," in *Proc. 38th Symp. Inter. Stat., Comput. Sci., Appl.*, 2006, pp. 1–24.
[25] H. Hotelling, "The generalization of student's ratio," *Ann. Math. Stat.*, vol. 2, no. 3, pp. 360–378, Nov. 1931.
[26] A. Lung-Yut-Fong, C. Lévy-Leduc, and O. Cappé, "Robust changepoint detection based on multivariate rank statistics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 3608–3611.
[27] A. Singh, "Review article: Digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.
[28] L. I. Kuncheva, "Change detection in streaming multivariate data using likelihood detectors," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 99, p. 1, Oct. 2011, DOI: 10.1109/TKDE.2011.226.
[29] B. S. Everitt, *A Handbook of Statistical Analyses Using S-Plus*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2001.
[30] A. Asuncion and D. Newman. (2007). *UCI Machine Learning Repository* [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

Ludmila Kuncheva (M'99) received the M.Sc. degree from the Technical University of Sofia, Sofia, Bulgaria, in 1982, and the Ph.D. degree from the Bulgarian Academy of Sciences, Sofia, in 1987.

She was with the Central Laboratory of Biomedical Engineering, Bulgarian Academy of Sciences, until 1997. She is currently a Professor with the School of Computer Science, Bangor University, Gwynedd, U.K. Her current research interests include pattern recognition and classification, machine learning and classifier ensembles. She has published two books and above 200 scientific papers.

William Faithfull graduated in computer science from Bangor University, Gwynedd, U.K., in 2011, where he is currently pursuing the Ph.D. degree with the School of Computer Science.

His current interests include pattern recognition, change detection in streaming data, video data processing and visual analytics.