An Effective Just-in-Time Adaptive Classifier for Gradual Concept Drifts

Cesare Alippi, Giacomo Boracchi and Manuel Roveri

***Abstract* –Classification systems designed to work in nonstationary conditions rely on the ability to track the monitored process by detecting possible changes and adapting their knowledge-base accordingly. Adaptive classifiers present in the literature are effective in handling abrupt concept drifts (i.e., sudden variations), but, unfortunately, they are not able to adapt to gradual concept drifts (i.e., smooth variations) as these are, in the best case, detected as a sequence of abrupt concept drifts. To address this issue we introduce a novel adaptive classifier that is able to track and adapt its knowledge base to gradual concept drifts (modeled as polynomial trends in the expectations of the conditional probability density functions of input samples), while maintaining its effectiveness in dealing with abrupt ones. Experimental results show that the proposed classifier provides high classification accuracy both on synthetically generated datasets and measurements from real sensors.**

I. Introduction

Real world industrial and environmental processes are prone to non-stationary phenomena induced by ageing effects, drifts, soft or hard faults inducing a change over time of the probability density function of acquired measurements [1]. As a consequence, classification systems built over these processes cannot be granted to work properly since the stationarity hypothesis assumed during the parameter configuration phase a priori does not hold any more.

The changes, or *concept drift*, might degrade the accuracy of the classification system up to a point that the expected quality of service of the envisaged application is impaired. As stated in [2], concept drifts can be grouped into two main families: *abrupt* and *gradual*. The former type refers to situations where changes can be modeled as step-like changes affecting the environment in which the classification system is deployed. The latter models situations where the process slowly evolves over time, for example, due to ageing effects or degradation of the sensors, e.g., due to temperature and humidity.

The need to deal with concept drifts [1], [2] has pushed the research toward the development of classification systems able to work in nonstationary environments by adapting their knowledge base (e.g., the training set, the parameters or the model family) to track the process evolution. In this direction, FLORA and FLORA2 [3] include additional supervised samples in stationary conditions, while remove a fixed percentage of the oldest training pairs from the knowledge base when a change is suspected (i.e., the accuracy of the classifier decreases below a user-defined threshold). Similarly, [4] suggests to adapt the knowledge base by weighting old samples according to their age or their relevance in terms of classification accuracy (computed on supervised samples). The classifier presented in [5] assesses variations in the classification accuracy to adapt to changes in the data generating process and treats concept drifts as sequences of stationary states

Multiple Classification Systems [6]-[12] rely on an ensemble of classifiers whose decisions are combined to form the final output (e.g., with voting or weighting mechanisms). These also exploit management techniques to add, remove, and reactivate working classification systems.

The work [1] introduces the Just-in-Time (JIT) adaptive classifier, which integrates a change-detection test to identify variations in the distribution of the data generating process and remove obsolete training samples. This approach allows the classifier for automatically improving the accuracy in stationary conditions (by introducing supervised samples during the operational life) and promptly reacting to changes in nonstationary ones. Finally, [13] extends [1] by proposing an adaptive weighted *k*-NN classifier providing a fine-grain adaptation to smooth drifts; [14] suggests a method for identifying a suitable training set to be considered after detecting a change.

It was demonstrated that JIT classifiers naturally, and effectively, address the abrupt concept drifts which imply a transition from a stationary state to a new one. Unfortunately, gradual concept drifts are seen as a sequence of nonstationarities, due to the resolution of the change-detection test. In turn, this behavior induces frequent removal of supervised samples from the knowledge base of the classifier.

We propose a novel JIT adaptive classifier for gradual concept drifts that extends [14] by introducing:

- a novel change-detection test that deals with processes whose expectation follow a polynomial trend, and that reveals a change when such trend varies, as well as when other statistical properties of the i process change (e.g., the variance of the detrended process).

Cesare Alippi, Giacomo Boracchi and Manuel Roveri are with Politecnico di Milano, Dipartimento di Elettronica e Informazione, Milano, Italy e-mail: cesare.alippi@polimi.it, giacomo.boracchi@polimi.it, manuel.roveri@polimi.it. This research has been funded by the European Commission's 7th Framework Program, under grant Agreement INSFO-ICT-270428 (iSense).

- a classifier that effectively handle gradual concept drifts affecting the process expectation, as it integrates an index estimating the process dynamics. This allows us for improving the accuracy of the classifier when observations follow such drifts, by exploiting both supervised and unsupervised samples. The obtained classifier outperforms other adaptive classifiers that model and treat instead the drift as a sequence of stationary states.

In addition, the suggested classification system relies on a knowledge-base management procedure that, in the case of detected variations, updates the knowledge-base to keep the classification system coherent with the current state of the process. Experiments show that the suggested classification system outperforms state of the art adaptive classifiers working in nonstationary environments by guaranteeing high detection accuracies both in case of gradual and abrupt concept drifts.

The paper is organized as follows: Section II introduces the problem statement. The JIT classification system for these gradually drifting data is presented in Section III. Section IV specifies the methodology by presenting the ICI-based change-detection test. Experimental results are finally given in Section V.

II. Problem Statement

For sake of simplicity we initially consider a two-class classification problem and mono-dimensional observations. Extensions to multi-class classification problems and multi-dimensional observations will be discussed later.

The operational framework can be formalized as follows. Let $X \in \mathbb{R}$ be the input sample and $Y \in \{\omega_1, \omega_2\}$ the associated binary classification output. The probability density function (pdf) of the inputs at time t

$$p(X|t) = p(\omega_1|t)p(X|\omega_1, t) + p(\omega_2|t)p(X|\omega_2, t)$$

depends on the pdfs of the outputs $\mathrm{p}(\omega_1|\mathrm{t})$ and $p(\omega_2|t) = 1 - p(\omega_1|t)$, and the conditional probability distributions $\mathrm{p}(\mathrm{X}|\omega_1, \mathrm{t})$ and $\mathrm{p}(X|\omega_2, \mathrm{t})$. We focus on gradual concept drifts that can be represented as a -possibly slow- time-varying stochastic process whose expectation $\mathrm{E}[\mathrm{p}(\mathrm{X}|\mathrm{t})]$ follows a piecewise polynomial function $f_\theta(t)$. The parametric description of $f_\theta(t)$ is given by $\{(\theta_i, U_i)\}$ where θ_i is a set of coefficients defining the polynomial $f_{\theta_i}(t)$ defined on the i^{th} time interval U_i (i.e., a subsequence of consecutive time instants); M is the number of intervals (e.g. see Fig. 1). The expectations of the conditional probability distributions can be expressed as

$$\begin{aligned} \mathrm{E}[p_i(X|\omega_1, t)] &= f_{\theta_i}(t) + \mathrm{q}_{1,i} \\ \mathrm{E}[p_i(X|\omega_2, t)] &= f_{\theta_i}(t) + \mathrm{q}_{2,i} \quad t \in I_i \end{aligned} \tag{1}$$

where $\mathrm{q}_{1,i}$ and $\mathrm{q}_{2,i}$ are the means of the two classes ω_1, ω_2 in the stationary condition. As a consequence, the process generating observation $X(t)$ at time t becomes:

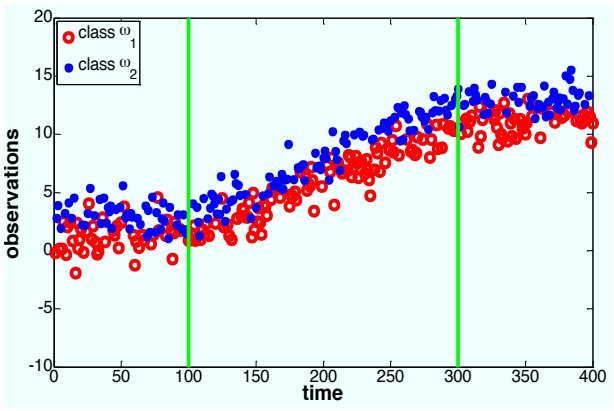

Fig. 1. An example of considered gradual concept drifts. The data sequence presents a transition ($U_2 = [100,299]$) between two stationary states ($U_1 = [1,99]$ and $U_3 = [300,400]$) and is characterized by a linear trend of the expected values for both classes. The green lines delimitate the samples affected by the concept drift, i.e., time instants $t = 100$ and $t = 300$. While the JIT adaptive classifiers proposed in [1], [13], [14] treat the gradual concept drift as a sequence of stationary states and, as such, would detect a continuous sequence of changes in stationarity between $t = 100$ and $t = 300$. The proposed classifier, by estimating an index associated with the dynamics of the process under monitoring, detects only two changes in this dataset at $t = 100$ and $t = 300$.

$$X(t) = \begin{cases} f_{\theta_i}(t) + \phi_{1,i}, & y(t) = \omega_1 \\ f_{\theta_i}(t) + \phi_{2,i}, & y(t) = \omega_2 \end{cases} \tag{2}$$

where $\phi_{1,i}$ and $\phi_{2,i}$ are the pdfs characterizing the distributions of their respective classes ω_1, ω_2 in stationary conditions with $\mathrm{E}[\phi_{1,i}] = \mathrm{q}_{1,i}$ and $\mathrm{E}[\phi_{2,i}] = \mathrm{q}_{2,i}$.

We further assume that the probabilities $p_i(X|\omega_1, t)$ and $p_i(X|\omega_2, t)$ do not change within each interval defining the piecewise polynomial function, thus, the pdf of $X(t)$ is

$$p_i(X|t) = p_i(\omega_1)p(X|\omega_1, t) + p_i(\omega_2)p(X|\omega_2, t), \ t \in I_i.$$

The pdf of the inputs, the conditional distributions and the output distributions are *unknown*. The piecewise-polynomial function within each interval U_i, i.e., $f_{\theta_i}(t)$, is also unknown, but common between the two classes, as expressed in (2). We emphasize that the considered framework is an extension of the traditional one that assumes

$$f_{\theta_i}(t) = const, \qquad i = 1, \dots, M.$$

Fig. 1 shows an example where observations have been generated by a gradual drifting process corresponding to a smooth transition between two stationary states.

While a multi-class classification problem can be easily accommodated by the *k*-NN classifier considered Section III.D, handling multi-dimensional observations is more critical due to the need to consider multivariate change-detection tests. However, to a first approximation, this can be addressed by considering each dimension independently [15].

III. JIT Adaptive Classifiers for Gradual Concept Drifts

A. The General Approach

The key point of the proposed approach is to extend observation model traditionally assumed in classification problems, by allowing the expectation of the conditional probability density functions to evolve over time as a piecewise polynomial function, as expressed in (1). Under such a hypothesis, we develop a change-detection test to assess variations in the (polynomial) trend of the process under monitoring, rather than in the value of its expectation. If the test does not detect variations, we perform a polynomial regression of the input samples and use the regression coefficients to modify on-line the knowledge base of an adaptive classifier. Differently, when a change is detected, the obsolete samples are removed from the knowledge base and the change-detection test is restarted.

Thus, the proposed classification system follows the JIT approach [1], [13], [14], which combines a change-detection test that identifies variations in the data generating process, and a knowledge-base management procedure that provide the classifier with the knowledge-base that correctly interpret the current state of the process. JIT classifiers allow for promptly detecting variations in the pdf of X, and react consequently by removing obsolete training samples from the knowledge base. Other JIT classifiers (e.g., [1], [13], [14]) well accommodate for abrupt changes but are not optimal for handling gradual concept drifts, which would be seen as sequences of stationary states hence inducing the JIT classifier to continuously detect changes and reset its knowledge-base accordingly.

To overcome such a shortcoming we introduce a novel JIT adaptive classifier, which combines a change-detection test able to deal with changes in stationary processes and concept drifts inducing polynomial trends in the processes expectation, and an adaptive *k*-NN classifier able to operate in these conditions, compensating such gradual concept drifts.

As presented in Algorithm 1, the proposed approach exploits polynomial regression to estimate, for each input (both supervised and unsupervised) sample, how the expectation of the data generating process varies over time (line 4), thus estimating $f_{\theta_i}(t)$, the deterministic additive component that represents the drift in (2). This allows the classifier (line 9) for properly updating the training samples during the classification phase according to the time instant in which each training sample has been acquired.

When the proposed change-detection test, which analyzes both supervised and unsupervised samples (line 5), detects a change in the process distribution or a change in the expectation trend, both the test and the classifier are re-configured (lines 6, 7). On the contrary, when no change is detected, the classification system improves the regression estimates by exploiting both unsupervised samples to be classified and supervised samples. Moreover, as it happens in stationary conditions, supervised samples available during the operational life are integrated in the knowledge-base of the classifier (line 8) to improve its classification accuracy.

Algorithm 1: General JIT Adaptive Classifier for Gradual Concept Drifts

1. Configure the classifier and the change detection test;
2. **while** (1){
3. New observation arrives;
4. Estimate the process expectation by polynomial regression;
5. **if** (change-detection test detects a change in the process distribution or a change in the expectation trend) {
6. Characterize the new process state;
7. Configure the classifier and the test on the new process state; }
8. **else** integrate the new information (if available) in the knowledge base;
9. Classify the input sample by exploiting the output of the polynomial regression;
10. }

It is worth noting that the proposed approach represents an extension of the traditional JIT classifier since, in stationary conditions or in case of abrupt changes, the classification system behaves as [1], [13], [14], while it differs only when the process undergoes the considered gradual concept drifts. In fact, an abrupt change represents a particular case where the functions $f_{\theta_i}(t)$ in (2) are constants. Note also that when concept drift induces non-polynomial trends, the test would reveal a sequence of nonstationarities.

We now detail the three main components of the proposed approach: the change-detection test, polynomial regression method and the adaptive classifier.

B. The Change Detection Test for Gradual Concept Drifts

The literature about change-detection tests is very well-established e.g., see [17], [18], and in the classical formulation, most of these tests aim at assessing stationarity of the data-generating process. However, ad-hoc techniques aim at detecting variations in observations generated according to linear models [18], [19]. Here we exploit the ICI-based change-detection test (ICI CDT) [20], which is natively able to deal with polynomial trends in the process under monitoring and provides high detection accuracy, promptness in detecting changes, and low computational complexity. Furthermore, it preserves good detection ability when reduced training sets are available and this makes the ICI CDT a particularly appealing candidate for the suggested JIT adaptive classifier working in gradual concept drifts (see Section IV for its use in a specific classification system). Nevertheless, other tests providing similar abilities could be considered as well.

C. Polynomial Regression

In principle, any regression technique can be used for fitting a polynomial to the observations, thus estimating the trend of the expectations $f_{\theta_i}(t)$ in (2). However, since the ICI-based change-detection test exploits least square regression, we use, at each time instant T a least square

estimator for computing the polynomial coefficients $\hat{\theta}(T)$ to be used in Algorithm 2, line 7.

D. Extended k-NN Classifier for Gradual Concept Drifts

As stated in [1], among the classification families present in the literature, *k*-NN classifiers [16] are the most suited for being embedded in JIT adaptive classification systems because they do not need a proper training phase and their knowledge-base can be easily managed. The proposed JIT adaptive classifier encloses a modified *k*-NN classifier that exploits polynomial estimates of the process under monitoring (which are obtained through a regression phase) to remove the deterministic additive component $f_{\theta_i}(t)$ in (2): in such a way, the classification system is able to handle, at each time instant t, both the observation $X(t)$ and the classifier' knowledge-base as if these were generated by the same process, thus considering only the terms $\phi_{1,i}$ and $\phi_{2,i}$ in (2).

In more detail, let $Z_T = \{(X(t), Y(t), t) | t \in I_T\}$ be the sequence of all the supervised couples $\big(X(t), Y(t)\big)$ available at time T, together with their acquisition time t: Y(t) is the classification label associated with the sample $X(t)$ acquired at time t, and I_T the set of arrival times. Let us set $m > 0$ the maximum polynomial order that is used to compensate for the gradual concept drift, and let $\hat{\theta}(T)$ be the set of coefficients of the polynomial function that provides the best fit of samples $\{X(t), t < T\}$, which are estimated with a polynomial regression on both supervised and unsupervised samples up to time instant T.

The suggested extended *k*-NN classifier for gradual concept drifts is presented in Algorithm 2. It is easy to see that the only difference w.r.t. the traditional *k*-NN classifier is the computation of the distance between the input sample and the training samples (line 4): here we correct each term in the traditional ℓ^2-norm with the value assumed by the fitted polynomial. In particular, the distance between the current sample $X(t)$ and the training sample $X(t_i)$ is computed after subtracting the values of the (estimated) polynomial having coefficients $\hat{\theta}$(T) in their corresponding time instants (i.e.$f_{\hat{\theta}(T)}(T)$ and $f_{\hat{\theta}(T)}(t_i)$). Note that $f_{\hat{\theta}(T)}(T)$ and $f_{\hat{\theta}(T)}(t_i)$ are indeed the estimates of the expectation of the data generating process at the current time instant *T,* and at t_i, when the i^{th} training sample has been received, respectively. This procedure allows the classifier for removing a polynomial trend from all the training samples that are hence brought back to a common expectation whose value is

$$p(\omega_1)\mathrm{q}_{1,i} + p(\omega_2)\mathrm{q}_{2,i}.$$

Then, the traditional *k*-NN classifier can be applied (line 6 and 7).

Algorithm 2: Extended *k*-NN Classifier for Gradual Concept Drift ($X(T)$, k, $\hat{\theta}$(T),Z_T)

1. $N = |Z_T|$;
2. $i = 1$;
3. **while** $(i < N)\{$
4. $d_i = \Big(\big(X(T) - f_{\hat{\theta}(T)}(T)\big) - \big(X(t_i) - f_{\hat{\theta}(T)}(t_i)\big)\Big)$;
5. $i = i + 1;\}$
6. Identify the nearest k training samples according to the distances $\{d_i\}_{i=1,..,N}$.
7. Classify $X(T)$ as the most represented class among the k nearest training samples.

IV. A Specific Solution: the ICI-based Classifier

This section presents the JIT adaptive classifier for gradual concept drifts, which combines the extended *k*-NN classifier of Section III.D, and an ICI-based change-detection test introduced in [20]. Process stationarity is thus monitored by means of the Intersection of Confidence Intervals (ICI) rule which embeds a polynomial fitting operator [21], [22]; thus, the ICI CDT is natively able to assess variations in the polynomial trend of the process expectation . Without loss of generality, we handle in the following a gradual concept drift by means of 1st order polynomials, i.e., we approximate the process expectation with a piecewise linear function (i.e., $m = 1$). Any gradual concept drift characterized by a 2nd or higher order polynomial would indeed result in a sequence of detections, as the regression model paired with the ICI rule cannot properly fit the observations. Details concerning the change-detection test are discussed in Section IV.A, while the JIT adaptive classifier is formulated in Algorithm 3 and detailed afterwards.

Let $Z_0 = \{(X(t), Y(t), t) | t \in I_0\}$ with $I_0 = \{1, \dots, T_0\}$ be the initial training set used for configuring both the change-detection test and the extended *k*-NN classifier. The training phase (lines 2-4) includes the estimation of the regression parameters of the process within the training set: since we use 1st order polynomials, the regression coefficients at time T_0 are denoted by $\hat{m}(T_0)$, $\hat{q}(T_0)$. The training samples are also used to compute the initial value of *k* in the extended *k*-NN classifier by means of leave-one-out procedure (LOO, line 4). Even during the training phase, the distances in the *k*-NN classifier are computed as expressed in Algorithm II (line 4), using the regression estimates (for the LOO procedure we use $\hat{m}(T_0)$, $\hat{q}(T_0)$).

After the initial training phase, the suggested classification system works on line by introducing, whenever available, additional supervised samples or classifying the input samples, otherwise. In particular, when new knowledge is available, this is inserted in the knowledge base of the classifier (lines 8 and 9), and the parameter *k* is updated according to Equation (3) of [1] (line 10). Then, the ICI CDT verifies possible occurrences of changes in the process w.r.t. the configuration phase (line 14). As detailed in Section IV.A, such variations are both changes in the process trend (i.e., changes in the piecewise polynomial function f_{θ_i} which rules the expectation of *X*), as well as changes in the variance of the de-trended process $X - f_{\theta_i}$.

The change-detection test works on sub-sequences of observations (line 14): whenever a change is detected in the subsequence containing the input data $X(t)$, the ICI-based

knowledge management procedure (presented in Algorithm 3 of [14]) is executed to identify the time instant T_{ref} in which the variation begun (line 15). This estimate is then used to reconfigure both the test from the observations arrived within $[T_{\text{ref}}, t]$ (line 16) and the classifier by removing the obsolete training samples, i.e., those acquired before T_{ref} (lines 17 and 18). The new value of k is then estimated from the new training set with LOO (line 20) by using the new regression parameters estimated from the new training set (line 19).

The classification phase (lines 21 and 22) consists of computing the regression parameters $\hat{m}(t)$, $\hat{q}(t)$ (line 21) from all the observations generated by the process in the current conditions (i.e., all the samples received since T_{ref}), and classifying $X(t)$ with the k-NN classifier described in Algorithm 2, by relying on the updated knowledge-base Z_t, the current value of k, and the regression coefficients $\hat{\theta}(t) = \{\hat{m}(t), \hat{q}(t)\}$ (line 22).

A. Details

The ICI CDT requires a feature-extraction phase, which is followed by the ICI rule for assessing the process stationarity by monitoring the feature values. This general approach can be customized by defining particular features to be employed, which determine the nature of detectable changes in X. Features have to provide values $z(t)$ that are Gaussian distributed as:

$$z(t) \sim \mathcal{N}(\mu(t), \sigma^2), \quad (3)$$

where $\mu(t)$ is its expectation, which is time-dependent, and σ indicates the feature standard deviation, which is indeed constant.

In particular, [17] details a solution for change-detection that exploits two features: the sample mean and the sample variance transformed according to a power-law, which guarantees the transformed values (i.e., the second feature) to satisfy the (3). In what follows we discuss how to modify this test to cope with observations distributed as in (2): in fact the original test has been devised for solving a *classic* change-detection problem, where observations, in stationary conditions, are i.i.d.

The first feature, the sample mean computed on disjoint subsequences of observations, has a distribution that approaches to (3), even when the expectation of the observations follows a polynomial trend as in (2). The test has to be slightly modified w.r.t. the one presented in [17], since a 1st order polynomial function has to fit values of the sample mean: thus, the ICI rule determines the largest neighborhood where $\mu(t)$ can be considered as linear. Note that the regression coefficients obtained from the sample mean are indeed estimates of the regression coefficients on the observations, and can be rightly used instead of $\hat{m}(t)$, $\hat{q}(t)$. It follows that the change-detection test provides the regression estimates required by the adaptive classifier, hence reducing the processing and memory requirements.

The transformed sample variance is not Gaussian distributed when observations are distributed as in (2). In fact, Gaussian distribution holds solely when observations are i.i.d. Therefore, in order to compute the sample variance (and the coefficient of the power-law transform) we perform a preliminary de-trend of the observations. De-trending is accomplished by convolving the observations with a high-pass filter having coefficients [-1 , 1], which removes the linear component in the observations, followed by a downsampling. Further details concerning the detection test can be found in [17], Section IV.

Algorithm 3: ICI-based JIT Adaptive Classifier

1. $I_0 = \{1, \dots, T_0\}, Z_0 = \{(X(t), Y(t), t) | t \in I_0\}$;
2. configure the ICI change detection test using $\{X(t), t \in I_0\}$;
3. estimate $\hat{m}(T_0)$, $\hat{q}(T_0)$the regression coefficients from $\{X(t), t < T_0\}$;
4. estimate k with the extended k-NN by means of LOO on Z_0, using $\hat{m}(T_0)$, and $\hat{q}(T_0)$;
5. $t = T_0 + 1, T_{\text{ref}} = 1$;
6. **while** (1) {
7. **if** (new knowledge on $X(t)$is available) {
8. $I_t = I_{t-1} \cup \{t\}$;
9. $Z_t = Z_{t-1} \cup \{(X(t), Y(t), t)\}$;
10. update k using Equation (3) of [1]}
11. **else** {
12. $I_t = I_{t-1}$;
13. $Z_t = Z_{t-1}$;}
14. **if** (ICI test (sub-sequence containing $X(t)$) detects a variation) {
15. Run the ICI-based knowledge-base management procedure (Algorithm 3 of [14]) to identify T_{ref};
16. Configure the ICI change detection test using the observations in $[T_{\text{ref}}, t]$;
17. Set $I_t = \{t \in I_t, t > T_{\text{ref}}\}$
18. Set $Z_t = \{(X(t), Y(t), t) | t \in I_t\}$;
19. Estimate the regression parameters $\hat{m}(t)$ and $\hat{q}(t)$ from $\{X(t), t > T_{\text{ref}}\}$.
20. Estimate k with the extended k-NN by means of LOO on Z_t using $\hat{m}(t)$ and $\hat{q}(t)$}.
21. Estimate $\hat{m}(t)$ and $\hat{q}(t)$ from $\{X(t), t > T_{\text{ref}}\}$.
22. Classify using the extended k-NN on Z_t, using $\hat{m}(t)$ and $\hat{q}(t)$.
23. $t = t + 1$;}

B. Comments

A peculiarity of the JIT classification systems is their ability to adapt the classifier to evolving processes, without need to inspect the classification performance. As such, any change that does not alter the distribution of X cannot be perceived. For example, the change-detection test embedded in the JIT classifier is not able to identify situations where two classes having $p(\omega_1) = p(\omega_2)$ swap their pdfs. Approaches that exploit the classification accuracy (e.g., [3]-[8]) are instead able to correctly deal with these situations, but require several supervised samples to effectively estimate the classifier performance.

Nevertheless, JIT approaches are preferable in certain circumstances as they do not rely on supervised samples to detect variations in the operating conditions. Furthermore, in case of the considered gradual concept drifts, when the two classes undergo a common trend, a straightforward analysis of the process trend (as in Algorithm 2), is beneficial for

classification performance, as shown in the experimental section.

Note, finally, that the proposed system can easily include, in the process monitoring (line 14 of Algorithm 3), additional change-detection tests that consider only supervised samples of a specific class each. These would allow the proposed system for reacting even to classes' swaps, which otherwise would not be detected.

V. Experiments

The performance of the proposed JIT adaptive classification system for gradual concept drift has been compared with those of JIT [1], JIT soft [13], and the ICI-based Adaptive Classifier [14] in the case of synthetically generated data (Application D1) and measurements coming from X-ray sensors (Application D2).

Application D1 contains three classification datasets each of which presents a different change in stationarity: *abrupt, drift,* and *transient*. Each dataset is composed of 200 sequences of 24000 real valued observations drawn from two equiprobable Gaussian distributed classes ω_1 and ω_2, that, in the initial stationary state, are distributed as $p(X|\omega_1) = N(0,3)$ and $p(X|\omega_2) = N(4,3)$. Each sequence of the *abrupt* dataset presents a change at sample 12000, which increases the mean of both classes by 15. In *drift*s sequences, the change starts at time 12000, increasing the means of both classes linearly, reaching +15 at the end of the sequence. Finally, each sequence of the *transient* dataset is characterized by a change occurring at 8000, which produces a linear trend increasing both classes' means of 15 at sample 16000. Fig. 2- 4 show sequences taken from each dataset.

Application D2 refers to a dataset composed of 250 sequences of measurements taken from couples of photodiodes. Each sequence is composed of 5500 16-bit measurements (2750 per sensor) that, after sample 2000, undergo a gradual concept drift. The distribution of the samples both before and after the change-point may however vary within the dataset: we have manually aligned the sequences to guarantee that the change points coincide. The considered classifiers have been used to classify the observations according to the sensor. An example of such a sequence is shown in Fig. 5.

In both applications *D1* and *D2*, the length of the initial training set is $T_0 = 500$ samples; after time T_0 we provide each classifier with 1 supervised observation out of 5 to update the knowledge base. We imposed a minimum size of 80 observations to the training sets adaptively identified by ICI-based knowledge-base management procedure; JIT soft has been configured with a minimum training set size of 80 observations for the classifier and 400 for the test (as required in [13]), while the JIT requires 400 observations both for the classifier and the test (as stated in [1]). The parameters of the ICI CDT and the ICI knowledge-base management procedure have been set as in [14], with $\Gamma = 2$, $\Gamma_{\text{ref}} = 3$ and $\lambda = 2$.

The classification performance is measured by the classification error at time *t*, averaged over the available experiments. Fig. 2-5 present these percentages averaged over a window of the 1000 previous values (for *D1*) and 50 (for *D2*) for the four considered classifiers.

A. Discussions

In stationary conditions (i.e., before the change), the classification error typically decreases, asymptotically approaches the Bayes one, thanks to the introduction of additional supervised samples. Thus, any detection (false positive) results in an unnecessary removal of new training samples, which may reduce the classifier accuracy. Here, the JIT soft shows the highest classification error due to the occurrence of false positives that reduce the training set size (and hence the classification accuracy). Classifiers enforcing the ICI CDT perform better, as this provides less false positives then the CI-CUSUM test (see [17]), which is used in JIT [1]and JIT soft.

The *abrupt* dataset (Fig. 2) shows the promptness of the considered classification systems in detecting occurred changes. When an abrupt change occurs, the classifier has to promptly remove the obsolete knowledge-base: thus, detection delays decrease the accuracy since the classification phase relies on an obsolete knowledge-base. The ICI-based classification systems provide similar performance, while the JIT shows the highest classification error because the training set identified after detecting the change has fixed size, and therefore could include training samples generated by the process in the previous state.

The *drift* dataset (Fig. 3) shows the effectiveness of the considered classification system in adapting to data undergoing a linear trend. As expected, the proposed classification system presents the lowest errors as it compensates the drift and, at the same time, does not detect further changes allowing to achieve the same classification accuracies of the previous stationary conditions. Instead, the ICI-based classifier enforcing a 0^{th} order polynomial approximation treats the drift as a sequence of abrupt changes: this motivates the performance gap between the two classifiers. Similarly, JIT and JIT soft classifiers suffer from the continuous detection of changes (hence inducing the continuous editing of the knowledge base) that induces an increasing classification error.

The *transient* case (Fig. 4) shows results coherent with the previous ones: the proposed classifier shows two *bumps* in the classification error plot in correspondence with the time instants where the process expectation changes its trend. At these time instants, the model assumed by the considered classification systems becomes obsolete and it is not able for explaining the current state of the process. This results in an immediate increase of the classification error, which decreases only when the system is able to explain the new operating conditions. In this case (as in the *drift* dataset), only the proposed classifier is able to achieve the classification accuracies provided in stationary conditions. When the data generating process returns in stationary

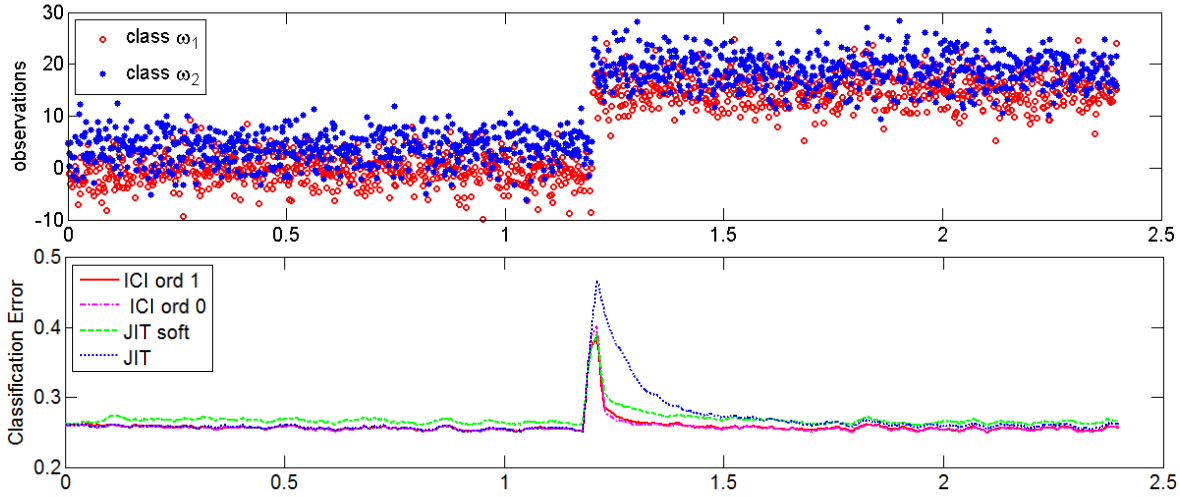

Fig. 2. Application *D1* - Abrupt: an example of sequence and classification error.

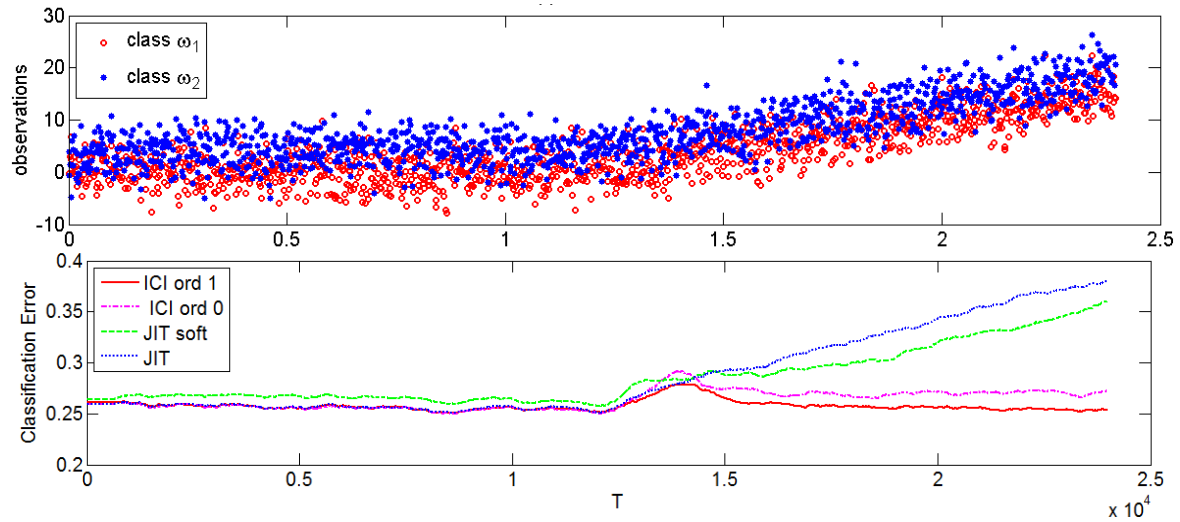

Fig. 3. Application *D1* - Drift: an example of sequence and classification error.

conditions a second change is perceived with a consequent increase in the classification error. Note that after the second change, the 0^{th} order ICI-based classification system provides best performance as the corresponding ICI CDT presents lower detection delays since during the drift it has continuously detected changes (emptying the knowledge base, without reach the performance of the proposed classifier).

Experimental results on X-ray sensor measurements (Fig. 5) are consistent with synthetically generated datasets of *D1*. Note that, since the drift here is nonlinear, the proposed classifier is not able to recover, during drift, the same performance as in stationary conditions. The process expectation in these (possibly non polynomial) drifts is approximated by a means of a piecewise linear function, and both the change-detection test and the classifier act accordingly. However, the proposed solution provides higher classification accuracy than the others.

VI. Conclusions

This paper presents a novel JIT adaptive classifier that, differently from traditional adaptive classifiers working in nonstationary environments, is able to work with gradual concept drifts by enforcing a change-detection test able to detect variations in the polynomial trend of the expectation of the data generating process, and a classifier that effectively deals with gradual concept drifts by integrating estimates of the drift dynamics. Experiments both on simulated and real data show the advantages of the proposed approach both in case of drift and abrupt changes.

References

[1] C. Alippi and M. Roveri, "Just-in-Time Adaptive Classifiers--Part II: Designing the classifier," in *IEEE Trans.on Neural Networks,* vol. 19, no. 11, pp. 2053-2064, December 2008

[2] A. Tsymbal, "Technical Report: The problem of concept drift: definitions and related work," Trinity College, Dublin, Ireland, TCD-CS-2004-15, 2004.

[3] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden contexts," in *Machine Learning*, vol. 23, no. 1, pp. 69-101,1996.

[4] R. Klinkenberg, "Learning drifting concepts: example selection vs. example weighting," in *Intelligent Data Analysis,* Special Issue on Incremental Learning Systems Capable of Dealing with Concept Drift, 8 (3), 2004

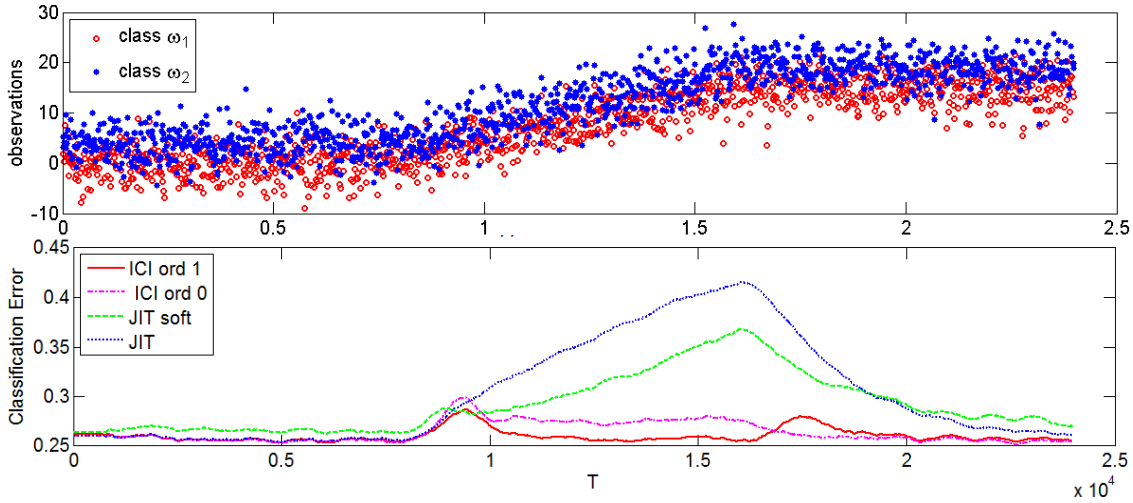

Fig. 4. Application *D1* - Transient: an example of sequence and the classification error.

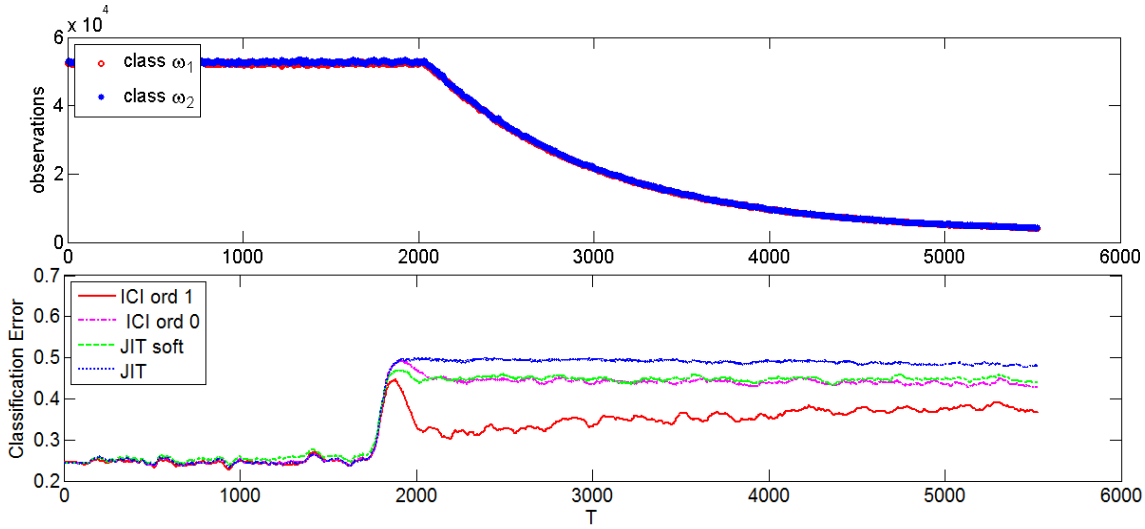

Fig. 5. Application *D2* - An example of sequence and the classification error.

[5] J. Gama and G. Castillo, "Learning with local drift detection," in *Advaced Data Mining and Applications*, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 42–55, 2006

[6] R. Elwell and R. Polikar. "Incremental learning in nonstationary environments with controlled forgetting," in *Proc IJCNN* 2009, pp.771-778, 14-19 June 2009

[7] K. Nishida and K.Yamauchi, "Learning, detecting, understanding, and predicting concept changes," *Proc IJCNN* 2009, pp. 2280-2287.

[8] W. N. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," *in Proc. of ACM SIGKDD Int. Conf. on Knowledge discovery and data mining,* 2001, pp.377-382.

[9] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proc. of ACM SIGKDD Int. Conf. on Knowledge discovery and data mining*, 2003, pp. 226-235.

[10] P. Wang, H. Wang, X. Wu, W. Wang, and B. Shi, "A Low-Granularity Classifier for Data Streams with Concept Drifts and Biased Class Distribution,"in IEEE *Trans. on Knowledge and Data Engineering*, vol.19 n.9, pp.1202-1213, 2007

[11] J. Z. Kolter and M. A. Maloof, "Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift," in *Proc. of 3rd IEEE Int. Conference on Data Mining*, p.123, 2003.

[12] J. Z. Kolter and M. A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts," in *The Journal of Machine Learning Research*, vol. 8, pp.2755-2790, 2007

[13] C. Alippi, G. Boracchi, and M. Roveri, "Just in time classifiers: Managing the slow drift case," in *Proc of IJCNN,* 2009, pp.114-120.

[14] C. Alippi, G. Boracchi, and M. Roveri, "Adaptive Classifiers with ICI-based Adaptive Knowledge Base Management," in *Proc of ICANN* 2010,Lecture Notes on Computer Science. Springer Berlin / Heidelberg, vol 6353, pp. 458-467.

[15] A. G. Tartakovsky, B. L. Rozovskii, R. B. Blazek, and H. Kim, "Detection of intrusions in information systems by sequential change-point methods", Statistical Methodology, Volume 3, Issue 3, July 2006, pp. 252-293.

[16] T. M. Cover and R. E. Hart, "Nearest neighbor pattern classification," in *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967

[17] T.L. Lai, "Sequential analysis: some classical problems and new challenges," in *Statistica Sinica*, vol. 11, n. 2, pp. 303-350, 2001.

[18] M. Basseville and I. V. Nikiforov *Detection of Abrupt Changes: Theory and Application*.Prentice-Hall,Englewood Cliffs,N.J, 1993.

[19] L. Horváth, M. Hušková, P. Kokoszka, and J. Steinebach, "Monitoring changes in linear models," in *Journal of Statistical Planning and Inference* vol. 126, pp. 225–251, 2004.

[20] C. Alippi, G. Boracchi, and M. Roveri, "Change Detection Tests Using the ICI rule", in *Proc. of IJCNN*, pp. 1-7, 2010.

[21] A. Goldenshluger and A. Nemirovski, "On spatial adaptive estimation of nonparametric regression," in *Mathematical Methods of Statistics*, vol. 6, pp. 135-170, 1997.

[22] V. Katkovnik, "A new method for varying adaptive bandwidth selection," in *IEEE Trans. on Signal Processing,* vol. 47, no. 9, pp. 2567-2571, 1999.