Adaptive Ensemble of Classifiers

Cesare Alippi, Giacomo Boracchi, Manuel Roveri
Politecnico di Milano, Dipartimento di Elettronica e Informazione, Milano, Italy
{alippi,boracchi,roveri}@elet.polimi.it

I. Introduction

Classification applications where the probability density function of classes evolve over time are referred as *concept drifts*. Abrupt concept drifts refer to situations where the data-generating process suddenly changes from a stationary state to another one, e.g., due to a permanent or a transient fault. Differently, gradual concept drifts refer to cases where the process continuously evolves over time, a situation typically caused by aging effects or thermal drifts.

Input samples (observations) are vectors generated from process X according to an unknown distribution. Denote by $x \in \mathbb{R}^n$ the observation, x_t the observation at time t, and y_t the associated class label. Without loss of generality, we consider a two-class classification problem, i.e., $y(t) \in \{\omega_1, \omega_2\}$. The probability density function of the inputs at time t can be thus defined as

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

where $p(\omega_1|t)$ and $p(\omega_2|t) = 1 - p(\omega_1|t)$ are the probabilities of getting a sample of class ω_1 and ω_2, respectively, while $p(x|\omega_1, t)$, $p(x|\omega_2, t)$ are the conditional probability distributions at time t. Both the probabilities of the classes and the conditional pdfs are assumed to be unknown and may evolve over time, whenever a non-stationarity occurs.

Monitoring the classification error allows a generic consistent classifier for reacting to changes when these directly influence its accuracy, and this is in principle preferable. However, assessing the stationarity by means of the online classification error becomes critical in applications where obtaining supervised information is difficult or costly. A viable option at low supervised-sample rates consists in monitoring the distribution of the unlabeled input observations. Unfortunately, this solution does allow us for detecting only changes that affect the distribution of observations, but not the conditional classification probabilities $p(x|\omega_1, t)$ and $p(x|\omega_2, t)$ such as a swap of classes.

We suggest a solution which combines the two mechanisms by monitoring both the distribution of observations and the classification error to detect changes in the data-generating process. The designed adaptive classifier exploits both supervised and unsupervised samples to adapt to changes within an ensemble framework: each time a change is detected, the previously trained classifiers are tested to identify if the novel concept has been already envisaged or not. If the concept is recurrent [1], the existing classifier is re-activated; otherwise the obsolete classifier is replaced with a new one.

The training sequence consists in the first T_0 observations that are assumed to be generated in stationary conditions; supervised pairs (x_t, y_t) are provided both within the training sequence and during the operational life asynchronously w.r.t. the inputs.

II. Adaptive Ensamble of Classifiers

The key elements to design an adaptive classifier evolving when a non-stationarity occurs are: CDT_X, the Change-Detection Test (CDT) observing changes in stationarity of X; CDT_ϵ, the CDT for assessing the stationarity of the classification error; K the current classifier used to classify inputs, and C_i the i-th concept, which has to be considered as a set of observations (together with supervised labels when applicable) associated to a specific state of the data-generating process.

In the proposed adaptive classifier we adopt the ICI-based CDT [2] as CDT_X, which exploits the Intersection of Confidence Intervals (ICI) rule [3], for its effectiveness in detecting both abrupt and gradual concept drifts. Thus, CDT_X relies on specific features extracted from disjoint subsequences of data which, in stationary conditions, are i.i.d. and follow a Gaussian distribution. Examples of such features are the sample mean and variance computed on disjoint subsequences (the former can be approximated with a Gaussian distribution from the Central Limit Theorem, the latter follows the Gaussian distribution thanks to an ad-hoc transformation). Then, the ICI-rule is used to assess, on-line and sequentially, if the feature values have been generated from the same Gaussian distribution.

Differently, CDT_ϵ consists in a customization of the ICI-based CDT to assess the constant value hypothesis for the classification error. Let (x_t, y_t) be a supervised couple and let $K(x_t)$ be the outcome of classifier K on x_t. The element-wise classification error of K is

$$\epsilon_t = \begin{cases} 0, & \text{if } y_t = K(x_t); \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

which over time can be modeled as a sequence of i.i.d. Bernoulli random variables. The parameter p of the Bernoulli distribution corresponds to the expected error of K and, in stationary conditions (when both X and K do not change), is constant. To identify changes in stationarity (both in the classes' probability or in the conditional pdfs), the ICI-rule monitors the average classification

Train K, CDT$_X$ and CDT$_\epsilon$ on observations in $[0, T_0]$;
while *(1)* **do**
  input receive new data x_t and, when available, y_t;
  if *(Either CDT$_X$ or CDT$_\epsilon$ detects a nonstationarity at time t)* **then**
    Characterize the current concept C_i;
    Check if C_i is coherent with any of other C_j, $j \neq i$;
    if *(C_i is recurrent)* **then**
      Reactivate the corresponding classifier;
      Integrate recent supervised couples (if any);
    else
      Drop K and save the corresponding concept C_i;
      Reconfigure and activate K from C_i;
      Reconfigure both CDT$_X$ and CDT$_\epsilon$ on C_i;
    end
  end
  if *(Supervised label y_t is provided)* **then**
    Insert (x_t, y_t) in the knowledge-base of K and update K;
  else
    Assign label $K(x_t)$ to x_t.
  end
end

Algorithm 1: The high-level algorithm for the Adaptive Ensemble Classifier.

error computed on disjoint subsequences of ν supervised observations which, indeed, follow a Binomial distribution $\mathcal{B}(p, \nu)$ approximable with a Gaussian one whenever ν is sufficiently large,

$$\mathcal{B}(p, \nu) \sim \mathcal{N}\big(p\nu, p(1-p)\nu\big). \tag{2}$$

The Gaussian approximation allows us for directly applying the ICI-rule to verify sequentially if the average classifier error is constant as new observations arrive. For monitoring purposes we leverage an auxiliary classifier K_0, which is configured on the initial training set and is never updated; as such, its expected error p_0 remains constant as far as the data-generating process is stationary. Thus, CDT$_\epsilon$ assesses if the sample mean of ν classification errors (1) – computed from K_0 – is constant over time.

Whenever CDT$_X$ or CDT$_\epsilon$ detects a non stationarity at time $\hat{T}$, a refinement procedure is executed yielding T_{ref}, a more accurate estimate of the change-time instant. It is thus possible to define two subsequences of observations: those from $[0, T_0]$ and those from $[T_{\text{ref}}, \hat{T}]$ that are representative of X before and after the non-stationarity, respectively. Each non-stationarity needs to be validated, to prevent a false detection to result in a classifier reconfiguration. The change-validation procedure is performed by assessing if both the features of CDT$_X$ and the average classification error of K_0 computed from the observations in $[0, T_0]$ (i.e., before the suspected change) coincide with those in $[T_{\text{ref}}, \hat{T}]$ (i.e., after the suspected change). The change validation on the features of CDT$_X$ is performed with a multivariate hypothesis test relying on the Hotelling T^2 statistics (the features follow a Gaussian distribution) as suggested in [4], while the change validation on the classification error of K_0 can be formulated as an (univariate) inference problem on the proportions of two populations, which is ruled by a Gaussian distribution. In both cases the null hypothesis can be rejected according to a defined significance level α.

Any validated change requires the classifier K to be retrained (see Algorithm 1), otherwise the change is discarded and only the CDT providing the false detection is reconfigured to continue its monitoring activity.

Recurrent concepts are identified by testing both the observations and classification errors similarity. More specifically, when comparing two concepts C_i and C_j, $i \neq j$, we assess if both the averages of the features and the classification errors computed in $[T_{\text{ref},i}, \hat{T}_i]$ correspond to those computed in $[T_{\text{ref},j}, \hat{T}_j]$. When C_i and C_j satisfy a stochastic similarity condition we consider concept C_i to be recurrent and the supervised samples in $[T_{\text{ref},j}, \hat{T}_j]$ can be safely paired with those in $[T_{\text{ref},i}, \hat{T}_i]$ to retrain K.

ACKNOWLEDGEMENTS

This research has been funded by the European Commissions 7th Framework Program, under grant Agreement INSFO-ICT-270428 (iSense).

REFERENCES

[1] R. Elwell and R. Polikar, "Incremental learning in nonstationary environments with controlled forgetting," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, june 2009, pp. 771 –778.

[2] C. Alippi, G. Boracchi, and M. Roveri, "A just-in-time adaptive classification system based on the intersection of confidence intervals rule," *Neural Networks*, vol. 24, no. 8, pp. 791 – 800, 2011, artificial Neural Networks: Selected Papers from ICANN 2010.

[3] A. Goldenshluger and A. Nemirovski, "On spatial adaptive estimation of nonparametric regression," *Math. Meth. Statistics*, vol. 6, pp. 135–170, 1997.

[4] C. Alippi, G. Boracchi, and M. Roveri, "A hierarchical, nonparametric, sequential change-detection test," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, 31 2011-aug. 5 2011, pp. 2889 –2896.