A Bayesian Approach to Abrupt Concept Drift

Andrés Cano, Manuel Gómez-Olmedo, Serafín Moral
Dpto. Ciencias de la Computación e IA
Universidad de Granada
18071 - Granada, Spain
emails: (acu,mgomez,smc)@decsai.ugr.es

Abstract

This paper proposes a model for detecting probabilities in the presence of abrupt concept drift. This proposal is based on a dynamic Bayesian network. As the exact estimation of the parameters is unfeasible we propose an approximate procedure based on discretazing both the possible probability values and the parameter representing the probability of change. The result is a procedure which is quite efficient in time and space (with a complexity directly related to the number of points used in the discretization) and providing very accurate prediction as well. These benefits are checked with a detailed comparison with other standard procedures based on variable size windows or forgetting rates. The procedure is presented for a binomial variable but it will be explained how it can be extended to more general settings.

Key Words.- concept drift, dynamic Bayesian networks, change detection, propagation algorithms.

1. Introduction

In many situations we have a stream of observations for a certain set of variables which comes continuously over time. In many cases, the stationarity of the model generating the data is assumed, i.e. it does not change as long as time goes by. However this hypothesis is unrealistic in certain environments that evolve with time. An example can be the probability of appearance of a specific word in a junk e-mail. When a change in probabilities occurs we say that we have a *concept drift* [5]. When considering a model for a supervised classification problem it is usual to distinguish between changes in the prior probabilities of the class, its posterior probabilities, or the conditional probabilities relating the class to the attributes [5, 14, 13]. In this paper this discrimination is not considered: it will focus on detecting probability changes without taking into account if this parameter is a conditional, prior, or posterior probability. We can also distinguish between abrupt and gradual changes [5, 13]. In the first case a period of no-change is followed by a sudden change (for example a machine that stops working). In the second case there are small and gradual changes (a machine producing more defective items as long as it gets older). In this paper we will only consider abrupt change. The approaches for dealing with this problem can be categorized as follows:

- Methods based on a sliding window of fixed size: the probabilities are estimated taking into account the last N observations where N is a fixed parameter. This is perhaps the simplest approach. An example of this strategy is the procedure for updating decision trees proposed in [8]. The main problem is the determination of N: a low value will imply that probabilities are always estimated with small samples and will be subject to non negligible errors; a large value of N will increase the complexity of the model and will limit its capability to adapt to changes.

- Methods based on a sliding window of variable size: in this case the probabilities are estimated with a window containing the last N observations, but now N is not a fixed parameter. Each new observation is added to the window and then some statistical tests (or more 'ad hoc' decision procedures) are computed in order to determine whether there are differences between the distribution of the first N_1 observations and the last N_2 ($N = N_1 + N_2$). If a difference is detected the samples in N_1 are removed (forgotten) from the sliding window. Usually the stability of the accuracy is considered as the basis to detect changes when the sample is employed to estimate a supervised classification model [4]. However, *ADWIN* ([2]) is perhaps the best known procedure for another kind of problems. It is based on a generic statistical procedure for deciding about changes in different sub-windows of the last N observations (its performance is guaranteed).

- Methods based on gradual forgetting: these methods consider that old observations are less important than new ones, being the loss of relevance gradual with respect to their age [9]. In general, these procedures keep a set S_i of sufficient statistics for estimating parameters; for each observation X_{i+1} a new set of sufficient statistics is computed as a function $G(\alpha_i S_i, X_{i+1})$ where $\alpha_i \in (0,1)$ is the forgetting factor which usually is constant [5]. This idea seems to be more appropriate for gradual changes than for abrupt ones.

- Methods based on Bayesian updating: in this case a full probabilistic model is specified and updated each time an observation or a bunch of them arrives. In general, this will produce an estimation of the probabilities of both change and next observation. Honkela, Vanpola [7] proposes a probabilistic model in which the probabilities evolve according to a forgetting factor using variational inference. A similar approach is followed by Masegosa et al. [10] but in this case the forgetting factor can be estimated with the model as well. This work and the one presented in Cabañas et al. [3] employ hidden variables to monitor the presence of changes.

In this paper we propose a Bayesian approach based on probabilistic graphical models [12] of a dynamic nature [11] and specific for abrupt changes in the probabilities. We will give an exact expression for computing the estimation of the probabilities, but it will be difficult to use in practice. So, our approach will employ an approximate computation based on discretization.

The paper is organized as follows: Section 2 presents the basic model for the binomial case and the approximate computation procedure; it also explains how the basic model can be extended to monitor the evolution of conditional probabilities; Section 3 is devoted to the experiments; and finally Section 4 considers the conclusions and future work.

2. The Basic Model: Binomial Case

Consider a sequence of variables $\{X_i\}_{i=1}^n$ where each variable X_i takes values on the set $\{0,1\}$ with probabilities $\mu_i = P(X_i = 1)$ that may change with i. The problem under consideration can be formulated as follows: given a certain set of observations $X_i = x_i, i = 1, \ldots, n$, how to compute an estimation of μ_i ($\hat{\mu}_i$) for each i assuming that we have observed $X_1 = x_1, \ldots, X_i = x_i$.

The frequencies of 0s and 1s in a certain set of observations $x_j, \ldots, x_i$, will be denoted as $N_{ji} = \sum_{k=j}^{k=i} x_k$ and $\overline{N}_{ji} = (i - j + 1) - N_{ji}$ respectively. In the case of abrupt change μ_i may present a change at each moment i (the value of μ_i will be randomly selected in $[0,1]$ with an uniform distribution) or remain stable (and then $\mu_i = \mu_{i-1}$). The variable C_i will be used to represent the occurrence of change: its value will be 1 in the case of change and 0 otherwise. It is assumed that the probability of change, $P(C_i = 1)$, is constant and denoted by ρ. In order to get an estimation of μ_i it is needed to model the relation between μ_i and μ_{i+1}. This is the purpose of the dynamic Bayesian network presented in Figure 1. As we employ a Bayesian approach all the elements involved in this problem will be considered as random variables.

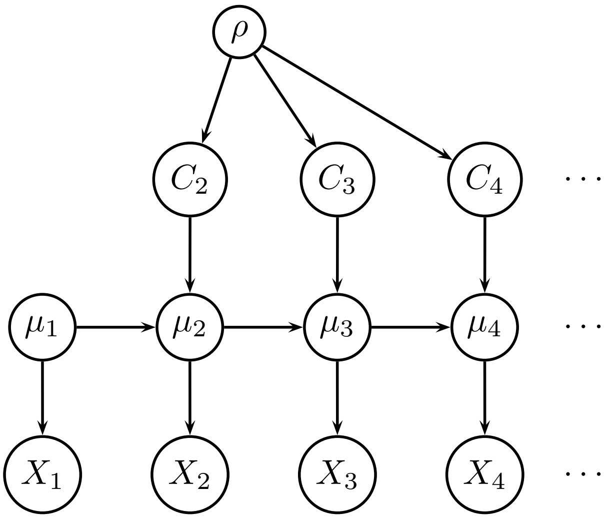


Figure 1: Dynamic Bayesian network modeling the problem

2.1. Exact Estimation

To simplify the notation, let us call A_i to the event $X_1 = x_1, \ldots, X_i = x_i$. $f_i(\mu_i, \rho|A_i)$ denotes the posterior density of (μ_i, ρ) given the observations A_i. Analogously $f_i(\mu_i|A_i)$ and $f_i(\rho|A_i)$ will denote the marginal densities conditioned to A_i. $f_i(\mu_i, \rho, A_i)$ and $f_i(\mu_i, A_i)$ represent the joint densities of the parameters and A_i, i.e. the densities conditioned to the observations, but without normalization (the prior density of the parameters multiplied by the probability of A_i):

$$\begin{gathered} f_i(\mu_i, \rho, A_i) = P(A_i|\mu_i, \rho) f_i(\mu_i, \rho) \\ f_i(\mu_i, A_i) = P(A_i|\mu_i) f_i(\mu_i) \end{gathered} \tag{1}$$

The computation of $f_i(\mu_i, \rho|A_i)$ can be done from $f_i(\mu_i, \rho, A_i)$ by normalization (dividing the density by the integral of $f_i(\mu_i, \rho, A_i)$ on parameters μ_i and ρ). The density for μ_1 and A_1, i.e. $f_1(\mu_1, A_1)$ is equal to $\mu_1^{N_{11}}(1-\mu_1)^{\overline{N}_{11}}$. This is immediate, as the prior information for μ_1 is uniform $(f_1(\mu_1) = 1)$ and this is exactly the likelihood on the space $[0, 1]$, associated to the observation $X_1 = x_1$. The joint density $f_1(\mu_1, \rho, A_1)$ can be computed as:

$$f_1(\mu_1, \rho, A_1) = \mu_1^{N_{11}}(1 - \mu_1)^{\overline{N}_{11}} f(\rho) \tag{2}$$

where $f(\rho)$ is the prior density of ρ. This is immediate just because μ_1 and ρ are independent according to the model of Figure 1.

Theorem 1. *The density $f_i(\mu_i, \rho, A_i)$ can be computed as*

$$f(\rho)\sum_{j=1}^{i} P(A_{j-1})(1-\rho)^{i-j}\rho^{r(i)}\mu_i^{N_{ji}}(1-\mu_i)^{\overline{N}_{ji}}$$
$$r(i) = \begin{cases} 0 & if \;\; i = 1 \\ 1 & otherwise \end{cases} \tag{3}$$
$$P(A_0) = 1$$

Proof. Let us prove that

$$f_i(\mu_i, A_i|\rho) = \sum_{j=1}^{i} P(A_{j-1})(1-\rho)^{i-j}\rho^{r(i)}\mu_i^{N_{ji}}(1-\mu_i)^{\overline{N}_{ji}} \tag{4}$$

Then the proof of the theorem will be obtained multiplying by the prior density of ρ, $f(\rho)$, in order to get $f_i(\mu_i, \rho, A_i)$.

Eq. 4 will be proved by induction on i. For $i = 1$, the result was established before this theorem (Eq. 2), considering that A_0 is the empty set of observations and $P(A_0) = 1$. Assuming that it is true for i then it must be proved for $i + 1$.

As the density of ρ is not considered, it is needed the computation of the distribution $f_{i+1}(\mu_{i+1}, A_{i+1})$. This can be expressed as:

$$f_{i+1}(\mu_{i+1}, A_{i+1}) = f_{i+1}(\mu_{i+1}, A_i)P(X_{i+1} = x_{i+1}|\mu_{i+1}, A_i) \tag{5}$$

Taking into account the fact that X_{i+1} is conditionally independent of A_i given μ_{i+1}, then Eq. 5 can be expressed as:

$$f_{i+1}(\mu_{i+1}, A_{i+1}) = \underbrace{f_{i+1}(\mu_{i+1}, A_i)}_{\text{(a)}}\underbrace{P(X_{i+1} = x_{i+1}|\mu_{i+1})}_{\text{(b)}} \tag{6}$$

As $N_{(i+1)(i+1)}$ refers to a single observation its value will be 1 if $x_{i+1} = 1$ and 0 otherwise, being $\overline{N}_{(i+1)(i+1)} = 1 - N_{(i+1)(i+1)}$. Therefore the right part of Eq. 6 (labeled as (b)) can be written as:

$$P(X_{i+1} = x_{i+1}|\mu_{i+1}) = \mu_{i+1}^{N_{(i+1)(i+1)}}(1-\mu_{i+1})^{\overline{N}_{(i+1)(i+1)}} \tag{7}$$

On the other hand, the left part of Eq. 6 (labeled as (a)) can be expressed as:

$$\begin{gathered} f_{i+1}(\mu_{i+1}, A_i) = f_{i+1}(\mu_{i+1}, A_i|C_{i+1} = 1)P(C_{i+1} = 1) \\ +f_{i+1}(\mu_{i+1}, A_i|C_{i+1} = 0)P(C_{i+1} = 0) = \\ f_{i+1}(\mu_{i+1}, A_i|C_{i+1} = 1)\rho + f_{i+1}(\mu_{i+1}, A_i|C_{i+1} = 0)(1-\rho) \end{gathered} \tag{8}$$

just because $P(C_{i+1} = 1) = \rho$ and $P(C_{i+1} = 0) = 1 - \rho$. As $C_{i+1} = 1$ refers to the situation when a change occurs, then μ_{i+1} is independent of A_i given ρ and having an uniform density in $[0, 1]$. Therefore

$$f_{i+1}(\mu_{i+1}, A_i | C_{i+1} = 1) P(C_{i+1} = 1) = P(A_i)\ \rho \tag{9}$$

When $C_{i+1} = 0$ there is no change and $\mu_{i+1} = \mu_i$ and $f_{i+1}(\mu_{i+1}, A_i | C_{i+1} = 0) = f_i(\mu_i, A_i)$. Then

$$f_{i+1}(\mu_{i+1}, A_i | C_{i+1} = 0) P(C_{i+1} = 0) = (1 - \rho) f_i(\mu_i, A_i) \tag{10}$$

Assuming the induction hypothesis expressed by Eq. 11, $f_i(\mu_i, A_i)$ can be written as:

$$f_i(\mu_i, A_i) = \sum_{j=1}^{i} P(A_{j-1})(1-\rho)^{i-j} \rho^{r(i)} \mu_i^{N_{ji}} (1 - \mu_i)^{\overline{N}_{ji}} \tag{11}$$

With these Eq. 8 is now:

$$f_{i+1}(\mu_{i+1}, A_i) = \rho P(A_i) + (1-\rho) \left[\sum_{j=1}^{i} P(A_{j-1})(1-\rho)^{i-j} \rho^{r(i)} \mu_i^{N_{ji}} (1 - \mu_i)^{\overline{N}_{ji}} \right] \tag{12}$$

This expression must be multiplied by $\mu_{i+1}^{N_{(i+1)(i+1)}} (1 - \mu_{i+1})^{\overline{N}_{(i+1)(i+1)}}$ in order to get the complete expression for $f_{i+1}(\mu_{i+1}, A_{i+1})$:

$$\begin{aligned} f_{i+1}(\mu_{i+1}, A_{i+1}) = \rho P(A_i) \mu_{i+1}^{N_{(i+1)(i+1)}} (1 - \mu_{i+1})^{\overline{N}_{(i+1)(i+1)}} + \\ \sum_{j=1}^{i} P(A_{j-1})(1-\rho)^{i+1-j} \rho^{r(i)} \mu_i^{N_{j(i+1)}} (1 - \mu_i)^{\overline{N}_{j(i+1)}} \end{aligned} \tag{13}$$

The right term in Eq. 13 is obtained considering that $N_{ji} + N_{(i+1)(i+1)} = N_{j(i+1)}$ and $\overline{N}_{ji} + \overline{N}_{(i+1)(i+1)} = \overline{N}_{j(i+1)}$. Finally the desired expression follows by noting that Eq. (13) is equal to

$$\sum_{j=1}^{i+1} P(A_{j-1})(1-\rho)^{i+1-j} \rho^{r(i)} \mu_i^{N_{j(i+1)}} (1 - \mu_i)^{\overline{N}_{j(i+1)}} \tag{14}$$

□

One problem with the density of Eq. 3 is that values $P(A_j)$ $(j = 1, \ldots, i-1)$ are required. Furthermore, in order to compute the posterior density, it is needed to integrate in μ_{i+1} and ρ in order to obtain $P(A_i)$ (the normalization constant). However, if the density of ρ is a Beta distribution, we can derive a recursive expression for these probabilities. This is stated by the following theorem for the case of an uniform density in $[0, 1]$ for ρ parameter.

Theorem 2. *If $f(\rho)$ is the uniform distribution, then*

$$\begin{gathered}P(A_i) = \sum_{j=1}^{i} P(A_{j-1}) \frac{(i-j)!r(i)!}{(i-j+r(i)+1)!} \frac{N_{ji}!\overline{N}_{ji}!}{(i-j+2)!} \\ r(i) = \begin{cases} 0 & if \;\; i = 1 \\ 1 & otherwise \end{cases} \\ P(A_0) = 1 \end{gathered} \tag{15}$$

Proof. The expression in Eq. 15 can be easily obtained from Eq. (3) taking into account that:

$$\int_0^1 x^\alpha (1-x)^\beta dx = \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+2)} = \frac{\alpha!\beta!}{(\alpha+\beta+1)!} \tag{16}$$

where Γ is the Gamma function, and $\alpha, \beta \geq 0$. It is also important to remark that $N_{ji} + \overline{N}_{ji} = i - j + 1$. □

Finally, we can compute the estimation $\hat{\mu}_i$ as the expected value of μ_i with respect to the posterior density of μ_i given A_i. The final expression is simply the integral of the density (3) multiplied by μ_i, divided by $P(A_i)$. This produces the following result:

Theorem 3. *If $f(\rho)$ is the uniform distribution, then $\hat{\mu}_i$ (expected value of μ_i with respect to $f_i(\mu_i)$) is equal to:*

$$\begin{gathered}\hat{\mu}_i = \frac{\sum_{j=1}^{i} P(A_{j-1}) \frac{(i-j)!r(i)!}{(i-j+r(i)+1)!} \frac{(N_{ji}+1)!\overline{N}_{ji}!}{(i-j+3)!}}{\sum_{j=1}^{i} P(A_{j-1}) \frac{(i-j)!r(i)!}{(i-j+r(i)+1)!} \frac{N_{ji}!\overline{N}_{ji}!}{(i-j+2)!}} \\ r(i) = \begin{cases} 0 & if \;\; i = 1 \\ 1 & otherwise \end{cases} \\ P(A_0) = 1 \end{gathered} \tag{17}$$

Proof. The result is obtained just because the numerator is the integral $\int_0^1 \int_0^1 f_i(\mu_i, \rho, A_i)\mu_i d\mu_i d\rho$ and the denominator is the normalization constant $P(A_i)$ (see Eq. 15). □

Expressions (15) and (17) provide recursive equations allowing an exact estimation of the parameters μ_i by the expectation of its posterior density. However, the complexity of these computations increases with the length of the sequence. Moreover, all the observations $X_1 = x_1, \ldots, X_i = x_i$ must be stored (we could keep in memory N_{ji} values but this is equivalent to keep all the observations just because $x_j = N_{ji} - N_{(j+1)i}$). As a consequence this exact procedure is unfeasible for very large data streams where a fast computation of $\hat{\mu}_i$ is necessary each time a new observation arrives. For this reason, we have to develop an approximate procedure.

2.2. Approximate Computation

The approximate computation of $f_i(\mu_i, \rho|A_i)$ will be based on discretization. We will consider two different cases: fixed value of parameter ρ and uncertainty for both parameters (ρ and μ).

2.2.1. Fixed value of ρ

In general the estimation of a density $f(x), x \in [a, b]$ will be based on setting a value for an integer parameter K and the selection of a set of $K + 1$ points ($r_0 = a, \ldots, r_K = b$) with a certain probability distribution for them ($p_0, \ldots, p_K$). Given a subset $A \subseteq [a, b]$ then $P(A)$ and $E[X]$ will be approximated as:

$$\begin{aligned} P(A) &= \sum_{r_k \in A} p_k \\ E[X] &= \sum_{k=0}^{K} r_k p_k \end{aligned} \tag{18}$$

The points will be selected according to the following expression:

$$r_k = F^{-1}(\frac{1}{k}) \tag{19}$$

where F is the cumulative distribution function of f and F^{-1} its generalized inverse (assuming F is a continuous function). The probability distribution assigns a probability given by $p_k = 1/(K + 1)$ to each point r_k. This approximation can be quite different from the true one if, for example, A is a subset which does not contain any of the points selected for the discretization. The following proposition states that the error for the probability is limited according to K.

Proposition 1. *Given a certain interval $[c, d] \subseteq [a, b]$ then:*

$$|P([c,d]) - P'([c,d])| \leq \frac{2}{K+1} \tag{20}$$

where P represents the true probability computed according to f and P' the approximate one.

Proof. The proof for this relation is simple. Let us assume that $P([c,d]) \geq P'([c,d])$. Let I denote the greatest interval containing $[c,d]$ but without including new points $r_k \notin [c,d]$ (this interval can be open). Then

$$|P([c,d]) - P'([c,d])| \leq P(I) - P'(I) \tag{21}$$

Note that enlarging the interval (I enlarges $[c,d]$) will increase its true probability. However $P'(I)$ will not change because I contains the same set of points that $[c,d]$. The *infimum* and *supremum* points in I will be noted r_l and r_u respectively. Then the points contained in $[c,d]$ and I will be $r_{l+1}, \ldots, r_{u-1}$. Therefore:

$$\begin{aligned} P(I) &= \frac{u-l}{K} \\ P'(I) &= \frac{(u-1)-(l+1)+1}{K+1} = \frac{u-l-1}{K+1} \end{aligned} \tag{22}$$

Then the difference between these values is:

$$P(I) - P'(I) = \frac{u-l+K}{K+1} \leq \frac{2}{K+1} \tag{23}$$

With this last result and according to expression 21 Proposition 1 is demonstrated.

The proof is analogous if $P([c,d]) \leq P'([c,d])$ but now I will be a reduced interval respect to $[c,d]$. □

The result presented in Eq. 23 shows that the quality of the approximation can be improved increasing the number of points, that is, the value of K. A similar result can be obtained for the expected value.

Proposition 2. *Having a set of points r_k selected as indicated above, then the values of the true average $\overline{x}$ and the approximate one, $\overline{x}'$ (related to A), hold the following relation*

$$|\overline{x} - \overline{x}'| \leq \frac{3b-a}{K+1} \tag{24}$$

Proof. The true average $\overline{x}$ can be computed as:

$$\overline{x} = \int_a^b xf(x)dx = \sum_{k=0}^{K-1} \int_{r_k}^{r_{k+1}} xf(x)dx \tag{25}$$

The approximate value, $\overline{x}'$, is given by the following expression:

$$\overline{x}' = \sum_{k=0}^{K} r_k \frac{1}{K+1} \tag{26}$$

For any $k = 0, \dots, K-1$, we have that

$$\begin{gathered} \int_{r_k}^{r_{k+1}} xf(x)dx \geq r_k \frac{1}{K} \geq r_k \frac{1}{K+1} \\ \int_{r_k}^{r_{k+1}} xf(x)dx \leq r_{k+1} \frac{1}{K} \end{gathered} \tag{27}$$

Taking into account Eq. 27 the difference $|\overline{x} - \overline{x}'|$ can be expressed as:

$$\begin{gathered} |\overline{x} - \overline{x}'| = \sum_{k=0}^{K-1} \int_{r_k}^{r_{k+1}} xf(x)dx - \sum_{k=0}^{K} r_k \frac{1}{K+1} \leq \\ \sum_{k=0}^{K-1} \int_{r_k}^{r_{k+1}} xf(x)dx - \sum_{k=0}^{K-1} r_k \frac{1}{K+1} + r_K \frac{1}{K+1} \end{gathered} \tag{28}$$

As $r_K = b$ then

$$\begin{gathered} |\overline{x} - \overline{x}'| = \sum_{k=0}^{K-1} r_{k+1} \frac{1}{K} - \sum_{k=0}^{K-1} r_k \frac{1}{K+1} + b \frac{1}{K+1} = \\ \sum_{k=0}^{K-1} \left[r_{k+1} \frac{1}{K} - r_k \frac{1}{K+1} \right] + b \frac{1}{K+1} = \\ \sum_{k=0}^{K-1} \frac{r_{k+1}(K+1) - K r_k}{K(K+1)} + b \frac{1}{K+1} = \\ \sum_{k=0}^{K-1} \left[\frac{K(r_{k+1} - r_k)}{K(K+1)} + \frac{r_{k+1}}{K(K+1)} \right] + b \frac{1}{K+1} \end{gathered} \tag{29}$$

Observing that $\sum_{i=0}^{K-1} (r_{k+1} - r_k) = (b - a)$ and that $r_{k+1} \leq r_K = b$, we obtain

$$|\overline{x} - \overline{x}'| \leq \frac{b-a}{K+1} + K\frac{b}{K(K+1)} + \frac{b}{K+1} =$$
$$\frac{b-a}{K+1} + \frac{2b}{K+1} = \frac{3b-a}{K+1}$$

□

With these results, it is clear that the initial approximation for probability and expected value can be improved to a desired quality by increasing the number of points K. It is important to notice that it would have been even simpler to consider K points instead of $K+1$ with values at the center of intervals $[a + k(b-a)/K, a + (k+1)(b-a)/K]$. However we have selected the option of $K+1$ points including the intervals extremes (a and b) just because we wanted a and b were the *infimum* and *supremum* values of the expected value considering all the possible probability distributions p_k.

Once performed the initial approximation, the variables are considered discrete (a finite set of points in $[a, b]$ with positive probability) and updated accordingly each time a new observation X_i arrives. Let us call p_k^i to the joint probability of r_k and the observations $X_1 = x_1, \ldots, X_i = x_i$. According to this,

$$p_k^i = \left[(1-\rho)p_k^{i-1} + (\sum_{k=0}^{K} p_k^{i-1})\frac{\rho}{K+1}\right] r_k^{N_{ii}}(1-r_k)^{\overline{N}_{ii}}. \qquad (30)$$

This allows to update the probabilities p_k^i, starting with $p_k^0 = p_k = 1/(K+1)$. With them we can compute the approximate expected value of the parameter $\hat{\mu'}_i$ after the first i observations as:

$$\hat{\mu}'_i = \frac{\sum_{k=0}^{K} r_k p_k^i}{\sum_{k=0}^{K} p_k^i}. \qquad (31)$$

These formulas allow an efficient updating of the estimations, both in time and space (with linear complexity in the number of intervals K).

2.2.2. Uncertainty about μ and ρ

Now, for the case of handling uncertainty about both parameters μ and ρ, we also discretize ρ, by selecting $L+1$ points $s_0, s_1, \ldots, s_L$, and assigning to them a probability $1/(L+1)$. To select the points we consider a prior interval for the parameter $[\rho_1, \rho_2]$ and then a density in this interval which is discretized with the same procedure used for the discretization of μ. Two cases are considered: an uniform density as in the case of μ and an alternative density which is concentrated in the smaller values of the interval, with the

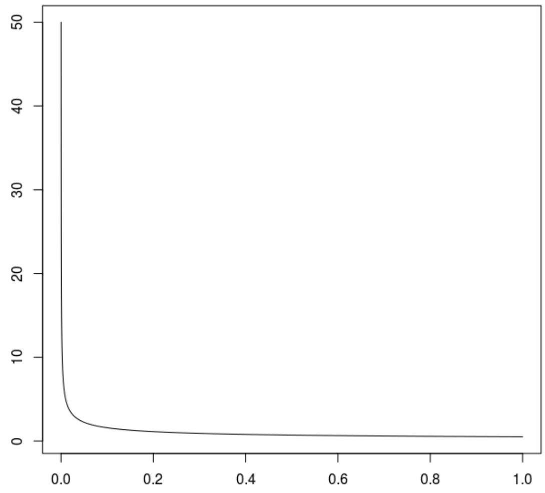


Figure 2: Prior density for ρ.

idea that if the prior interval is for example $[0.001, 0.1]$, then we assume that the density is more concentrated around the lower bound, 0.001, than in the upper bound 0.1. In concrete we assume that ρ follows a density in $[\rho_1, \rho_2]$:

$$f(\rho) = \frac{1}{2(\rho_2 - \rho_1)} \left(\frac{\rho - \rho_1}{\rho_2 - \rho_1} \right)^{-1/2} \tag{32}$$

The shape of the prior density for ρ is shown in Figure 2 for the case of the interval $[0, 1]$. Given L this density is discretized following the same criteria. It is very simple to prove that the points that are obtained are $s_l = \rho_1 + (l/L)^2(\rho_2 - \rho_1)$. It is important to remark that with this density, we do not only concentrate the probability in the low values, but also that the points s_l are concentrated in the lower part of the interval, and the approximation there is finer.

Initially, ρ and μ_1 are independent variables, even after observing X_1, but ρ and μ_i are not longer independent if $i \geq 2$. For this reason, we must keep an approximation of the joint density $f_i(\mu_i, \rho, A_i)$. This would be done by keeping a grid of points (r_k, s_l) $(k = 0, \ldots, K; \, l = 0, \ldots, L)$ with probabilities p^i_{kl}, which are the joint probabilities of (r_k, s_j) and A_i approximating $f_i(\mu_i, \rho, A_i)$. These probabilities are updated according to an expression similar to (30), but considering different value of ρ $(s_0, \ldots, s_L)$:

$$p_{kl}^{i} = \left[(1-s_l)p_{kl}^{i-1} + (\sum_{k=0}^{K} p_{kl}^{i-1})\frac{s_l}{K+1} \right] r_k^{N_{ii}}(1-r_k)^{\overline{N}_{ii}}. \tag{33}$$

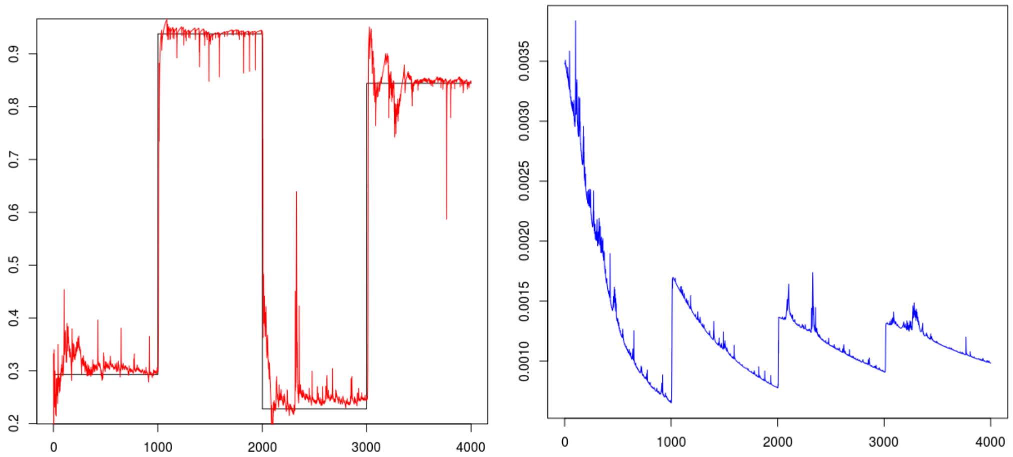
The estimations of μ_i and ρ'_i can be computed according to these probabilities:

$$\hat{\mu'_i} = \frac{\sum_{l=0}^{L}\sum_{k=0}^{K} r_k p_{kl}^{i}}{\sum_{l=0}^{L}\sum_{k=0}^{K} p_{kl}^{i}} \quad \hat{\rho'_i} = \frac{\sum_{l=0}^{L}\sum_{k=0}^{K} s_l p_{kl}^{i}}{\sum_{l=0}^{L}\sum_{k=0}^{K} p_{kl}^{i}}. \tag{34}$$

3. Experiments

The general scheme for our experiments will be to generate a series of 4000 observations for X_i. Each X_i is an independent binary variable with $P(X_i = 1) = \mu_i$. A change in μ_i is carried out after each 1000 observations and therefore a new value for μ_i is uniformly selected in the interval $[0, 1]$.

First, we will illustrate how our method works by showing the real values for μ_i and the estimations of the parameters computed according to Eq. (34). Both estimations can be seen in Figures 3a and 3b, respectively.

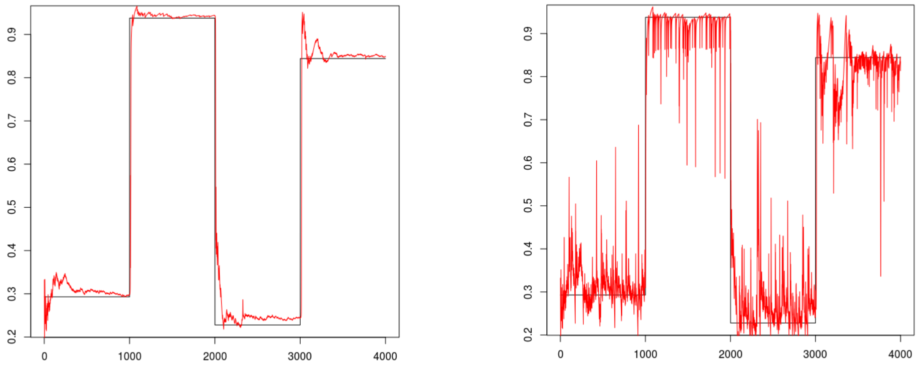


(a) Estimation of μ_i

(b) Estimation of ρ_i

Figure 3: Example: parameter estimation on a series of observations

The black line in Figure 3a represents the true values for $\mu_i, i = 1, \ldots, 4000$ and the red one shows the estimation obtained with our method and these parameters: $K = 300, L = 30, [\rho_1, \rho_2] = [0.0001, 0.1]$. We can observe that there is a fast adaption to changes in μ_i. Although there are oscillations in the periods in which μ_i is stable we can observe two features: oscillations



(a) Estimation with $\rho = 0.00001$

(b) Estimation with $\rho = 0.1$

Figure 4: Comparing the estimations of μ_i with different values of ρ

decrease as long as the periods goes on; and oscillations are bigger for extreme probabilities (close to 0 or 1) than for intermediate ones (near to 0.5). This last fact is significant as extreme probabilities are more risky and then the estimation tries to reduce this risk (Bayesian procedures do this in an automatic way).

In Figure 3b we can see that the estimation of ρ starts in the expected value of the prior probability and decreases in periods without changes, increasing abruptly when changes occur. Moreover, oscillations are less important when the series progress approaching the true value (0.001).

In this problem, there is a trade-off between the capability of adaptation to changes and the stability of the estimations in periods without changes. In our second example, Figure 4 shows how to control the balance point by selecting a fixed ρ. This figure includes two estimations computed with different values for ρ: a low value (left part) and a high one (right part). Higher values of ρ produce bigger oscillations and better capability of adaptation. When using a low value for ρ the procedure requires 66 observations to reach a value under 0.3 (sample 2001 of the sequence) but only 44 are required in the case of a high value.

To compare with other procedures we have carried out an extensive experiment in which we have repeated 100 times the generation and estimation of the probabilities in series of 4000 observations with 3 changes, as it was described above. In each situation we have computed the averaged *Kullback-Leibler* distance between the true parameter and the approximations:

$$KL = \frac{1}{4000} \sum_{i=1}^{4000} \left(\mu_i \log \left(\frac{\mu_i}{\hat{\mu_i}} \right) + (1-\mu_i) \log \left(\frac{1-\mu_i}{1-\hat{\mu_i}} \right) \right) \tag{35}$$

Many classic methods are based on keeping a sliding window with the last m observations and computing an statistical test to check whether there was a change in the last k ($k < m$) observations. In the case of a positive result the window is reduced in order to keep the last k observations only. To set the notation, let us consider that the current observation is X_n and that $l_1 = n - k + 1$ and $l_2 = n - m + 1$ are the starting points of the last k and m observations respectively. The last m observations are separated into two subwindows: one with the last k observations and the other with the remaining $m - k$ ones. Let us consider that W is a variable with two values indicating the subwindow to which each observation belongs.

X will denote the variable taking as values the last observation. To test the homogeneity of X in both subwindows an independence test of X and W is computed. If these variables are independent, then the knowledge about the subwindow does not provide information about X, and therefore X should be considered homogeneous. If the result of the test points out dependence, then the subwindow gives information about X and this implies the existence of a change in the last k observations. To carry out the test we compute the contingency table of X and W:

	$X = 0$	$X = 1$
$W = 0$	$\overline{N}_{l_2(l_1-1)}$	$N_{l_2(l_1-1)}$
$W = 1$	$\overline{N}_{l_1 n}$	$N_{l_1 n}$

We have carried out two different procedures for the independence test:

- A frequentist Chi-square test when the absolute frequencies in the table are greater than 5, and a Fisher exact test otherwise (R implementation).

- A Bayesian independence test [1, 6]. This test computes a score for the case of dependence, S_{Dep}:

$$S_{Dep} = \frac{\Gamma(N_{l_2(l_1-1)} + s/4)}{\Gamma(s/4)} \frac{\Gamma(\overline{N}_{l_2(l_1-1)} + s/4)}{\Gamma(s/4)} \frac{\Gamma(N_{l_1 n} + s/4)}{\Gamma(s/4)} \frac{\Gamma(\overline{N}_{l_1 n} + s/4)}{\Gamma(s/4)} \frac{\Gamma(s/2)}{\Gamma(k + s/2)} \frac{\Gamma(s/2)}{\Gamma((m-k) + s/2)}, \tag{36}$$

  where Γ is the Gamma function and s is the global sample size parameter ($s = 4$ in our case). Then, it computes a score for the case of independence, S_{Indep}:

$$S_{Indep} = \frac{\Gamma(N_{l_1n} + s/2)}{\Gamma(s/2)} \frac{\Gamma(\overline{N}_{l_1n} + s/2)}{\Gamma(s/2)} \frac{\Gamma(s)}{\Gamma(m+s)}. \tag{37}$$

Then, the ratio S_{Indep}/S_{Dep} is compared with the significance level (α). If the ratio is less than the significance level, the test determines dependence and the window is reduced.

The estimation has been done with the following algorithms (all of them implemented in R). The chosen parameters are based on preliminary experiments except for our procedure (only an initial selection is carried out) and ADWIN (the parameters are fixed to the values suggested by authors [2]).

- **BAF01**: Our approximate Bayesian approach with fixed values for the parameters: $\rho = 0.01$ and $K = 100$.
- **BAF001**: Our approximate Bayesian approach with fixed values: $\rho = 0.001$ and $K = 100$.
- **BAF0001**: Our approximate Bayesian approach with fixed values: $\rho = 0.0001$ and $K = 100$.
- **BFV1**: Our approximate Bayesian approach with values for ρ in the interval $[0.0001, 0.01], K = 100, L = 10$ and an uniform prior density for ρ.
- **BFV2**: Our approximate Bayesian approach with values for ρ in the interval $[0.0001, 0.01], K = 100, L = 10$ and the density concentrated in the low values of ρ.
- **FW**: A method with a fix window which only takes into account the last 68 observations.
- **FF**: A method based on a fix forgetting factor of 0.97 (the past frequencies are multiplied by 0.97 before new observations are added).
- **SWB**: Method based on a sliding window with the last m observations, in which a Bayesian statistical test is carried out for the last $k = 28$ observations in order to reduce the current window. The significance value is 0.04.
- **SWF**: Method based on a sliding window with the last m observations, in which a Bayesian statistical test is carried out for the last $k = 28$ observations in order to reduce the current window. The significance value is 0.006.

- **ADWIN**: Method proposed by Bifet and Gavalda [2] with suggested parameter $\delta = 0.2$ based on a sliding window.
- **SWMTF**: This procedure uses a sliding window as well, but following the strategy of Bifet and Gavalda [2], performing multiple tests instead of only one test. More precisely, it uses the procedure described in Alg. 1 (the current window is denoted by $[X_{l_1} \dots X_n]$ and therefore X_{l_1} will be the first sample of the sliding window).

Algorithm 1 SWMTF strategy

```
 1: $end \leftarrow false$
 2: $C = [X_{l_1} \dots X_n]$ (current window)
 3: $l_2 \leftarrow l_1 + 1$
 4: while ($l_2 \leq n$ and $end == false$) do
 5:     $C_1 = [X_{l_1} \dots X_{l_2} - 1]$, $C2 = [X_{l_2} \dots X_n]$
 6:     Test homogeneity of $C_1$ and $C_2$ with independence test
 7:     if dependence is detected then
 8:         $l_1 \leftarrow l_1 + 1$ (move forward sliding window)
 9:         $end \leftarrow true$ (no needed to check more partitions)
10:     else
11:         $l_2 \leftarrow l_2 + 1$ (consider next partition)
12:     end if
13: end while
```

It carries out much more tests than **SWF** approach by considering all the partitions of the current window (loop from lines 4 to 13 in Alg. 1). Each time the test determines dependence only one observation is removed from the current window (line 8). Then the process starts again by considering all the partitions of the new current window. The significance level of the test has been set to $0.01/\log(m)$, where m is the number of observations included in the current window. The division is done to correct the significance level taking into account the number of statistical tests.

- **SWMTFIn**: This is a modification of **SWMTF** with the idea of making it faster. It considers a parameter $l = 5$ involved in the determination of the possible partitions of the current window $[X_{l_1} \dots X_n]$ in subwindows $[X_{l_1} \dots X_{l_2-1}], [X_{l_2} \dots X_n]$. Therefore this procedure only considers partitions obtained by increasing the cutting point (l_2) by l (with this change m/l partitions are considered instead of m). Accordingly, the significance level is set to $0.01/\log(m/l)$. In addition

l observations are removed when the result is dependence by making $l_1 \leftarrow l_1 + l$.

- **SWMTB**: Similar to **SWMTF**, with a single difference in the significance level of the Bayesian test (0.06 in this case).
- **SWMTBIn**: Similar to **SWMTFIn** but changing the significance level of the Bayesian test (0.08).

Table 1 shows the averages of Kullback-Leibler distance and computation time (in seconds) for 100 repetitions of the experiment. The box plot in Figure 5 represents the average of Kullback-Leibler distances.

Method	KL distance	Computation time
BAF01	0.010639284	4.955781e-01
BAF001	**0.005528802**	4.085955e-01
BAF0001	0.006380337	4.807440e-01
BFV1	0.006033974	4.561622
BFV2	0.005858679	4.538627
FW	0.013487189	1.512650e-02
FF	0.012317891	**6.546898e-02**
SWB	0.009522752	3.447633e-01
SWF	0.009597048	2.713945
ADWIN	0.007932044	1.783343e+01
SWMTF	0.007322510	1.310755e+03
SWMTFIn	0.007361513	1.155341e+02
SWMTB	0.0078110880	1.757771e+02
SWMTBIn	0.009178993	2.744234e+01

Table 1: Averages of Kullback-Leibler distance and computation time

A Friedman test for differences in error (*Kullback-Leibler* distance) is highly significant with *p-value* $< 2.2e-16$. We have also carried out a post hoc Friedman Nemenyi test finding the following facts:

- **BAF001** (our approximate method with fixed $\rho = 0.001$) is the best procedure and it is significantly better than any other one, except **BFV2**. When the value of ρ is not the true one, then the performance deteriorates, specially when it is too high ($\rho = 0.01$). To observe the performance of this method as a function of ρ we have carried out a complementary experiment in which we have measured the average

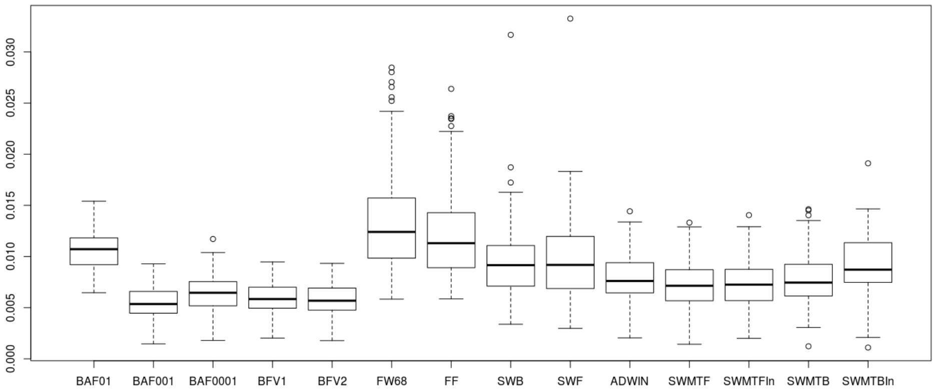


Figure 5: Averages of Kullback-Leibler distance.

error as a function of ρ. The results can be seen in Figure 6. This graphic shows that the minimum is related to $\rho = 0.001$ and how the performance declines specially for high values. This method is also very fast. The procedures **FF** (fixed size sliding window) and **FW** (constant forgetting factor) are faster, but their performance is much worse.

- Our methods with variable ρ, **BFV1** and **BFV2**, show a very good performance, even with a low number of discretization points ($L = 10$). For example, **BFV2** is not significantly worse than the best one **BAF001** and it is significantly better than the rest of methods except **BAF001** and **BAF0001** (both of them present lowest p-value in the comparisons: $6.9e-06$). Considering **BFV1** and **BFV2**, the last one shows lower error although the difference is not significant. With respect to time, both **BFV1** and **BFV2** are slower than **BAF01**, **BAF001** and **BAF0001**; anyway, they faster than **ADWIN** and its variants with Bayesian and frequentist methods.

  We have carried out a complementary experiment to analyze the impact of changing the number of discretization points ($L = 20$) for ρ density. The results in average error are 0.005995166 and 0.005852139 for **BFV1** and **BFV2**, respectively. We can see that error decreases using more discretization points, but at the cost of increasing the average times: 5.58744716 and 5.59621713 respectively. Again better performance is obtained when considering prior densities concentrated in lower values.

- Methods based on a sliding window of fixed size or a constant forgetting factor (**FW** and **FF**) are the fastest ones, but their performance

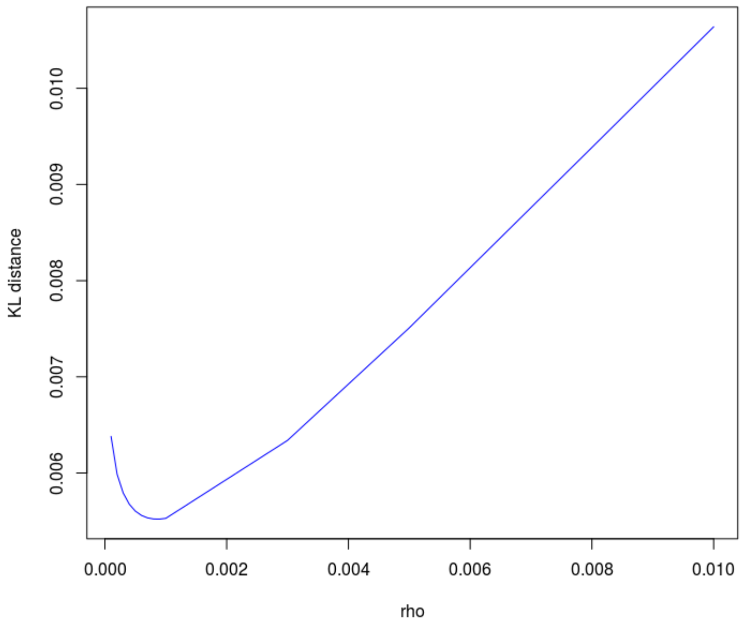
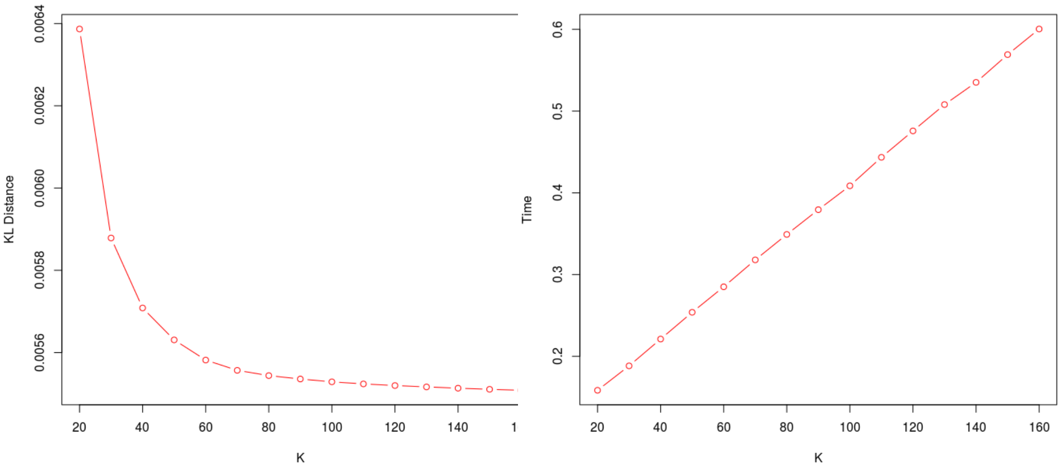


Figure 6: Our procedure as a function of a fixed ρ

is worse (significant difference with respect to the rest of procedures except **BAF01**).

- Procedures based on a sliding window of variable size using statistical tests, **SWF** and **SWB**, are better than the method based on a sliding window of constant size, but the time is greater. The use of a Bayesian statistical test produces a lower error (with non significant difference) and lower computation times.
- **ADWIN** is even slower than **SWF** and **SWB**, but the performance is better (significant difference with **SWF** and non-significant with **SWB**).
- Our implementations of the **ADWIN** strategy with Bayesian and frequentist tests (**SWMTF**, **SWMTFIn**, **SWMTB** and **SWMTBIn**) are not significantly better than **ADWIN**, being much slower at the same time, specially in the case of frequentist tests.

A last experiment tests the evolution of Kulback-Leibler distance and computation times as a function of K in our approximate method with a fix value $\rho = 0.001$ (100 repetitions). The results are represented in Figures 7a and 7b for error and time respectively. Respect to the error it is observed an exponential decay to a lower threshold (somewhat about 0.0055) even if the differences are not very important in absolute terms. The cost in time is linear with respect to K as expected.



(a) Kulback-Leibler distance

(b) Time

Figure 7: Our procedure with fixed ρ as a function of the number of points for discretization K

4. Conclusions

In this paper we have proposed a Bayesian approach for detecting abrupt concept drift, which has been solved, both exactly and approximately. Only the approximate method was implemented and compared with other existing approaches. The results show an excellent performance with respect to error of the estimations and computation time. We have considered two procedures: one based on a fix rate of change ρ and the other based on an unknown value for ρ. The former is faster, but it is recommended only when the value of ρ is known in advance (at least a close value). In other case, we should use the later one. It also has the chance of providing a real time estimation of the unknown value of ρ. The parameters of our methods (K and L) allow to control the performance of the algorithm considering error versus computation time. In the case of the method using a fix value of ρ, the values of the parameters can graduate the capacity of adaption to new changes against the stability in absence of changes.

In the future, we will adapt our procedure to gradual changes in the probabilities, by using a similar method but estimating a linear trend, which can have abrupt changes. The proposals in this paper have been described for the case of a binomial probability and we plan to extend them for multinomial probabilities and general conditional probability distributions.

Acknowledgments

This research was supported by the Spanish Ministry of Education and Science under project TIN2016-77902-C3-2-P, and the European Regional Development Fund (FEDER).

[1] Joaquín Abellán, Manuel Gómez-Olmedo, and Serafín Moral. Some variations on the PC algorithm. In M. Studeny and J. Vomlel, editors, *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM' 06)*, pages 1–8, 2006.

[2] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 443–448. SIAM, 2007.

[3] Rafael Cabañas, Andrés Cano, Manuel Gómez-Olmedo, Andrés R. Masegosa, and Serafín Moral. *Virtual Subconcept Drift Detection in Discrete Data Using Probabilistic Graphical Models*, volume 885 of *Communications in Computer and Information Science.* Springer-Verlag, 2018.

[4] Joao Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Brazilian symposium on artificial intelligence*, pages 286–295. Springer, 2004.

[5] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):44, 2014.

[6] David Heckerman. A tutorial on learning with Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research, 1995.

[7] Antti Honkela and Harri Valpola. On-line variational Bayesian learning. In *4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 803–808, 2003.

[8] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD*

international conference on Knowledge discovery and data mining, pages 97–106. ACM, 2001.

[9] Ivan Koychev. Gradual forgetting for adaptation to concept drift. Proceedings of ECAI 2000 Workshop on Current Issues in Spatio-Temporal Reasoning,, 2000.

[10] Andrés Masegosa, Thomas D Nielsen, Helge Langseth, Dario Ramos-Lopez, Antonio Salmerón, and Anders L Madsen. Bayesian models of data streams with hierarchical power priors. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning; PMLR*, volume 70, pages 2334–2343, 2017.

[11] Kevin Patrick Murphy and Stuart Russell. Dynamic Bayesian networks: representation, inference and learning. 2002.

[12] Judea Pearl. *Probabilistic Reasoning with Intelligent Systems.* Morgan & Kaufman, San Mateo, 1988.

[13] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: current status and future directions. *Neurocomputing*, 239:39–57, 2017.

[14] Moamar Sayed-Mouchaweh. *Learning from data streams in dynamic environments.* Springer, 2016.