Sequential Dynamic Classification
Using Latent Variable Models

SEUNG MIN LEE AND STEPHEN J. ROBERTS*

Pattern Analysis & Machine Learning Research Group, Department of Engineering Science, University of Oxford, Oxford, UK
*Corresponding author: sjrob@robots.ox.ac.uk

Adaptive classification is an important online problem in data analysis. The nonlinear and nonstationary nature of much data makes standard static approaches unsuitable. In this paper, we propose a set of sequential dynamic classification algorithms based on extension of nonlinear variants of Bayesian Kalman processes and dynamic generalized linear models. The approaches are shown to work well not only in their ability to track changes in the underlying decision surfaces but also in their ability to handle in a principled manner missing data. We investigate both situations in which target labels are unobserved and also where incoming sensor data are unavailable. We extend the models to allow for active label requesting for use in situations in which there is a cost associated with such information and hence a fully labelled target set is prohibitive.

Keywords: dynamic classification; sequential Bayesian inference; partially observed data; nonstationary decision processes

Received 11 August 2009; revised 30 November 2009
Handling editor: Nick Jennings

1. INTRODUCTION

Many data analysis problems require online learning. In this paper we consider the problem of sequential classification in environments in which the data are by nature noisy, nonlinear and nonstationary. Furthermore, we consider scenarios in which components of the data are delayed and/or missing.

We use the framework of dynamic models to tackle this problem domain, in particular we use generic state-space models. For a binary classification problem, a logistic regression model is trained to produce posterior class probabilities. By modelling the parameters (weights) of the logistic regression classifier as time-varying parameters, the nonstationary nature of the problem can be captured. In [1, 2] the extended Kalman filter (EKF) is employed to solve the nonlinear dynamic classification model. In this paper we extend this approach and demonstrate how the unscented Kalman filter (UKF) [3, 4] can be applied to the adaptive classification problem. Alternatively, this problem can be regarded as a special case of dynamic generalized linear models [5]. Binary outcomes are known to have a binomial distribution, which is then modelled by a generalized linear model. By allowing the parametric weights of the model to evolve over time, we can extend canonical generalized linear models to the dynamic classification problem. The goal of dynamic streaming classification is intrinsically similar to that of classification in the presence of concept drifts, normally handled using ensembles of weak classifiers that are adaptively combined so as to form robust decisions in the presence of drifts [6–8]. This paper departs somewhat from this approach in that we aim to adapt the base learning model itself. Furthermore, we aim to do this in an optimal (Bayesian) framework. A combined approach in which not only the base learners, but also the ensemble mixture are adaptive is explored in [8].

In Section 2 we briefly review the logistic regression model and its nonstationary version using the framework of the standard EKF. In Section 3 we extend the algorithm for parameter inference in the nonstationary logistic regression model by use of the UKF. Section 4 reviews the dynamic generalized linear model and we examine how dynamic classification can be dealt with in that context. In Section 5 we apply the proposed models to synthetic and real data and carry out experiments to evaluate the performance of these models in cases of incompletely observed information. Section 6 concludes the paper with discussions about future work.

THE COMPUTER JOURNAL, Vol. 53 No. 9, 2010
2. DYNAMIC LOGISTIC REGRESSION

2.1. Logistic regression

We consider a set of sequential ‘sensor’ observations which we denote by the vector $h$. Associated with each of these observations (or feature) vectors is a target classification label which is denoted by the variable $y$. For the binary decision problem (which we concentrate on in this paper), these target labels are represented by $y = 0$ and $y = 1$. Our objective is to infer the class probability defined by

$$\pi(h) \triangleq \Pr(y = 1|h),$$

(1)

where $h$ is the given input vector of, for example, sensor observations. By definition, $\Pr(y = 0|h) = 1 - \pi(h)$. Logistic regression models the class probability as follows:

$$\pi(h) = l(h^T w),$$

(2)

where $w$ is a vector of weights and $l(\cdot)$ is a logistic function defined by

$$l(h^T w) = \frac{\exp(h^T w)}{1 + \exp(h^T w)}.$$

(3)

Instead of using the input vector per se, we can consider an extended form as follows:

$$\varphi(h) = \left\{ \begin{array}{c} h \\ \phi(h) \\ 1 \end{array} \right\},$$

(4)

where $\phi$ represents a set of nonlinear kernel functions such as Gaussian radial basis functions. We may hence see this model as a dynamic Radial Basis Function classifier. The addition of the unity component in the above equation allows for a time varying offset (decision bias) to be modelled [2]. Throughout this report we consider, unless otherwise stated, a logistic function with activation $a = \varphi(h)^T w$, not $a = h^T w$. With this kernel input vector it is possible to find a nonlinear boundary between two classes.

2.1.1. Bayesian approach

Suppose that the parameter or weight vector $w$ has a probability distribution $p(w)$. The class probability given in Equation (2) ignores any uncertainty associated with the weight vector, being in effect conditioned on a most likely point value. A Bayesian approach takes into account the full distribution, and accordingly in cases where uncertainty is intrinsically high, moderates the class probability such that less confident decisions are made. If we take the distribution of $w$ is a multivariate Gaussian with mean $\hat{w}$ and covariance $P_w$, then the distribution of activation $a$ is also a Gaussian with mean $\hat{a}$

and variance $c^2$ defined by

$$\hat{a} = \varphi(h)^T \hat{w},$$

(5)

$$c^2 = \varphi^T(h) P_w \varphi(h).$$

(6)

By integrating $a$ out, the posterior class probability $\pi$ is moderated; that is,

$$\pi = \int l(a) p(a) \, da.$$ 

(7)

Since it is impossible to compute the probability analytically, it is approximated by [9]:

$$\pi \approx l(\kappa(c^2) \hat{a}),$$

(8)

where

$$\kappa(c^2) = \left(1 + \frac{\pi c^2}{8}\right)^{-1/2}.$$

(9)

Here $\pi$ represents the standard mathematical constant approximately equal to 3.14159. If the variance of activation is close to zero, the extent of moderation is negligible. However, as the variance becomes very large, the posterior class probability (for binary systems) is moderated to lie closer to the prior at 0.5.

2.2. Nonstationary logistic regression

In the stationary logistic regression model given in Equation (2), the weight vector $w$ is assumed to be static. In [1] a dynamic logistic regression model is proposed in which the weight vector evolves according to a Gaussian random walk which has the following form:

$$\pi_t = l(\varphi(h_t)^T w_t),$$

(10)

$$w_t = w_{t-1} + v_t,$$

(11)

where $v_t$ is a diffusion variable assumed to be a Gaussian with mean $0$ and covariance $q I$. For notational convenience we will use the notation $\varphi(h_t)$ to $\varphi$. Also we note that the weight vector of the above equation can be seen as the state variable of a dynamical systems model.

In the dynamic logistic regression the nonstationarity is captured by the time-evolving state variable $w_t$, and the binary decision is made according to the logistic function. To perform adaptive classification, we have to estimate the state variable $w_t$ in an online manner, and this inference problem is dealt with in the following subsections.

2.2.1. Dynamic classification using EKF

The EKF is the most commonly used filter for handling nonlinear systems [10]. It is based on linearizing a nonlinear function; the nonlinear function is approximated by the first two
make it suitable for the classification context by maximizing the ‘evidence’ of model predictions instead. For example, Lowne et al. [2] suggests updating state noise variance \( q_t \) according to

\[ q_t = \max \{ u_{t|t} - u_{t|t-1}, 0 \}, \]

where

\[ u_{t|t} = \tilde{\pi}_{t|t} (1 - \tilde{\pi}_{t|t}) \]

is the uncertainty (variance) in the predictive Bernoulli distribution.

3. Dynamic Classification using the UKF

3.1. The UKF

The UKF is proposed in [3, 4]. The UKF is based on the unscented transformation, which is a method for calculating the statistics of a random variable that undergoes a nonlinear transformation and is founded on the principle that it is easier to approximate a probability distribution than an arbitrary nonlinear function [3]. The unscented transformation resembles Monte Carlo sampling methods in that a number of points are selected and the points are propagated through a nonlinear function. However, the unscented transformation selects a set of points (referred to as sigma points) not randomly but deterministically, so that they preserve the statistics (e.g. mean and variance) of their underlying distribution.

Suppose that we have two random variables \( \mathbf{a} \) and \( \mathbf{c} \) and a nonlinear function \( \mathbf{b}(\cdot) \). We wish to infer the probability density function for \( \mathbf{c} = \mathbf{b}(\mathbf{a}) \); we denote the means and covariances of \( \mathbf{a} \) and \( \mathbf{c} \) by \( \tilde{\mathbf{a}} \), \( \tilde{\mathbf{c}} \), \( \mathbf{P}_\mathbf{a} \) and \( \mathbf{P}_\mathbf{c} \), respectively. We first select \( 2L + 1 \) (\( L \) is the dimension of \( \mathbf{a} \)) sigma vectors \( A_i \) in the following deterministic way:

\[ A_0 = \tilde{\mathbf{a}}, \]
\[ A_i = \tilde{\mathbf{a}} + \sqrt{(L + \lambda) \mathbf{P}_\mathbf{a}} i, \quad i = 1, \ldots, L, \]
\[ A_i = \tilde{\mathbf{a}} - \sqrt{(L + \lambda) \mathbf{P}_\mathbf{a}} i, \quad i = L + 1, \ldots, 2L, \]

where \( \lambda = \alpha^2(L + \kappa) - L \) and \( \sqrt{(L + \lambda) \mathbf{P}_\mathbf{a}} \) represents the \( i \)th column of the matrix square root. The variables \( \alpha \) and \( \kappa \) are parameters of the method: the former is a scaling parameter that determines the spread of the sigma points around \( \tilde{\mathbf{a}} \) and is normally set to a small positive value, and the latter is a secondary scaling parameter and is normally set to 0 [11].

To compute the statistics of the transformed vector \( \mathbf{c} \), the sigma vectors \( A_i \) are propagated through the nonlinear function to provide another set of sigma vectors:

\[ C_i = \mathbf{b}(A_i), \quad i = 0, 1, \ldots, 2L. \]

The unscented transformation approximates the mean and covariance of \( \mathbf{c} \) as weighted sums of the propagated sigma
vectors according to
\[ \hat{c} \approx \sum_{i=0}^{2L} \omega_i^{(m)} C_i, \]  
\[ P_c \approx \sum_{i=0}^{2L} \omega_i^{(c)} (C_i - \hat{c})(C_i - \hat{c})^T. \]  

The weights \( \omega_i \) are given by
\[ \omega_0^{(m)} = \frac{\lambda}{L + \lambda}, \]
\[ \omega_0^{(c)} = \frac{\lambda}{L + \lambda} + 1 - \alpha^2 + \beta, \]
\[ \omega_i^{(m)} = \omega_i^{(c)} = \frac{1}{2(L + \lambda)}, \quad i = 1, \ldots, 2L, \]

where \( \beta \) is another parameter that incorporates prior knowledge of the distribution of \( a \). For Gaussian distributions and arbitrary smooth nonlinearities, \( \beta = 2 \) is optimal [11].

In the UKF, the prior and posterior means and covariances of the state variable of a dynamic nonlinear model are approximated by the unscented transformation. There are two major advantages of the UKF over the EKF. First, whereas the linearization method of the EKF approximates the true mean and covariance up to the first order, the mean and covariance obtained by the UKF, for Gaussians, are accurate up to third order for all smooth monotone nonlinearities, and for non-Gaussians up to at least the second order is guaranteed [12]. Secondly, despite the improvement in approximation accuracy, the computational cost required by the unscented transformation is equal to the linearization method of the EKF. A comparison between the linearization approach, the unscented transformation and a sampling method is excellently illustrated in [11].

### 3.2. Dynamic logistic regression using the UKF

We repeat the dynamic logistic regression model of the previous section, namely
\[ \pi_t = l(\varphi(h_t)^T w_t), \]
\[ w_t = w_{t-1} + v_t, \]

where \( w_t \) is a Gaussian state noise with mean 0 and covariance \( q_t I \). In Section 2.2.1 we derived an algorithm for solving this nonlinear system using the EKF. We here present an algorithm in which the UKF is applied to the dynamic classification problem. Assuming that the posterior mean and covariance of the state variable \( w_{t-1} \) are \( \hat{w}_{t-1} \) and \( P_{t-1} \), respectively, the two moments of the prior state distribution at time \( t \) are
\[ \hat{w}_{t|t-1} = \hat{w}_{t-1} + K (y_t - \hat{q}_{t|t-1}), \]
\[ P_{t|t-1} = P_{t-1} + q_t I, \]
in which \( q_t \) is inferred using Equations (23) and (24).

To compute the mean and variance of the one-step ahead predictive distribution, the unscented transformation method is employed. We first draw \( 2L + 1 \) sigma vectors for the distribution of \( w_t \) from its estimated mean and covariance, \( \hat{w}_{t|t-1} \) and \( P_{t|t-1} \):
\[ W_{t|t-1} = \{ W_{t|t-1}^0, \ldots, W_{t|t-1}^L \}, \]
\[ \approx \{ \hat{w}_{t|t-1}, \hat{w}_{t|t-1} + (\sqrt{(L + \lambda)P_{t|t-1}}) \}, \]
\[ \hat{w}_{t|t-1} + (\sqrt{(L + \lambda)P_{t|t-1}}) \] is the \( i \)th column of the matrix square root.

The sigma vectors are propagated through the logistic function:
\[ \Pi_{t|t-1} = l(\varphi_i^T (\hat{W}_{t|t-1})), \quad i = 0, 1, \ldots, 2L. \]

Using these propagated sigma vectors we can compute a one-step ahead predicted class probability and its variance, which are given by
\[ \hat{\pi}_{t|t-1} = \sum_{i=0}^{2L} C_{i|t-1}^{(m)} \Pi_{t|t-1}, \]
\[ P_{jy} = \sum_{i=0}^{2L} C_{i|t-1}^{(c)} (\Pi_{t|t-1} - \hat{\pi}_{t|t-1})^2 + r_j. \]

The prior observation noise variance \( r_j \) can be either fixed to a static small value or estimated according to \( r_j = \hat{\pi}_{t|t-1} (1 - \hat{\pi}_{t|t-1}) \). In addition, we compute the covariance between the state variable and the class probability according to
\[ P_{wy} = \sum_{i=0}^{2L} C_{i|t-1}^{(c)} (\Pi_{t|t-1} - \hat{\pi}_{t|t-1}) (\Pi_{t|t-1} - \hat{\pi}_{t|t-1})^T. \]

The weights \( C_{i|t-1}^{(m)} \) and \( C_{i|t-1}^{(c)} \) are given by
\[ C_{0|t-1}^{(m)} = \frac{\lambda}{L + \lambda}, \]
\[ C_{0|t-1}^{(c)} = \frac{\lambda}{L + \lambda} + 1 - \alpha^2 + \beta, \]
\[ C_{i|t-1}^{(m)} = C_{i|t-1}^{(c)} = \frac{1}{2(L + \lambda)}, \quad i = 1, \ldots, 2L. \]

After observing a new label \( y_t \), the state variable distribution is updated. The posterior mean and covariance of \( w_t \) are computed according to
\[ \hat{w}_{t|t} = \hat{w}_{t|t-1} + K (y_t - \hat{q}_{t|t-1}), \]
\[ P_{t|t} = P_{t|t-1} - K_t P_{yy} K_t^T, \]
where the Kalman gain is
\[ K_t = P_{wy} P_{yy}^{-1}. \]

In Section 2.1.1 we introduced a Bayesian approach to the logistic regression model in order to take into account the
uncertainty associated with the state variable \( w \). We note that the UKF represents the distribution of \( w \) by selecting sigma points from the distribution. Once the probability density function over \( w \) has been inferred using the unscented filter, the dynamic classification scheme proceeds identically as for the EKF.

4. Dynamic Generalized Linear Models

Generalized linear models are a powerful method for analysing distributions in the exponential family. They can be generalized to capture the nonstationary nature of data such as we consider in this paper by allowing the model parameters to evolve over time, giving rise to a family of approaches which we refer to as dynamic generalized linear models. Since a binary label sequence follows a Bernoulli distribution, which is a special case of a binomial distribution and in the exponential family, the dynamic classification problem can be solved in the framework of dynamic generalized linear models. In this section we review the generalized linear model itself and extend it to allow for dynamics using a Bayesian perspective.

4.1. Generalized Linear Models

In an ordinary linear regression model an observation random variable \( y \) is modelled by

\[
y = \mu + n,
\]

(43)

where \( \mu = h^T w \). Note that, as in previous sections, \( h \) represents a known input variable and \( w \) is a model parameter vector. In addition, \( n \) is a Gaussian noise variable with mean 0 and variance \( \sigma^2 \). The observation variable is also taken to have a multivariate Gaussian distribution with the following mean and variance:

\[
E(Y) = \mu, \quad \Var(Y) = \sigma^2.
\]

(44)

Note that the mean is a linear function of the predictor variable, \( h \).

In a generalized linear model two generalizations are made from the standard linear model above: (i) the probability distribution of the observation variable is not restricted to be a Gaussian but is allowed to have any distribution in the exponential family; (ii) it is not the mean of the variable but a function of the mean that is linearly related to the predictor variable.

An observation random variable, \( y \), that has an associated distribution in the exponential family takes the form

\[
p(y|\theta, \phi) = \exp \left( \frac{d(y) \theta - b(\theta)}{a(\phi)} + c(y, \phi) \right),
\]

(45)

where \( \theta \) is referred to as the canonical parameter, \( \phi \) the dispersion parameter and \( a, b, c, d \) are all known functions for a given parametric distribution. The mean and variance of the variable are related to the canonical and dispersion parameters as follows:

\[
E(d(y)|\theta, \phi) = \mu = b'(\theta),
\]

(46)

\[
\Var(d(y)|\theta, \phi) = b''(\theta)a(\phi).
\]

(47)

Members of the exponential family include Gaussian, binomial, Poisson, exponential, gamma, inverse Gaussian distributions etc. In the ordinary Gaussian linear model the mean is linearly related to the predictor variable, \( h \); that is, \( \mu = h^Tw \).

(48)

The right-hand side of the equation is denoted by \( \eta \triangleq h^Tw \), and \( \eta \) is referred to as the linear predictor. Note that the linear predictor and the mean of a Gaussian variable can take any real values. However, this may or may not be true for other distributions in the exponential family; for instance, the mean of a Poisson distribution must be non-negative. A link function \( g(.) \) is used to relate the linear predictor to the mean, that is

\[
g(\mu) = \eta = h^Tw.
\]

(49)

The mean can therefore be computed simply by inverting the link function:

\[
\mu = g^{-1}(\eta) = g^{-1}(h^Tw).
\]

(50)

Comparing Equation (48) with Equation (49), we can see that the link function for the standard Gaussian linear model is an identity function. A particular form of the link function is of importance. A link function that makes the linear predictor \( \eta \) equal to the canonical parameter \( \theta \) is referred to as canonical link:

\[
g(\mu) = \eta = \theta.
\]

(51)

The canonical link is important due to the existence of minimal sufficient statistics for the model parameters [13], thereby being used in most cases.

4.2. Exponential Family Dynamic Models

Suppose that an observation variable \( y \), that has a distribution in the exponential family, is observed over time. The observation model of a dynamic generalized linear model is of the form

\[
p(y_t|\theta_t, \phi_t) = \exp \left( \frac{d(y_t) \theta_t - b(\theta_t)}{a(\phi)} + c(y_t, \phi_t) \right),
\]

(52)

with the following link equation

\[
g(\mu_t) = \eta_t = h_t^Tw_t.
\]

(53)

where \( \mu_t = E(y_t) \). If \( y_t \) has a Gaussian distribution, then the observation model and link equation can be merged into a single observation equation as follows:

\[
y_t = h_t^Tw_t + n_t,
\]

(54)

where \( n_t \sim N(0, r_t) \) and \( \mu_t = h_t^Tw_t \). This is equivalent to the observation equation of a standard dynamic linear model.
The state variable (or model parameter vector) $\mathbf{w}_t$ is assumed to not be static but to evolve over time via the state process as

$$\mathbf{w}_t = \mathbf{F}_t \mathbf{w}_{t-1} + \mathbf{v}_t,$$  \hspace{1cm} (55)

where $\mathbf{v}_t \sim (\mathbf{0}, \mathbf{Q}_t)$. Note that the notation ‘$a \sim (b, c)$’ represents that the mean and variance of $a$ are $b$ and $c$, respectively, without specifying the probability density function of the variable $a$.

We first consider the sequential inference method for the state variable in the Bayesian setting [5]. It is an approximate method in the sense that distributions are only specified in terms of their first and second moments without full distributional information. Suppose that, given $D_{t-1}$, state variable $\mathbf{w}_{t-1}$ has mean $\hat{\mathbf{w}}_{t-1}$ and covariance $\mathbf{P}_{t-1}$; that is,

$$\mathbf{w}_{t-1} | D_{t-1} \sim (\hat{\mathbf{w}}_{t-1}, \mathbf{P}_{t-1}).$$  \hspace{1cm} (56)

Since the state process of Equation (55) is linear, the prior distribution of the state variable at time $t$ can be readily computed as

$$\mathbf{w}_t | D_{t-1} \sim (\hat{\mathbf{w}}_{t}, \mathbf{P}_{t})$$  \hspace{1cm} (57)

where

$$\hat{\mathbf{w}}_{t} = \mathbf{F}_t \hat{\mathbf{w}}_{t-1},$$  \hspace{1cm} (58)

$$\mathbf{P}_{t} = \mathbf{F}_t \mathbf{P}_{t-1} \mathbf{F}^T_t + \mathbf{Q}_t.$$  \hspace{1cm} (59)

As the linear predictor (i.e. $\eta_t = g(\mu_t) = \mathbf{h}_t^T \mathbf{w}_t$) is linearly related to the state variable, we can readily compute the prior distribution of $\eta_t$, given $D_{t-1}$. Therefore, may evaluate the joint distribution of $\eta_t$ and $\mathbf{w}_t$ as

$$\begin{bmatrix} \eta_t \\ \mathbf{w}_t \end{bmatrix} \bar{D}_{t-1} \sim \begin{bmatrix} \hat{\mathbf{w}}_{t-1} \\ \mathbf{P}_{t-1} \end{bmatrix},$$  \hspace{1cm} (60)

\[ \begin{bmatrix} \eta_t \\ \mathbf{w}_t \end{bmatrix} \bar{D}_{t-1} \sim \begin{bmatrix} \hat{\mathbf{w}}_{t-1} \\ \mathbf{P}_{t-1} \end{bmatrix}, \]

where

$$\hat{\mathbf{w}}_{t-1} = \mathbf{h}_t^T \hat{\mathbf{w}}_{t-1},$$  \hspace{1cm} (61)

$$S_{t-1} = \mathbf{h}_t^T \mathbf{P}_{t-1} \mathbf{h}_t.$$  \hspace{1cm} (62)

The one-step ahead forecast distribution (i.e. the distribution of $y_t$ given $D_{t-1}$) can be obtained by marginalizing over the canonical parameter assuming that the dispersion parameter is known:

$$p(y_t | D_{t-1}) = \int p(y_t | \theta_t) p(\theta_t | D_{t-1}) \, d\theta_t.$$  \hspace{1cm} (63)

To solve the equation we need to specify the prior distribution of the canonical parameter, $p(\theta_t | D_{t-1})$. A conjugate prior is widely used because it allows us to derive the forecast distribution analytically. The conjugate prior distribution is of form

$$p(\theta_t | D_{t-1}) = \omega(k_t, m_t) \exp(k_t \theta_t - m_t b(\theta_t)),$$  \hspace{1cm} (64)

where $\omega$ is a known function providing the normalizing constant, and $k_t$ and $m_t$ are hyperparameters of the conjugate prior.

The complete form of the one-step ahead forecast distribution may hence be obtained in the following way:

$$p(y_t | D_{t-1}) = \int p(y_t | \theta_t) p(\theta_t | D_{t-1}) \, d\theta_t$$

$$= \int \omega(k_t, m_t) \exp(k_t \theta_t - m_t b(\theta_t)) \times \exp \left( \frac{d(y_t) \theta_t - b(\theta_t)}{a(\phi_t)} + c(y_t, \phi_t) \right) \, d\theta_t$$

$$= \omega(k_t, m_t) \exp(c(y_t, \phi_t)) \times \exp \left( c(y_t, \phi_t) \right).$$  \hspace{1cm} (65)

To employ the one-step ahead forecast distribution we need to determine the values of the hyperparameters of the conjugate prior, $k_t$ and $m_t$. If a canonical link is used, which ensures $\theta_t = \eta_t$, then we know that

$$E(\theta_t | D_{t-1}) = E(\eta_t | D_{t-1}) = \hat{\eta}_{t},$$  \hspace{1cm} (66)

$$\text{Var}(\theta_t | D_{t-1}) = \text{Var}(\eta_t | D_{t-1}) = S_{t-1}.$$  \hspace{1cm} (67)

From the form of the conjugate prior in Equation (64), we can obtain $E(\theta_t | D_{t-1})$ and $\text{Var}(\theta_t | D_{t-1})$ in terms of the hyperparameters. In addition, we know the values of $\hat{\eta}_{t-1}$ and $S_{t-1}$, which are given in Equation (62). We can, therefore, calculate the values of the hyperparameters.

Via Bayes’ theorem, the posterior distribution of the canonical parameter $\theta_t$ can be represented by

$$p(\theta_t | D_t) \propto p(\theta_t | D_{t-1}) p(y_t | \theta_t),$$  \hspace{1cm} (68)

which we may evaluate as

$$p(\theta_t | D_t) = \omega \left( k_t + \frac{d(y_t)}{a(\phi_t)}, m_t + \frac{1}{a(\phi_t)} \right) \times \exp \left( \left( k_t + \frac{d(y_t)}{a(\phi_t)} \right) \theta_t - \left( m_t + \frac{1}{a(\phi_t)} \right) b(\theta_t) \right).$$  \hspace{1cm} (69)

Since using the canonical link makes the posterior distribution of $\eta_t$ equivalent to that of $\theta_t$, we obtain the posterior distribution of $\eta_t$ as

$$\eta_t | D_t \sim (\hat{\eta}_t, S_t),$$  \hspace{1cm} (70)

where

$$\hat{\eta}_t = E(\eta_t | D_t) = E(\theta_t | D_t),$$  \hspace{1cm} (71)

$$S_t = \text{Var}(\eta_t | D_t) = \text{Var}(\theta_t | D_t).$$  \hspace{1cm} (72)

We note that $E(\theta_t | D_t)$ and $\text{Var}(\theta_t | D_t)$ can be computed from Equation (69).
The mean and covariance of the state variable may therefore be rewritten as

\[
p(w_t|D_t) = \int p(\eta_t, w_t|D_t) \, d\eta_t
\]

\[
\propto \int p(\eta_t, w_t|D_{t-1}) \, p(y_t|\eta_t) \, d\eta_t
\]

\[
\propto \int p(w_t|\eta_t, D_{t-1}) p(\eta_t|D_{t-1}) p(y_t|\eta_t) \, d\eta_t
\]

\[
\propto \int p(w_t|\eta_t, D_{t-1}) p(\eta_t|D_t) \, d\eta_t.
\] (73)

The first two moments of \(p(\eta_t|D_t)\) are given in Equation (72). The mean and covariance of \(p(w_t|\eta_t, D_{t-1})\) are estimated using linear Bayesian estimation [5], and are given by

\[
\hat{E}(w_t|\eta_t, D_{t-1}) = \hat{w}_{t|t-1} + \frac{P_{t|t-1}^T h}{S_{t|t-1}} (\eta_t - \hat{\eta}_{t|t-1}),
\] (74)

\[
\text{Var}(w_t|\eta_t, D_{t-1}) = P_{t|t-1} - \frac{P_{t|t-1}^T h h^T P_{t|t-1}}{S_{t|t-1}}.
\] (75)

Using these first two moments of \(p(w_t|\eta_t, D_{t-1})\) and \(p(\eta_t|D_t)\), we can finally compute the posterior mean and covariance of \(w_t\) as follows:

\[
\hat{w}_{t|t} = E(w_t|D_t) = E[E(w_t|\eta_t, D_{t-1})|D_t]
\]

\[
= E\left[ \hat{w}_{t|t-1} + \frac{P_{t|t-1}^T h}{S_{t|t-1}} (\eta_t - \hat{\eta}_{t|t-1}) \right| D_t]
\]

\[
= \hat{w}_{t|t-1} + \frac{P_{t|t-1}^T h}{S_{t|t-1}} (E(\eta_t|D_t) - \hat{\eta}_{t|t-1})
\]

\[
= \hat{w}_{t|t-1} + \frac{P_{t|t-1}^T h}{S_{t|t-1}} (\hat{\eta}_{t|t} - \hat{\eta}_{t|t-1})
\] (76)

and

\[
P_{t|t} = \text{Var}(w_t|D_t)
\]

\[
= \text{Var}[E(w_t|\eta_t, D_{t-1})|D_t] + \text{Var}[E(\eta_t|D_t)|D_t]
\]

\[
= \text{Var}\left[ \hat{w}_{t|t-1} + \frac{P_{t|t-1}^T h}{S_{t|t-1}} (\eta_t - \hat{\eta}_{t|t-1}) \right| D_t]
\]

\[
+ E\left[ P_{t|t-1} - \frac{P_{t|t-1}^T h h^T P_{t|t-1}}{S_{t|t-1}} \right| D_t]
\]

\[
= \frac{P_{t|t-1}^T h}{S_{t|t-1}} h^T P_{t|t-1} \text{Var}(\eta_t|D_t) + P_{t|t-1}
\]

\[
- \frac{P_{t|t-1}^T h h^T P_{t|t-1}}{S_{t|t-1}}
\]

\[
= P_{t|t-1} - \frac{P_{t|t-1}^T h}{S_{t|t-1}} h^T P_{t|t-1} \left( 1 - \frac{S_{t|t}}{S_{t|t-1}} \right).
\] (77)

4.3. Dynamic classification using the dynamic binomial model

The total number of ‘successes’ \(y\) in \(n\) independent experiments follows a binomial distribution with probability of success \(\pi\). The binomial distribution is a member of the exponential family defined by

\[
p(y|n, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y},
\] (78)

where \(0 \leq \pi \leq 1\), \(n = \{0, 1, 2, \ldots\}\) and \(y = \{0, 1, 2, \ldots, n\}\). We can rearrange the above density function in the following way:

\[
p(y|n, \pi) = \exp \left( n \left[ \frac{y}{n} \log \frac{\pi}{1 - \pi} - \log \frac{1}{1 - \pi} \right] \right)
\]

\[
+ \log \binom{n}{y}. \]
(79)

Comparing this with the general form of distributions in the exponential family, \(p(y|\theta, \phi)\), we can readily find that \(\theta = \log(\pi/(1 - \pi))\), \(a(\phi) = 1\), \(b(\theta) = \log(1 + \exp(\theta))\), \(c(y, \phi) = \log(\frac{y}{n})\) and \(d(y) = y/n\).

The mean and variance of the binomial variable are

\[
E \left( \frac{y}{n} | n, \pi \right) = \mu = \pi,
\] (81)

\[
\text{Var} \left( \frac{y}{n} | n, \pi \right) = \frac{1}{n} \pi (1 - \pi).
\] (82)

For the binomial distribution, the canonical link, which makes the linear predictor \(\eta\) equal to the canonical parameter, \(\theta\), is the \textit{logit function} (the inverse of the logistic function) defined by

\[
\text{logit}(p) = \log \left( \frac{p}{1 - p} \right).
\] (83)

To solve the dynamic binary classification problem we consider the dynamic binomial model in which both \(y\) and \(\pi\) in Equation (78) are functions of time. The link equation is defined by

\[
g(\mu_t) = \eta_t = h_t^T w_t.
\] (84)

In addition, we assume that the state variable \(w_t\) evolves according to a random walk

\[
w_t = w_{t-1} + v_t,
\] (85)

where \(v_t \sim (0, q_t I)\).

Suppose that we have the posterior mean and covariance of \(w_{t-1}\), given a set of observations \(D_{t-1} = \{y_1, \ldots, y_{t-1}\},\)
denoted by \( \mathbf{w}_{t-1|t-1} \) and \( \mathbf{P}_{t-1|t-1} \), respectively. Owing to the linear state process, the prior distribution of the state variable is

\[
\mathbf{w}_{t|D_{t-1}} \sim (\mathbf{w}_{t-1|t-1}, \mathbf{P}_{t|t-1}) ,
\]

where

\[
\mathbf{w}_{t|t-1} = \mathbf{w}_{t-1|t-1} ,
\]

\[
\mathbf{P}_{t|t-1} = \mathbf{P}_{t|t-1} + q_t \mathbf{I} ,
\]

where, as before, \( q_t \) is estimated via Equations (23) and (24). The mean and variance of the prior distribution of the linear predictor can be easily computed by

\[
\hat{\eta}_{t|t-1} = \mathbf{h}^T \mathbf{w}_{t|t-1} ,
\]

\[
r_{t|t-1} = \mathbf{h}^T \mathbf{P}_{t|t-1} \mathbf{h} .
\]

To derive the one-step ahead forecast distribution given by

\[
p(y_t|D_{t-1}) = \int p(y_t|\theta_t) p(\theta_t|D_{t-1}) \, d\theta_t ,
\]

we need to specify the conjugate prior, \( p(\theta_t|D_{t-1}) \). Using the derivation given in Appendix A.1, we find that the prior conjugate distribution is

\[
p(\theta_t|D_{t-1}) \propto \Gamma(m_t) \exp(k_t \theta_t) / \Gamma(k_t) \Gamma(m_t - k_t) (1 + \exp(\theta_t))^m_t ,
\]

and its first two moments are approximately

\[
E(\theta_t|D_{t-1}) \approx \log \left( \frac{k_t}{m_t - k_t} \right) ,
\]

\[
\text{Var}(\theta_t|D_{t-1}) \approx \frac{1}{k_t} + \frac{1}{m_t - k_t} .
\]

We know that the mean and variance of the canonical parameter are equivalent to that of the linear predictor because a canonical link is used. Hence, we can calculate the values of the hyperparameters as follows:

\[
k_t = \frac{1}{r_{t|t-1}} (1 + \exp(\hat{\eta}_{t|t-1})) ,
\]

\[
m_t = \frac{1}{r_{t|t-1}} (1 + \exp(-\hat{\eta}_{t|t-1})) (1 + \exp(\hat{\eta}_{t|t-1})) .
\]

With these hyperparameter values, the one-step ahead forecast distribution can be analytically obtained as

\[
p(y_t|D_{t-1}) = \int_{-\infty}^{\infty} p(y_t|\theta_t) p(\theta_t|D_{t-1}) \, d\theta_t ,
\]

\[
= \frac{\Gamma(m_t) \Gamma(k_t + y_t) \Gamma(m_t - k_t + n - y_t)}{\Gamma(k_t) \Gamma(m_t - k_t) \Gamma(m_t + n)} \times \left( \frac{n}{y_t} \right) .
\]

A full derivation is given in Appendix A.2. This is a beta-binomial distribution with parameters \( k_t \) and \( m_t - k_t \), and accordingly the mean and variance of the forecast distribution are

\[
E(y_t|D_{t-1}) = \frac{n k_t}{m_t} ,
\]

\[
\text{Var}(y_t|D_{t-1}) = \frac{n k_t (m_t - k_t) (m_t + n)}{m_t^2 (1 + m_t)} .
\]

After observing a new datum \( y_t \), we update the distributions of \( \theta_t \) and \( \eta_t \). The posterior distribution of \( \theta_t \) is of the form

\[
p(\theta_t|D_t) \propto p(\theta_t|D_{t-1}) p(y_t|D_t)
\]

\[
= \exp ((k_t + y_t) \theta_t - (m_t + n) \log(1 + \exp(\theta_t))) .
\]

The approximate values of the mean and variance of the posterior distribution are

\[
E(\theta_t|D_t) \approx \log \left( \frac{k_t + y_t}{m_t - k_t + n - y_t} \right) ,
\]

\[
\text{Var}(\theta_t|D_t) \approx \frac{1}{k_t + y_t} + \frac{1}{m_t - k_t + n - y_t} .
\]

They are equivalent to the mean and variance of the posterior distribution of \( \eta_t \), which are represented by \( \hat{\eta}_{t|t} \) and \( r_{t|t} \), respectively. Finally, the posterior mean and covariance of \( \mathbf{w}_t \) are

\[
\hat{\mathbf{w}}_{t|t} = \hat{\mathbf{w}}_{t|t-1} + \frac{\mathbf{P}_{t|t-1} \mathbf{h}_t}{r_{t|t-1}} (\hat{\eta}_{t|t} - \hat{\eta}_{t|t-1}) ,
\]

\[
\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \frac{\mathbf{P}_{t|t-1} \mathbf{h}_t \mathbf{h}^T \mathbf{P}_{t|t-1}}{r_{t|t-1}} (1 - \frac{r_{t|t}}{r_{t|t-1}}) .
\]

The dynamic classification problem can be solved by a special case of the dynamic binomial model in which the number of trials, \( n_t \), is set to 1 and the number of success, \( y_t \), is either 0 or 1. Hence, the expected value of the observation, \( E(y_t) \), is equal to the success probability, \( \pi_t \). In the alternate approach using nonlinear Kalman filters, as discussed in Sections 2.2.1 and 3.2, the class (or success) probability \( \pi_t \) is directly predicted from inputs via the logistic function. However, in the dynamic binomial model approach it is predicted in a full Bayesian manner, rather than the requisite link functions being approximated, as is the case with the EKF and the UKF detailed earlier. This allows us to work with the explicit distribution form of the class variable and its conjugate prior distribution. The drawback is however, that we impose a linear model to describe the dynamics of the sufficient statistics of the distributions. We compare the performances of these methods later in this paper.

5. RESULTS

In this section we investigate the performance of three dynamic classification models developed in this paper: (i)
dynamic logistic regression using the EKF, (ii) dynamic logistic regression using the UKF and (iii) the dynamic binomial model. We refer to these as ‘DLR-E’, ‘DLR-U’ and ‘DBM’, respectively. As we are ultimately interested in making classifications sequentially, not only when input and label information is completely given, but also when part of the information is missing or incomplete, and experiments investigating the performance of the models to missing labels and sensor data are reported here. Finally, we develop a method that allows an adaptive classification model to actively request labels. This is of importance in domains in which obtaining labels continuously may be too costly.

5.1. Synthetic data

As a simple example of the approaches, we consider two overlapping Gaussian distributions rotating in a circular fashion around a central point at $[0, 0]$, with the two distributions out of phase by $\pi$ radians. Target labels are interleaved, i.e. $\{0, 1, 0, 1, \ldots\}$ (Fig. 1). Three data sets are constructed with different Bayes errors (Fig. 2). The data are as presented in [2]. Owing to the nonstationary nature of the data any static classifier fails to classify the data. In these example experiments the kernel input vector is set to

$$\varphi(h_t) = \begin{bmatrix} 1 \\ h_t \end{bmatrix},$$ (105)

i.e. with no nonlinear basis functions, because we know that the underlying boundary is linear. The vector $h_t$ represents an input at time $t$. The state evolution noise variance $q_t$ is set to 0.1, and for DLR-E and DLR-U the observation noise variance $r_t$ is adaptively estimated using the scheme advocated by [1, 2, 10].

We computed one-step ahead predictions of the class probability (i.e. $\hat{\pi}_{t-1}$) with all three adaptive classifiers using all observations (both input and label) up to time $t - 1$. The predicted label class $\hat{y}_t$ is determined with a fixed decision threshold set to 0.5, i.e. if $\hat{\pi}_{t-1} > 0.5$, then $\hat{y}_t$ is 1 or 0 otherwise. Comparing the predicted labels with the true labels, all three dynamic classifiers achieved performances near the Bayes errors of each data set (Table 1). A static classifier is unable to separate the data and achieves a performance close to random at 0.5 as expected.

5.1.1. Missing labels

In practice we may only partially observe the class decision information i.e. some labels are missing. We require,
nevertheless, to make a decision forecast. Bayesian dynamic models can cope easily with missing observations. When class labels are not known, we may infer the missing label via the predictive distribution of the dynamic model. In this case we have

\[ \tilde{y}_t = \Pr(y_t = 1 | h_t) = \tilde{\pi}_{t|t-1}, \]  

(106)

which we may use as a ‘quasi-target’ in place of a true target label \( y_t \). This approach allows us not to discard information contained within the input vector \( h_t \), such as slow drift, by discarding the entire observation event. However, as Lowne et al. [2] pointed out, treating the quasi-targets as if they were ‘true’ labels runs the risk of a classifier becoming excessively confident in its predictions. To compensate for this illusory knowledge when external labelled feedback is unavailable, we may augment the state variable diffusion process using the uncertainty in the label. If the latter is fully observed, then this uncertainty is zero. If we rely on imputing the missing label, then we have

\[ q_t = q + \tilde{\pi}_{t|t-1}(1 - \tilde{\pi}_{t|t-1}), \]  

(107)

for the dynamic logistic regression-based algorithms, and

\[ q_t = q + \text{Var}(y_t | D_{t-1}), \]  

(108)

for the dynamic binomial model. The quantity \( q \) is the fixed value for the stationary state noise variance and \( \text{Var}(y_t | D_{t-1}) \) is the variance of the one-step ahead forecasting distribution given in Equation (99). We thereby ensure that the classifiers do not become overly confident too quickly based on fictitious feedback.

We carried out experiments in which label information is successively removed, at random, from 0 to 100% within each data stream. The performances of the classifiers were evaluated over 100 runs (Fig. 3). We can see that the classification performances of the models did not worsen rapidly with the fraction of unobserved labels. Performances with only 50% labelling were degraded by 0.13–0.31% for the data with 0% Bayes error, by 0.99–1.17% for the data with 4% Bayes error and by 0.6–3.8% for the data with 22% Bayes error. With only 20% class labels available the classifiers’ performances were degraded by only 1.5–5.2% in comparison with the 100% labelling case. This result indicates that the adaptive classifiers can deal effectively with sparsely observed labels while maintaining good classification performance.

5.2. Mountain fire scenario

We consider in this section a more realistic scenario in which there is a forest fire on a wooded mountain with several villages in the nearby region, as schematically depicted in Fig. 4. To allocate limited fire-fighting resources effectively we must classify each of the villages as being in potential danger or not based on local weather conditions measured by a set of weather sensors. For example, consider the village in the bottom left of the figure (with a ring). The stronger the wind speed and the higher the air temperature, the more likely it is for a village to be in danger. In addition, unless the wind blows in a northerly direction, the village is more vulnerable to the spread of fire.

### Table 1. Proportion of correct classifications for the three dynamic classification approaches.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>0% Bayes error</th>
<th>4% Bayes error</th>
<th>22% Bayes error</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLR-E</td>
<td>0.9980</td>
<td>0.9750</td>
<td>0.7740</td>
</tr>
<tr>
<td>DLR-U</td>
<td>0.9980</td>
<td>0.9680</td>
<td>0.7930</td>
</tr>
<tr>
<td>DBM</td>
<td>0.9960</td>
<td>0.9650</td>
<td>0.7720</td>
</tr>
</tbody>
</table>

![Figure 3](https://example.com/figure3.png)

**Figure 3.** Missing label experiments: the fractions of observed labels ranged from close to 0 to 100% for the data sets with (a) 0%, (b) 4% and (c) 22% Bayes errors. The classifiers maintained good classification performances with up to 70% of missing labels.
FIGURE 4. Mountain fire scenario: a fire starts in a wooded mountain with five villages (denoted by dots), which are in potential danger according to local weather conditions.

With weather data collected by a network of weather sensors [14], we created a realistic data set consisting of three input variables (wind speed, wind direction and air temperature) and a binary class variable (‘danger’ or ‘no danger’). The latter was obtained by running a simple diffusive fire spread model with the anisotropic diffusion kernel being a function of wind direction and wind speed.

The manner in which inputs and target labels of the data are distributed means that a static classifier is unable to separate the two classes with high accuracy (see Fig. 5a). The true labels alternate between ‘danger’ and ‘no danger’ at irregular intervals as the weather variables change (Fig. 5b). We employed three nonlinear Gaussian kernels to form the extended basis set \( \psi \) as in Equation 4, located at random within the data space (this number was not optimized and adapted to the location and number of basis functions remains a current research focus). For the sake of comparison, we ran a standard logistic regression model, with inputs being the extended basis set, \( \psi \); 100 input–output (basis–target label) pairs were randomly sampled without replacement from the full data set to use as a test set (needed as this baseline static algorithm is non-sequential). The performance of this static model was averaged over 100 runs of different training data samples. The three adaptive classifiers significantly outperformed the static classifier, achieving an improvement of between 19.2 and 22.5% (Table 2). Furthermore, when label information is successively randomly removed from the observed data, the classification performances of the models degraded slowly as the proportion of missing labels increased (Fig. 6).

5.2.1. Active label requesting

In many scenarios we can conceive of situations in which true class labels are costly to obtain. In this section we propose a mechanism for actively requesting a label whereby the adaptive classifier decides whether to request a label or whether to make missing value inference. This problem is closely related to the problem of active data selection [15, 16], as both methods ultimately use uncertainty as a guideline to enable requests of information. Consider at each time step a classifier making a one-step ahead label prediction. If the uncertainty associated with the prediction is high, the classifier requests a label for the next data point to improve its prediction accuracy. If the uncertainty is low, the classifier infers the label using the currently available information.

TABLE 2. Mountain fire scenario: correct classification proportions. All three adaptive models outperformed a static logistic regression.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.7390</td>
</tr>
<tr>
<td>DLR-E</td>
<td>0.9636</td>
</tr>
<tr>
<td>DLR-U</td>
<td>0.9313</td>
</tr>
<tr>
<td>DBM</td>
<td>0.9623</td>
</tr>
</tbody>
</table>

FIGURE 5. Mountain fire scenario: (a) a 3D plot of the three input variables and the class variable. (b) A time plot of the class variable. The true labels alternate between ‘danger’ and ‘no danger’ at irregular intervals as the weather variables change. Note that we deal with a single village located at the lower left in Fig. 4.
The threshold is intimately linked to the cost associated with labels. The quantity \( q_t \) is the constant value for the stationary noise variance as previously defined. To evaluate how the adaptive classifiers perform with missing inputs, we randomly removed input information from the mountain fire data set. The performances of the three adaptive classifiers were evaluated over 100 runs (Fig. 9). We note in particular the performance of the DLR-E: with only 40% of inputs available it still classified labels with a more than 90% accuracy.

### Table 3. Active label requesting: proportion of correct classifications on the mountain fire data when the classifiers actively request labels.

<table>
<thead>
<tr>
<th>Labelling</th>
<th>DLR-E</th>
<th>DLR-U</th>
<th>DBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>0.9636</td>
<td>0.9313</td>
<td>0.9623</td>
</tr>
<tr>
<td>90%</td>
<td>0.9528</td>
<td>0.9299</td>
<td>0.9407</td>
</tr>
<tr>
<td>70%</td>
<td>0.9528</td>
<td>0.9259</td>
<td>0.9407</td>
</tr>
<tr>
<td>50%</td>
<td>0.9394</td>
<td>0.9191</td>
<td>0.9178</td>
</tr>
<tr>
<td>30%</td>
<td>0.9272</td>
<td>0.8935</td>
<td>0.7278</td>
</tr>
</tbody>
</table>

**Figure 6.** Mountain fire scenario: classification performance of adaptive models as a function of the fraction of observed labels. Note that the classification performances of the models did not degrade in proportion to the number of unobserved labels.

with this prediction is higher than a pre-defined threshold, then it requests a label, otherwise, it proceeds without labelling. The threshold is intimately linked to the cost associated with obtaining a label and the cost associated with iterating the model with an imputed label and making a wrong decision. If these costs are known then this scheme allows for an adaptive minimum cost decision process. We present proof of concept results in this paper and more full exposition of this approach is relegated to future work.

For dynamic logistic classifiers, the prediction uncertainty is the variance of the associated Bernoulli distribution, namely \( \tilde{\pi}_{t-1}(1 - \tilde{\pi}_{t-1}) \). For the dynamic generalized linear binomial model it is the variance of a beta-binomial predictive distribution, \( \text{Var}(y_t|D_{t-1}) \), as given in Equation (99). By varying the uncertainty threshold which the algorithms use to request labels, we may obtain performance measures for different proportions of requested labels.

We carried out such a set of experiments on the mountain fire data set, leading to different proportions of labelling from 100 to approximately 30% (Table 3). With 50% requested labelling, for example, classification accuracy was degraded by 1.2–4.5% (Fig. 7). In this figure the grey regions represent time steps when a classifier requested a label. We can see that when the models were confident enough about their predictions they did not request a label. Note that despite the relatively low overall classification error, the dynamic binomial model failed to detect a change of label on two occasions, and the dynamic logistic regression using the EKF on one occasion. We note that our models for active selection do not at present include measures of wide-sense correlation between requested samples, in that they look instantaneously at the expected uncertainty and use this as a criterion for active requests. It is possible that, for some applications, a Markov-style look back or a multistep request look-forward may offer an improvement. This is an area of active research.

#### 5.2.2. Missing inputs

We finally consider the scenario in which not only labels but also input sensor data can be missing, for example, as a result of a failure in the sensors or in transmitting subsequent information. Our decision models are defined by

\[
\pi_t = l(\varphi(h_t)_{1:T}w_t),
\]

for the DLR models, or

\[
y_t \sim \text{Binomial with } \eta_t = \varphi(h_t)_{1:T}w_t,
\]

for the DBM. Here \( l(\cdot) \) is the logistic function given in Equation (3) and \( \eta_t \) is the link equation given in Equation (53). When the input, \( h_t \), is not available, we can impute it as \( \hat{h}_t \). With a set of observed inputs up to time \( t - 1 \), i.e. \( \{h_1, \ldots, h_{t-1}\} \), we can predict the missing input using, for example, a Gaussian Process model [17] or a nonstationary autoregressive model, such as the sequential Bayesian multivariate autoregressive model detailed in [18].

The data streams of the three inputs in the mountain fire scenario (i.e. wind speed, wind direction, air temperature) are illustrated in Fig. 8a. We removed at random 70% of the input data (Fig. 8b) and predicted the missing input values with a sequential dynamic multivariate autoregressive model [18] of order 2 (Fig. 8c). We can see that the predicted inputs are very close to the true ones. To compensate for using these predicted inputs, we fold the uncertainty in the predicted values of missing inputs (denote by \( u_t \)) into the state noise. We therefore update \( q_t \) using

\[
q_t = q + u_t.
\]

The quantity \( q \) is the constant value for the stationary noise variance as previously defined. To evaluate how the adaptive classifiers perform with missing inputs, we randomly removed input information from the mountain fire data set. The performances of the three adaptive classifiers were evaluated over 100 runs (Fig. 9). We note in particular the performance of the DLR-E: with only 40% of inputs available it still classified labels with a more than 90% accuracy.
FIGURE 7. Active label requesting on the mountain fire data when 50% of total labels is requested by (a) DLR-E, (b) DLR-U and (c) DBM. The one-step ahead predictions for the class probability are represented by the innermost of the three traces in each subplot, the label predictions by the middle traces and the observed labels by the outermost traces. The grey shaded regions represent time steps when the classifiers requested a label.

Combined with active label requesting, it is possible for a classifier to decide which label to request even when input sensor information is not available. We carried out an experiment in which 50% of inputs are missing and the classifiers requests 50% of labelling. We computed the average proportion of correct classification and the corresponding standard deviation for the three adaptive classifiers over 10 runs: 0.8846/0.0229 (DLR-E), 0.7344/0.0221 (DLR-U) and 0.8274/0.0840 (DBM).

FIGURE 8. Time plot of (a) true inputs, (b) given inputs (70% missing) and (c) predicted inputs. A dynamic multivariate autoregressive model of AR order 2 was used.
As the models are developed in a fully Bayesian framework,
We have demonstrated how a classification decision can be made
The performance of all algorithms detailed in this paper
The performance of the DLR-E model, in particular, degraded slowly
these avenues of research constitute active ongoing research.

In comparison to corresponding results when given complete
input information provided in Table 3, we can see that for
DLR-E classification accuracy was degraded by only 5%.

6. CONCLUSIONS AND FUTURE WORK

6.1. Conclusions

We have demonstrated how a classification decision can be made sequentially in highly nonlinear, nonstationary and incomplete environments. We have studied two different approaches to the problem: (i) dynamic logistic regression using nonlinear variants of the Kalman filter and (ii) a dynamic generalized linear model. The latter model provides a mathematically elegant Bayesian framework for dynamic models so long as the distribution of an observation variable is a member of the exponential family. Since in binary classification an observation variable has a simple Bernoulli distribution, a dynamic binomial model has been developed as an adaptive binary classifier. All three models achieved considerably better performance compared with static (fixed parameter) classification for both the simple rotating data example and the mountain fire scenario. As the models are developed in a fully Bayesian framework, handling missing labels and input (sensor) data can be managed. The performance of all algorithms detailed in this paper did not degrade rapidly as the amount of unobservable data increased. Labels, and indeed sensor observations, can be actively requested by all the algorithms if the uncertainty (and hence cost) of making a predictive decision is high.

We note that the approaches discussed in this paper may be naturally extended to multiclass decision making and for multistep decision forecasting. Along with the co-ordination of sequential adaptive decisions from multiple autonomous agents, these avenues of research constitute active ongoing research.

FUNDING

This work was undertaken as part of the ALADDIN (Autonomous Learning Agents for Decentralised Data and Information Systems) project and was jointly funded by a BAE Systems and EPSRC (Engineering and Physical Research Council) strategic partnership.

REFERENCES

APPENDIX: DERIVATIVES

A.1. Conjugate prior distribution for the binomial distribution

The conjugate prior for the binomial distribution is of the form

\[ p(\theta_t | D_{t-1}) = \omega(k_t, m_t) \exp(k_t\theta_t - m_t \log(1 + \exp(\theta_t))). \]  \hspace{1cm} (A.1)

Using \( \theta_t = \log(\pi_t/(l - \pi_t)), \) the probability density function of \( \pi_t \) is of form

\[ p(\pi_t | D_{t-1}) = p_{\theta_t | D_{t-1}}(\pi_t) \bigg|_{\theta_t = \log(\pi_t/(l - \pi_t))} = \frac{\omega(k_t, m_t)\pi_t^{k_t-1}(1 - \pi_t)^{m_t-k_t-1}}{\Gamma(k_t)\Gamma(m_t-k_t)}. \]  \hspace{1cm} (A.2)

This is a beta distribution with parameters \( k_t \) and \( m_t-k_t \), thereby we know that \( \omega(k_t, m_t) = \Gamma(k_t)/\Gamma(m_t-k_t). \)

The mean and variance of the conjugate prior are computed in the following way. The moment generating function of \( \theta_t \) is

\[ M_{\theta_t}(z) = \mathbb{E}(\exp(z\theta_t)) = \int_{-\infty}^{\infty} \exp(z\theta_t) p(\theta_t | D_{t-1}) d\theta_t = \frac{\Gamma(k_t + z)\Gamma(m_t - k_t - z)}{\Gamma(k_t)\Gamma(m_t - k_t)} (m_t - k_t)^{-z}. \]  \hspace{1cm} (A.3)

and thus the cumulant generating function is

\[ C_{\theta_t}(z) = \log(M_{\theta_t}(z)) = \log \Gamma(k_t + z) - \log \Gamma(m_t - k_t - z) - \Gamma(k_t) - \Gamma(m_t - k_t). \]  \hspace{1cm} (A.4)

From the cumulant generating function the two moments of the conjugate prior can be computed:

\[ \text{E}(\theta_t | D_{t-1}) = \frac{d}{dz} C_{\theta_t}(z) \bigg|_{z=0} = \psi(k_t) - \psi(m_t - k_t) \approx \log \left( \frac{k_t}{m_t - k_t} \right), \]  \hspace{1cm} (A.5)

\[ \text{Var}(\theta_t | D_{t-1}) = \frac{d^2}{dz^2} C_{\theta_t}(z) \bigg|_{z=0} = \psi'(k_t) + \psi'(m_t - k_t) \approx \frac{1}{k_t} + \frac{1}{m_t - k_t}. \]  \hspace{1cm} (A.6)

The function \( \psi(\cdot) \) denotes the digamma function defined by

\[ \psi(x) = \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}. \]  \hspace{1cm} (A.7)

A.2. One-step forecast distribution for the binomial distribution

When the observation variable \( Y_t \) is a binomial, the one-step ahead forecast distribution is computed as follows:

\[ p(y_t | D_{t-1}) = \int_{-\infty}^{\infty} p(y_t | \theta_t) p(\theta_t | D_{t-1}) d\theta_t, \]

\[ = \int_{-\infty}^{\infty} \exp \left( n \left[ \frac{y_t}{n} - \log(1 + \exp(\theta_t)) \right] + \log \left( \frac{n}{y_t} \right) \right) \times \frac{\Gamma(m_t)}{\Gamma(k_t)\Gamma(m_t - k_t)} \left( 1 + \exp(\theta_t) \right)^{m_t} d\theta_t, \]

\[ = \frac{\Gamma(m_t)}{\Gamma(k_t)\Gamma(m_t - k_t)} \left( \frac{n}{y_t} \right) \times \frac{\exp(k_t \theta_t)}{(1 + \exp(\theta_t))^{m_t+n}} d\theta_t, \]

\[ = \frac{\Gamma(m_t)}{\Gamma(k_t)\Gamma(m_t - k_t)} \left( \frac{n}{y_t} \right) \times \frac{\Gamma(k_t + y_t)\Gamma(m_t - k_t + n - y_t)}{\Gamma(m_t + n)} \left( \frac{n}{y_t} \right). \]  \hspace{1cm} (A.8)