Neural Networks 22 (2009) 1286–1294

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Adaptive classification for Brain Computer Interface systems using Sequential Monte Carlo sampling

Ji Won Yoon [a,*], Stephen J. Roberts [a], Matt Dyson [b], John Q. Gan [b]

[a] Pattern Analysis & Machine Learning Research Group, Department of Engineering Science, University of Oxford, Oxford, UK

[b] Department of Computing and Electronic Systems, University of Essex, UK

ARTICLE INFO

Article history:
Received 30 September 2008
Received in revised form 21 May 2009
Accepted 24 June 2009

Keywords:
Online Brain Computer Interfacing (BCI)
Nonlinear dynamics
Rao-Blackwellised Sequential Monte Carlo
Dynamic probit model
Missing labels
Auxiliary labels

ABSTRACT

Adaptive classification is a key function of Brain Computer Interfacing (BCI) systems. This paper proposes robust mathematical frameworks and their implementation for the on-line sequential classification of EEG signals in BCI systems. The proposed algorithms are extensions to the basic method of Andrieu et al. [Andrieu, C., de Freitas, N., and Doucet, A. (2001). Sequential bayesian semi-parametric binary classification. In *Proc. NIPS*], modified to be suitable for BCI use. We focus on the inference and prediction of target labels under a non-linear and non-Gaussian model. In this paper we introduce two new algorithms to handle missing or erroneous labeling in BCI data. One algorithm introduces auxiliary labels to process the uncertainty of the labels and the other modifies the optimal proposal functions to allow for uncertain labels. Although we focus on BCI problems in this paper, the algorithms can be generalized and applied to other application domains in which sequential missing labels are to be imputed under the presence of uncertainty.

1. Introduction

The goal of Brain Computer Interfacing (BCI) is to enable computer operation by manipulation of the brain's signals rather than by physical means (Pfurtscheller, Flotzinger, & Kalcher, 1993; Wolpaw, McFarland, Neat, & Forneris, 1991). Many classification approaches have been studied to process and analyze non-stationary signals with subject training (Babiloni et al., 2000; Dornhege, Blankertz, Curio, & Muller, 2004; Muller, Anderson, & Birch, 2003; Sugiyama, Krauledat, & Muller, 2007). However, it is known that the generation and control of brain activity for BCI use, when recorded using non-invasive methods (typically via the electroencephalogram or EEG), often requires extensive subject training before a reliable communication channel is formed. In order to reduce the overheads of training, tracking non-stationarity and to cope rapidly with new subjects, adaptive approaches to the core data modeling have been developed for BCI systems. Such adaptive approaches differ from the typical methodology, in which an algorithm is trained *off-line* on retrospective data, in so far that the process of *learning* is continuously taking place rather than being confined to a section of *training data*. In the signal processing and machine learning communities, this is referred to as *sequential classification*. Previous research in this area applied to BCI data has used *state space modeling* of the time series (Lowne, Haw, & Roberts, 2006; Penny, Roberts, Curran, & Stokes, 2000; Sykacek, Roberts, & Stokes, 2004). In particular, we have investigated the use of the Extended Kalman Filter (EKF) for BCI systems (Yoon, Roberts, Dyson, & Gan, 2008a, 2008b). The EKF offers one framework for approximating the non-linear linkage between observations and decisions (via the logistic function). There are several alternative ways to approach this non-linear state space mapping, in particular the Unscented Kalman Filter (UKF) (Julier & Uhlmann, 1997), based on an alternate approximation scheme, and the Particle Filter (PF) (Chen & Liu, 2000; Doucet, Godsill, & Andrieu, 2000; Liu, 2001) based on Sequential Monte Carlo sampling. There has been some significant research addressing adaptive classification using Sequential Monte Carlo (SMC) filters (Andrieu, de Freitas, & Doucet, 2001; Hojen-Sorensen, de Freitas, & Fog, 2000). Although the SMC algorithm based on Rao-Blackwellization by Andrieu et al. (2001) is regarded as the most efficient sampling algorithm for online classification, it cannot be directly applied for use in BCI systems for two major reasons. First, the output function takes the target labels (the decision classes) to be time independent, i.e. there is no explicit Markov dependency from one decision to the next. This is a poor model for the characteristics of a BCI interface, at the sample rate we operate at (over 0.2 s intervals), in which sequences of labels of the same class persist for several seconds before making a transition. What we require, therefore, is a model similar to a *Markov* process in which decisions persist

* Corresponding author. Tel.: +44 7771964170.
E-mail addresses: jwyoon@robots.ox.ac.uk (J.W. Yoon), sjrob@robots.ox.ac.uk (S.J. Roberts), mdyson@essex.ac.uk (M. Dyson), jqgan@essex.ac.uk (J.Q. Gan).

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.06.005

until evidence for a *change point* accumulates. Secondly, standard SMC approaches assume that an observed label is certain, i.e. there are no 'bit errors' in the label stream. This assumption provides SMC methods efficient proposal functions by using a truncated distribution (Andrieu et al., 2001).

We do not, however, know the ground truth of the labels in BCI systems — we have only imperfect labels, such as those obtained in this study from electromyography (EMG) signals, which may be used as a reference. The assumption of error-free labels is hence invalid. However, sampling using the SMC method without this assumption is far less efficient since the proposal function without truncation is not designed efficiently (Andrieu et al., 2001). In order to address these two problems, we have developed two different algorithms. One algorithm introduces an auxiliary latent variable which corresponds to the underlying labels. The other algorithm samples from a proposal distribution which has probability over both label classes.

This paper is organized as follows. Section 2 presents the background of state space modeling in time series. In Section 3 the mathematical models which we propose for BCI systems are described. The proposed algorithms are described in Section 4. Both synthetic and real experimental data sets are then tested to compare performances in Sections 5 and 6 respectively. In Section 7, we also discuss the practicality of a SMC classifier in BCI systems with several views: implementation, computational efficiency and scalability with respect to class labels and EEG channels.

2. Time series state space models in a Bayesian framework

We first consider a Bayesian sequential estimation framework for state space models. We describe the general framework for a dynamic model with state space $\mathbf{h}_t$ and observation space $\mathbf{o}_t$, where t denotes the discrete time index. The distribution of interest is the posterior, $p(\mathbf{h}_t|\mathbf{o}_{1:t})$ where $\mathbf{o}_{1:t}$ represents a shorthand notation for $(\mathbf{o}_1, \ldots, \mathbf{o}_t)$. In the Bayesian sequential estimation framework, the posterior distribution is obtained by the following two step recursion:

prediction step

$$p(\mathbf{h}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{h}_t|\mathbf{h}_{t-1})p(\mathbf{h}_{t-1}|\mathbf{o}_{1:t-1})\mathrm{d}\mathbf{h}_{t-1} \tag{1}$$

filtering step

$$p(\mathbf{h}_t|\mathbf{o}_{1:t}) = \frac{p(\mathbf{o}_t|\mathbf{h}_t)p(\mathbf{h}_t|\mathbf{o}_{1:t-1})}{p(\mathbf{o}_t|\mathbf{o}_{1:t-1})}. \tag{2}$$

This recursion requires the specification of two models; a dynamic update model for the hidden states $p(\mathbf{h}_t|\mathbf{h}_{t-1})$ and a model for the state likelihood given the current measurement $p(\mathbf{o}_t|\mathbf{h}_t)$. The recursion is initialized with some prior distribution for the initial state $p(\mathbf{h}_0)$. The dynamic update and likelihood models are given by, $\mathbf{h}_t = \mathbf{F}_t(\mathbf{h}_{t-1}, U_t)$ and $\mathbf{o}_t = \mathbf{G}_t(\mathbf{h}_t, V_t)$ where $\mathbf{F}_t$ and $\mathbf{G}_t$ can be regarded as either non-linear or linear functions corrupted by noise, U_t and V_t at time t. However, if $\mathbf{F}_t(\cdot)$ and $\mathbf{G}_t(\cdot)$ are non-linear and U_t and V_t are non-Gaussian noise, we may approximate the non-linearities using the first and second derivatives of the function in a linear model. This approximation enables us to use the standard Kalman filter framework and the approach is commonly referred to as the *Extended Kalman Filter* (EKF). Alternatively, we can avoid the linearization approximations by using Monte Carlo paradigms. These are referred to as *Sequential Monte Carlo* (SMC) methods, and are detailed in Appendix A.

3. Mathematical model for BCI classification with label uncertainty

Our proposed algorithms are based on a kernel probit classifier (regression) model. Given an observed input $\mathbf{x}_t$ at time $t = 1, 2, \ldots, T$, we observe an associated target, $z_t \in \{0, 1\}$. The observation label is defined $z_t \in \{0, 1\}$, such that $Pr(z_t = 1|\mathbf{x}_t, \mathbf{w}_t) = \Phi(f(\mathbf{x}_t, \mathbf{w}_t))$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and the vector $\mathbf{w}_t$ represents the time-varying regression weights. The state space model for the BCI system can hence be regarded as a hierarchical model since the noise of observations influences the model indirectly through the probit regression model. Suppose that we have radial basis functions $\varphi_t(\mathbf{x}_t)$ from an input vector $\mathbf{x}_t$ at time t. To simplify notation, we use φ_t instead of $\varphi_t(\mathbf{x}_t)$ i.e. $\varphi_t = \varphi_t(\mathbf{x}_t) = \left[\mathbf{x}_t' \quad \{\phi_t^1(\mathbf{x}_t)\}' \quad \{\phi_t^2(\mathbf{x}_t)\}' \quad \cdots \quad \{\phi_t^{N_b}(\mathbf{x}_t)\}' \quad 1\right]'$ where $\phi_t^i(\mathbf{x}_t)$ denotes the ith Gaussian basis function, N_b is the number of basis functions and $'$ denotes a transpose operator. Such an indirect influence makes for a more complicated model, but we can circumvent much of this complexity by forming a three stage state space model and by introducing an auxiliary variable y_t. The latter variable acts so as to link the indirect relationships between observations and the probit regression model given by

$$\begin{aligned} p(z_t|y_t) &= \Phi(y_t)^{z_t}(1 - \Phi(y_t))^{1-z_t} \\ y_t &= \varphi_t^{\mathrm{T}}(\mathbf{x}_t)\mathbf{w}_t + v_t \\ \mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{u}_t \end{aligned} \tag{3}$$

where v_t and $\mathbf{u}_t$ respectively explain the uncertainty in the system and regression coefficients with $v_t \sim \mathcal{N}(0, \kappa_t)$ and $\mathbf{u}_t \sim \mathcal{N}(0, \tau_t\mathbf{I})$; the hyper-parameters κ_t and τ_t denote the covariances of states y_t and $\mathbf{w}_t$. We set up $\kappa_t = 1$ and $\tau_t = 1$ in this paper as used in Andrieu et al. (2001). $\mathcal{N}(\cdot, \cdot)$ represents normal distribution with mean and covariance. Eq. (3) is formed from our previous work (Yoon et al., 2008a, 2008b). To allow for mis-labeling, we modify the model by introducing an underlying hidden set of labels, $\bar{z}_t$ giving a modified hierarchical model, namely:

$$\begin{aligned} p(z_t|\bar{z}_t, \rho) &= \rho^{\delta(z_t-\bar{z}_t)}(1 - \rho)^{1-\delta(z_t-\bar{z}_t)} \\ p(\bar{z}_t|y_t) &= \Phi(y_t)^{\bar{z}_t}(1 - \Phi(y_t))^{1-\bar{z}_t} \\ y_t &= \varphi_t^{\mathrm{T}}(\mathbf{x}_t)\mathbf{w}_t + v_t \\ \mathbf{w}_t &= \mathbf{w}_{t-1} + \mathbf{u}_t \end{aligned} \tag{4}$$

where ρ is the matching probability between z_t and $\bar{z}_t$ and $\delta(\cdot)$ denotes the *Dirac delta* function. In this mathematical model, we also introduce a variable π which parameterises the Markov transition probability between $\bar{z}_t$ and $\bar{z}_{t-1}$. The Markov property of $\bar{z}_t$ is hence $p(\bar{z}_t|\bar{z}_{t-1}, \pi) = \pi^{\delta(\bar{z}_t-\bar{z}_{t-1})}(1 - \pi)^{1-\delta(\bar{z}_t-\bar{z}_{t-1})}$. This probability can explain the continuity and dependency of the labels since the labels are not independent of the time series as assumed in other literature (Andrieu et al., 2001; Lowne et al., 2006; Yoon et al., 2008a).

3.1. Priors

In Bayesian statistics, we need to choose prior distributions carefully (Bernardo & Smith, 1994). The prior distributions for the hidden states are described by $p(y_0) = \mathcal{N}(\cdot; 0, \Sigma_{y_0})$, $p(\mathbf{w}_0) = \mathcal{N}(\cdot; 0, \Sigma_{\mathbf{w}_0})$, $p(\rho) = \mathcal{B}(\rho; \alpha_\rho, \beta_\rho)$ and $p(\pi) = \mathcal{B}(\pi; \alpha_\pi, \beta_\pi)$ where $\mathcal{B}$ denotes beta distribution. We infer the hyper-parameters α_ρ, β_ρ, α_π and β_π by EKF simulation with a short length of *burn-in* training data (500 samples with 5 events). The prior transition kernel of the hidden labels is defined as: