Computational Statistics and Data Analysis 56 (2012) 2375–2387

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Variational Bayes approach for model aggregation in unsupervised classification with Markovian dependency

Stevenn Volant [a,b,*], Marie-Laure Martin Magniette [a,b,c,d,e], Stéphane Robin [a,b]

[a] AgroParisTech, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
[b] INRA UMR MIA 518, 16 rue Claude Bernard, 75231 Paris Cedex 05, France
[c] INRA UMR 1165, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France
[d] UEVE, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France
[e] CNRS ERL 8196, URGV, 2 rue Gaston Crémieux, CP5708, 91057, Evry Cedex, France

ARTICLE INFO

Article history:
Received 11 April 2011
Received in revised form 23 January 2012
Accepted 23 January 2012
Available online 15 February 2012

Keywords:
Model averaging
Variational Bayes inference
Markov chain
Unsupervised classification

ABSTRACT

A binary unsupervised classification problem where each observation is associated with an unobserved label that needs to be retrieved is considered. More precisely, it is assumed that there are two groups of observation: normal and abnormal. The 'normal' observations are coming from a known distribution whereas the distribution of the 'abnormal' observations is unknown. Several models have been developed to fit this unknown distribution. An alternative based on a mixture of Gaussian distributions is proposed. The inference is performed within a variational Bayesian framework and the aim is to infer the posterior probability of belonging to the class of interest. To this end, it makes little sense to estimate the number of mixture components since each mixture model provides more or less relevant information to the posterior probability estimation. By computing a weighted average (named aggregated estimator) over the model collection, Bayesian Model Averaging (BMA) is one way of combining models in order to account for information provided by each model. An aim is then the estimation of the weights and the posterior probability for a specific model. Optimal approximations of these quantities from the variational theory are derived; other approximations of the weights are also proposed. It is assumed that the data are dependent (Markovian dependency) and hence a Hidden Markov Model is considered. A simulation study is carried out to evaluate the accuracy of the estimates in terms of classification performance. An illustration on both epidemiologic and genetic datasets is presented.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Binary unsupervised classification. We consider an unsupervised classification problem where each observation is associated with an unobserved label that we want to retrieve. Such problems occur in a wide variety of domains, such as climate, epidemiology (see Sun and Cai (2009)), or genomics (see McLachlan et al. (2002)) where we want to distinguish 'normal' observations from abnormal ones or, equivalently, to distinguish pure noise from signal. In such situations, some prior information about the distribution of 'normal' observations, or about the distribution of the noise is often available and we want to take advantage of it.

[*] Corresponding author at: AgroParisTech, 16 rue Claude Bernard, 75231 Paris Cedex 05, France.
E-mail addresses: stevenn.volant@agroparistech.fr (S. Volant), marie_laure.martin@agroparistech.fr (M.-L. Martin Magniette), stephane.robin@agroparistech.fr (S. Robin).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.01.027

More precisely, based on observations $X = \{X_t\}$, we want to retrieve the unknown binary labels $S = \{S_t\}$ associated with each of them. We assume that 'normal' observations (labelled 0) have distribution ϕ, whereas 'abnormal' observations (labelled 1) have distribution f. We further assume that the null distribution ϕ is known, whereas the alternative distribution f is not. From a classification perspective, we want to compute

$$T_t = \Pr\{S_t = 0|X\}. \tag{1}$$

Bayesian model averaging (BMA). The probability T_t depends on the unknown distribution f. Many models can be considered to fit this distribution and we denote $\mathcal{M} = \{f_m; m = 1, \dots, M\}$ a finite collection of such models. As none of these models is likely to be the true one, it seems more natural to gather information provided by each of them, rather than to try to select the 'best' one. The Bayesian framework is natural for this purpose, as we have to deal with model uncertainty.

Bayesian model averaging (BMA) has been mainly developed by Hoeting et al. (1999) and provides the general framework of our work. It has been demonstrated that BMA can improve predictive performance and parameter estimation in Madigan and Raftery (1993), Madigan and Hutchinson (1995), Volinsky et al. (1997), Raftery and Zheng (2003) or Ruggieri and Lawrence (2011). Jaakkola and Jordan (1998) also demonstrated that model averaging often provides a gain in terms of classification and fitting. The determination of the weight α_m associated with each model m when averaging is a key ingredient of all these approaches.

Weight determination. As shown in Hoeting et al. (1999) the standard Bayesian reasoning leads to $\alpha_m = \Pr\{M = m|X\}$, where M stands for the model. In a classical context, the calculation of α_m requires one to integrate the joint conditional distribution $P(M, \Theta|X)$, where Θ is the vector of model parameters, and several approaches can be used. The BIC criterion (Schwarz, 1978) is based on a Laplace approximation of this integral, which is questionable for small sample sizes. One other commonly used method is MCMC (Monte Carlo Markov Chain, Andrieu (2003)) which samples the distribution and can provide an accurate estimation of the joint conditional distribution, but at the cost of huge (sometimes prohibitive) computational time.

In the unsupervised classification context, the problem is even more difficult as we need to integrate the conditional $P(M, \Theta, S|X)$ since the labels are unobserved. This distribution is generally not tractable but, for a given model, Beal and Ghahramani (2003) developed a variational Bayes strategy to approximate $P(\Theta, S|X)$. Variational techniques aim at minimising the Kullback–Leibler (KL) divergence between $P(\Theta, S|X)$ and an approximated distribution $Q_{\Theta,S}$ (Corduneanu and Bishop, 2001; Wainwright and Jordan, 2008; Ren and Hodges, 2011). Jaakkola and Jordan (1998) proved that the variational approximation can be improved by using a mixture of distributions rather than factorised distribution as the approximating distribution. A mixture distribution Q_{mix} is chosen to minimise the KL-divergence with respect to $P(\Theta, S|X)$. Unfortunately, their method averages the log of Q_{mix} over all the configurations which leads to untractable computation and a costly algorithm involving a smoothing distribution.

Our contribution. In this article, we propose variational-based weights for model averaging, in presence of a Markov dependency between the unobserved labels. We prove that these weights are optimal in terms of KL-divergence from the true conditional distribution $P(M|X)$. To this end, we optimise the KL-divergence between $P(\Theta, S, M|X)$ and an approximated distribution $Q_{\Theta,S,M}$ (Section 2). This optimisation problem differs from that of Jaakkola and Jordan (1998) (see their Eq. (14)). Based on the approximated distribution of $P(\theta, S|M, X)$, we derive other estimations of the weights.

We then reconsider the case of unsupervised classification and consider a collection $\mathcal{M}$ of mixtures of parametric exponential family distributions (Section 3). We propose a complete inference procedure that does not require any specific development in terms of an inference algorithm. In order to assess our approach, we propose a simulation study which highlights the gain of model averaging in terms of binary classification (Section 4). We also present two illustrations on epidemiologic and genomic datasets (Section 5). An R package named VBMA4HMM (Variational Bayes Models Averaging for hidden Markov models) is available on the CRAN.

2. Variational weights

The aim of model averaging is to account for the information in each model of a collection of M models. To do so, we need to calculate the weight of each model. In this section, we propose three different weights based on the variational Bayes theory.

2.1. A two-step optimisation problem

In a Bayesian Model Averaging context, we focus on averaged estimators to account for model uncertainty. It implies evaluating the conditional distribution:

$$P(M|X) = \int P(H, M|X)dH, \tag{2}$$

where H stands for all hidden variables, that is $H = (S, \Theta)$, and M denotes the model.

In order to calculate this distribution, we need to compute the joint posterior distribution of H and M. Due to the latent structure of the problem, this is not feasible. However, the mean field/variational theory allows an approximation of this

Download English Version:

https://daneshyari.com/en/article/417581

Download Persian Version:

https://daneshyari.com/article/417581

Daneshyari.com