CLASSIFICATION IN THE SAMPLING PARADIGM: A PREDICTIVE APPROACH TOWARDS A SIESTA SLEEP ANALYZER

Sykacek P.[1], Roberts S. J.[2], Rezek I.[2], Flexer A.[1], Dorffner G.[1]
[1] Austrian Research Institute for Artificial Intelligence,
Schottengasse 3, A-1010 Vienna, Austria
[2] Robotics Research Group, Department of Engineering Science,
University of Oxford, Oxford OX1 6PJ, UK

peter@ai.univie.ac.at

Abstract: **The SIESTA project aims at defining a new description of human sleep. The basic idea of the algorithm described below is that extreme states of the sleeping subject are classified with higher reliability. Hence we use labels of the extreme states wake, deep sleep as well as rapid eye movement sleep as well as all other data without labels. In order to have some benefit from unlabeled data, the classifier has to model class conditional densities and class priors. This allows to use Bayes theorem and predict the a-posteriori probability of class. We embed this classifier into a Bayesian framework and use Gibbs sampling to infer model coefficients.**

Introduction

SIESTA is an EC funded project with several clinical and technical partners. We aim at designing a new sleep analyzer that overcomes the shortcomings of the current Rechtschaffen & Kales (R&K) standard. R & K uses an *ordered* set of labels to describe different sleep stages. The design of the SIESTA sleep analyzer assumes that assigning intermediate stages is difficult and the most reliable information clinicians can provide are labels for epochs of wake, deep sleep and rapid eye movement (REM) sleep. The scorings of the R & K hypnograms are based on polysomnographic recordings which are also the basis of the SIESTA sleep analyzer. The main idea of the SIESTA sleep analyzer is to interpret the entire night in the light of these extreme states.

Methods

This contribution describes a classifier in the sampling paradigm that is a classifier which is based on a generative model of class conditional densities together with an appropriate training algorithm. The generative model is the Gaussian mixture model, (1).

$$\begin{aligned} p(\underline{x}|C_k) &= \sum_{l=1}^{L} w_{kl} p(\underline{x}|\underline{\Theta}_l) \\ p(\underline{x}) &= \sum_{k=1}^{K} P_k p(\underline{x}|C_k) \end{aligned} \tag{1}$$

Using P_k for the class priors and $p(\underline{x}|C_k)$ for the class conditional densities, we may express the posterior probabilities for classes as $P(C_k|\underline{x}) = P_k p(\underline{x}|C_k)/p(\underline{x})$. The component densities $p(\underline{x}|\underline{\Theta}_l)$ can be any parametrised density function. For sake of convenience we will use normal densities. Such an architecture has been used by [3], who used maximum likelihood methods for training.

We perform inference in the Bayesian framework using Gibbs sampling. The likelihood function using labeled as well as unlabeled data reads as:

$$p(\mathcal{T}, \mathcal{X}|\underline{\Theta}) = \Pi_{k=1}^{K} \Pi_{n_k=1}^{N_k} P_k p_k(\underline{x}_{n_k}|\underline{\Theta}_k) \Pi_{m=1}^{M} p(\underline{x}_m|\underline{\Theta}), \tag{2}$$

where $\mathcal{T}$ denotes labeled and $\mathcal{X}$ unlabeled training data. In (2) $\underline{\Theta}_k$ are all coefficients the k-th class conditional density depends on. We further use $\underline{\Theta}$ for all model coefficients together, n_k as number of samples belonging to class k and m as index for unlabeled samples. In order to allow sampling from the full conditional, we have to choose priors over coefficients from their *conjugate family* (see [1] for a detailed discussion of conjugate analysis):

- Each component mean, $\underline{m}_d$, is given a Gaussian prior: $\underline{m}_d \sim \mathcal{N}_d(\underline{\xi}, \underline{\kappa})$.
- The inverse variance is given a Gamma prior: $\sigma^{-2} \sim \Gamma(\alpha, \beta)$.

- The hyperparameter, β, gets a Gamma hyperprior: $\beta \sim \Gamma(g, h)$.
- The mixing coefficients, $\underline{w}_k$, get a Dirichlet prior: $\underline{w}_k \sim \mathcal{D}(\delta_w, ..., \delta_w)$.
- Class priors, $\underline{P}$, also get a Dirichlet prior: $\underline{P} \sim \mathcal{D}(\delta_P, ..., \delta_P)$.

The quantitative settings are similar to those used in [2]. The Gibbs sampler uses updates from the full conditional distributions in (3). For notational convenience we use $\underline{\Theta}_k$ for the parameters that determine class conditional densities. We use m as index over unlabeled data and c_m as latent class label. The index for all data is n, d_n are the latent kernel allocations and n_d the number of samples allocated by the d-th component. One distribution does not occur in the prior specification. That is $\mathcal{M}n(1, ...)$ which is a multinomial-one distribution. Finally we need some counters: m_1 ... m_K are the counts per class and m_{1k} .. m_{Dk} count kernel allocations of class-k-patterns. The full conditional of the kernel variance contains I to denote the number of model inputs.

$$
\begin{aligned}
p(c_m|\ldots) &= \mathcal{M}n\left(1, \left\{\frac{P_k\, p(\underline{x}_m|\underline{\Theta}_k)}{\sum_k P_k\, p(\underline{x}_m|\underline{\Theta}_k)}, k = 1..K\right\}\right) \qquad (3)\\
p(d_n|\ldots) &= \mathcal{M}n\left(1, \left\{\frac{w_{t_n d}\, p(\underline{x}_n|\underline{\Phi}_d)}{\sum_l w_{t_n d}\, p(\underline{x}_n|\underline{\Phi}_d)}, d = 1..D\right\}\right)\\
p(\underline{\beta}|\ldots) &= \Gamma\left(g + D\alpha, h + \sum_d \sigma_d^{-2}\right)\\
p(\underline{w}_k|\ldots) &= \mathcal{D}\left(\delta_w + m_{1k}, \ldots, \delta_w + m_{Dk}\right)\\
p(\underline{P}|\ldots) &= \mathcal{D}\left(\delta_P + m_1, \ldots, \delta_P + m_K\right)\\
p(\underline{m}_d|\ldots) &= \mathcal{N}\left((n_d\sigma^{-2}\underline{I} + \underline{\kappa})^{-1}(n_d\sigma^{-2}\underline{I}\underline{\bar{x}}_d + \underline{\kappa}\underline{\xi}), (n_d\sigma^{-2}\underline{I} + \underline{\kappa})\right)\\
p(\sigma^{-2}|\ldots) &= \Gamma(\alpha + I\frac{n}{2},\\
&\qquad \beta + \frac{1}{2}\sum_d \sum_{\underline{x}_n \forall n|d_n = d} (\underline{x}_n - \underline{m}_d)^T \underline{I}(\underline{x}_n - \underline{m}_d))
\end{aligned}
$$

Experiments

We used the classifier in the sampling paradigm to perform some preliminary experiments on our SIESTA data. The results reported below were obtained using the following features:

Inputs for classification
reflection coefficient at C3, 1 st. coeff.
reflection coefficient at C3, 3 rd. coeff.
Hjorth: cmplx./(act. + mob.) at C3

Training was done using data from 1 recording. We resampled the labels to equal class priors and added all other data unlabeled. Then we draw 10000 samples from the a-posteriori distribution over model coefficients. The first 5000 samples of the Markov chain are regarded as "burn in" and not used for predictions. In order to test the sleep analyzer, we generated a sleep profile for an independent test recording. We obtain three probability traces for stages wake, deep sleep and REM sleep as shown in figure 1.

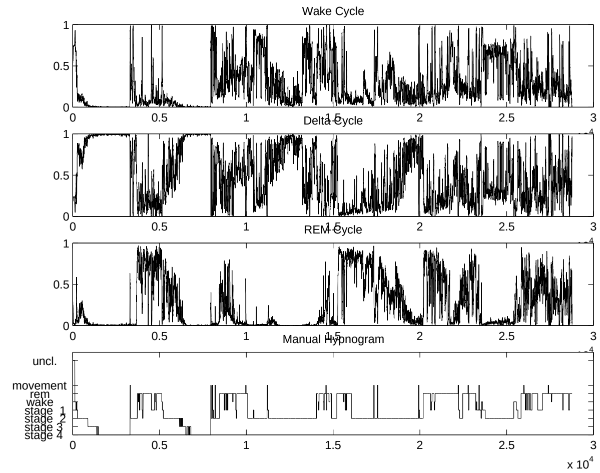

Figure 1: Probabilities for states wake, REM and deep sleep and the corresponding Rechtschaffen & Kales scoring. Note that we smoothed using a 120 seconds sliding window.

Conclusion

The results obtained with the classifier proposed in this contribution suggest that the approach is promising. However some problems need to be solved: We have to think about proper presentation of the results. The probability plots are technically correct but not easily comprehensible for the user. Some further evaluation is needed to test how the method can be used to discriminate different sleep disorders.

Acknowledgements

This work has been done within the project SIESTA, funded by the EC Dg. XII grant BMH4-CT97-2040. The authors want to express gratitude to the clinical partners for providing the R & K labels and all partners who contributed preprocessed features.

References

[1] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 1994.

[2] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *Journal Royal Stat. Soc. B*, 59:731–792, 1997.

[3] H. G. C. Tråvén. A neural network approach to statistical pattern classification by "semiparametric" estimation of probability density functions. *IEEE Transactions on Neural Networks*, 2:366–377, 1991.