BAYESIAN WRAPPERS VERSUS CONVENTIONAL FILTERS: FEATURE SUBSET SELECTION IN THE SIESTA PROJECT

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 sleep analyzer is inferred by semi-supervised techniques. Using good features is very important as we can not improve on them in subsequent processing stages. Hence we looked at various preprocessing techniques and applied feature subset selection techniques to select the most promising ones. This paper presents results obtained with two entirely different techniques: Classical feature subset selection based on feature evaluation criteria and search algorithms versus a methodology using Bayesian ideas: integrating over the *entire* space of probable feature subsets. We call this second approach the *Bayesian wrapper*.

Introduction

In the EC funded project SIESTA, we aim at defining a reliable and automatic description of human sleep. The analysis system is based on polysomnographic recordings of EEG, EMG and EOG channels, as well as other biosignals. Our target information are the most reliable labels for epochs of wake, deep sleep and rapid eye movement sleep. In order to avoid losing some important information, we decided to look at a range of different preprocessing methods, leading to a large number of features - 234 from 6 EEG channels alone. Statistics tells us that these are too many to use them all in the subsequent analysis and we have to use feature subset selection (FSS). Conventional FSS is done by searching for *one* promising subset. There are two ways to evaluate feature subsets: to minimize a general impurity measure or to measure performance of the final classifier. Since FSS is a problem of model selection, we have to use a significance test to decide how many features to use. Such a filter approach is our benchmark solution. However we think that in many problems different subsets explain the targets equally well and to prefer *one* of them cannot be justified. The answer to this problem is the Bayesian wrapper: We find the entire a-posteriori distribution over the space of all feature subsets and predict by marginalizing over this distribution.

Feature subset selection

The large initial number of features available here forces us to use sequential forward selection (sfs), a suboptimal search algorithm. This search algorithm is combined with two different feature evaluation criteria: We used the likelihood function of logistic regression and the so called gini index ([2]):

$$i(\nu) = \sum_{k \neq l} \hat{P}(k|\nu)\hat{P}(l|\nu) = 1 - \sum_{k} \hat{P}(k|\nu)^2. \qquad (1)$$

Evaluation of (1) needs estimates of the a-posteriori probabilities for class, $\hat{P}(k|\nu)$. We used a k nearest neighbor classifier to provide these. After having found a new promising feature, we use a statistical test that checks whether the classification accuracy increases significantly. We count the errors made by one classifier and not by the other (n_a) and vice versa (n_b). The difference is significant if (n_a,n_b) can not be explained as probable observation of a binomial $\mathcal{B}_n(0.5, n_a + n_b)$ distribution.

The Bayesian wrapper

The Bayesian approach to feature subset selection is to avoid it. The Bayesian wrapper samples the entire a-posteriori distribution over feature subsets and predicts by integrating over this distribution. In [3] such an approach has been applied to variable selection for

Table 1: Results from conventional FSS

Features from gini index
stochastic complexity at C3
Hjorth coefficient at Fp2, cmplx/(act*mob)

Features from logistic regression
reflection coefficient at C4, 1 st. coeff.
power spectral density at Fp1, Beta (12.5-30.0 Hz)
AR coefficient at C3, 2 nd. coeff.

linear regression. Our problem is slightly more complicated: the classifier is an architecture with a nonlinear dependence of predictions on some model coefficients. Hence the Bayesian model evidence can not be derived analytically and we have to resort to the reversible jump Monte Carlo sampling algorithm (reversible jump MC) recently proposed by [1]. This algorithm is an extension of Metropolis Hastings updates that allows sampling across different dimensional parameter spaces. The Bayesian wrapper uses two different moves. One consists of a dimension increasing and a matching dimension reducing step. The second move exchanges two inputs which allows "tunneling" through low likelihood regions. Fixed dimension sampling is carried out with Gibbs updates similar to those used in [4].

Experiments

We used data from four selected recordings of the SIESTA database. The targets are 30 seconds based Rechtschaffen & Kales (R & K) labels of states wake, REM and deep sleep. The following features were calculated for segments of one second duration: Coherence functions, power spectral densities, Kalman AR-coefficients, Hjorth coefficients, stochastic complexity measures and finally static reflection coefficients. The difficulty of this FSS was that the algorithms used different window lengths. Together with the 30 seconds based R&K scorings, this means that features can not be compared. Features with longer windows will be preferred. In order to avoid that longer windows are an advantage, we decided to run the FSS with the median segment of each 30 seconds epoch. Resampling to equal priors, we get 546 samples. The results of both algorithms are summaried in table 1.

The Bayesian wrapper was used with data from electrode C3 only. This reduces the total number of available features to 43. After drawing 10000 samples from the a-posteriori distribution of model coefficients and different dimensions, we discarded the first 5000 samples as burn in. The probabilities of feature subsets observed in the remaining samples are plotted in figure 1.

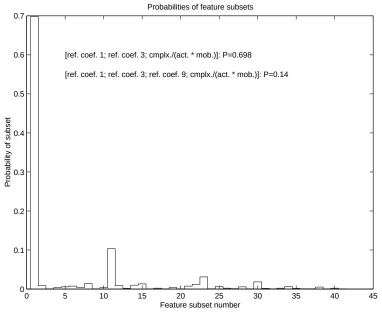

Figure 1: Probabilities of input subsets measuring their relevance.

Conclusion

The results obtained from feature subset selection suggest that the subsequent analysis should use groups of two or at most three features. The most promising features are static reflection coefficients, the Hjorth coefficients and the stochastic complexity measures. We also find power spectral estimates and Kalman AR coefficients to be of some importance.

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 provided their preprocessed features.

References

[1] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[2] D. J. Hand. *Construction and Assessment of Classsification Rules.* John Wiley and Sons, New York, 1997.

[3] D. B. Phillips and A. F. M. Smith. Bayesian model comparison via jump diffusions. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 215–239, London, 1996. Chapman & Hall.

[4] 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.