Classification of Clinical Time Series with Constrained estimation of Mixtures of HMMs

Ivan G. Costa[1], Alexander Schönhuth[2], Christoph Hafemeister[3] Alexander Schliep[3]

[1] Center of Informatics, Federal University of Pernambuco, Recife, Brazil
[2] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[3] Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

Abstract The use of molecular aspects of diseases for clinical diagnosis, has become increasingly popular. There are several difficulties in analyzing clinical gene expression data; high dimension of feature spaces vs. few examples, noise and missing data. We use constrained estimation of mixtures of hidden Markov models as a methodology to classify Multiple Sclerosis patient response to IFNβ treatment. The approach models the temporal nature of the data, is robust to noise and mislabeled examples. It also allows finding sub-groups of patients. Our method outperforms all previously method, and indicate the existence of biologically interesting sub-groups of patients.

Keywords: contrained mixture estimation, hidden Markov Models, gene expression time courses, clinical diagnosis

1 Introduction

The use of molecular aspects of diseases for clinical diagnosis, the so called personalized medicine, has become increasingly popular. One faces multiple challenges when analyzing clinical gene expression data; most of the well-known theoretical issues such as high dimension of feature spaces vs. few examples, noise and missing data apply. Special care is needed when designing classification procedures that support personalized diagnosis and choice of treatment. We analyse here the classification of interferon-β (IFNβ) treatment response in Multiple Sclerosis (MS) patients. Half of the patients remain unaffected by IFNβ treatment, which is still the standard. For them the treatment should be timely ceased to mitigate the side effects.

2 Method

We investigate the problem of classification of Multiple Sclerosis (MS) patients with respect to their response to Interferon-beta (IFNβ) treatment based on their gene expression profiles alone. IFNβ can still be considered to be the standard treatment in MS [Baranzini 2005]. To classify and further explore clinical differences between the groups of good and bad responders [Baranzini 2005], followed fifty-two patients for two years after initiation of IFNβ therapy. Every three months expression profiles of 70 genes were measured. Patients were divided into good and bad responders based on clinical

criteria such as relapse rate and disability status. They demonstrated that the patients' response could be predicted by studying gene expression profiles of the first time point after treatment alone [Baranzini 2005].

We propose constrained estimation of mixtures of hidden Markov models as a methodology to classify patient response to IFNβ treatment. By using HMMs with linear topology, our method takes the temporal nature of the data into account and allows the modelling of patient specific response rate [Schliep 2005]. Moreover, constraint based mixture estimation enables to explore the presence of response sub-groups of patients based on their expression profiles and allows the detection of misllabelled samples.

3 Results

We perform a 5 replications 4 fold application procedure and measure the classification accuracy in the test sets. Our method had a accuracy of 92% [Costa 2009]. It outperformed all prior approaches [Baranzini 2005, Borgwardt 2006, Lin 2008], which had a maximun acuracy of 88%. Additionally, we were able to identify potentially mislabeled samples, which was latter confirmed by the authors of the original paper [Baranzini 2005]. Furthermore, our results subdivide the good responders into two subgroups that exhibited different transcriptional response programs. This is supported by recent findings on MS pathology and therefore may raise interesting clinical follow-up questions.

References

[Baranzini 2005] Baranzini, S. E., Mousavi, P., Rio, J., Caillier, S. J., Stillman, A., Villoslada, P., Wyatt, M. M., Comabella, M., Greller, L. D., Somogyi, R., Montalban, X., and Oksenberg, J. R. (2005). Transcription-based prediction of response to ifnbeta using supervised computational methods. *PLoS Biol*, **3**(1), e2.

[Borgwardt 2006] Borgwardt, K. M., Vishwanathan, S. V. N., and Kriegel, H.-P. (2006). Class prediction from time series gene expression profiles using dynamical systems kernel. *Pacific Symposium on Biocomputing*, **11**, 547–558.

[Costa 2009] Costa, I. G., Schönhuth A., Hafemeister C., Schliep A. (2009) Constrained Mixture Estimation for Analysis and Robust Classification of Clinical Time Series *Bioinformatics*, To appear.

[Lin 2008] Lin, T. H., Kaminski, N., and Bar-Joseph, Z. (2008). Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, **24**(13), i147–i155.

[Schliep 2005] Schliep, A., Costa, I. G., Steinhoff, C., and Schönhuth, A. (2005). Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **2**(3), 179–193.