Learning Hidden Markov Model Topology

Alexander Schliep

schliep@zpr.uni-koeln.de
ZAIK/ZPR, Universität zu Köln

Abstract

Hidden-Markov-Models (HMMs) are a widely and successfully used tool in statistical modeling and statistical pattern recognition with gene finding being one of the prime examples in computational biology. One fundamental problem in the application of Hidden Markov Models is finding the HMMs underlying architecture or topology especially when there is no strong evidence towards a specific topology from the application domain (e.g., when doing black box modeling).

Topology is important with regard to good parameter estimates and with regard to performance: A model with "too many" states – and hence too many parameters – requires too much training data while an model with "not enough" states prohibits the HMM from capturing subtle statistical patterns.

To determine the "optimal" topology either knowledge from the application domain is used or a trial and error procedure using ad-hoc methods (i.e., model surgery) are employed; systematic procedures have been rarely considered (e.g., Bayesian Model merging, Stolcke and Omohundro).

We have developed a novel algorithm that will infer an HMM representation of the (ergodic) process generating a sequence, without prespecifying the topology of the model. That is, we infer the number of hidden states, and their transition and emission probabilities. We use a Bayesian approach where a suitable prior forces generalization (and necessarily reduces data likelyhood) from the maximum likelyhood model.

We will present the algorithm, some of our theoretical results and results from numerical experiments on biological sequence data.

Keywords: Hidden Markov Models, Bayesian statistics, Sequence Analysis