Strategies to avoid overfitting of MCMC Bayesian learning in some biological applications

Diego García
School of Systems Engineering and Computation, Universidad del Valle, Cali, Colombia
diego.mauricio.garcia@correounivalle.edu.co

Irene Tischer
School of Systems Engineering and Computation, Universidad del Valle, Cali, Colombia
irene.tischer@correounivalle.edu.co

Abstract

Model learning from observed data is typically affected by overfitting, because in order to find the model's best parameter set, all relations between data are used indifferently whether they represent relevant or noisy interactions.

Bayesian networks are widely used in biological modeling (e.g. networks of gene interactions), given that they allow representing graphically and determining statistically the dependence /independence relations between considered variables. A frequent approach in Bayesian learning is Markov Chain Monte Carlo simulation (MCMC), where a set of viable networks are explored by a random walk which converges to a network fitted optimally to data with respect to the likelihood or similar evaluation function.

Here we propose various strategies to mitigate overfitting in Bayesian learning by MCMC in order to reduce the resulting models' complexity. They either apply constraints inside the MCMC simulation or consider post-optimal operations. We show the effectiveness of these strategies in some biological applications.

Key words

Bayesian networks, Bayesian learning, MCMC simulation, overfitting.