Looking for applications of mixtures of Markov trees in bioinformatics

François Schnitzler[1], Pierre Geurts[1], Louis Wehenkel[1]

1. University of Liège, Montefiore Institute & GIGA research

Abstract

The probabilistic graphical model (PGM) framework provides efficient tools to model a probability distribution defined on a large set of variables using models composed of a graph and a set of parameters. The nodes of the graph are often in a one-to-one correspondence with the variables of the problem, and the edges present in the graph are related to the relationships between them, e.g. causal relationships or direct dependency. Parameters are usually grouped in local functions quantifying interactions between variables. As an example, bayesian networks use a directed graph and each local function is a conditional probability distribution of one variable given the variables associated to its parent-nodes in the graph. The product of those conditional distributions is the joint probability distribution over all variables.

Those models are largely used in computational biology, since they provide a visual representation that can be easily understood, can be used to answer a query using the distribution and can be learned automatically from a data set. From a set of gene expression and/or polymorphism data, PGM learning algorithms can infer a regulation network. PGMs have also been trained on databases (with or without expert intervention for specifying the structure) to predict gene position or function, protein structure or interaction loci, or the effect of a molecule on a given pathology. PGMs can also be employed for clustering of biomolecules, or to model time series data.

While PGMs have had already several successful applications in biology, their poor scaling to high dimensional problems (in terms of the number variables) may make them unfit to tackle problems of increasing size. Indeed, both inference and learning are NP-hard on general models, and in practice dealing with thousands of variables is already problematic. The complexity of those two operations is in fact linked by the tree-width of the graphical structure underlying the PGM, i.e. the size of the largest fully connected subgraph of a chordal graph containing that graphical structure, minus one.

It is therefore possible to ensure good algorithmic properties for PGMs by constraining their underlying structures to trees (graph without cycles, tree-width of one), but this constraint limits also the class of problems they are able to model faithfully. Using mixtures of trees, which model a distribution by a weighted sum of tree-structured models, each one defining a distribution on the whole set of variables, leads to improved modeling power while retaining the attractive algorithmic properties, but sacrifices the interpretability of the model.

Our research has so far focused on the development of new methods for learning such mixtures of trees from a data set, for problems with many variables and much fewer samples. Those methods were developed in the perturb and combine framework, where mixtures are constructed by averaging many models built by a randomized learning algorithm, allowing for variance reduction and further improving algorithmic complexity.

Our experiments on synthetic data have shown the interest of these methods, and we now wish to apply them to relevant problems in bioinformatics.