Advanced Statistics Project

Physics of Data

July 11, 2020

Learning the topology of a Bayesian network from a database of cases using the K2 algorithm.

A Bayesian belief-network structure is a directed acyclic graph in which nodes represent domain variables and arcs between nodes represent probabilistic dependencies [1].

Given a database of records, it is interesting to construct a probabilistic network which can provide insights into probabilistic dependencies existing among the variables in the database. Such network can be further used to classify future behavior of the modelled system [1]. Although researchers have made substantial advances in developing the theory and application of belief networks, the actual construction of these networks often remains a difficult, time-consuming task. An efficient method for determining the relative probabilities of different belief-network structures, given a database of cases and a set of explicit assumptions is described in [1] and [2]. The K2 algorithm [2] can be used to learn the topology of a Bayes network [1], i.e. of finding the most probable belief-network structure, given a database. After having studied the problem in the suggested literature ([1]-[2]), implement the algorithm in R and check its performances with the test dataset given in [2]. After having implemented and tested the K2 algorithm, investigate if it is possible to code it inside the `bnstruct` R package [3]-[4].

References

[1] G. F. Cooper and E. Herskovits, *A Bayesian Method for the Induction of Probabilistic Networks from Data*, Machine Learning **9**, (1992) 309

[2] C. Ruiz, *Illustration of the K2 algorithm for learning Bayes Net Structures*, http://web.cs.wpi.edu/~cs539/s11/Projects/k2_algorithm.pdf

[3] A. Franzin et al, *bnstruct: an R package for Bayesian Network structure learning in the presence of missing data*, Bioinformatics **33(8)** (2017) 1250

[4] F. Sambo and A. Franzin, *bnstruct: an R package for Bayesian Network Structure Learning with missing data*, December 12, 2016