Graphical and numerical methods for insight into BNs

Felix Andrews
The Australian National University

While Bayesian Networks make clear which variables are involved in each step of a causal chain, it is by no means obvious what effect each variable has. The effects of changing a variable state can be investigated interactively in an ad-hoc way, but this quickly becomes a hopeless exercise when interactions are considered. In a complex Bayesian Network, which has been built up from various data and information sources, it is important, but non-trivial, to summarise the behaviour of the model. This involves formulating the primary results of interest and extracting and presenting these effectively.

For instance, the direction and scale of causal relationships can be assessed graphically for face validity (Rykiel, 1996). Such checks are essential before going into more detailed analysis.

To generate a holistic summary, one method is to generate a large set of simulations from the BN, and then fit a tree model (Breiman et al, 1984), as a meta-model. Tree models can be presented graphically, allowing easy interpretation. However, one should be aware that the variables chosen at each step of the Decision Tree fitting process may themselves be somewhat sensitive to the precise dataset used. This is particularly true when input variables are correlated. To get a handle on such effects, a sequence of trees can be produced, with each identified primary variable excluded from the next tree. This can reveal the set of important variables and their substitutability for each other (due to correlation).

A more quantitative approach can be built on Random Forests (Breiman, 2001). A Random Forest is a large collection of tree models, where each tree is built from a random sub-sample of the data, and using a random sub-sample of the input variables. This makes it an extremely general model, and unlike single tree models, is not sensitive to the specific data set. In this context a variable’s importance can be defined as Permutation Importance: the reduction in predictive accuracy when a given variable is excluded from the model (by randomly permuting its values, thus destroying its information content). Recent advances have refined the method to account for heterogeneous input variables and correlated input variables (Strobl et al, 2007).

Several approaches will be applied in a case study, and their consistency and complementarity assessed.