Mark Paluta

DMU Project 1: Bayesian Network Structure Learning

Introduction

This report details a K2 structure learning algorithm. The algorithm was applied to three data sets of different sizes, using four, ten, and fifty nodes respectively. The algorithm generates a structure with an associated log Bayesian score, and compares this to the best known structure, from which the data was generated.

Algorithm

In one iteration of the K2 algorithm as designed, the nodes are first shuffled in order. Starting with the first node in this order, a single parent is assigned to maximize the resulting score. Next, another parent is assigned to the same child node in the same manner. This operation is terminated when the optimal parent to add decreases the score (which could be immediately, i.e. no parents for the child in consideration). The second node in the shuffle is then considered and the process of adding parents is repeated. This sequence continues until all nodes have been considered for parent addition.

K2 structure runs in polynomial time. The small data set took 1.8 seconds in JuliaBox for a single iteration (one shuffling of the nodes). The medium size took about 13 seconds, and the large size took around 2 hours. The long computational time is mostly in computing the log Bayesian score, which is computationally intensive for many nodes. Additionally, running on JuliaBox may have slowed the computations.

Structures

The best known log Bayesian score for the small network was -25672 which was found by this algorithm. Figure 1 shows the Bayesian network which gives this score.

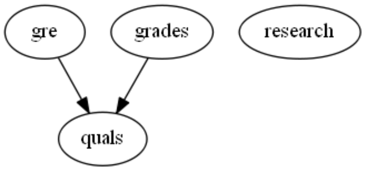

Figure 1. Discovered structure for small data set.

For the medium size network, again, the best known score was achieved. This score is -52120. The network is shown in Figure 2.

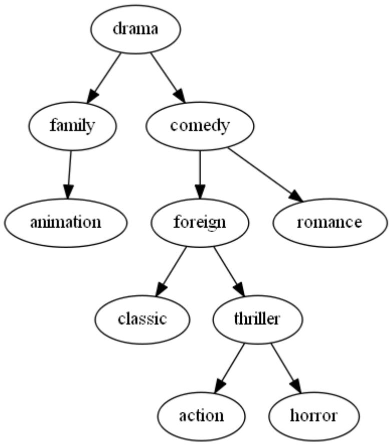

Figure 2. Discovered structure for medium data set.

The structure computed for the large network, on the other hand, did not yield the best known score of -259898. Instead a score of -263309 was achieved. Figure 3 shows the discovered network.

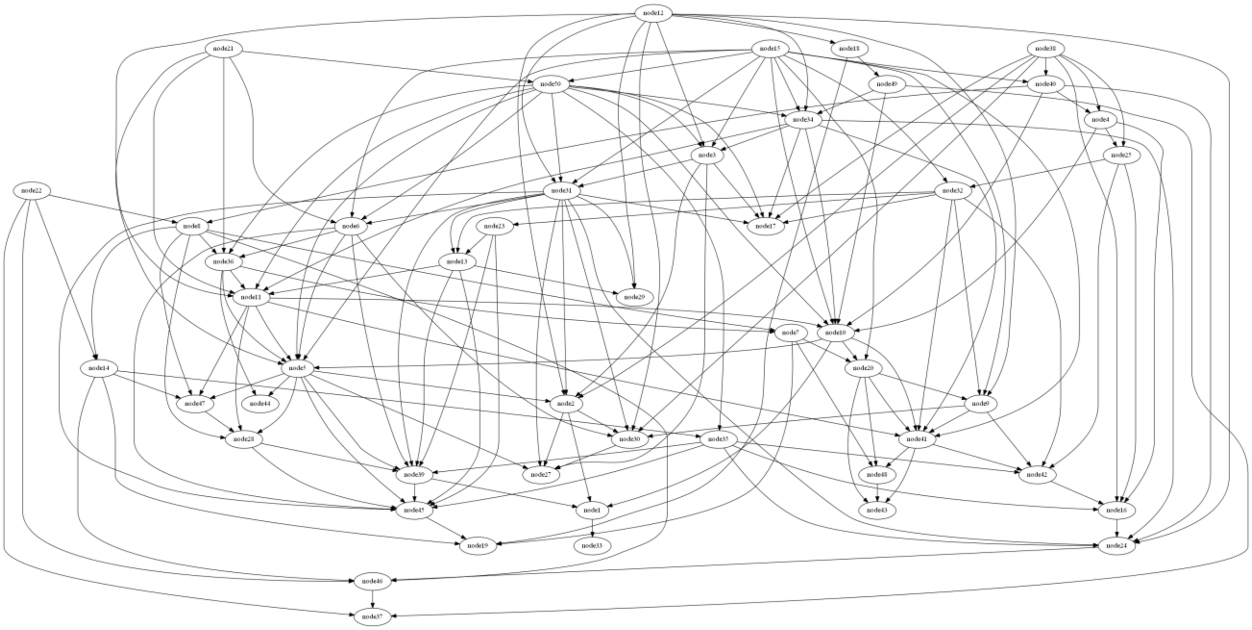

Figure 3. Discovered structure for large data set.

Conclusion

This algorithm could be improved by calculating differences in log Bayesian scores only on the node under consideration for the addition of parents. This is the only node whose score would change, so the algorithm could run much more quickly on this subsection of the entire network (that node and its parents). Despite this drawback in computation time, for Bayesian networks of ten or fewer nodes, the algorithm has proved effective at finding optimal structures.