Optimal Partitioning of Directed Acyclic Graphs with Dependency Between Clusters

Paul Pao-Yen Wu

School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

Directed Acyclic Graphs (DAGs) such as Bayesian Networks (BNs) can be used to infer parameter values or predict behaviour in complex systems with high dimensional data. Oftentimes, this inference process could be made more computationally efficient by partitioning (i.e. mapping) the DAG into clusters. In BNs, inference could refer to computation of posterior marginal probabilities given observations (evidence), or to update conditional probabilities of one or more nodes given data. Inference could also involve finding clusters such as homogeneous portions of a non-homogeneous system. The cost function to optimise for is arbitrary, and could include computational cost, AIC, likelihood or variance given a dataset. Computational cost is important as BN and Dynamic BN (DBN) inference, for instance, is NP-hard or worse. In addition, optimal partitioning is NP-hard, and the challenge is exacerbated by statistical inference as the cost to be optimised is dependent on both nodes within a cluster, and the mapping of clusters connected via parent and/or child nodes, which we call dependent clusters.

We discuss a novel algorithm called DCMAP which can, given an arbitrarily defined, positive cost function, iteratively and rapidly find near-optimal, then optimal cluster mappings. Shown analytically to converge to optimal solutions using dynamic programming, we use a simple example and a complex systems seagrass DBN to demonstrate the algorithm. For this 25 (one time-slice) and 50-node (two time-slices) DBN, the search space size was 9.91×10^{9} and 1.51×10^{21} possible cluster mappings, respectively, but near-optimal solutions with 88% and 72% similarity to the optimal solution were found at iterations 170 and 865, respectively. The first optimal solution was found at iteration 934 (95% CI 926,971), and 2256 (2150,2271) with a cost that was 4% and 0.2% of the naive heuristic cost, respectively. DCMAP opens up new research opportunities for combining optimisation with inference to support prediction and learning of DAGs with high dimensional data.