University at Buffalo The State University of New York

SCoRe group

High Performance Algorithms for Learning Exact Bayesian Networks

Subhadeep **KARAN**, Jaroslaw **ZOLA** · SCoRe – Scalable Computing Research Group
{skaran,jzola}@buffalo.edu Department of Computer Science and Engineering

Abstract

Bayesian Networks (BN) are probabilistic graphical models often used in biomedical applications. In these applications, BN structure is a critical component learned from data (e.g. medical records, gene markers, environmental factors). The problem of BN exact structure learning is to find a network structure that is optimal under certain scoring criteria. This problem has been shown to be NP-hard and the existing solutions are computationally and memory intensive.

We investigate new high-performance and parallel algorithms for learning globally optimal large-scale Bayesian Networks. In the process, we develop novel scalable solutions to sub-problems shared by various machine learning algorithms.

Why Bayesian Networks

- Clear and intuitive interpretation
- Explicit support to handle uncertainty in data
- Ability to answer complex and speculative queries
- Ability to incorporate expert knowledge

Challenges

- Irrespective of the scoring function, the search space of possible BN structures is super-exponential
- Existing approaches based on dynamic programming and memoization are infeasible even for small input data

References

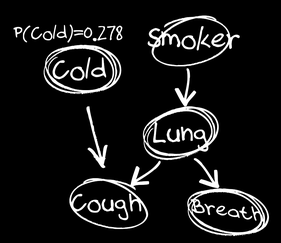
1. S. Karan, J. Zola, Exact Structure Learning of Bayesian Networks by Optimal Path Extension, In Proc. of IEEE BigData, 2016.

2. S. Karan, J. Zola, Scalable Exact Parent Sets Identification in Bayesian Networks Learning with Apache Spark, In Proc. of IEEE HiPC, 2017.

Bayesian Network

Pair (G,P) where G is a DAG over a set of random variables, and P is the corresponding probability distribution. G encodes efficient factorization of the joint probability P.

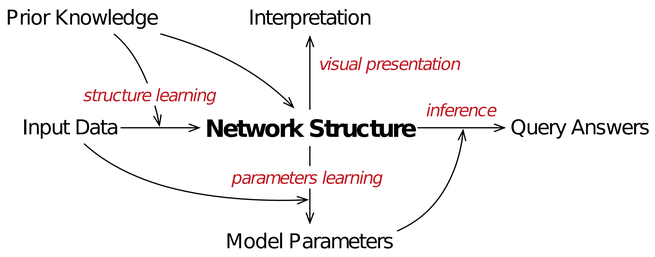
BN provides compact and easy to interpret representation of conditional independencies.



Introduction

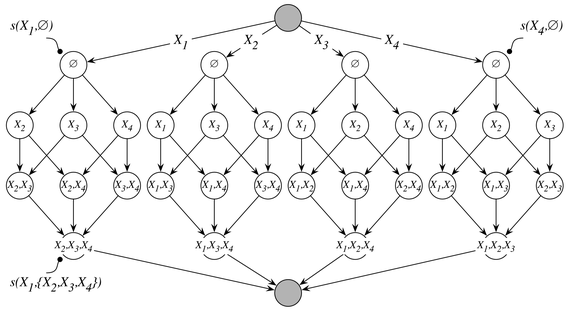
- **Exact BN Learning**: Given input data of observations for a set of random variables, and certain scoring function, find globally optimal BN explaining the data
- Two-step approach: identify optimal parent set for each variable given scoring function, find optimal structure by solving shortest path problem with distances defined by the choice of optimal parent sets
- By decomposing the problem, we can develop more efficient and scalable algorithms

Why Network Structure?

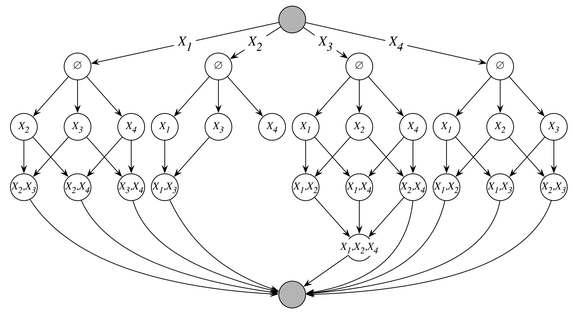


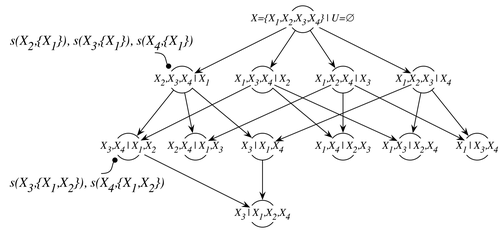
Exact Parent Set Identification (EPSI)

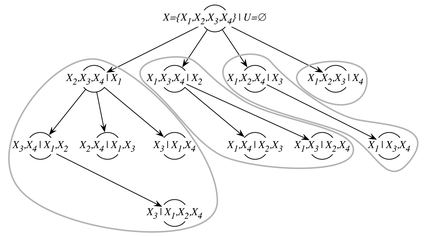
- Efficient parallel algorithm for distributed memory systems using Apache Spark [2]
- New strategy to constraint and reorganize dynamic programming computations such that computational grain is improved and fine-grained synchronization is avoided
- Currently, the only method capable of analyzing data sets with more than 40 variables



Example dynamic programming lattice for EPSI with four variables and the resulting transformations





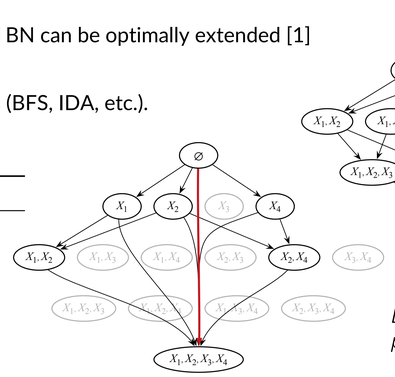


Optimal Path Extension (OPE)

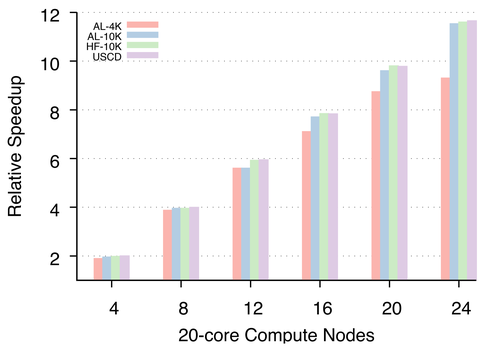
- Novel strategy to constraint the number of ways in which partial BN can be optimally extended [1]
- Negligible overhead, on-the-fly compaction of DP lattice
- Can be seamlessly combined with different shortest path solvers (BFS, IDA, etc.).

Dataset	n	BFS	URLearning	**SABNA**
Mushroom	23	2m12s	1m29s	1m
Autos	26	3m54s	37s	13s
Insurance	27	8m14s	7m25s	2m28s
Water	32	M	M	2m8s
Soybean	36	M	M	1h36m
Alarm	37	M	T	1h3m
Bands	39	M	T	1h10m

Comparison of SABNA with other structure searching methods



Effect of applying OPE to problem with four variables



Parallel scalability of our EPSI solver

SABNA Toolkit

- We are developing SABNA – Scalable Accelerated Bayesian Network Analytics Toolkit: *https://gitlab.com/SCoRe-Group/SABNA-Release*
- The toolkit is open source, and provides our optimized search strategies with support for different scoring functions

Acknowledgment

Authors wish to acknowledge support provided by the Center for Computational Research at the University at Buffalo.