Curriculum Learning of Bayesian Network Structures

Yanpeng Zhao1, Yetian Chen2, Kewei Tu1 and Jin Tian2

1School of Information Science and Technology, ShanghaiTech University, Shanghai, China
2Department of Computer Science, Iowa State University, Ames, IA, USA

Introduction

- Bayesian Network (BN)
  - A directed acyclic graph (DAG) where nodes are random variables and directed edges represent probability dependencies among variables
- BN Structure Learning
  - Firstly construct the topology (structure) of the network
  - Then estimate the parameters (CPDs) given the fixed structure
- Curriculum Learning (CL)
  - **Ideas:** learn with the simpler samples or easier tasks as the start
  - **Definition:** a curriculum is a sequence of weighting schemes of the training data \( (W_1, W_2, \ldots, W_n) \), where \( W_1 \) assigns more weight to easier samples, then each next scheme assigns more weight to harder samples, at last \( W_n \) assigns uniform weight to all samples.

Learn BN Structure via CL

- **Motivation:**
  - Given a set of variables, human rarely try to find the dependency relations between all variables by looking at all the training samples at once
  - Instead, human learn in a more organized way, starting with more common samples that involve dependency relations between only a small subset of variables
- **Curriculum in BN Structure Learning:**

  We define the curriculum as \( (X_{(1)}, \ldots, X_{(n)}) \), a sequence of selected subsets of the random variables \( X_{(i)} \), over which the corresponding subnet \( G_i \) is learnt.

  Where \( X = (X_1, \ldots, X_n) \) is a variable set, \( X_{(i)} \subseteq X \) and \( X_{(i)} = (X_1, \ldots, X_i) \).

  \( X_{(i)} \) is a sequence of intermediate learning targets.

Scoring Function

- **Bayesian Score Function**

  \[
  \log P(G_i, D_j) = C + \sum_{q=1}^{n} \log P(G_i, D_{j,q})
  \]

 Limitation and Solution

- **Limitation**
  - We only used a small fraction of the dataset at each learning stage
- **Solution**

  Let \( X'_{(i)} \) take value from \( \{X'_{(1)}, \ldots, X'_{(n)}\} \), the set of data segments \( D_i \) = \( \{D_{i,1}, \ldots, D_{i,k}\} \) by grouping samples based on the values of \( X_{(i)} \).

  - **Important Observation:** when we fix \( X_{(i)} \) to different values, our learning target is actually the same DAG structure \( G_i \) but with different parameters (CPDs).

  - **Assumption:** \( D_{i,1}, \ldots, D_{i,k} \) are generated by the same \( G_i \) but with independent CPDs.

  - We can revise the scoring function to take into account multiple versions of parameters.

Algorithm

**Algorithm 1: Curriculum Learning of BN Structures**

```plaintext
input: Variable Set X, Training Data D, Curriculum (X_{(1)}, \ldots, X_{(n)}). G_0 is initialized to a network containing variables in X_{(1)} with no edge.
for i = 1 \ldots m do
  Generate the set of data segments \( D_i = \{D_{i,1}, \ldots, D_{i,k}\} \) based on the values of \( X_{(i)} \).
  \( G_i \leftarrow \text{search}(D_i, X_{(i)}, G_{i-1}) \)
end
return \( G_m \)
```

Theorems

**Theorem 1.** For any \( i, j, k \) s.t. \( 1 \leq i < j < k \leq n \), we have

\[
\Delta d_{TV}(G_i, G_k) \geq \Delta d_{TV}(G_i, G_j)
\]

**Theorem 2.** For any \( i, j, k \) s.t. \( 1 \leq i < j < k \leq n \), we have

\[
\Delta d_{TV}(G_i, G_k) \geq \Delta d_{TV}(G_i, G_j)
\]

Experiments

- 10 benchmark BNs from the balearn repository (alarm,andes,asia,child,heartfinder,hepar2,insurance,sachs,water,win95pts)
- Comparisons with MMHC under metrics of BDeu, BIC, KL and SHD

**Figure 1:** Variables, \( S, B, D, L, E, X, A, T \) correspond to each column of the dataset respectively. **Left:** at stage 1, learn a subnet \( G_1 \) over \( (S, B, D) \) from scratch with the rest variables fixed at \((0, 1, 0)\). **Right:** at stage 2, learn a larger subnet \( G_2 \) over \( (S, B, D, L, E, X) \) with \( G_1 \) as the start point. Each query is a sequence of selected subsets over which the corresponding subnet \( G \) is learnt.

**Figure 2:** Left: using the previous method, learn a subnet \( G_i \) over \( X_{(i)} = \{S, B, D\} \). We only used samples with \( X_{(i)} = \{L, E, X, A, T\} \) fixed at \((1, 1, 1, 0, 0)\), the samples with a strikeout is NOT used. Right: using the new method, learn a subnet \( G_{(i)} \) over \( X_{(i)} = \{S, B, D, L, E, X\} \) when \( A \) takes value from \((0, 0, 0, 1, 0, 0)\) and \((1, 1, 1, 0, 0, 0)\), we divide the dataset into three partitions by grouping samples based on the values of \( X' \) and use all of them.

**Figure 3:** SHD between the intermediate learning result at each stage and the target BN.

Conclusions

- We propose a novel heuristic algorithm for BN structure learning.
- We tailored the Bayesian scoring function for our algorithm.
- We proved two theorems that show theoretical properties of our algorithm.
- We have empirically shown that our algorithm outperformed the state-of-the-art MMHC algorithm in learning BN structures.