Bayesian Network Structure Learning: The Two-Step Clustering-Based Algorithm

Yikun Zhang[1], Jiming Liu[2], and Yang Liu[2]

[1]School of Mathematics, Sun Yat-sen University (yikunzhang@foxmail.com) [2]Department of Computer Science, Hong Kong Baptist University

Problem

Structure learning is a fundamental and challenging issue in dealing with Bayesian networks. As the number of variables (nodes) increases, traditional structure learning algorithms are time-consuming and could yield some network structures that hardly describe the original data sets.

Although prior information helps to ameliorate time efficiency and accuracies, practitioners are still facing two problems during the applications of existing algorithms.

1. Availability of prior information
2. Effectiveness of prior information

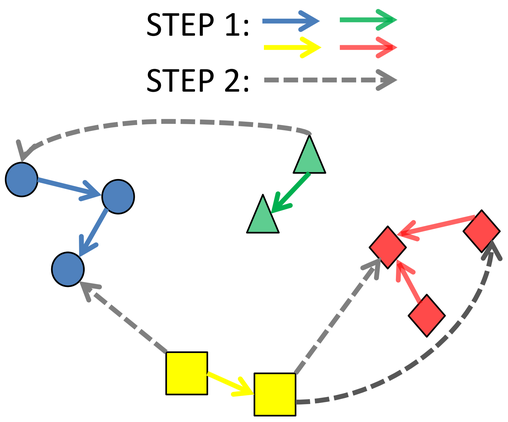
Contributions

We proposed a two-step clustering-based (TSCB) strategy, which divides the network variables into small clusters and generates some strongly prospective arcs in the first step as prior information.

Our main contributions fall in two categories,

1. Propose an automatic approach to generate prior information directly from data
2. Improve the performances of a wide range of conventional structure learning algorithms in terms of time efficiency and accuracies

Method



Dissimilarity Metric for Clustering:

- Data sets with only discrete variables: negative *mutual information*.
- Data sets with only continuous variables: *1-(Pearson's) correlation*.
- Hybrid data sets:
  1. **Conversion**: Label attributes of discrete (or categorical) variables by nonnegative integers
  2. **Centralization**: Shift the variables such that their attributes are central at 0
  3. Apply *1-(Pearson's) correlation*.

Algorithm

Input:

- Data set $\mathcal{D} = \{X_1, X_2, ..., X_N\}$ with N variables
- The number of clusters: K (Parameter)

Step 1:

1. Compute the dissimilarity matrix.
2. Carry out clustering analysis via *average linkage agglomerative clustering method* and cut the dendrogram into K groups (clusters).
3. Learn Bayesian network structures within each cluster using a traditional algorithm A[1].

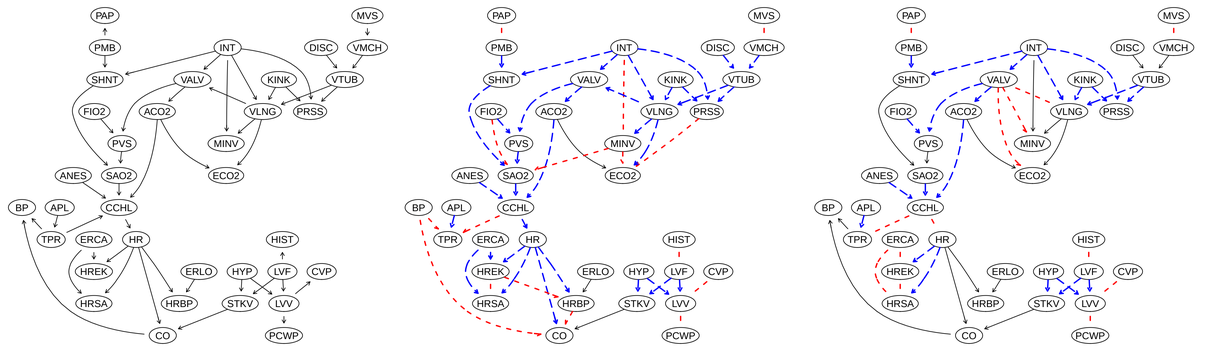
Step 2:

1. Apply the algorithm A again on all variables with the retained arcs to combine clusters.

Output: Bayesian network structure learned from the data set $\mathcal{D}$.

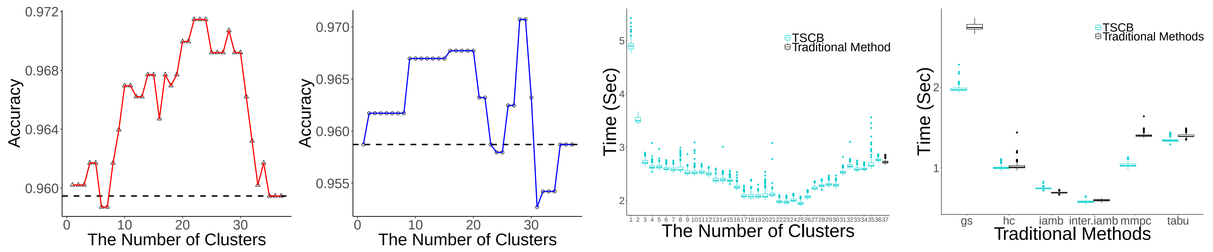
[1]This could be any traditional structure learning algorithm, like the Grow-Shrink algorithm[1].

Results



(a) Actual Network ("alarm") (b) Learned by the GS Algorithm* (c) Learned by the TSCB Algorithm*

*The red dashed arcs represent false positive arcs while the blue ones are false negative arcs.



(a) Embed GS† (b) Embed HC† (c) Time Variations (d) Time Comparisons

†The horizontal dashed line indicates the accuracy of the embedded traditional algorithm without TSCB strategy.

Table 1: Synthetic Benchmark Data Sets for Experiments

Network Data	Number of nodes	Number of arcs	Average degrees
"asia"	8	8	2.00
"insurance"	27	52	3.85
"alarm"	37	46	2.49
"hepar2"	70	123	3.51

Accuracy Analysis

Table 2: Accuracies of Traditional Algorithms *With and Without* TSCB Strategy. (The records inside round brackets are accuracies of the embedded traditional algorithms without TSCB strategy.)

Methods	"asia"	"insurance"	"alarm"	"hepar2"
GS	**0.9096** (0.8918)	**0.9309** (0.9263)	**0.9662** (0.9602)	**0.9763** (0.9753)
IAMB	**0.9084** (0.8896)	**0.9287** (0.9218)	**0.9715** (0.9686)	**0.9747** (0.9741)
Inter-IAMB	**0.9082** (0.8936)	**0.9281** (0.9208)	**0.9716** (0.9689)	**0.9748** (0.9742)
MMPC	**0.8557** (0.8546)	**0.9259** (**0.9259**)	**0.9649** (0.9646)	**0.9732** (0.9728)
HC	**0.9766** (**0.9766**)	**0.9328** (0.9293)	**0.9768** (0.9724)	**0.9824** (0.9822)
TABU	**0.9664** (0.9657)	**0.9422** (0.9312)	**0.9788** (0.9744)	**0.9814** (0.9810)

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population}$$

Traditional structure learning algorithms:

- Constraint-based methods: Grow-Shrink (GS); Incremental Association Markov Blanket (IAMB); Interleaved Incremental Association (Inter-IAMB)
- Score-based methods: Hill-climbing (HC); Tabu greedy search (TABU)
- Hybrid methods: Max-Min Parents and Children (MMPC)

Time Efficiency Analysis

Table 3: Average Elapsed Times Comparison (Embedded Grow-Shrink algorithm on "alarm")

Records (in Secs)	"asia"	"insurance"	"alarm"	"hepar2"
Clustering	0.00230	0.00788	0.01076	0.04432
Within Clusters	0.00464	0.01670	0.05012	0.04744
Between Clusters	0.00962	0.16420	0.24640	1.46168
TSCB	0.01656	**0.18878**	**0.30728**	**1.55344**
Traditional	**0.01010**	0.19362	0.35900	1.65584

References

[1] Margaritis, D. 2003. *Learning Bayesian Network Model Structure from Data.* Ph.D. Dissertation, Pittsburgh, USA.

[2] Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network Motifs: Simple Building Blocks of Complex Networks. *Science* 298(5594):824-827

Future Directions

The optimal state of our TSCB strategy always attains when nearly all the clusters contain no more than three variables, which might be the primitive units of the network structure. Therefore, we are interested in the intrinsic connections between each detected cluster and the concept of "network motifs"[2].

Additionally, the robustness of our TSCB strategy in the presence of latent variables is a possible direction for future research.

Acknowledgement

Yikun Zhang is grateful to the unswerving support and liberal nurture from his family. He also thanks Department of Computer Science, Hong Kong Baptist University for a precious summer research opportunity.