

A Comparison of Bayesian Network Learning Algorithms from Continuous Data

Lawrence Fu, Ioannis Tsamardinos PhD

Department of Biomedical Informatics, Vanderbilt University, Nashville, TN

Abstract. Learning a Bayesian network from data is an important problem in biomedicine for the automatic construction of decision support systems and inference of plausible causal relations. Most Bayesian network learning algorithms require discrete data; however discretization may impact the quality of the learned structure. In this project, we present a comparison of different approaches for learning from continuous data to identify the most promising one and to quantify the impact of discretization in Bayesian network learning.

Problem Description. Despite the wide applicability of Bayesian networks in biomedicine, the fact that most Bayesian network structure learning algorithms require discrete data is a limitation since biomedical and biological data are routinely continuous. Studies usually employ simple discretization techniques such as frequency-based partitions. By neglecting to adequately address the ramifications of discretization, researchers unknowingly may lose information such as interactions and dependencies between variables and impact the learned structure. Unfortunately, there is no consensus on a standard procedure for discretization. Consequently, it is still an unresolved research question as how to best handle continuous data.

There are three typical approaches to learning network structure with continuous data. First, data can be discretized prior to and independent from the application of the learning algorithm. Second, the discretization can be integrated into the learning phase in an effort to exploit the synergies. Algorithms following this approach output a discretization of the input variables and the network structure. Third, learning can be done directly with continuous data without committing to a specific discretization for the variables.

Purpose. This project has two major components. First, it comprehensively compares the three different approaches in order to ascertain the relative strengths and weaknesses of each and to quantify the impact of discretization in network learning. Secondly, it presents a toolkit of discretization and learning techniques for use by biomedical researchers. The specific algorithms that are compared are:

1) Prediscretization methods: equal-frequency/equal-width discretization, Hartemink's method¹

2) integrated methods: Monti's method², Friedman's method³, Steck's method⁴

3) direct handling of continuous data methods: Bach's method⁵, Margaritis' method⁶, Davies's method⁷, PC algorithm⁸ with Fisher's Z-test

Despite the lack of well-characterized gold standards for continuous Bayesian networks, two types of data have been used. First, continuous data were simulated from existing discrete networks used in biomedical systems and research. In this case, the metric of comparison will be the number of added and missing edges of the learned structure relative to the true known structure. Second, real biological data with unknown structure were used. For methods that commit to a discretization, the BDeu score⁹ was used as a measure of the fit of learned structure to the data. For the methods in (3) above that do not discretize the data, other scoring functions for continuous data have been employed^{5,7}.

Acknowledgement. This work was supported by a Training Grant from the National Library of Medicine (T15 LM 007450-03).

References

- ¹Hartemink A, Gifford D, et al. Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Networks. In: PSB, 2002.
- ²Monti S, Cooper GF. A Multivariate Discretization Method for Learning Bayesian Networks from Mixed Data. In: UAI, 1998.
- ³Friedman N., Goldszmidt M. Discretizing Continuous Attributes While Learning Bayesian Networks. In: ICML, 1996.
- ⁴Steck H, Jaakkola T. (Semi)-Predictive Discretization during Model Selection. AI Memo AIM-2003-002, 2003.
- ⁵Bach FR, Jordan MI. Learning Graphical Models with Kernel Methods. In: NIPS, 2002.
- ⁶Margaritis D. Distribution-Free Learning of Graphical Model Structure in Continuous Domains. TR-ISU-CS-04-06, Iowa State University, 2004.
- ⁷Davies S, Moore A. Interpolating Conditional Density Trees. In: UAI, 2002.
- ⁸Spirtes P, Glymour C, Scheines R. Causation, Prediction, Search. Cambridge: MIT Press; 2000.
- ⁹Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Machine Learning 1995; 20:197-243