Noname manuscript No.
(will be inserted by the editor)

Learning Discrete-valued Bayesian Network from Mixed Data

First Author · Second Author

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords Discretization · Bayesian Network · More

1 Introduction

Bayesian networks (Pearl 1988; Koller and Friedman 2009) are an increasingly popular methods for modeling uncertainty and causality in science and engineering. They provide an efficient factorization of the joint probability distribution over a set of random variables. Bayesian networks first emerge from artificial intelligence research and have been applied to a wide variety of problems, ranging from decision-making systems (Kochenderfer 2015) to medical diagnoses. In most cases, we assume that all random variables in Bayesian networks are discrete, since many algorithms on Bayesian networks are unable to deal with continuous variables efficiently. However, this assumption is often too restrictive. For example, in the decision-making system of autonomous cars, it is inevitable to deal with continuous variables such as position and velocity.

There are two solutions around this assumption. The first one is to model conditional probability density of each continuous variable by specific families of parametric distributions, then redesign algorithms on Bayesian networks based on these parameters. One successful example is the belief propagation

F. Author
first address
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: fauthor@example.com

S. Author
second address

in Gaussian graphical models (Weiss and Freeman 2001). Nevertheless, for other shapes of particles (Ihler and McAllester 2009) or non-linear functions, algorithms are computationally expensive and do not perform well.

The second solution is to discretize continuous variables. Discretization that learns from data has developed well and discussed (Dougherty et al. 1995) in the fields of machine learning and statistics for many year. Most of these discretization methods are designed for classification problems. They search the best discretization of a continuous attribute by considering its interation with the class variable of interest. However, these discretization methods do not apply to continuous variables in Bayesian network. In Bayesian network, interations and dependencies between variables are determined by graph structure. Therefore, if a discretization method only uses the interation between a continuous variable and its class variable of interest, instead of considering the graph structure, then it is not an appropriate discretization method for Bayesian networks. There are some reseach on discretizating continuous variables in naive Bayesian network and tree augmented network (Friedman et al. 1997). Nevertheless, only few discretization methods on general Bayesian netowrk have been proposed (Friedman and Goldszmidt 1996; Kozlov and Koller 1997; Monti and Cooper 1998; Stech and Jaakkola 2007).

The discretization technique proposed by N. Friedman and M. Goldszmidt Friedman and Goldszmidt (1996) is the most well-known one among these methods. It is based on minimum description length principle (MDL): optimal discretization policy minimizes the description length of Bayesian network and data information. If there is only one continuous variable in Bayesian network and other variables are discrete, MDL discretization method takes running time $O(N^3 + n_c {v_{max}}^{(n_c^p)_{max}} \cdot N^2 + {v_{max}}^{n_p} \cdot N^2)$, where N is the number of data instances for learning the discretization, n_p and n_c are the numbers of parent and children variables for the continuous variable, v_{max} is the largest cardinality number over all variables in Markov blanket, and $(n_c^p)_{max}$ is the largest number of parent variables of the continuous variable's children.

In this paper, we propose a new discretization method for continuous variables in Bayesian network by learning from mixed data. This method is a generalized version of Boullé (2006) and Lustgarten et al. (2011), which are both discretization methods of a continuous attribute with one class variable. We begin our method with the situation that only one variable in a given Bayesian network is continuous and other variable are discrete. Furthermore we assume that the network structure is known in advance. Under this situation, we look for the most possible discretization policy M given the data on other discrete variables. That is to say, the desired policy is $\arg_M P(M|D)$. With Bayes rule, $P(M|D) = P(M) \cdot P(D|M)/P(D) \propto P(M) \cdot P(D|M)$. Usually, $P(D|M)$ increases as number of discretized intervals increases, since more intervals can provide more accuracy. Therefore, we design probability priors of $P(M)$ to decrease as number of intervals rises. As a result, we obtain a nature trade-off to determine the number of intervals after discretization. Besides, the $P(D|M)$ can be factorized according to Bayesian network structure. On the other hand, with the proposed priors, we are able to restrict the number of discretized intervals not exceeding the largest cardinality number of variable in Markov blanket too much. This is important, since for most algorithms on Bayesian network, their running times exponentially depend on the cardinality of variables. Another advantage of our algorithm is running time. If there is only one continuous variable, the running time is $O(n_c v_{max} \cdot N^2 + {v_{max}}^{n_p} \cdot N^2)$, which beats the running time of MDL principle discretization (Friedman and Goldszmidt 1996). With an effective approximation proposed in this paper, we can further reduce running time to $O(v_{max}(n_p + n_c)N^2)$.

For a Bayesian network with multiple continuous variables, we apply the discretization method discussed above iteratively. In the beginning, we discretize each continuous variable by equal-width method. Number of intervals for equal-width discretization is equal to largest cardinality of discrete variables. Then we iterate the one-variable discretization on each continuous variable by the following order: from the variable with highest topological order (leaves) to the variable with lowest topological order (root). Everytime when we finish a discretization procedure on one variable, we store the discretization result and this discretization result will be used for later discretization on another variable. Experiments on real data show that with same iteration order, our discretization provides better discretization result than MDL principle method in terms of likelihood. The latter is easily catched by local minima and thus produce non-necessarily low number of intervals after discretization.

Finally, we can combine the new discretization method with K2 structure learning alorithm (Cooper and Herskovits 1992). We first prediscretize all continuous variables before K2 learning, since K2 algorithm requires all variables are discrete. Everytime when an edge is added by K2 algorithm procedure, we rediscretize the relevant variables in Bayesian network and store the result for next K2 algorithm iteration. By this principle, we are able to learn a discrete-valued Bayesian network from mixed data.

The rest of the paper is organized as follows: we prepare the preliminaries of Bayesian network in Section 2. Related work, including MDL principle discretization (Friedman and Goldszmidt 1996) and MODL discretization by Boullé (2006), are summarized in Section 3. In Section 4, we introduce the new discretization method, including formulation of objective function and algorithms. One continuous variable case and multiple continuous variable case are also included in this section. Finally, in Section 5, we apply the new discretization method to real-world data and show the result.

2 Preliminaries

In this section we provide a brief review of Bayesian network, including the factorization of joint probability distribution, sampling from a given network, and strucutre learning. These concepts will be used in later sections.

2.1 Bayesian Network

A Bayesian network B is defined by a pair (G, Θ), where $G = (X, E)$ is a directed acyclic graph whose nodes correspond to a set of random variable $X = \{X_1, X_2, \cdots, X_n\}$, and whose edges E represents probabilistic dependencies among nodes. The graph structure G encodes the Markov property: each node X_i is independent of its non-descendants given its parents in G. The second component of B, namely Θ, contains a set of parameter that qualifies the network. Elements of Θ take the form $\theta_{x_i|\Pi_{x_i}} = P(x_i|\Pi_{x_i})$ for each possible value x_i of X_i, and Π_{x_i} of Π_{X_i} (the set of parents of X_i in G). Due to Markov property, we can represent the multivariate joint distribution over X as

$$P_B(X_1, \cdots, X_n) = \prod_{i=1}^{n} P_B(X_i|\Pi_{X_i}) = \prod_{i=1}^{n} \theta_{x_i|\Pi_{x_i}}$$

For a given Bayesian network, we can generate its data instances by forward sampling, i.e., sampling variables one by one in a topological order (see Cormen et al. 2009, chap. 22). Given a instance of X_i's parent set Π_{X_i}, value of X_i can be sampled according to the conditional probability table $P(X_i|\Pi_{X_i})$. Furthermore, this parent-child sampling order can be reversed if we know the marginal probability of child variable $P(X_i)$. By Bayes rule,

$$P(X_i|\Pi_{X_i}) \cdot \prod_{j=1}^{Pa(X_i)} P(\{\Pi_{X_i}\}_j) = P(X_i) \cdot P(\Pi_{X_i}|X_i),$$

where $Pa(X_i)$ is the number of parents of X_i, and $\{\Pi_{X_i}\}_j$ is the jth parent of X_i. We can first sample X_i by $P(X_i)$, then sample all the parent varaibles simultaneously by $P(\Pi_{X_i}|X_i)$. That is to say, X_i becomes the starting point for sampling.

2.2 K2 Structure Learning

Roughly speaking, there are three approaches to learn a Bayesian network structure from data (see Koller and Friedman 2009, chap. 18). They are constraint-based structure learning, score-based structure learning, and Bayesian model averaging. Here we review K2 structure learning algorithm (Cooper and Herskovits 1992), which is one of the most successful score-based structure learning method. Same as most structure learning algorithms, K2 algorithm requires all variables are discrete. The score of a learned network is defined as $\prod_i f(X_i, \Pi_{X_i})$, where

$$f(X_i, \Pi_{X_i}) = \prod_{j=1}^{|\Pi_{X_i}|} \frac{(|X_i| - 1)!}{(N_{ij} + |X_i| - 1)!} \prod_{k=1}^{|X_i|} \alpha_{ijk}!. \tag{1}$$

$|X_i|$ is the cardinality of variable X_i. $|\Pi_{X_i}|$ is number of all possible instantiations of the parent variables of X_i, i.e., $|\Pi_{X_i}| = \prod_{Y \in \Pi_{X_i}} |Y|$. α_{ijk} is number of instances in data set that variable X_i is instantiated with its kth value, and the parent of X_i are instantiated with the jth value of Π_{X_i}. $N_{ij} = \sum_{k=1}^{|X_i|} \alpha_{ijk}$. K2 algorithm searchs the network structure with highest score, which can be interpreted as the most probable network with the given data. Besides, K2 requires the topological order of variables be known before scoring can start. This constraint can prevents cycles from being introduced. The searching for high-score network is a iterative process. There is no way to find the optimal network directly, since there are $2^{O(n^2)}$ possible structures, where n is number of variables. In order to compensate for this, we run K2 algorithm for many times, and each time we start with a different order of variables. The network with highest score in iterations is the desired one.

3 Related Work

In these section we review two related works: MDL principle discretization (Friedman and Goldszmidt 1996) and MODL discretization (Boullé 2006). The former is the most famous discretization method for continuous variables in Bayesian network, and we will compare it with our proposed method in Section 6. The latter is a discretization method for one continuous attribute according to a target class. Our proposed method is a generalization of this method.

3.1 MDL Principle Discretization

3.2 MODL Discretization

4 Discretize One Continuous Variable

4.1 Notations

4.2 Objective Function

4.3 Algorithm

4.4 Approximation

5 Discretizing Multiple Continuous Variable with Structure Learning

6 Experiments

References

M. Boullé. Modl: A bayes optimal discretization method for continuous attributes. *Machine Learning*, pages pp131–165, 2006.

G. F. Cooper and E. Herskovits. A bayes method for the induction of probabilistic network from data. *Machine Learning*, 9:pp.309–147, 1992.

T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stain. *Introduction to algorithms*. The MIT Press, 3 edition, 2009.

J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *12th International Conference on Maching Learning (ICML)*. Morgan Kaufman Publishers, San Francisco, CA, 1995.

N. Friedman and M. Goldszmidt. Discretizing continuous attributes while learning bayesian networks. In *13th International Conference on Maching Learning (ICML)*. Morgan Kaufman Publishers, San Francisco, CA, 1996.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:pp.131–163, 1997.

A. Ihler and D. McAllester. Particle belief propagation. In *Artificial Intelligence and Statistics*, 2009.

M. J. Kochenderfer. *Decision making under uncertainty*. MIT Lincoln Laboratory Series, 2015.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

A.V. Kozlov and D. Koller. Nonuniform dynamic discretization in hybrid networks. In *13th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1997.

J. L. Lustgarten, S. Viswewaran, Gopalakrishnan V., and G. F. Cooper. Application of an efficient bayesian discretization method to biomedical data. *BMC Bioinformatics*, 2011.

S. Monti and G.F. Cooper. A multivariate discretization method for learning bayesian networks from mixed data. In *14th International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. *Morgan Kaufman Publishers, Inc.*, 1988.

H. Stech and T. Jaakkola. Predictive discretization during model selection. In *11th International Conference on Artificial Intelligence and Statistics*, 2007.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, 2001.