Knowledge-Based Systems 150 (2018) 95–110

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Novel binary encoding water cycle algorithm for solving Bayesian network structures learning problem

Jingyun Wang, Sanyang Liu*

School of Mathematics and Statistics, Xidian University, Xi'an 710126, China

ARTICLE INFO

Article history:
Received 30 October 2017
Revised 28 February 2018
Accepted 2 March 2018
Available online 7 March 2018

Keywords:
Water cycle algorithm
Binary encoding
Heuristic algorithm
Bayesian network
Structure learning

ABSTRACT

Constructing Bayesian network structures from data is a computationally hard task. One important method to learn Bayesian network structures uses the meta-heuristic algorithms. In this paper, a novel binary encoding water cycle algorithm is proposed for the first time to address the Bayesian network structures learning problem. In this study, the sea, rivers and streams correspond to the candidate Bayesian network structures. Since it is a discrete problem to find an optimal structure, the logic operators have been used to calculate the positions of the individuals. Meanwhile, to balance the exploitation and exploration abilities of the algorithm, the ways how rivers and streams flow to the sea and the evaporation process have been designed with the new strategies. Experiments on well-known benchmark networks demonstrate that the proposed algorithm is capable of identifying the optimal or near-optimal structures. In the comparison to the use of the other algorithms, our method performs well and turns out to have the better solution quality.

1. Introduction

Bayesian networks (BNs) combine graph and probability theories to obtain a comprehensible representation of the joint probability distribution. Bayesian networks have been seen as one of the best way to represent causal knowledge and used in reasoning and decision making tasks in uncertain domains. Due to their powerful representation, inference and learning abilities, Bayesian networks have become increasingly popular in large number of research areas, such as risk analysis [21,32,41], bioinformatics research [38,39], medical problem [29,35] and image processing [23], etc.

The learning task in Bayesian networks can be grouped into two subtasks: structure learning and parameter estimation. The first subtask is to identify the best topology for a network, and the second subtask is to learn the parameters that define the conditional probability distribution for a given network topology. Throughout the last decade, there has been a growing interesting in the area of learning the structure of Bayesian networks from data. The structure learning can be considered the problem of selecting a probabilistic model that explains the given data. A wealth of literature are available to provide methods of learning BN structures. Roughly speaking, there are three methods to learning the structures of Bayesian network: constraint-based [22,33] , score-based [1,15] and hybrid learning [34] algorithms. The constraint-based algorithms construct BNs by analyzing conditional independent relations between nodes in network. The score-based algorithms learn BNs by maximizing the scores of candidate structures with some heuristic search algorithms. The hybrid algorithms combine aspects of both constraint-based and score-based algorithms, they use conditional independence tests and network scores at the same time. The focus of this paper is on the score-based algorithms rather than the constraint-based methods. Learning the Bayesian network structures from data is an NP-hard problem [8], the number of possible structures grows super-exponentially by the number of nodes [25]. To efficiently search the optimum or near-optimum in the space of possible solutions, the meta-heuristic algorithms have been used for finding out optimal structures, such as genetic algorithm (GA) [6,18], evolutionary programming (EP) [36], ant colony optimization (ACO) [10,24], artificial bee colony (ABC) [17], bacterial foraging optimization (BFO) [40] and particle swarm optimization (PSO) [2,14,37].

Several heuristic algorithms have been shown that they are promising candidates for solving BN structures learning problem. For instance, an artificial bee colony algorithm for learning Bayesian networks (ABC-B) [17]; structural learning of Bayesian networks by bacterial foraging optimization (BFO-B) [40]; BNC-PSO: structure learning of Bayesian networks by particle swarm optimization (BNC-PSO) [14]. However, some drawbacks still exist in these algorithms. ABC-B and BFO-B algorithms generate candidate solutions based on neighbor searching. When ABC-B is imple-

* Corresponding author.
E-mail addresses: ygbai@stu.xidian.edu.cn (J. Wang), liusanyang@126.com (S. Liu).

https://doi.org/10.1016/j.knosys.2018.03.007
0950-7051/

mented, the inductive pheromone transmission mechanism guides the search for global solution, the risk of the algorithm be trapped in the local optimum is increased. Although BFO-B has the certain probability of jumping out the local optimum, it generally spends much time in finding out the optimal solution. BNC-PSO generates a new solution based on mutation and crossover operators, the algorithm converges fast and obtains the optimum in a small number of generations, it works well for small databases but may be trapped in the local optimum for large databases.

Water cycle algorithm (WCA) is a novel meta-heuristic algorithm developed by Eskandar et al. in 2012 [12], which is inspired from nature and based on the observation of water cycle process and the ways how rivers and streams flow to the sea in the real world. The basic WCA and its modified versions have been applied to constrained optimization and engineering design problems, the algorithms have good performance in convergence rate and quality of optimized designs [26–28]. However, to our knowledge, the WCA algorithm has not been applied to learning BN structures.

In this paper, we propose a novel binary encoding water cycle algorithm for structures learning of Bayesian networks (BEWCA-BN). Since the solution space of the problem is binary-structured, **xor**, **and** and **or** operators are used to generate solutions in the new optimization technology. Unlike classical water cycle algorithm where the individuals are updated only based on the best (sea) and better (rivers) solutions, each individual in the proposed method also learns from randomly selected individuals in the current population. In addition, we define a method to calculate the distance of two individuals, and the evaporation process is performed by mutation operator. The new strategies are introduced in the proposed algorithm to enhance the search performance and improve the quality of the solutions. The theorem of Markov chain is used to prove the convergence of the proposed algorithm. BEWCA-BN algorithm is then implemented on several well-known benchmark networks and compared with other state of the art algorithms.

The rest of this paper is organized as follows: we begin in Section 2 with the concepts of BNs and the scoring metric related to BNs. In Section 3 we discuss the WCA algorithm. In Section 4 we develop the novel binary encoding algorithm for structures learning of BNs. In Section 5 we present the experimental evaluation of the proposed algorithm. Finally, Section 6 contains the conclusions and possible future directions.

2. Bayesian networks and scoring metric

2.1. Bayesian networks

Let $\mathbb{G} = (X, E)$ be a directed acyclic graph (DAG), where $X = (x_1, x_2, \cdots, x_n)$ is the set of nodes representing the system random variables, $E = \{e_{ij}\}$ is the set of edges representing the direct dependence relationships between the variables. If there is a directed edge from node x_j to node x_i, we say x_j is a parent of x_i. $Pa(x_i)$ is defined as the set containing the parents of x_i in the graph. Let P be a joint probability distribution of random variables in set V. If $(\mathbb{G}, P)$ satisfies the Markov condition, then $(\mathbb{G}, P)$ is called a Bayesian network (BN). Together with the graph structure, the joint probability distribution of the domain can be decomposed into a product of local conditional probability distributions according to Eq. (1), and each conditional probability distribution involves a node and its parents only.

$$P(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(x_i | Pa(x_i)). \tag{1}$$

The learning task in BNs consists of two subtasks: structure learning and parameter estimation. The first subtask aims at identifying the best topology for a network. The second subtask is to learn the parameters that define the conditional probability distribution for a fixed network structure. In this work, we focus on the problem of structure learning, and learn the structure from observed data based on score-based method.

2.2. k2 scoring metric of the Bayesian network

Score-based approach is one of the most popular method of inducing BNs from data. Generally, given a scoring metric, approaches to BN learning concentrate on finding one or more structures that fit the observed data well according to the scoring metric. Assuming a structure $\mathbb{G}$, the score as shown in Eq. (2) is the posterior probability of $\mathbb{G}$ given the data set $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$ with N cases. Since $P(\mathcal{D})$ does not depend on the structure, the joint distribution $P(\mathbb{G}, \mathcal{D})$ can be used as scoring metric.

$$P(\mathbb{G}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathbb{G})P(\mathbb{G})}{P(\mathcal{D})} = \frac{P(\mathbb{G}, \mathcal{D})}{P(\mathcal{D})}. \tag{2}$$

The $k2$ metric is one of the widely used Bayesian scoring methods. Since the metric is first used in $k2$ algorithm [9], so it adopts the name of the algorithm and refers to as the $k2$ metric. The $k2$ metric is used to measure the joint probability of a BN $\mathbb{G}$ and a data $\mathcal{D}$, it can be expressed as Eq. (3):

$$P(\mathbb{G}, \mathcal{D}) = P(\mathbb{G}) \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}!, \tag{3}$$

where r_i is the number of different states of the variable x_i, q_i is the number of possible configurations for $Pa(x_i)$, N_{ijk} is the number of samples in $\mathcal{D}$, in which x_i is in its kth state and $Pa(x_i)$ is in its jth configuration, and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

The score decomposability is an important property for a scoring metric. If the score of the whole network can be written as the sum of scores that depend only on one node and its parents, we say that the score metric is decomposable. The major benefit of this property is that a local change in a BN does not alter the scores of other parts. When assuming a uniform prior for $P(\mathbb{G})$ and using $\log(P(\mathbb{G}, \mathcal{D}))$ instead of $P(\mathbb{G}, \mathcal{D})$, the $k2$ metric can be written as Eq. (4), the score of each node x_i is defined as Eq. (5):

$$f(\mathbb{G}, \mathcal{D}) = \log(P(\mathbb{G}, \mathcal{D})) \approx \sum_{i=1}^{n} f(x_i, Pa(x_i)), \tag{4}$$

$$f(x_i, Pa(x_i)) = \sum_{j=1}^{q_i} \left(\log \left(\frac{(r_i - 1)!}{(N_{ij} + r_i - 1)} \right) + \sum_{k=1}^{r_i} \log(N_{ijk}!) \right). \tag{5}$$

3. Water cycle algorithm

3.1. Representation of initial population

Water cycle algorithm (WCA), a new population-based algorithm, is inspired by the observation of water cycle process and the ways how rivers and streams flow to the sea in the real world. Similar to other meta-heuristic methods, WCA algorithm requires an initial population of individuals so called raindrops. An initial population is randomly generated by first assuming the rain or precipitation phenomena, then the quality of each raindrop is evaluated to determine the sea, the rivers and the streams. The raindrop with the best quality is chosen as a sea, a number of good raindrops are considered as rivers and the rest of raindrops are chosen as streams that flow to the rivers and sea.

In D dimensional optimization problem, each raindrop is represented as an array of $1 \times D$, the total population is represented as