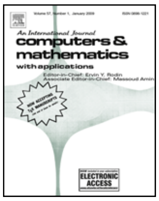
Contents lists available at ScienceDirect

Computers and Mathematics with Applications

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



Bayesian network modeling for evolutionary genetic structures

Lisa Jing Yan [*], Nick Cercone

Department of Computer Science and Engineering, York University, Toronto, ON, Canada M3J 1P3

ARTICLE INFO

Article history:
Received 23 July 2009
Accepted 30 December 2009

Keywords:
Bayesian network modeling
Evolutionary computing
Genetic algorithm
Structure learning
Constraint based
Score based

ABSTRACT

Evolutionary theory states that stronger genetic characteristics reflect the organism's ability to adapt to its environment and to survive the harsh competition faced by every species. Evolution normally takes millions of generations to assess and measure changes in heredity. Determining the connections, which constrain genotypes and lead superior ones to survive is an interesting problem. In order to accelerate this process,we develop an artificial genetic dataset, based on an artificial life (AL) environment genetic expression (ALGAE). ALGAE can provide a useful and unique set of meaningful data, which can not only describe the characteristics of genetic data, but also simplify its complexity for later analysis.

To explore the hidden dependencies among the variables, Bayesian Networks (BNs) are used to analyze genotype data derived from simulated evolutionary processes and provide a graphical model to describe various connections among genes. There are a number of models available for data analysis such as artificial neural networks, decision trees, factor analysis, BNs, and so on. Yet BNs have distinct advantages as analytical methods which can discern hidden relationships among variables. Two main approaches, constraint based and score based, have been used to learn the BN structure. However, both suit either sparse structures or dense structures. Firstly, we introduce a hybrid algorithm, called "the E-algorithm", to complement the benefits and limitations in both approaches for BN structure learning. Testing E-algorithm against a standardized benchmark dataset ALARM, suggests valid and accurate results. BAyesian Network ANAlysis (BANANA) is then developed which incorporates the E-algorithm to analyze the genetic data from ALGAE. The resulting BN topological structure with conditional probabilistic distributions reveals the principles of how survivors adapt during evolution producing an optimal genetic profile for evolutionary fitness.

1. Introduction

Bayesian Network (BN) modeling for evolutionary genetic structure, uses BN to analyze genotype data derived from evolutionary processes and provides a graphical model to describe hidden dependencies among genes. According to evolutionary theory, stronger genetic characteristics reflect the organism's ability to adapt to its environment and to survive the harsh competition faced by every species [1–3]. Each individual's traits and characteristics are coded into cellular information called genes. Genes evolve to be strong, fit genes; that is, nature selects the best genes and reproduces them using inheritance through generations of survivors. Such evolution normally takes millions of generations. But what are the hidden connections which constrain genotypes, yet lead to superior characteristics which promote survival is rather interesting. In order to explore this problem, we accelerate this process significantly, so that we can evaluate the genetic change much more rapidly. We then analyze the hidden evolutionary relationships. Having revealed these connections, we can determine which precise factors and connections promote fitness in an individual population or species.

[*] Corresponding author. Tel.: +1 4169975002.
E-mail addresses: jingyan@cse.yorku.ca (L.J. Yan), ncercone@yorku.ca (N. Cercone).

0898-1221/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2009.12.039

There are a number of models available for data analysis such as artificial neural networks, decision trees, factor analysis, BNs, and so on. Yet BNs have distinct advantages as computational tools. BN is an analytical tool which can discern hidden relationships among variables [4]. BN can handle incomplete datasets just as well as complete ones, and it can discover dependencies among all variables by representing them in a comprehensible graphical model.

BNs have been widely used in bioinformatics (gene regulatory networks, protein structure), medicine, document classification, information retrieval and image processing [5–10,24–26]. As probabilistic models, BNs have been used to replace traditional variation of genetic and evolutionary algorithm in evolutionary computing [11]. In [11], Pelikan segments chromosomes to different traps as variables and build a probabilistic model based on this; after that, only use this model to sample the solutions and generate new candidates population. BN has provided a more promising solution population, however, the real reason why this method can bring out the optimal candidates population more efficiently is the discovery of the hidden relationship among the genes. Thus, our work is undertaken as a response to reveal the discovery of this hidden relationship among the genes by applying BN as an analytical tool for a population solution space, rather than a probabilistic sampling tool.

We therefore propose to apply BNs to analyze data arising in genetic research. We demonstrate our idea on a simulated genetic dataset, which mimics a biology-driven artificial life (AL) environment [12]. This AL simulation, Artificial Life Genetic Algorithm Expression (ALGAE), provides a useful and unique set of meaningful data, which can not only describe the characteristics of genetic data, but also simplify its complexity for our BN analysis. BAyesian Network ANAlysis (BANANA) is then developed to analyze the genetic data from ALGAE. BANANA incorporates a BN structure learning algorithm: the E-algorithm, first proposed by Yan et al. [13] and has been proven to be an efficient and accurate algorithm for constructing BN structure by later adaptations, applied to a business model [10,14].

The goal of our research is to reveal the hidden connections among genetic characteristics. Each chromosome in the AL species contains a coded gene sequence representing particular species characteristics. These characteristics appear random, but after generations of evolution, certain genetic attributes will emerge as dominant. However, this hidden information is not apparent from the raw data, and the meaning needs to be extracted and interpreted. BN analysis of the genetic data can produce a graphical and statistical representation showing the dependencies between genotypes among populations.

The significance of the analysis of the hidden dependencies between genetic descriptors is that two important outcomes are produced as a result of research. Firstly, we generate an interesting and unique genetic dataset using the AL model, which extends the versatility and utility of the Genetic Algorithm (GA) so that it becomes a remarkable instrument for creating hypotheses for any given entities. Secondly, using BN to analyze the hidden dependencies among AL genetic data is a unique methodology. It provides a new approach for problem solving by combining evolutionary principles and BN modeling, based upon generating unique and expressive data.

This paper is organized as follows: Section 2 provides background regarding Bayesian network learning and the E-algorithm; Section 3 introduces the design of ALGAE, and experiments to obtain artificial genetic data; Section 4 explains the process called BANANA, and the modeling for AL genetic data structure, and discusses the experimental results of genotype characteristic hidden connections; Section 5 summarizes our contribution and provides some open questions for further research.

2. Bayesian network learning

Bayesian networks are a graphical representation of probabilistic causal relationships among random variables (factors). A BN has two components: a topological structure and its conditional probability distribution (CPD). The BN structure is an acyclic directed graph in which each vertex i corresponds to a random variable X_i. An arc $X_i \rightarrow X_j$ describes the dependency between variable i and j. This dependency also states the causal relationship between them, thus, variable i is the parent node of j, and variable j is the descendant node of i. In this graph, each vertex i is attached with its conditional probabilistic distribution $p(X_i|\Pi_i)$ of X_i given its parents Π_i. We assume that each variable is probabilistic independent of its non-descendants given its parent states. Thus, the joint probability distribution $\mathcal{P}(X)$ for all the variables X [15], can be described as follows in Eq. (1):

$$\mathcal{P}(X) = \prod_{i=1}^{n} p(X_i|\Pi_i). \tag{1}$$

The advantage of a BN is that it can describe data in both qualitative and quantitative aspects. Qualitatively, a BN structure gives data a graphical interpretation which can be understood easily; and quantitatively, CPD describes strength of the causal relationships among the factors. Thus, learning Bayesian networks can be examined as the combination of parameter learning and structure learning. Parameter learning is to estimate the conditional probabilities (dependencies) in the network, whereas, structural learning is to estimate the topology (arcs) of the network. This following section discusses how to learn Bayesian network structures from data.

2.1. Basic approaches for BN structure learning

Given a set of variables and a dataset composed of all these variables' values, the problem is to build a structure to present the connections among the variables. This structure learning process needs to select the arcs between them and estimate