A Memetic Approach to Bayesian Network Structure Learning

Alberto Tonda[1], Evelyne Lutton[2],
Giovanni Squillero[3], and Pierre-Henri Wuillemin[4]

[1] INRA UMR 782 GMPA, 1 Av. Brétignières, 78850, Thiverval-Grignon, France
alberto.tonda@grignon.inra.fr
[2] INRIA Saclay-Ile-de-France, AVIZ team, Bâtiment 650, 91405, Orsay Cedex, France
evelyne.lutton@grignon.inria.fr
[3] Politecnico di Torino, DAUIN, Corso Duca degli Abruzzi 124, 10129, Torino, Italy
giovanni.squillero@polito.it
[4] LIP6 - Département DÉSIR, 4, place Jussieu, 75005, Paris
pierre-henri.wuillemin@lip6.fr

Abstract. Bayesian networks are graphical statistical models that represent inference between data. For their effectiveness and versatility, they are widely adopted to represent knowledge in different domains. Several research lines address the NP-hard problem of Bayesian network structure learning starting from data: over the years, the machine learning community delivered effective heuristics, while different Evolutionary Algorithms have been devised to tackle this complex problem. This paper presents a Memetic Algorithm for Bayesian network structure learning, that combines the exploratory power of an Evolutionary Algorithm with the speed of local search. Experimental results show that the proposed approach is able to outperform state-of-the-art heuristics on two well-studied benchmarks.

Keywords: Memetic Algorithms, Evolutionary Algorithms, Local Optimization, Bayesian Networks, Model Learning.

1 Introduction

Bayesian networks are probabilistic graphical models that represent a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). They are widely used to encode knowledge and perform predictions in many different fields, ranging from medicine to document classification, from computational biology to law.

It is theoretically possible to learn the optimal structure for a Bayesian network from a dataset. However, the number of possible structures is superexponential in the number of variables of the model [1] and the problem of Bayesian network learning is proved to be NP-hard [2].

The machine learning community delivered fast heuristic algorithms that build the structure of a Bayesian network on the basis of conditional independence

A.I. Esparcia-Alcázar et al. (Eds.): EvoApplications 2013, LNCS 7835, pp. 102–111, 2013.

evaluations between variables [3] [4]. On the other hand, several attempts have been made in evolutionary computation to tackle this complex issue [5] [6] [7]. Interestingly, many evolutionary approaches also feature local search techniques to improve the quality of the results.

This paper presents a memetic approach to Bayesian network structure learning. The proposed technique exploits an evolutionary framework evolving initial conditions for a state-of-the-art heuristic that efficiently explores the search space. The fitness function is based on the Akaike information criterion, a metric taking into account both the accuracy and the complexity of a candidate model.

An additional objective of this work is to link the community facing the complex Bayesian network structure learning problem, to the community of memetic computing. While combinations of heuristics and evolutionary optimization are prominently featured in the literature related to structure learning, to the authors' knowledge the methods presented are almost never ascribed to the field of memetic algorithms. In the authors' opinion, an explicit interaction between the two communities could lead to extremely beneficial results.

2 Background

In order to introduce the scope of the present work, some necessary concepts about Bayesian networks and memetic algorithms are summarized in the following.

2.1 Bayesian Networks

A Bayesian Network (BN) is defined as a *graph-based model of a joint multivariate probability distribution that captures properties of conditional independence between variables* [8]. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. The network could thus be used to compute the probabilities of the presence of various diseases, given the symptoms.

Formally, a Bayesian network is a directed acyclic graph (DAG) whose nodes represent variables, and whose arcs encode conditional dependencies between the variables. This graph is called the *structure* of the network and the nodes containing probabilistic information are called the *parameters* of the network. Figure 1 reports an example of a Bayesian network.

The set of parent nodes of a node X_i is denoted by $pa(X_i)$. In a Bayesian network, the joint probability distribution of the node values can be written as the product of the local probability distribution of each node and its parents:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))$$

2.2 The Structure Learning Problem

Learning the structure of a Bayesian network starting from a dataset is proved to be a NP-hard problem [2]. The algorithmic approaches devised to solve this