Information Sciences 225 (2013) 18–34

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Signal propagation in Bayesian networks and its relationship with intrinsically multivariate predictive variables

David C. Martins Jr. [a,*], Evaldo A. de Oliveira [b], Ulisses M. Braga-Neto [e], Ronaldo F. Hashimoto [c], Roberto M. Cesar Jr. [c,d]

[a] Center for Mathematics, Computation and Cognition, Federal University of ABC, R. Santa Adélia 166, Santo André, SP 09210-170, Brazil
[b] Department of Earth and Exact Sciences, Federal University of São Paulo, R. Arthur Ridel 275, Diadema, SP 09972-270, Brazil
[c] Institute of Mathematics and Statistics, University of São Paulo, R. do Matão 1010, São Paulo, SP 05508-090, Brazil
[d] Brazilian Bioethanol Science and Technology Laboratory, Campinas, SP 13083-970, Brazil
[e] Genomic Signal Processing Lab, Texas A&M University, College Station, TX 77843-3128, USA

ARTICLE INFO

Article history:
Received 1 June 2011
Received in revised form 9 October 2012
Accepted 14 October 2012
Available online 23 November 2012

Keywords:
Bayesian network
Feature selection
Intrinsically multivariate prediction

ABSTRACT

A set of predictor variables is said to be intrinsically multivariate predictive (IMP) for a target variable if all properly contained subsets of the predictor set are poor predictors of the target but the full set predicts the target with great accuracy. In a previous article, the main properties of IMP Boolean variables have been analytically described, including the introduction of the IMP score, a metric based on the coefficient of determination (CoD) as a measure of predictiveness with respect to the target variable. It was shown that the IMP score depends on four main properties: logic of connection, predictive power, covariance between predictors and marginal predictor probabilities (biases). This paper extends that work to a broader context, in an attempt to characterize properties of discrete Bayesian networks that contribute to the presence of variables (network nodes) with high IMP scores. We have found that there is a relationship between the IMP score of a node and its territory size, i.e., its position along a pathway with one source: nodes far from the source display larger IMP scores than those closer to the source, and longer pathways display larger maximum IMP scores. This appears to be a consequence of the fact that nodes with small territory have larger probability of having highly covariate predictors, which leads to smaller IMP scores. In addition, a larger number of XOR and NXOR predictive logic relationships has positive influence over the maximum IMP score found in the pathway. This work presents analytical results based on a simple structure network and an analysis involving random networks constructed by computational simulations. Finally, results from a real Bayesian network application are provided.

1. Introduction

Bayesian networks [2,8] have been used as a useful approach to model systems composed of components that communicate by local interaction, i.e. each component directly depending on a small number of elements. Biological systems, for instance, present such property [4]. Bayesian networks are mathematically defined in terms of probabilities and conditional independence properties and can be employed to infer direct "causal" influence (connections between variables) [13,18,5,8,14].

* Corresponding author.
E-mail address: david.martins@ufabc.edu.br (D.C. Martins Jr.)

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.10.027

The concept of intrinsically multivariate predictive (IMP) variables was introduced in [9], in which a target variable strongly depends on a set of other variables, but such dependence is weak or absent when one considers properly contained subsets of variables. The IMP score was introduced as a metric based on the coefficient of determination (CoD) [12,3] as a measure of predictiveness with respect to the target variable. It was shown in [9] that the IMP score of a target variable is affected by four properties: logic of prediction, predictive power, covariance between the predictors, and the marginal probabilities of each individual predictor. It was demonstrated that IMP variables (i.e., variables with large IMP score) tend to occur for large predictive power, small correlation between predictors, and certain specific predictor logics—2-minterm logics (*XOR* and *NXOR*) lead to larger IMP scores than 1-and 3-minterm logics ($AND, OR, NOR, NAND, x_1 \wedge \bar{x}_2$ and $x_1 \vee \bar{x}_2$). Based on these results, we hypothesized that large proportions of nodes with *XOR* logic of prediction in the networks could improve the chance for the appearance of nodes with large IMP score. We show in this paper that this is indeed the case; the larger the number of *XOR*/*NXOR* logics in the network is, the larger the maximum IMP score in the network is.

The study of the IMP phenomenon can be useful in feature selection for pattern recognition, since it is one of the main reasons for the occurrence of the *nesting effect*. Basically, the nesting effect is a feature selection issue that occurs when some features included in the partial subset solution by some algorithm are not present in the optimal solution and never discarded, leading to a suboptimal solution [15]. Another application of IMP is that it seems to be associated with variables that possess *canalizing functions* [7,6,10], an important concept in Systems Biology—canalizing genes exhibit key roles on gene regulatory networks [16]. Martins et al. showed that DUSP1 gene, which is canalizing gene exhibiting control over a central, process-integrating signaling pathway, displays the largest number of IMP predictors in melanoma expression data [9]. Besides, Bayesian networks are often applied to financial risk analysis in order to model conditional multivariate dependence among variables [19,11].

In this paper, we analyze the intrinsically multivariate prediction phenomenon in networks with three or more nodes, an extension of the study presented in [9] which considers only one target and its set of predictors (two or three). In particular, we analyze how the territory size of a target node (a graph-theoretical property defined in Section 4) impacts the probability of occurrence of IMP nodes in Bayesian networks with Boolean variables. We show that a target with large territory can achieve larger IMP scores with its predictors than a target with small territory. This finding is in agreement with the hypothesis, advanced in [9], that subsets with high IMP score are more susceptible to be responsible for regulation of several metabolic pathways or subsystems as observed in microarray data analysis of melanoma experiments. We also show that the absolute value of the covariance between predictors is negatively correlated with the territory size. It is worth mentioning that, although these results are given in the context of logical functions, they can be easily extended to other types of functions. In summary, this paper contributes to theoretical advances in the analysis of the intrinsically multivariate prediction phenomenon in the context of Bayesian networks.

This work is organized as follows. Section 2 reviews fundamental concepts. Section 3 describes the network model used to analyze the IMP score behavior as a function of the territory size of a given target. Section 4 presents analytical results based on a simple structure network. In order to generalize the analytical results, Section 5 presents an analysis of the IMP score in random networks constructed by computational simulations, as well as a real example from the Bayesian networks Repository (http://www.cs.huji.ac.il/site/labs/compbio/Repository). Finally, conclusions are given in Section 6.

2. Background

2.1. Bayesian networks

Here we review fundamental concepts of Bayesian networks to aid the comprehension of the paper; for more details, see [4]. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\} \in D^n$ be a set of random variables, each one defined in a given domain D. A Bayesian network is a representation of the joint probability distribution of $\mathbf{X}$ consisting of a directed acyclic graph (DAG) G whose vertices are the random variables $x_1, \ldots, x_n$ and a component that describes a conditional probability distribution for each variable, given its parents in the graph. These two components describe a unique joint probability distribution on $X_1, \ldots, X_n$.

Such graph allows a decomposition of the joint distribution that leads to a reduced number of parameters. Each variable is independent of its non-descendants, given its parents. The joint distribution can be decomposed into the product form:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \mathbf{Pa}^G(X_i)), \tag{1}$$

where $\mathbf{Pa}^G(X_i)$ is the set of parents of X_i in G.

Considering $X_1, \ldots, X_n$ as Boolean variables (i.e. $X_i \in \{0,1\}$), the conditional probability $P(X_i|\mathbf{Pa}^G(X_i))$ can be represented by a table that specifies the probability of values for X_i considering all possible observations (or configurations) of the values of $\mathbf{Pa}^G(X_i)$. So, if the set $\mathbf{Pa}^G(X_i)$ has size k, the number of rows in the table is given by 2^k, where each row corresponds to a distribution of X_i given a specific configuration of $\mathbf{Pa}^G(X_i)$.

2.2. Intrinsically multivariate prediction

Here we recall from Ref. [9] fundamentals behind the intrinsically multivariate prediction concept. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\} \in \{0,1\}^n$ be a set of predictor random variables and $Y = y \in \{0,1\}$ be the target random variable. The coefficient of determination (CoD) of $\mathbf{X}$ with respect to Y [3] is given by: