Artificial Intelligence in Medicine 57 (2013) 219–229

Contents lists available at SciVerse ScienceDirect

Artificial Intelligence in Medicine

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

Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers

Hanen Borchani [a,*], Concha Bielza [a], Carlos Toro [b], Pedro Larrañaga [a]

[a] Computational Intelligence Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Boadilla del Monte 28660, Spain
[b] Department of Microbiology, Hospital Carlos III, Madrid 28029, Spain

ARTICLE INFO

Article history:
Received 26 October 2011
Received in revised form
14 December 2012
Accepted 16 December 2012

Keywords:
Multi-dimensional Bayesian network classifiers
Markov blanket
Human immunodeficiency virus
Protease inhibitors
Reverse transcriptase inhibitors

ABSTRACT

Objective: Our aim is to use multi-dimensional Bayesian network classifiers in order to predict the human immunodeficiency virus type 1 (HIV-1) reverse transcriptase and protease inhibitors given an input set of respective resistance mutations that an HIV patient carries.
Materials and methods: Multi-dimensional Bayesian network classifiers (MBCs) are probabilistic graphical models especially designed to solve multi-dimensional classification problems, where each input instance in the data set has to be assigned simultaneously to multiple output class variables that are not necessarily binary. In this paper, we introduce a new method, named MB-MBC, for learning MBCs from data by determining the Markov blanket around each class variable using the HITON algorithm. Our method is applied to both reverse transcriptase and protease data sets obtained from the Stanford HIV-1 database.
Results: Regarding the prediction of antiretroviral combination therapies, the experimental study shows promising results in terms of classification accuracy compared with state-of-the-art MBC learning algorithms. For reverse transcriptase inhibitors, we get 71% and 11% in mean and global accuracy, respectively; while for protease inhibitors, we get more than 84% and 31% in mean and global accuracy, respectively. In addition, the analysis of MBC graphical structures lets us gain insight into both known and novel interactions between reverse transcriptase and protease inhibitors and their respective resistance mutations.
Conclusion: MB-MBC algorithm is a valuable tool to analyze the HIV-1 reverse transcriptase and protease inhibitors prediction problem and to discover interactions within and between these two classes of inhibitors.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

The multi-dimensional classification problem is an extension of the classical one-dimensional classification problem, where we have to deal with multiple output class variables rather than a single output class variable [1]. Formally, the multi-dimensional classification problem consists of finding a function f that predicts for each input instance, given by a vector of m features $\mathbf{x}=(x_1, \ldots, x_m)$, a vector of d class values $\mathbf{c}=(c_1, \ldots, c_d)$:

$$f : \Omega_{X_1} \times \cdots \times \Omega_{X_m} \longrightarrow \Omega_{C_1} \times \cdots \times \Omega_{C_d}$$
$$\mathbf{x} = (x_1, \ldots, x_m) \longmapsto \mathbf{c} = (c_1, \ldots, c_d)$$

where Ω_{C_i} and Ω_{X_j} denote the sample spaces of each class variable C_i, for all $i \in \{1, \ldots, d\}$, and each feature variable X_j, for all $j \in \{1, \ldots, m\}$, respectively. Note that, we consider that all class and feature variables are discrete random variables such that $|\Omega_{C_i}|$ and $|\Omega_{X_j}|$ are greater than 1.

When $|\Omega_{C_i}| = 2$ for all $i \in \{1, \ldots, d\}$, i.e., all class variables are binary, the multi-dimensional classification problem is known as a multi-label classification problem [2,3]. In general, a multi-label classification problem can be easily modeled as a multi-dimensional classification problem where each label corresponds to a binary class variable. However, modeling a multi-dimensional classification problem, that possibly includes non-binary class variables, as a multi-label classification problem may require a transformation over the data set to meet multi-label framework requirements.

In recent years, the concept of multi-dimensionality has been introduced in Bayesian network classifiers providing an accurate modeling of this emerging problem and ensuring interactions among all variables [1,4–8]. In these probabilistic graphical models, known as multi-dimensional Bayesian network classifiers (MBCs),

* Corresponding author. Tel.: +34 91 3363675; fax: +34 91 3524819.
E-mail addresses: hanen.borchani@upm.es (H. Borchani), mcbielza@fi.upm.es (C. Bielza), carlos.toro@salud.madrid.org (C. Toro), pedro.larranaga@fi.upm.es (P. Larrañaga).

0933-3657/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.artmed.2012.12.005

the graphical structure partitions the set of class and feature variables into three different subgraphs: class subgraph, feature subgraph and bridge subgraph, and the parameter set defines the conditional probability distribution of each variable given its parents.

In this paper, we introduce a novel MBC learning algorithm based on Markov blankets. Motivated by the fact that the classification is unaffected by parts of the structure that lie outside the Markov blankets of the class variables, we first build the Markov blanket around each class variable using the well-known HITON algorithm [9–11], and then we determine edge directionality over all three MBC subgraphs. Thanks to this filter and a local approach to MBC learning, we can lighten the computational burden of MBC learning using wrapper algorithms [1,4,5] and provide more accurate MBC structures.

We finally apply our Markov blanket MBC (`MB-MBC`) algorithm to the problem of predicting human immunodeficiency virus type 1 (HIV-1) reverse transcriptase and protease inhibitors given an input set of corresponding resistance mutations that an HIV patient carries. In general, a combination of several antiretroviral drugs should be repeatedly administered for each patient in order to prevent and treat the HIV infection.

We analyze both reverse transcriptase and protease data sets obtained from the Stanford HIV-1 database [12]. In the reverse transcriptase data set (respectively, protease data set), the class variables are ten reverse transcriptase inhibitors (respectively, eight protease inhibitors) and the feature variables are 38 predefined mutations [13] associated with resistance to reverse transcriptase inhibitors (respectively, 74 predefined mutations [13] associated with resistance to protease inhibitors).

In both data sets, all class and feature variables are binary, so that the problem of predicting HIV-1 reverse transcriptase and protease inhibitors can be also viewed as a multi-label classification problem. However, since our approach is general and can be applied to additional classification problems where class variables are not necessarily binary, we opt to use the term multi-dimensional classification as a more general concept. Moreover, contrary to multi-label classification methods, our approach presents the merit of explicitly modeling the relationships between all variables through their graphical structure component which, in our study, may be useful in further investigating the interactions among the different inhibitors and resistance mutations.

Experimental results on reverse transcriptase and protease inhibitors data sets were promising in terms of classification accuracy compared with state-of-the-art MBC and multi-label classification methods, as well as regarding the identification of interactions among inhibitors and resistance mutations, which were either consistent with the latest knowledge or not previously mentioned in the literature.

The remainder of this paper is organized as follows. Section 2 introduces Bayesian networks. Section 3 presents MBCs and briefly reviews state-of-the-art MBC learning algorithms. Section 4 describes our new MBC learning approach. Section 5 presents the experimental study on the HIV-1 reverse transcriptase and protease inhibitor data sets. Finally, Section 6 sums up the paper with some conclusions.

2. Background

A Bayesian network [14,15] over a set of discrete random variables $\mathbf{U} = \{X_1, \ldots, X_n\}$, $n \geq 1$, is a pair $\mathcal{B} = (\mathcal{G}, \boldsymbol{\Theta})$. $\mathcal{G} = (V, A)$ is a directed acyclic graph (DAG) whose vertices V correspond to variables in $\mathbf{U}$ and whose arcs A represent direct dependencies between the vertices. $\boldsymbol{\Theta}$ is a set of conditional probability distributions such that $\theta_{x_i|\mathbf{pa}(x_i)} = p(x_i|\mathbf{pa}(x_i))$ defines the conditional probability of each possible value x_i of X_i given a set value $\mathbf{pa}(x_i)$ of $\mathbf{Pa}(X_i)$, where $\mathbf{Pa}(X_i)$ denotes the set of parents of X_i in $\mathcal{G}$.

A Bayesian network $\mathcal{B}$ represents a joint probability distribution over $\mathbf{U}$ factorized according to structure $\mathcal{G}$ as follows:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i|\mathbf{Pa}(X_i)) \cdot \quad (1)$$

Definition 1 *(Conditional independence [14]).* Two set of variables $\mathbf{X}$ and $\mathbf{Y}$ are conditionally independent given some set of variables $\mathbf{Z}$, denoted as $I(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$, iff $P(\mathbf{X}|\mathbf{Y}, \mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})$ for any assignment of values $\mathbf{x}, \mathbf{y}, \mathbf{z}$ of $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, respectively, such that $P(\mathbf{Z} = \mathbf{z}) > 0$.

Definition 2 *(Markov blanket [14]).* A Markov blanket of a variable X, denoted as $MB(X)$, is a minimal set of variables with the following property: $I(X, \mathbf{S}|\, MB(X))$ holds for every variable subset $\mathbf{S}$ with no variables in $MB(X) \cup X$.

In other words, $MB(X)$ is a minimal set of variables conditioned by which X is conditionally independent of all the remaining variables. Under the faithfulness assumption, ensuring that all the conditional independencies in the data distribution are strictly those entailed by $\mathcal{G}$, $MB(X)$ consists of the union of the set of parents, children, and parents of children (i.e., spouses) of X [16].

3. Multi-dimensional Bayesian network classifiers

In this section we present MBCs, then briefly review the state-of-the-art methods for learning these models from data.

Definition 3 *(Multi-dimensional Bayesian network classifier [1]).* An MBC is a Bayesian network $\mathcal{B} = (\mathcal{G}, \boldsymbol{\Theta})$ where the structure $\mathcal{G} = (V, A)$ has a restricted topology. The set of n vertices V is partitioned into two sets: $V_C = \{C_1, \ldots, C_d\}$, $d \geq 1$, of class variables and $V_X = \{X_1, \ldots, X_m\}$, $m \geq 1$, of feature variables ($d + m = n$). The set of arcs A is partitioned into three sets A_C, A_X and A_{CX}, such that:

- $A_C \subseteq V_C \times V_C$ is composed of the arcs between the class variables having a subgraph $\mathcal{G}_C = (V_C, A_C)$ – *class subgraph* – of $\mathcal{G}$ induced by V_C.
- $A_X \subseteq V_X \times V_X$ is composed of the arcs between the feature variables having a subgraph $\mathcal{G}_X = (V_X, A_X)$ – *feature subgraph* – of $\mathcal{G}$ induced by V_X.
- $A_{CX} \subseteq V_C \times V_X$ is composed of the arcs from the class variables to the feature variables having a subgraph $\mathcal{G}_{CX} = (V, A_{CX})$ – *bridge subgraph* – of $\mathcal{G}$ induced by V [4].

Depending on the graphical structures of the class and feature subgraphs MBCs can be divided into several families. These families can be denoted as <`class subgraph structure`>-<`feature subgraph structure`> MBCs, where the possible structures of each subgraph are: empty, tree, polytree, or DAG [4]. In this paper, we do not consider any constraints on the subgraph structures of the learned MBCs, i.e., any possible structure type is allowed for either class or feature subgraphs.

Classification with an MBC under a 0–1 loss function is equivalent to solving the most probable explanation (MPE) problem, which consists of finding the most likely instantiation of the vector of class variables $\mathbf{c}^* = (c_1^*, \ldots, c_d^*)$ given an evidence about the input vector of feature variables $\mathbf{x} = (x_1, \ldots, x_m)$. More formally, for a given observed evidence $\mathbf{x}$, we have to determine

$$\mathbf{c}^* = (c_1^*, \ldots, c_d^*) = arg \max_{c_1, \ldots, c_d} \; p(C_1 = c_1, \ldots, C_d = c_d|\mathbf{x}) \cdot \quad (2)$$

Download English Version:

https://daneshyari.com/en/article/10320520

Download Persian Version:

https://daneshyari.com/article/10320520

Daneshyari.com