Variable selection for dynamic treatment regimes: a reinforcement learning approach

Raphael Fonteneau RAPHAEL.FONTENEAU@ULG.AC.BE
Louis Wehenkel L.WEHENKEL@ULG.AC.BE
Damien Ernst[†] DERNST@ULG.AC.BE
Department of Electrical Engineering and Computer Science and GIGA-Research, University of Liège, Grande Traverse 10, 4000 Liège, Belgium. [†] Research Associate FNRS.

1. Introduction

Nowadays, many diseases as for example HIV/AIDS, cancer, inflammatory or neurological diseases are seen by the medical community as being chronic-like diseases, resulting in medical treatments that can last over very long periods. For treating such diseases, physicians often adopt explicit, operationalized series of decision rules specifying how drug types and treatment levels should be administered over time, which are referred to in the medical community as Dynamic Treatment Regimes (DTRs). Designing an appropriate DTR for a given disease is a challenging issue. Among the difficulties encountered, we can mention the complex dynamics of the human body interacting with treatments and other environmental factors, as well as the often poor compliance to treatments due to the side effects of some of the administered drugs. While typically DTRs are based on clinical judgment and medical insight, since a few years the biostatistics community is investigating a new research field addressing specifically the problem of inferring in a well principled way DTRs directly from clinical data gathered from patients under treatment. Among the results already published in this area, we mention (Murphy, 2005) which uses statistical tools for designing DTRs for psychotic patients.

2. Problem formulation

One possible approach to infer DTR from the data collected through clinical trials is to formalize this problem as an optimal control problem for which most of the information available on the 'system dynamics' (the system is here the patient and the input of the system is the treatment) is 'hidden' in the clinical data. This problem has been vastly studied in Reinforcement Learning (RL), a subfield of machine learning (see e.g., (Ernst et al., 2005)). Its application to the DTR problem would consist of processing the clinical data so as to compute a closed-loop treatment strategy which takes as inputs all the various clinical indicators which have been collected from the patients. Using policies computed in this way may however be inconvenient for the physicians who may prefer DTRs based on an as small as possible subset of *relevant* indicators rather than on the possibly very large set of variables monitored through the clinical trial. In this research, we therefore address the problem of determining a small subset of indicators among a larger set of candidate ones, in order to infer by RL convenient decision strategies. Our approach is closely inspired by work on 'variable selection' for supervised learning.

3. Learning from a sample

We assume that the information available for designing DTRs is a sample of discrete-time trajectories of treated patients, i.e. successive tuples (x_t, u_t, x_{t+1}), where x_t represents the state of a patient at some time-step t and lies in an n-dimensional space X of clinical indicators, u_t is an element of the action space (representing treatments taken by the patient in the time interval $[t, t+1]$), and x_{t+1} is the state at the subsequent time-step.

We further suppose that the responses of patients suffering from a specific type of chronic disease all obey the same discrete-time dynamics:

$$x_{t+1} = f(x_t, u_t, w_t) \quad t = 0, 1, \ldots$$

where disturbances w_t are generated by the probability distribution $P(w|x, u)$. Finally, we assume that one can associate to the state of the patient at time t and to the action at time t, a reward signal $r_t = r(x_t, u_t) \in \mathbb{R}$ which represents the 'well being' of the patient over the time interval $[t, t+1]$. Once the choice of the function $r_t = r(x_t, u_t)$ has been realized (a problem known as preference elicitation), the problem of finding a 'good' DTR may be stated as an optimal control problem for which one seeks to find a policy which leads to a sequence of actions $u_0, u_1, \ldots, u_{T-1}$, which maximizes,

over the time horizon $T \in \mathbb{N}$, and for any initial state the criterion:

$$R_T^{(u_0,u_1,\ldots,u_{T-1})}(x_0) = \mathop{\mathbb{E}}_{\substack{w_t \\ t=0,1,\ldots,T-1}} \left[\sum_{t=0}^{T-1} r(x_t, u_t) \right]$$

One can show (see e.g., (Ernst et al., 2005)) that there exists a policy $\pi_T^* : X \times [0, \ldots, T-1] \rightarrow U$ which produces such a sequence of actions for any initial state x_0. To characterize these optimal T-stage policies, let us define iteratively the sequence of *state-action value functions* $Q_N : X \times U \rightarrow \mathbb{R}$, $N = 1, \ldots, T$ as follows:

$$Q_N(x,u) = \mathop{\mathbb{E}}_{w} \left[r(x,u) + \sup_{u' \in U} Q_{N-1}(f(x,u,w),u') \right] \quad (1)$$

with $Q_0(x,u) = 0$ for all $(x,u) \in X \times U$. Dynamic programming theory implies that, for all $t \in \{1, \ldots, T-1\}$ and $x \in X$, the policy

$$\pi_T^\star(t,x) = \arg\max_{u \in U} Q_{T-t}(x,u)$$

is a T-step optimal policy.

Exploiting directly (1) for computing the Q_N-functions is not possible in our context since f is unknown and replaced here by an ensemble of one-step trajectories $\mathcal{F} = \left\{(x_t^l, u_t^l, r_t^l, x_{t+1}^l)\right\}_{l=1}^{\#\mathcal{F}}$, where $r_t^l = r(x_t^l, u_t^l)$. To address this problem, we exploit the fitted Q iteration algorithm which offers a way for computing (approximations of) the Q_N-functions ($\hat{Q}_N$) from the sole knowledge of $\mathcal{F}$ (Ernst et al., 2005). Notice that when used with tree based approximators, as it is the case in this paper, this algorithm offers good inference performances. Furthermore, we exploit the particular structure of these tree-based approximators in order to identify the most relevant clinical indicators among the n candidate ones.

4. Selection of clinical indicators

As mentioned in Section 2, we propose to find a small subset of state variables (clinical indicators), the m ($m \ll n$) most relevant ones with respect to a certain criterion, so as to create an m-dimensional subspace of X on which DTRs will be computed. The approach we propose for this exploits the tree structure of the $\hat{Q}_N$-functions computed by the fitted Q iteration algorithm. More specifically, it evaluates the relevance of each state variable x^i, by the score function:

$$S(x^i) = \frac{\sum_{N=1}^{T} \sum_{\tau \in \hat{Q}_N} \sum_{\nu \in \tau} \delta(\nu, x^i) \Delta_{var}(\nu) |\nu|}{\sum_{N=1}^{T} \sum_{\tau \in \hat{Q}_N} \sum_{\nu \in \tau} \Delta_{var}(\nu) |\nu|}$$

where ν is a nonterminal node in a tree τ (used to build the ensemble model representing one of the $\hat{Q}_N$-functions), $\delta(\nu, x^i) = 1$ if x^i is used to split at node ν and 0 otherwise, $\Delta_{var}(\nu)$ is the variance reduction when splitting node ν, and $|\nu|$ is the cardinality of the subset of tuples residing at node ν.

The approach then sorts the state variables x^i by decreasing values of their score so as to identify the m most relevant ones. A DTR defined on this subset of attributes is then computed by running the fitted Q iteration algorithm again on a 'modified $\mathcal{F}$', where the state variables of x_t^l and x_{t+1}^l that are not among these m most relevant ones are discarded.

The algorithm for computing a DTR defined on a small subset of state variables is thus as follows:
(1) compute the $\hat{Q}_N$-functions ($N = 1, \ldots, T$) using the fitted Q iteration algorithm on $\mathcal{F}$,
(2) compute the score function for each state variable, and determine the m best ones,
(3) run the fitted Q iteration algorithm on $\widetilde{\mathcal{F}} = \left\{(\widetilde{x}_t^l, u_t^l, r_t^l, \widetilde{x}_{t+1}^l)\right\}_{l=1}^{\#\widetilde{\mathcal{F}}}$ where $\widetilde{x}_t = \widetilde{M} x_t$, and $\widetilde{M}$ is a $m \times n$ boolean matrix where $\widetilde{m}_{i,j} = 1$ if the state variable x^j is the i-th most relevant one and 0 otherwise.

5. Preliminary validation

The method has been tested on the 'car on the hill' problem, a classical benchmark in RL (Ernst et al., 2005). This problem, which has a (continuous) state space of dimension two (the position p and the speed s of the car), is originally a deterministic problem. We have added to these variables some non-informative components so as to set up an experimental protocol. In our trials, the algorithm described previously was able to identify s and p as the most informative variables, which is encouraging for our future work with real-life clinical data.

Acknowledgments

This paper presents research results of the Belgian Network BIOMAGNET (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

References

Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, *6*, 503–556.

Murphy, S. (2005). An experimental design for the development of adaptative treatment strategies. *Statistics in Medicine*, *24*, 1455–1481.