Variable selection for dynamic treatment regimes: a reinforcement learning approach

Raphael Fonteneau, Louis Wehenkel and Damien Ernst

Department of Electrical Engineering and Computer Science and GIGA-Research,
University of Liège, Grande Traverse 10, 4000 Liège, Belgium.
{raphael.fonteneau, L.Wehenkel, dernst}@ulg.ac.be

Abstract. Dynamic treatment regimes (DTRs) can be inferred from data collected through some randomized clinical trials by using reinforcement learning algorithms. During these clinical trials, a large set of clinical indicators are usually monitored. However, it is often more convenient for clinicians to have DTRs which are only defined on a small set of indicators rather than on the original full set. To address this problem, we analyse the approximation architecture of the state-action value functions computed by the fitted Q iteration algorithm - a RL algorithm - using tree-based regressors in order to identify a small subset of relevant ones. The RL algorithm is then rerun by considering only as state variables these most relevant indicators to have DTRs defined on a small set of indicators. The approach is validated on benchmark problems inspired from the classical 'car on the hill' problem and the results obtained are positive.

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 the 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 [1] which uses statistical tools for designing DTRs for psychotic patients.

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., [2]). 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.

The rest of this paper is organized as follows. In Section II we formalize the problem of inferring DTRs from clinical data as an optimal control problem for which the sole information available on the system dynamics is the one contained in the clinical data. We also briefly present the fitted Q iteration algorithm which will be used to compute from these data a good approximate of the optimal policy. In Section III, we present our algorithm for selecting the most relevant clinical indicators and computing (near-) optimal policies defined only on these indicators. Section IV reports our simulation results and, finally, Section V concludes.

2 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:

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

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 often known as preference elicitation, see e.g., [3]), 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:

$$R_T^{(u_0,u_1,\ldots,u_{T-1})}(x_0) = \mathop{\mathbb{E}}_{\substack{w_t \\ t=0,1,\ldots,T-1}} \left[\sum_{t=0}^{T-1} r(x_t, u_t) \right] \tag{2}$$

One can show (see e.g., [2]) 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:

$$Q_N(x,u) = \mathop{\mathbb{E}}_{w} \left[r(x,u) + \sup_{u' \in U} Q_{N-1}(f(x,u,w), u') \right] \tag{3}$$

with $Q_0(x,u) = 0$ for all $(x,u) \in X \times U$. By using results from the dynamic programming theory, one can write 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 (3) for computing the Q_N-functions is not possible in our context since f is unknown and replaced here by a sample 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 the Q_N-functions from the sole knowledge of $\mathcal{F}$ [2]. In a few words, this RL algorithm computes these functions by solving a T-length sequence of standard supervised learning problems. A $\hat{Q}_N$-function - approximation of the Q_N-function as defined by Eqn (3) - is computed by solving the Nth supervised learning problem of the sequence. The training set for this problem is computed from $\mathcal{F}$ and the $\hat{Q}_{N-1}$-function. Notice that when used with tree-based approximators and especially Extremely Randomized Trees [4], as it is the case in this paper, this algorithm offers good generalization 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.

3 Selection of clinical indicators

As mentioned in Section 1, 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. This approach will score each attribute by estimating the variance reduction it can be associated with by propagating the training sample over the different tree structures

(this criterion was originally proposed in the context of supervised learning for identifying relevant attributes in the context of regression tree induction [5]). In our context, 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 (one of those 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 no or equal to zero otherwise, $|\nu|$ is the number of samples at node ν, $\Delta_{var}(\nu)$ is the variance reduction when splitting node ν:

$$\Delta_{var}(\nu) = v(\nu) - \frac{|\nu_L|}{|\nu|} v(\nu_L) - \frac{|\nu_R|}{|\nu|} v(\nu_R)$$

where ν_L (resp. ν_R) is the left-son node (resp. the right-son node) of node ν, and $v(\nu)$ (resp. $v(\nu_L)$ and $v(\nu_R)$) is the variance of the sample at node ν (resp. ν_L and ν_R).

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 variables 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.

4 Preliminary validation

We report in this section simulation results that have been obtained by testing the proposed approach on a modified version of the classical 'car on the hill' benchmark problem [2].[1] The original 'car on the hill' problem has two state

[1] The optimality criterion of the car on the hill problem is usually chosen as being the sum of the discounted rewards observed over an infinite time horizon. We have chosen here to shorten this infinite time horizon to 50 steps and not use discount factors in order to have an optimality criterion in accordance with (2).

variables, the position p and the speed s of the car, and one action variable u which represents the acceleration of the car. The action can only take two discrete values (full acceleration or full deceleration).

For illustrating our approach, we have slightly modified the car on the hill problem by adding new "dummy state variables" to the problem. These variables take at each time t a value which is drawn independently from all other variable-values according to a uniform probability distribution over the interval $[0, 1]$ and do not affect the actual dynamics of the problem.

In such a context, our approach is expected to associate the highest scores $S(\cdot)$ to the variables s and p since these are the only ones that actually contain relevant information about the optimal policy of the system. Results obtained are presented in Table 1. As one can see, the approach consistently gives the two highest scores to p and s.

Table 1. Variance reduction scores of the different state variables for various experimental settings. The first column gives the cardinality of the sets $\mathcal{F}$ considered (the elements of these sets have been generated by drawing (x_t^l, u_t^l) at random in $X \times U$ and computing x_{t+1}^l from the system dynamics (1)). The second column gives the number of Non-Relevant Variables (NRV) added to the original state vector. The remaining columns report the different scores $S(\cdot)$ computed for the different (relevant and non-relevant) variables considered in each scenario.

$\#\mathcal{F}$	NB. OF NRV	p	s	NRV 1	NRV 2	NRV 3
5000	0	0.24	0.35	-	-	-
5000	1	0.27	0.30	0.08	-	-
5000	2	0.16	0.26	0.12	0.06	-
5000	3	0.15	0.18	0.07	0.07	0.09
10000	1	0.16	0.34	0.09	-	-
10000	2	0.20	0.19	0.08	0.12	-
10000	3	0.15	0.31	0.05	0.05	0.06
20000	1	0.18	0.27	0.10	-	-
20000	2	0.15	0.24	0.08	0.10	-
20000	3	0.15	0.21	0.08	0.08	0.07

5 Conclusion

We have proposed in this paper an approach for computing from clinical data DTR strategies defined on a small subset of clinical indicators. The approach is based on a formalisation of the problem as an optimal control problem for which the system dynamics is unknown and replaced to some extent by the information contained in the clinical data. Once this formalisation is done, the tree-based approximators computed by the fitted Q iteration algorithm used for inferring policies from the data are analyzed to identify the 'most relevant variables'. This identification is carried out by exploiting variance reduction concepts which are

determinant in our approach. Preliminary simulation results carried out on some academic examples have shown that the proposed approach for selecting the most relevant indicators is promising.

Techniques based on variance reduction for selecting the most relevant indicators have already been successfully used in supervised learning (SL) (see, e.g., [5]) and have inspired the work reported in this paper. But many other techniques for selecting relevant variables have also been proposed in the literature on supervised learning, such as for example those based on Bayesian approaches [6, 7]. In this respect, it will be interesting to investigate to which extent these other approaches could be usefully exploited in our reinforcement learning context.

A next step in our research is to test our variable selection approach for getting policies defined on a small subset of indicators on real-life clinical data. However, in such a context, one difficulty we will face is the inability to determine whether the indicators selected by our approach are indeed the right ones since no accurate model of the system will be available. This issue is closely related to the problem of estimating the quality of a policy in model-free RL. We believe it is made particularly relevant in the context of DTRs since it would probably be unacceptable to adopt some dynamic treatment regimes which would trade the use of a smaller number of decision variables at the expense of a significant deterioration of the health of patients.

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. Damien Ernst acknowledges the financial support of the Belgium National Fund of Scientific Research (FNRS) of which he is a Research Associate. The scientific responsibility rests with its authors.

References

1. Murphy, S.: An experimental design for the development of adaptative treatment strategies. Statistics in Medicine **24** (2005) 1455–1481
2. Ernst, D., Geurts, P., Wehenkel, L.: Tree-based batch mode reinforcement learning. Journal of Machine Learning Research **6** (2005) 503–556
3. Froberg, D., Kane, R.: Methodology for measuring health-state preferences–ii: Scaling methods. Journal of Clinical Epidemiology **42** (1989) 459471
4. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Machine Learning. **36**(Number 1) (2006) 3–42
5. Wehenkel, L.: Automatic learning techniques in power systems. Kluwer Academic, Boston (1998)
6. Cui, W.: Variable Selection: Empirical Bayes vs. Fully Bayes. PhD thesis, The University of Texas at Austin (2002)
7. George, E., McCulloch, R.: Approaches for Bayesian variable selection. Statistica Sinica **7, 2** (1997) 339–373