Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 826–834
Neural Networks
www.elsevier.com/locate/neunet

2005 Special Issue

Individualization of pharmacological anemia management using reinforcement learning☆

Adam E. Gaweda[a,*], Mehmet K. Muezzinoglu[b], George R. Aronoff[a], Alfred A. Jacobs[a], Jacek M. Zurada[b], Michael E. Brier[a,c]

[a]Department of Medicine, University of Louisville, Louisville, KY 40292, USA
[b]Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY 40292, USA
[c]Department of Veteran Affairs, Louisville, KY 40202, USA

Abstract

Effective management of anemia due to renal failure poses many challenges to physicians. Individual response to treatment varies across patient populations and, due to the prolonged character of the therapy, changes over time. In this work, a Reinforcement Learning-based approach is proposed as an alternative method for individualization of drug administration in the treatment of renal anemia. *Q*-learning, an off-policy approximate dynamic programming method, is applied to determine the proper dosing strategy in real time. Simulations compare the proposed methodology with the currently used dosing protocol. Presented results illustrate the ability of the proposed method to achieve the therapeutic goal for individuals with different response characteristics and its potential to become an alternative to currently used techniques.

Keywords: Reinforcement learning; Drug dosing; Anemia management

1. Introduction

Drug administration in chronic conditions is a process of trial and error within a feedback loop. An initial drug dose is first selected as recommended by a standard reference. The patient is then observed for specific physiologic responses or adverse events. Subsequently, the clinician adjusts the dose following the observed state of the patient. If toxicity occurs, the dose amount is decreased. If an inadequate response is observed, the dose is increased. The trial and error process continues until a desired response is achieved.

Oftentimes, the relationship between the drug dose and the patient's response is complex. To facilitate drug administration, practitioners attempt to use protocols. Such protocols are developed from average responses to treatment in populations of patients. Nevertheless, achieving a desired therapeutic response on an individual basis is complicated by the differences within the population, as well as other concurrent medications and comorbidities, specific for each patient.

Reinforcement Learning (RL) is a methodology based on ideas from psychology that serves for control theory and stochastic optimization. It has a potential to become an effective tool for support of clinical decision making in patient care. (Bellman, 1983) described a general framework for applying Dynamic Programming (DP), a cornerstone methodology to RL, to pharmacotherapeutic planning using Pharmacokinetic and Pharmacodynamic (PK/PD) models. A pioneering demonstration of DP in pharmacotherapy can be found in (Buell, Jeliffe, Kalaba, & Sridhar, 1970). Other examples of using DP for pharmacotherapeutic planning include the works (Hu, Lovejoy, & Shafer, 1994a,b). (Schaeffer, Bailey, Shechter, & Roberts, 2004) reviewed various instances of medical application of Markov Decision Processes (MDP), the underlying control setting in RL. Most recently, (Moore, Sinzinger, Quasny, & Pyeatt, 2004) demonstrated how RL can be successfully employed in closed-loop control of patient sedation in an Intensive Care Unit.

Our previous work (Gaweda et al., 2005), which constitutes the origin of this paper, was aimed at discovering a complete administration policy for proper drug dosing

☆An abbreviated version of some portions of this article appeared in (Gaweda et al., 2005), published under the IEEE copyright.

* Corresponding author.
E-mail address: adam.gaweda@kdp.louisville.edu (A.E. Gaweda).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.06.020

during pharmacotherapeutic management of renal anemia. This was achieved by an on-policy RL method, SARSA, in which a patient model was probed by possibly non-optimal policies during an episodic learning process. Construction of such a policy requires sufficiently many occurrences of all possible state transitions, potentially causing over- or under-dosing. As a result, we showed that on-policy episodic RL tools can discover a useful dosing policy, as a product of a learning process, which may be however unacceptably long in real-time pharmacotherapy.

In this paper, we view the control problem at a lower level of generality, where the goal is to stabilize the Hemoglobin level within the target range of 11–12 mg/dl by evaluating reinforcements derived from state transitions. Due to the partially known, monotonic character of the dose-response relationship, we were able to reduce the Markov chain representation of the patient to a few representative states. In this way, the learning phase to reach an acceptable control can be shortened. To avoid probing the system by Suboptimal dosing policies during long training episodes, we utilize here a Q-learning mechanism (Watkins & Dayan, 1992) for evaluation of the state/action pairs. The proposed learning system determines the optimal drug dose using reinforcements, which are produced immediately after state transitions occurring within the patient dynamics during treatment. In contrast to our previous work where an RBF network was used for Q-table approximation, we use the RBF network here as an interpolator to identify the policy on the entire continuous state space.

The organization of the paper is as follows. Modeling patient dynamics and the drug dosing problem are presented in Section 2. Section 3 describes the use of a Markovian finite-state Q-learning method to achieve the control of the continuous-state patient dynamics. Experimental evaluation of the proposed approach is presented in Section 4. The results are also compared to those obtained using a simulated clinical protocol for anemia management. Concluding remarks and observations are discussed in Section 5.

2. Drug dosing problem

2.1. Patient dynamics

Anemia management is a typical control problem under uncertainty. The controlled quantity is the Hemoglobin level (HgB) and the control signal is the amount of Erythropoietin (EPO) administered by the physician. Iron stores in the body, determined by Transferrin Saturation (TSat), have an impact on the process of red blood cell production and are considered as an auxiliary state component. In this setting, a patient is viewed as a discrete-time dynamic system with the state space $\mathcal{H} \times \mathcal{S}$, where $\mathcal{H}$ and $\mathcal{S}$ are sets of valid HgB and TSat levels, respectively. We denote the control space, i.e. the set of valid EPO amounts, by $\mathcal{E}$. As the HgB and TSat measurements are performed monthly, the time index k denotes a month.

In the classical pharmacological framework, a patient's response is analyzed using a PK/PD compartment model containing a set of differential equations. In the case of the red blood cell production, called erythropoiesis, regular measurement of EPO concentration would be required to acquire all the information necessary to build a PK/PD model. Due to the expensive character of this procedure, alternative modeling methods, such as Artificial Neural Networks become a feasible option. In (Gaweda, Jacobs, Brier, & Zurada, 2003), a population-based neural network was proposed for dose–response modeling in anemia management. For the purpose of this study, we developed a 'subpopulation' approach. The underlying principle for this approach was the existence of several distinct response groups within a patient population. Each one of these groups was assumed to bear a unique dose–response relationship. Using fuzzy rules, a patient's response is first classified and subsequently a prediction of HgB level one-step ahead is performed using the following second-order model:

$$x_1[k+1] = \theta_1 u[k-1] + \theta_2 u[k] + \theta_3 u[k+1] + \theta_4 x_1[k-1] + \theta_5 x_1[k] + \theta_6 x_2[k] + \theta_0 \quad (1)$$

where u is the control input (EPO), x_1 is the HgB, and x_2 is the TSat. The response is classified based on the six month average levels of HgB, TSat, and EPO. The proposed approach can be conveniently implemented using Takagi-Sugeno (TS) fuzzy model (Takagi & Sugeno, 1985).

Records of 186 patients at the Division of Nephrology, University of Louisville, were used to perform data-driven estimation of the TS model. The data were randomly divided into equally sized estimation (training) and evaluation (testing) sets, containing data of 93 patients each. For consistency, a total of 100 model estimations were performed using different patient selections for estimation and evaluation. Eventually, the following three-rule TS model was obtained:

R_1: If (*avg* $\text{EPO}_{6\text{m}}$, *target* $\text{HgB}_{6\text{m}}$, *norm* $\text{TSat}_{6\text{m}}$) Then $\text{HgB}[k+1] = \Theta_1 \zeta$

R_2: If (*avg* $\text{EPO}_{6\text{m}}$, *target* $\text{HgB}_{6\text{m}}$, *low* $\text{TSat}_{6\text{m}}$) Then $\text{HgB}[k+1] = \Theta_2 \zeta$

R_3: If (*high* $\text{EPO}_{6\text{m}}$, *low* $\text{HgB}_{6\text{m}}$, *low* $\text{TSat}_{6\text{m}}$) Then $\text{HgB}[k+1] = \Theta_3 \zeta$

In these rules, the subscript 6m denotes the six month average of the corresponding quantity, Θ_i are the parameter vectors of the predictive model (1), and ζ is the regressor vector:

$$\zeta = [\text{EPO}[k-1], \text{EPO}[k], \text{EPO}[k+1], \text{HgB}[k-1], \text{HgB}[k], \text{TSat}[k], 1]$$