REINFORCEMENT LEARNING APPROACH TO INDIVIDUALIZATION OF CHRONIC PHARMACOTHERAPY

Kerem Muezzinoglu

Electrical and Computer Engineering Department
University of Louisville

July 26, 2005
Autonomous Learning Lab.
University of Massachusetts, Amherst
1 Motivation
   - Anemia Management
   - The Problem

2 Learning Treatment
   - Patient Model
   - Reinforcement Learning Solutions

3 Summary and Future Work
Drug dosing is a control problem under uncertainty. Conventional control methods apply on dose/response models. Feasible models ignore relevant internal dynamics. Population data yields unacceptably low resolutions. Real-life drug dosing typically involves trial-and-error. This is what we want to formulate.
Anemia Management

Introduction

- Anemia is a universal consequence renal failure.
- State variables
  - Hemoglobin (HgB) - measured in g/dL
  - Transferrin Saturation (TSat) - in percent
- Input
  - Erythropoietin (EPO) - applied in 1,000 units
- Sampling period $k$ is a month.
Reformulated as a minimum time problem for the given patient

Extract the optimum deterministic (and algebraic) policy that drives HgB level into the target range 11-12 g/dL for each initial HgB level and maintains it therein.
Anemia Management
Constraints

- One state component, TSat, is a random variable.
- No trustable mathematical model of HgB dynamics.
- Need to store and evaluate past interactions with patient.
- Little tolerance for exploration and premature policies.
- Ideally, learning should be completed in early steps of treatment.
1. Motivation
   - Anemia Management
   - The Problem

2. Learning Treatment
   - Patient Model
   - Reinforcement Learning Solutions

3. Summary and Future Work
A model to respond upon valid actions is necessary.

Two typical tendencies in simulated patient data
- Responders
- Non-responders

An individual is located online based on observations.

The individual is assigned a linear combination of predefined prototype HgB iterations

\[
HgB[k + 1] = \mu_r f_r (HgB[k], TSat[k], EPO[k]) + \mu_{nr} f_{nr} (HgB[k], TSat[k], EPO[k])
\]

\[
TSat[k + 1] \sim N(TSat[k], \sigma_{TSat}^2)
\]

The model was initiated by real patient data.
Patient Model
Construction from real data

![Graph showing patient model construction from real data](image-url)
First, we assumed treatment in episodes.

An episode contained the history of interactions (EPO vs HgB/TSat) on a fake patient for two years.

Reward: \(- |HgB[k + 1] - 11.5|\)

State abstraction: \(HgB \in \{9, 10.5, 11.5, 12.5, 14\}\)

Trained using SARSA(\(\lambda\)) for 200 episodes.

Uniform random exploration, \(\gamma = 0.9, \lambda = 1, \nu = 0.1\)
SARSA(λ)
Policy Iteration Scheme

Sample Trajectory

\[ g(i_k, i_{k+1}) = |HGB[k+1] - 11.5| \]

Update Q-table

End of Episode? 
\[ k=23 \]

Yes

State/action Values \( Q^* \)

Update Policy

New Learning Episode

Learning the control law

Learning the values of actions
SARSA on Responder

Graphs showing changes in EPO and HGB over time.
SARSA on Non-Responder

![Graph showing EPC and HGB over time](image.png)
An online learning scheme.
Reward:

\[ g(s, s') = \begin{cases} 
2 & \text{if } 11.0 \leq s' \leq 12.0 \\
-1 & \text{if } \{s < 11.0 \text{ and } s' < s\} \\
1 & \text{otherwise}
\end{cases} \]

State abstraction: \( HgB \in \{9, 10.5, 11.5, 12.5, 14\} \)

Agent’s policy is generalized by an RBF interpolator.

Initialized with a best-guess policy.

Imposed less tendency to explore at extreme states.

\( \gamma = 0.9, \lambda = 1, \nu = 0.1 \)
Q-Learning on Responder

HgB (g/dL)

EPO (1,000 U)

month
Q-Learning on Non-Responder

---

Graphs showing the Hgb (g/dL) and EPO (1,000 U) levels over time in month.
### Simulation Statistics for Q-Learning

<table>
<thead>
<tr>
<th>Response Group</th>
<th>Normal</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>HgB Level</td>
<td>11.59 (11.12, 12.04)</td>
<td>11.16 (10.76, 11.55)</td>
</tr>
<tr>
<td>HgB Variability</td>
<td>0.29 (0.15, 0.42)</td>
<td>0.74 (0.52, 0.95)</td>
</tr>
<tr>
<td>Total EPO</td>
<td>589.29 (344.56, 834.02)</td>
<td>1145.25 (926.61, 1363.88)</td>
</tr>
</tbody>
</table>

### Simulation Statistics for AMP

<table>
<thead>
<tr>
<th>Response Group</th>
<th>Normal</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>HgB Level</td>
<td>11.66 (11.56, 11.78)</td>
<td>11.51 (11.35, 11.67)</td>
</tr>
<tr>
<td>HgB Variability</td>
<td>0.32 (0.22, 0.41)</td>
<td>0.67 (0.49, 0.84)</td>
</tr>
<tr>
<td>Total EPO</td>
<td>610.57 (356.91, 864.23)</td>
<td>1075.39 (942.50, 1208.28)</td>
</tr>
</tbody>
</table>
Use of RL in pharmacotherapy is rational when a reliable patient model is unavailable.

A crude implementation of $Q(\lambda)$ derives an acceptable dosing policy in reasonable time.

Current performance comparable to human experts on artificial patients.

Outlook

- Utilize prior knowledge on the optimal policy.
- Attempts to minimize the dose.
- Safer off-policy approaches.
Thank you!

These slides can be found at
http://www.ci.uofl.edu/kerem/all-talk.pdf

Any questions/comments?