Embedded Machine Learning
Papers from the AAAI 2015 Fall Symposium

Adaptive Treatment Allocation Using Sub-Sampled Gaussian Processes

Audrey Durand
Department of Electrical Engineering
and Computer Engineering
Université Laval, Québec, Canada
taudrey.durand.2@ulaval.ca

Joelle Pineau
School of Computer Science
McGill University
Montréal, Canada
jpineau@cs.mcgill.ca

Abstract

Personalized medicine targets the customization of treatment strategies to patients' individual characteristics. Here we consider the problem of optimizing personalized pharmacological treatment strategies for cancer. We focus primarily on developing effective strategies to collect the data necessary for the construction of personalized treatments. We formulate this problem as a contextual bandit and present a new algorithm based on repeated sub-sampling for robust data collection in this framework. We present a case study showing experiments on a simulation setting, built from real data collected in a previous animal experiments. Promising results in this case study have since lead us to deploy this strategy in a partner wet lab to allocate treatments for the next phase of animal experiments.

We consider the problem of designing an efficient data collection strategy during animal experiments investigating the effectiveness of cancer treatment medication. During an initial data collection phase, data was acquired by randomly assigning treatments twice a week to six mice with up to three induced cancer tumours each. This allowed to collect data on a total of 12 tumours. Tumour measurements were taken right before administering a treatment. We consider the administration of the following treatments: 1) none (n=42); 2) 5-FU (n=66); 3) imiquimod (n=24); and 4) simultaneous imiquimod and 5-FU (n=31). Given that some treatments may be more effective at different cancer stages, our goal is to use this data to determine an efficient treatment allocation policy to be applied in the next batch of experiments.

We model this problem as a categorical contextual bandit (Auer 2002; Langford and Zhang 2007) where the actions correspond to the available treatments, the context is the tumour volume (in mm^3) before the treatment, as calculated by $\frac{\pi}{6}(hv)^{\frac{3}{2}}$ where h and v are respectively the horizontal and vertical tumour measurements, and the reward is the tumour volume reduction following the treatment. We formalize the contextual bandit problem with categorical actions as an episodic game. At each episode $t > 0$, the player observes a context $\boldsymbol{x}(t) \in \mathcal{X}$ and must choose the next action $a(t) \in \mathcal{A}$ to play. The agent then observes a reward (perturbed by noise) $r(t) = f_{a(t)}(\boldsymbol{x}(t)) + \varepsilon(t)$, where $f_a : \mathcal{X} \rightarrow \mathbb{R}$ is an unknown function and $\varepsilon(t)$ is a zero mean random noise i.i.d. accross episodes. A typical performance metric in the stochastic bandit setting is the cumulative pseudo-regret, that is the loss in reward incurred for not knowing which arm is optimal in each context. For a given context $\boldsymbol{x}$, the optimal average reward $f^*(\boldsymbol{x}) = \max_{a \in \mathcal{A}} f_a(\boldsymbol{x})$. The goal is to minimize the cumulative pseudo-regret after T episodes, given by $\hat{R}(T) = \sum_{t=1}^{T} f^*(\boldsymbol{x}(t)) - f_{a(t)}(\boldsymbol{x}(t))$.

We assume that functions f_a are samples from *known* Gaussian process (GP) distributions. Rasmussen and Williams (2006) describe a GP as a generalization of the Gaussian probability distribution, where a stochastic process governs the properties of Gaussian distributions at every point of a space. A GP(μ, k) is completely described by its mean function $\mu : \mathcal{Z} \rightarrow \mathbb{R}$, $\mu(\mathbf{z}) = \mathbb{E}[g(\mathbf{z})]$ and covariance (kernel) function $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$, $k(\mathbf{z}, \mathbf{z}') = \mathbb{E}[(g(\mathbf{z}) - \mu(\mathbf{z}))(g(\mathbf{z}') - \mu(\mathbf{z}'))]$. Suppose we condition a GP(μ, k) on observed outputs $\mathbf{y} = [y_1, \ldots, y_N]^T$ associated with inputs $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, where $y_n = g(\mathbf{z}_n) + \varepsilon$ with i.i.d. Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The predictive distribution at test point $\mathbf{z}_*$ is estimated by

$$\hat{g}_* = \boldsymbol{k}_*^T(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}\mathbf{y}, \tag{1}$$

$$\mathbb{V}[g_*] = k(\mathbf{z}_*, \mathbf{z}_*) - \boldsymbol{k}_*^T(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}\boldsymbol{k}_* \tag{2}$$

where $\boldsymbol{k}_* = [k(\mathbf{z}_1, \mathbf{z}_*), \ldots, k(\mathbf{z}_N, \mathbf{z}_*)]^T$ and $\boldsymbol{K}$ is the positive semi-definite kernel matrix $[k(\mathbf{z}, \mathbf{z}')]_{\forall \mathbf{z}, \mathbf{z}' \in Z_N}$. Here, the space $\mathcal{Z}$ corresponds to the set of contexts $\mathcal{X}$.

Robust Estimation via Sub-Sampling

The recent bandit algorithm BESA (Baransi, Maillard, and Mannor 2014) uses a sub-sampling procedure to fairly compare actions. Given two actions a_1 and a_2 that have been played respectively $n_{a_1}(t)$ and $n_{a_2}(t) > n_{a_1}(t)$ times respectively up to episode t, comparing their emprirical averages is not fair because action a_1 has less samples than action a_2. To compensate for this situation, BESA sub-samples without replacement $n_{a_1}(t)$ data out of the $n_{a_2}(t)$ samples of action a_2 and computes the empirical average of action a_2 on this subset. We consider a generalization of BESA to the categorical contextual bandit problem.

In the categorical contextual bandit problem, we model each function f_a using a dedicated GP, denoted GP$_a$. The hyperparameters of each GP$_a$ are estimated using the initial available data (Krause and Ong 2011). When selecting one of two actions a_1 and a_2, a natural way of proceeding is to base the decision upon the posterior distributions of their reward functions by conditioning the GPs on their respective history of observations. Following the idea of BESA, this would not be fair because they might not have the same number of samples. To compensate, we propose to compute the posterior distribution by conditioning only on a sub-sample (without replacement) of the available samples, as shown by Algorithm 1. Similar to BESA, the algorithm can easily be extended to multiple arms by organizing a pair-wise tournament between randomly permuted arms (Baransi, Maillard, and Mannor 2014).

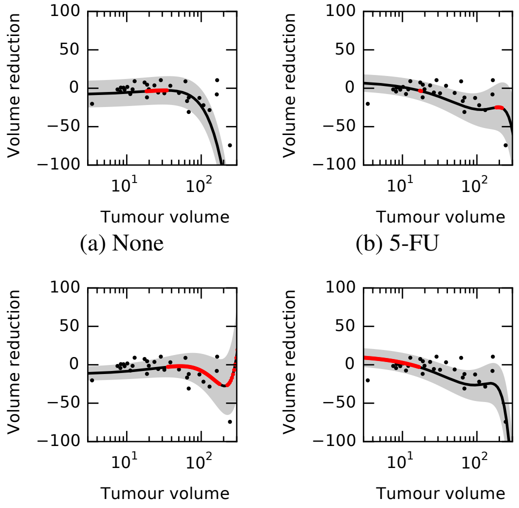
Algorithm 1 Sub-sampled GP for a contextual bandit with two categorical actions

```
Require: Actions $a_1$ and $a_2$ associated with Gaussian processes GP$_{a_1}$ and GP$_{a_2}$, the number of times $n_a(t)$ that action $a$ has been played up to episode $t$, the history of observations $\mathcal{D}_a(t)$ associated with action $a$ up to episode $t$, and the context $\boldsymbol{x}(t)$ received at episode $t$.
$n(t) = \min_{a \in \{a_1, a_2\}} n_a(t)$
Uniformely sample $n(t)$ observations without replacement from $\mathcal{D}_{a_1}(t)$ and $\mathcal{D}_{a_2}(t)$ as $\mathcal{S}_{a_1}(t)$ and $\mathcal{S}_{a_2}(t)$, respectively.
Define $\hat{f}_{a_1}(t)$ and $\hat{f}_{a_2}(t)$ as the posterior means of GP$_{a_1}$ and GP$_{a_2}$, respectively conditioned on observations $\mathcal{S}_{a_1}(t)$ and $\mathcal{S}_{a_2}(t)$, and evaluated at test point $\boldsymbol{x}(t)$.
Choose $a(t) = \operatorname{argmax}_{a \in \{a_1, a_2\}} \hat{f}_a(t)$, break tie by choosing $a(t) = \operatorname{argmin}_{a \in \{a_1, a_2\}} n_a(t)$.
Play $a(t)$ and observe $r(t)$.
```

Experiments

On each episode, a tumour volume is sampled from the exponential distribution $\lambda e^{-\lambda(x-\gamma)}$ with $\gamma = 3.42$ and scale $1/\lambda = 66.88$, fitted on the distribution of the available contexts. Tumour reductions are modelled using cubic regression on available data and noise is modelled by linear regression on the standard deviations of available data from the cubic model, as shown by Figure 1. Red indicates that the treatment is *optimal* (it has the largest expected value) in this context. A negative reduction indicates a grow. The logarithmic scale emphasizes the expected reward functions for small tumours, as they are more frequent than large tumours. This setting raises an important new challenge: the noise is not constant over the space of contexts. Moreover, the noise increases with the tumour volume, and so is the sub-optimality gap. This situation is especially hard on regret cumulation as each mistake leads to more regret and mistakes are more likely to occur when it is noisy.

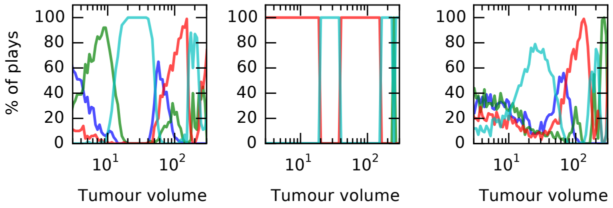
We compare the sub-sampling approach based on BESA with two other approaches for (non-linear) contextual bandits using GPs: CGP-UCB (Krause and Ong 2011) and Thompson sampling (Thompson 1933). All approaches use squared exponential kernels. Experiments are conducted over $T = 2\,000$ episodes, for 20 runs.



(a) None (b) 5-FU

(c) Imiquimod (d) 5-FU+Imiquimod

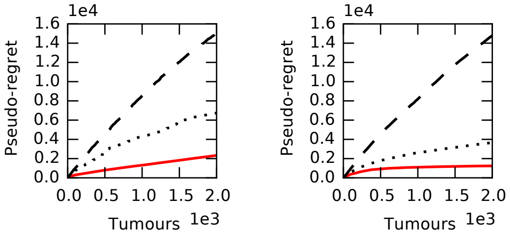
Figure 1: Experimental models of average reward functions using polynomial regression and linear noise (gray area).



(a) Sub-sampling (b) CGP-UCB (c) Thompson

Figure 2: Initial policies as probability of administering no treatement (dark blue), 5-FU (green), imiquimod (red), or 5-FU+imiquimod (light blue) given the context.

Figure 2 shows the initial recommended policies (using only the available data). Being randomized, the sub-sampling approach and Thompson recommend every treatments according to their estimated probabiblity of being optimal, while deterministic CGP-UCB recommends a single treatment for each context. The policy recommended by the sub-sampling is similar to Thompson, with emphasis on 5-FU for low tumour volume. Though this allocation policy does not yet match the optimal regions (red) of the average reward functions (Fig. 1), Figure 3a shows that the sub-sampling approach adapts faster to the model and at a lower regret cost than the other methods. In order to validate that the results not only due to a favorable modelling, experiments are also conducted using average rewards functions obtained with linear regression (instead of cubic regression). Figure 3b shows the average cumulative pseudo-regret with this linear model. One observes that, if Thompson performs better than with the polynomial model, the sub-sampling approach still outperforms all alternatives.



(a) Polynomial model (b) Linear model

Figure 3: Cumulative pseudo-regret using sub-sampling (red), CGP-UCB (dash), and Thompson sampling (dot).

Conclusion

We presented a problem that highlights several challenges of real-world applications such as how to deal with low amount of initial data and high noise. We formulate it as a (non-linear) contextual bandit and propose a new solution approach based on an easy-to-implement sub-sampling approach that requires few parameters. This approach is strongly supported by our experimental results in the simulation case. On the basis of these findings, we have since deployed the strategy in a partner wet lab to collect the next phase of experimental data. Results should be available in the coming months.

Acknowledgments We would like to thank Georgios Mitsis (McGill University), Katerina Strati (University of Cyprus), Demetris Iacovides (University of Cyprus), and Charis Achilleos (University of Cyprus) for providing the data used in this work. This work was supported through funding from NSERC (Canada) and REPARTI-FRQNT (Québec).

References

Auer, P. 2002. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3:397–422.

Baransi, A.; Maillard, O.-A.; and Mannor, S. 2014. Sub-sampling for multi-armed bandits. In *Proceedings of the European Conference on Machine Learning (ECML)*.

Krause, A., and Ong, C. S. 2011. Contextual Gaussian process bandit optimization. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2447–2455.

Langford, J., and Zhang, T. 2007. The epoch-greedy algorithm for contextual multi-armed bandits. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 1096–1104.

Rasmussen, C. E., and Williams, C. K. I. 2006. Gaussian processes for machine learning. *MIT Press* 2(4).

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika Trust* 25(3):285–294.