Online Contextual Multi-Armed Bandits: An Empirical Analysis Using News Article Recommendations

James Wang
Haas School of Business, UC Berkeley

November 26, 2014

Abstract

Hard to benchmark bandit algorithms due to "off-policy evaluation problem". Dataset aims to solve that problem, and take opportunity to empirically test, using Yahoo TODAY! data, a variety of context bandit algorithms and see their behavior against theoretical bounds.

Introduction

This paper evaluates a variety of contextual bandit algorithms and how empirical performance in a real-world dataset compares with theoretical results.

Most algorithms work far better than worst-case regret bounds, otherwise they wouldn't be useful in practical applications. Unfortunately, past work has largely focused on either simulated data, toy problems, or proprietary datasets whose results and underlying characteristics cannot be scrutinized.

This paper takes advantage of the fact that we have randomized policy results from a web service (Yahoo frontpage). Using a method described in a previous paper that reflects a process similar to rejection sampling, we can obtain an unbiased estimate of the how our algorithms would perform in a real-world dataset [6].

Problem Formulation

Online Contextual Bandit Problem

This is a contextual bandit problem, where we observe a context and reward.

This has to be online, because in the context of the web and other large-scale production applications, because it's inherently realtime and there's too much data to do offline.

Compared Algorithms

ϵ-greedy

This algorithm

UCB

This is another algorithm

Thompson Sampling

GLM-UCB

This is the main algorithm

Comparison of Theoretical Bounds

I can insert a fancy table here. Maybe I can talk about adversarial vs. stochastic bounds. Or not.

Experiments

Yahoo! Today Module and Webscope Data

This is a great dataset for this Describe how the data was collected

Experimental Setup

Data Description and Characteristics

Describe the random bucket thing, some basic features of the data. Varying pool of arms to choose from over time. Arms, number of arms, everything shifts over time. This is more in line with a real problem.

Describe how big the data is and how this necessitates an online solution.

Feature Construction

Conjoint analysis, k-means into 5 clusters that are interchangeable

Resolving the Off-Policy Problem

Of course, randomization doesn't solve the off-policy problem. We can still choose a policy that isn't covered by the actual data. So here: Talk about the unbiased estimation method. Analogous to rejection sampling. Since the articles are randomly assigned, we can expect that rejected samples will be unbiased and thus the ultimate estimate will be unbiased.

Performance Evaluation

CTR and regret vs. an omniscent policy that picks the best arm for each group every time. Picked within each hour to prevent issues of data sparsity (but the distribution is relatively stable anyway, perhaps show best arms each minute, half hour, hour).

Results

Algorithm Comparisons

Show tables and charts of results here

Remarks

There's something interesting here.

Conclusion

This was a good paper.

Reference List So Far

Offline Evaluation Method [6] A Contextual-Bandit Approach to Personalized News Article Recommendation [7] Parametric Bandits [3]. Further Optimal Regret Bounds for Thompson [1] Lai Robbins 1985 [4] Auer 2002 [2] Epoch Greedy Context [5]

References

[1]Agrawal, S. and Goyal, N. 2012. Further optimal regret bounds for thompson sampling. *CoRR.* abs/1209.3353, (2012).

[2]Auer, P., Cesa-Bianchi, N. and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47, 2-3 (May 2002), 235–256.

[3]Filippi, S., Cappe, O., Garivier, A. and Szepesvári, C. 2010. Parametric bandits: The generalized linear case. *Advances in neural information processing systems 23.* J. Lafferty, C. Williams, J. Shawe-taylor, R. Zemel, and A. Culotta, eds. 586–594.

[4]Lai, T.L. and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics.* 6, 1 (1985), 4–22.

[5]Langford, J. and Zhang, T. 2008. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems 20.* J. Platt, D. Koller, Y. Singer, and S. Roweis, eds. Curran Associates, Inc. 817–824.

[6]Li, L., Chu, W. and Langford, J. 2010. An unbiased, data-driven, offline evaluation method of contextual bandit algorithms. *CoRR.* abs/1003.5956, (2010).

[7]Li, L., Chu, W., Langford, J. and Schapire, R.E. 2010. A contextual-bandit approach to personalized news article recommendation. *CoRR.* abs/1003.0146, (2010).