A Large-scale Dataset for Decision Making Algorithms

Yuta Saito[1], Shunsuke Aihara[2], Megumi Matsutani[2] and Yusuke Narita[3]

[1]Hanjuku-kaso, Co., Ltd.
[2]ZOZO Technologies, Inc.
[3]Yale University

Abstract

We build and publicize the *Open Bandit Dataset and Pipeline* to facilitate scalable and reproducible research on bandit algorithms. They are especially suitable for *off-policy evaluation* (OPE), which attempts to predict the performance of hypothetical algorithms using data generated by a different algorithm. We construct the dataset based on experiments and implementations on a large-scale fashion e-commerce platform, ZOZOTOWN. The data contain the ground-truth about the performance of several bandit policies and enable the fair comparisons of different OPE estimators.

1. Introduction

Interactive bandit and reinforcement learning systems produce log data valuable for evaluating and redesigning the systems. For example, the logs of a news recommendation system record which news article was presented and whether the user read it, giving the system designer a chance to make its recommendation more relevant. Exploiting log data is, however, more difficult than conventional supervised machine learning: the result is only observed for the action chosen by the system but not for all the other actions the system could have taken. The logs are also biased in that the logs over-represent the actions favored by the system.

A potential solution to this problem is an A/B test that compares the performance of counterfactual systems in an online environment. However, A/B testing counterfactual systems is often difficult, since deploying a new policy is time- and money-consuming, and entails a risk of failure.

This leads us to the problem of *off-policy evaluation* (OPE), which aims to estimate the performance of a counterfactual policy using only log data collected by a past (or behavior) policy. Such an evaluation allows us to compare the performance of candidate counterfactual policies to decide which policy should be deployed. This alternative approach thus solves the above problem with the A/B test approach. Applications range from contextual bandits [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and reinforcement learning in the web industry [11, 12, 13, 14, 15, 16, 17] to other social domains such as healthcare [18] and education [19].

While the research community has produced theoretical breakthroughs, the experimental evaluation of OPE remains primitive. Specifically, it lacks a public benchmark dataset for comparing the performance of different methods. Researchers often validate their methods using synthetic simulation environments [12, 20, 17]. A version of the synthetic approach is to modify multi-class classification datasets and treat supervised machine learning methods as bandit policies to evaluate off-policy estimators [21, 22, 23, 10]. An obvious problem with these studies is that there is no guarantee that their simulation environment is similar to real-world settings. To solve this issue, [24, 25, 5, 13] use proprietary real-world datasets. Since these datasets are not public, however, it remains challenging to reproduce the results, and compare their methods with new ideas in a fair manner. This is in contrast to other domains of machine learning, where large-scale open datasets, such as the ImageNet dataset [26], have been pivotal in driving objective progress [27, 28, 29, 30, 31].

Our goal is to implement and evaluate OPE of bandit algorithms in realistic and reproducible ways. We release the *Open Bandit Dataset*, a logged bandit feedback collected on a large-scale fashion e-commerce platform, ZOZOTOWN.[1] ZOZOTOWN is the largest fashion EC platform in Japan with over 3 billion USD annual Gross Merchandise Value. When the platform produced the data, it used Bernoulli Thompson Sampling (Bernoulli TS) and Random policies to recommend fashion items to users. The dataset includes an A/B test of these policies and collected over 26 million records of users' clicks and the ground-truth about the performance of Bernoulli TS and Random. To streamline and standardize the analysis of the Open Bandit Dataset, we also provide the *Open Bandit Pipeline*, a series of implementations of dataset preprocessing, behavior bandit policy simulators, and

[1]https://corp.zozo.com/en/service/

Causality in Search and Recommendation (CSR) and Simulation of Information Retrieval Evaluation (Sim4IR) workshops at SIGIR, 2021
Editors of the proceedings (editors): Krisztian Balog, Xianjie Chen, Xu Chen, David Maxwell, Paul Thomas, Shuo Zhang, Yi Zhang, and Yongfeng Zhang.
saito@hanjuku-kaso.com (Y. Saito);
shunsuke.aihara@zozo.com (S. Aihara);
megumi.matsutani@zozo.com (M. Matsutani);
yusuke.narita@yale.edu (Y. Narita)

OPE estimators.

2. Setup

We consider a general multi-armed contextual bandit setting. Let $\mathcal{A} = \{0, ..., m\}$ be a finite set of $m + 1$ *actions* (equivalently, *arms* or *treatments*), that the decision maker can choose from. Let $Y(\cdot) : \mathcal{A} \to \mathbb{R}$ denote a potential reward function that maps actions into rewards or outcomes, where $Y(a)$ is the reward when action a is chosen (e.g., whether a fashion item as an action results in a click). Let X denote a *context* vector (e.g., the user's demographic profile and user-item interaction history) that the decision maker observes when picking an action. We denote the finite set of possible contexts by $\mathcal{X}$. We think of $(Y(\cdot), X)$ as a random vector with unknown distribution G. Given a vector of $(Y(\cdot), X)$, we define the mean reward function $\mu : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ as $\mu(x, a) = \mathbb{E}[Y(a)|X = x]$.

We call a function $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ a *policy*, which maps each context $x \in \mathcal{X}$ into a distribution over actions, where $\pi(a|x)$ is the probability of taking action a given a context vector x. Let $\{(Y_t, X_t, D_t)\}_{t=1}^T$ be historical logged bandit feedback with T rounds of observations. $D_t := (D_{t0}, ..., D_{tm})'$, where D_{ta} is a binary variable indicating whether action a is chosen in round t. If a is chosen in round t, $D_{ta} = 1$, otherwise $D_{ta} = 0$. $Y_t := \sum_{a=0}^m D_{ta} Y_t(a)$ and X_t denote the reward and the context observed in round t, respectively. We assume that a logged bandit feedback is generated by a *behavior policy* π_b as follows: (i) In each round $t = 1, ..., T$, $(Y_t(\cdot), X_t)$ is i.i.d. drawn from distribution G., (ii) Given X_t, an action is randomly chosen based on $\pi_b(\cdot|X_t)$, creating the action choice D_t and the associated reward Y_t.

- In each round $t = 1, ..., T$, $(Y_t(\cdot), X_t)$ is i.i.d. drawn from distribution G.
- Given X_t, an action is randomly chosen based on $\pi_b(\cdot|X_t)$, creating the action choice D_t and the associated reward Y_t.

Suppose that π_b is fixed for all rounds, and thus D_t is i.i.d. across rounds. Because $(Y_t(\cdot), X_t)$ is i.i.d. across rounds and $Y_t = \sum_{a=0}^m D_{ta} Y_t(a)$, each observation (Y_t, X_t, D_t) is i.i.d. across rounds. Note that D_t is independent of $Y_t(\cdot)$ conditional on X_t.

3. Off-Policy Evaluation

3.1. Prediction Target

We are interested in using the historical logged bandit data to estimate the following *policy value* of any given *counterfactual policy* π which might be different from π_b:

$$V^\pi := \mathbb{E}_{(Y(\cdot),X)\sim G}[\sum_{a=0}^m Y(a)\pi(a|X)] \quad (1)$$

$$= \mathbb{E}_{(Y(\cdot),X)\sim G,\ D\sim\pi_b}[\sum_{a=0}^m Y(a) D_a \frac{\pi(a|X)}{\pi_b(a|X)}] \quad (2)$$

where the last equality uses the independence of D and $Y(\cdot)$ conditional on X and the definition of $\pi_b(\cdot|X)$. We allow the counterfactual policy π to be degenerate, i.e., it may choose a particular action with probability 1. Estimating V^π before implementing π in an online environment is valuable because π may perform poorly and damage user satisfaction. Additionally, it is possible to select a counterfactual policy that maximizes the policy value by comparing their estimated performances.

3.2. Benchmark Estimators

There are several approaches to estimate the value of the counterfactual policy. A widely-used method, DM [32], first learns a supervised machine learning model, such as random forest, ridge regression, and gradient boosting, to predict the mean reward function. DM then uses it to estimate the policy value as

$$\hat{V}^\pi_{DM} = \frac{1}{T}\sum_{t=1}^T \sum_{a=0}^m \pi(a|X_t)\hat{\mu}(a|X_t).$$

where $\hat{\mu}(a|x)$ is the estimated reward function. If $\hat{\mu}(a|x)$ is a good approximation to the mean reward function, this estimator accurately predicts the policy value of the counterfactual policy V^π. If $\hat{\mu}(a|x)$ fails to approximate the mean reward function well, however, the final estimator is no longer consistent. The model misspecification issue is problematic because the extent of misspecification cannot be easily quantified from data [22].

To alleviate the issue with DM, researchers often use another estimator called IPW [33, 6]. IPW re-weights the rewards by the ratio of the counterfactual policy and behavior policy as

$$\hat{V}^\pi_{IPW} = \frac{1}{T}\sum_{t=1}^T \sum_{a=0}^m Y_t D_{ta} \frac{\pi(a|X_t)}{\pi_b(a|X_t)}.$$

When the behavior policy is known, the IPW estimator is unbiased and consistent for the policy value. However, it can have a large variance, especially when the counterfactual policy significantly deviates from the behavior policy.

The final approach is DR [21], which combines the above two estimators as

$$\hat{V}^\pi_{DR} = \frac{1}{T}\sum_{t=1}^T \sum_{a=0}^m (Y_t - \hat{\mu}(a|X_t)) D_{ta} \frac{\pi(a|X_t)}{\pi_b(a|X_t)}$$

Table 1
Statistics of the Open Bandit Dataset

Campaigns	Behavior Policies	#Data	#Items	Average Age	CTR (V^{π})	Relative-CTR
ALL	Random	1,374,327	80	37.93	0.35%	1.00
	Bernoulli TS	12,168,084			0.50%	1.43
Men's	Random	452,949	34	37.68	0.51%	1.48
	Bernoulli TS	4,077,727			0.67%	1.94
Women's	Random	864,585	46	37.99	0.48%	1.39
	Bernoulli TS	7,765,497			0.64%	1.84

Notes: Bernoulli TS stands for Bernoulli Thompson Sampling. **#Data** is the total number of user impressions observed during the 7-day experiment. **#Items** is the total number of items having a non-zero probability of being recommended by each behavior policy. **Average Age** is the average age of users in each campaign. **CTR** is the percentage of a click being observed in log data, and this is the ground-truth performance of behavior policies in each campaign. 95% confidence interval (CI) of CTR is calculated based on a normal approximation of Bernoulli sampling. **Relative-CTR** is CTR relative to that of the Random policy for the "All" campaign.

$$+ \pi(a|X_t)\hat{\mu}(a|X_t).$$

DR mimics IPW to use a weighted version of rewards, but DR also uses the estimated mean reward function as a control variate to decrease the variance. It preserves the consistency of IPW if either the importance weight or the mean reward estimator is accurate (a property called *double robustness*). Moreover, DR is *semiparametric efficient* [5] when the mean reward estimator is correctly specified. On the other hand, when it is wrong, this estimator can have larger asymptotic mean-squared-error than IPW [34] and perform poorly in practice [35].

4. Dataset

We apply and evaluate the above methods by using real-world data. Our data is logged bandit feedback data we call the *Open Bandit Dataset*.[2] The dataset is provided by ZOZO, Inc.[3], the largest Japanese fashion e-commerce company with a market capitalization of over 5 billion USD (as of May 2020). The company recently started using context-free multi-armed bandit algorithms to recommend fashion items to users in their large-scale fashion e-commerce platform called ZOZOTOWN.

We collected the data in a 7-days experiment in late November 2019 on three "campaigns," corresponding to "all", "men's", and "women's" items, respectively. Each campaign randomly uses either the Random algorithm or the Bernoulli Thompson Sampling (Bernoulli TS) algorithm for each user impression. In the notation of our bandit setups, action a is one of the possible fashion items, while reward Y is a click indicator. We describe some statistics of the dataset in Table 1. The data is large and contains many millions of recommendation instances. The number of actions is also sizable, so this setting is challenging for bandit algorithms and their OPE.

[2] https://research.zozo.com/data.html
[3] https://corp.zozo.com/en/about/profile/

5. Conclusion and Future Work

To enable realistic and reproducible evaluation of off-policy evaluation of bandit algorithms, we have publicized the Open Bandit Dataset–a benchmark logged bandit dataset collected on a large-scale fashion e-commerce platform.

In the near future, we plan to publicize the performance of the selected counterfactual policy in an online environment. Such an evaluation will produce additional log data generated by the contextual policy (while the current open dataset contains only log data generated by the old context-free policy). We aim to constantly expand and improve the Open Bandit Dataset to include more data and tasks.

References

[1] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, E. Snelson, Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, Journal of Machine Learning Research 14 (2013) 3207–3260.

[2] L. Li, W. Chu, J. Langford, T. Moon, X. Wang, An Unbiased Offline Evaluation of Contextual Bandit Algorithms with Generalized Linear Models, in: Journal of Machine Learning Research: Workshop

and Conference Proceedings, volume 26, 2012, pp. 19–36.

[3] L. Li, W. Chu, J. Langford, R. E. Schapire, A Contextual-bandit Approach to Personalized News Article Recommendation, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 661–670.

[4] L. Li, W. Chu, J. Langford, X. Wang, Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 2011, pp. 297–306.

[5] Y. Narita, S. Yasui, K. Yata, Efficient counterfactual learning from bandit feedback, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 4634–4641.

[6] A. Strehl, J. Langford, L. Li, S. M. Kakade, Learning from Logged Implicit Exploration Data, in: Advances in Neural Information Processing Systems, 2010, pp. 2217–2225.

[7] A. Swaminathan, T. Joachims, Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization, Journal of Machine Learning Research 16 (2015) 1731–1755.

[8] A. Swaminathan, T. Joachims, The Self-normalized Estimator for Counterfactual Learning, in: Advances in Neural Information Processing Systems, 2015, pp. 3231–3239.

[9] A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, I. Zitouni, Off-policy Evaluation for Slate Recommendation, in: Advances in Neural Information Processing Systems, 2017, pp. 3635–3645.

[10] Y.-X. Wang, A. Agarwal, M. Dudik, Optimal and Adaptive Off-policy Evaluation in Contextual Bandits, in: Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 3589–3597.

[11] N. Jiang, L. Li, Doubly robust off-policy value evaluation for reinforcement learning, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 652–661.

[12] Y. Liu, O. Gottesman, A. Raghu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, E. Brunskill, Representation balancing mdps for off-policy policy evaluation, in: Advances in Neural Information Processing Systems, 2018, pp. 2644–2653.

[13] Y. Narita, S. Yasui, K. Yata, Off-policy bandit and reinforcement learning, arXiv preprint arXiv:2002.08536 (2020).

[14] P. Thomas, E. Brunskill, Data-efficient Off-policy Policy Evaluation for Reinforcement Learning, in: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 2139–2148.

[15] P. Thomas, G. Theocharous, M. Ghavamzadeh, High confidence policy improvement, in: Proceedings of the 32th International Conference on Machine Learning, 2015, pp. 2380–2388.

[16] P. S. Thomas, G. Theocharous, M. Ghavamzadeh, High-Confidence Off-Policy Evaluation, AAAI (2015) 3000–3006.

[17] T. Xie, Y. Ma, Y.-X. Wang, Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling, in: Advances in Neural Information Processing Systems, 2019, pp. 9665–9675.

[18] S. A. Murphy, M. J. van der Laan, J. M. Robins, C. P. P. R. Group, Marginal mean models for dynamic regimes, Journal of the American Statistical Association 96 (2001) 1410–1423.

[19] T. Mandel, Y.-E. Liu, S. Levine, E. Brunskill, Z. Popovic, Offline policy evaluation across representations with applications to educational games, in: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems, 2014, pp. 1077–1084.

[20] C. Voloshin, H. M. Le, N. Jiang, Y. Yue, Empirical study of off-policy policy evaluation for reinforcement learning, arXiv preprint arXiv:1911.06854 (2019).

[21] M. Dudík, D. Erhan, J. Langford, L. Li, Doubly Robust Policy Evaluation and Optimization, Statistical Science 29 (2014) 485–511.

[22] M. Farajtabar, Y. Chow, M. Ghavamzadeh, More robust doubly robust off-policy evaluation, in: Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 1447–1456.

[23] N. Vlassis, A. Bibaut, M. Dimakopoulou, T. Jebara, On the design of estimators for bandit off-policy evaluation, in: International Conference on Machine Learning, 2019, pp. 6468–6476.

[24] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, S. Dollé, Offline a/b testing for recommender systems, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 198–206.

[25] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, B. Carterette, Offline evaluation to make decisions about playlist recommendation algorithms, in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 420–428.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[27] V. P. Dwivedi, C. K. Joshi, T. Laurent, Y. Bengio, X. Bresson, Benchmarking graph neural networks, arXiv preprint arXiv:2003.00982 (2020).

[28] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu,

M. Catasta, J. Leskovec, Open graph benchmark: Datasets for machine learning on graphs, arXiv preprint arXiv:2005.00687 (2020).

[29] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[32] A. Beygelzimer, J. Langford, The offset tree for learning with partial labels, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 129–138.

[33] D. Precup, R. S. Sutton, S. Singh, Eligibility Traces for Off-Policy Policy Evaluation, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 759–766.

[34] N. Kallus, M. Uehara, Intrinsically efficient, stable, and bounded off-policy evaluation for reinforcement learning, in: Advances in Neural Information Processing Systems, 2019.

[35] J. D. Kang, J. L. Schafer, et al., Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data, Statistical science 22 (2007) 523–539.