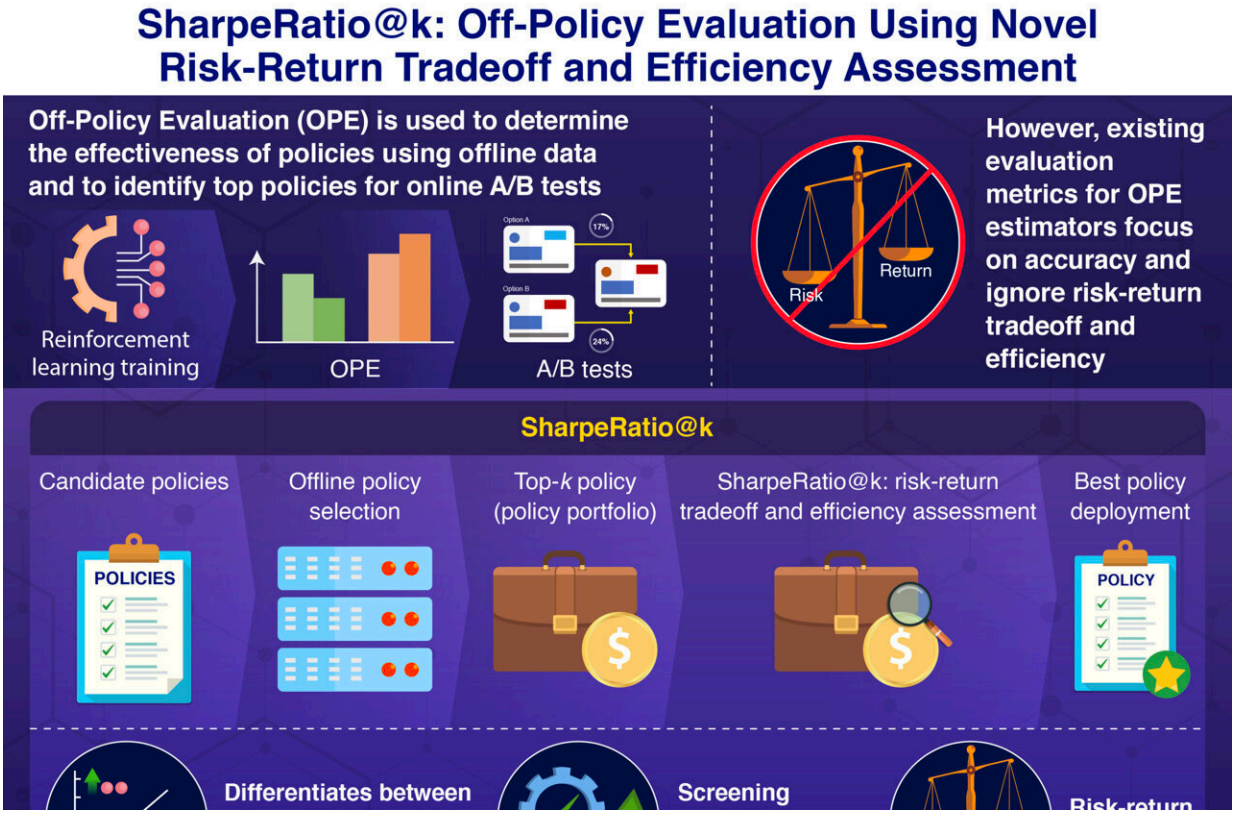
Research team develops novel metric for evaluation of risk-return tradeoff in off-policy evaluation

April 24 2024



SharpeRatio@k, a novel evaluation metric for Off-Policy Evaluation estimators, effectively measures the risk-return tradeoff of evaluating policies used in reinforcement learning and contextual bandits, which are typically ignored by conventional metrics, show scientists at Tokyo Tech. This novel metric, inspired from risk assessment in financial portfolio management, provides a more insightful evaluation of OPE, paving the way for improved policy selection.

Credit: Tokyo Institute of Technology

Reinforcement learning (RL) is a machine learning technique that trains software by mimicking the trial-and-error learning process of humans. It has demonstrated considerable success in many areas that involve sequential decision-making. However, training RL models with real-world online tests is often undesirable as it can be risky, time-consuming, and importantly, unethical. Thus, using offline datasets that are naturally collected through past operations is becoming increasingly popular for training and evaluating RL and bandit policies.

In particular, in practical applications, the Off-Policy Evaluation (OPE) method is used to first filter the most promising candidate policies, called "top-k policies," from an offline logged dataset, and then use more reliable real-world tests, called online A/B tests, to choose the final policy.

To evaluate the effectiveness of different OPE estimators, researchers have primarily focused on metrics such as the mean-squared error (MSE), RankCorr and Regret. However, these methods solely focus on the accuracy of OPE methods while failing to evaluate the risk-return tradeoff during online policy deployment.

Specifically, MSE and RankCorr fail to differentiate whether near-optimal policies are underestimated or poor-performing policies are overestimated, while Regret focuses only on the best policy and overlooks the possibility of harming the system due to sub-optimal policies in online A/B tests.

Addressing this issue, a team of researchers from Japan, led by Professor Kazuhide Nakata from Tokyo Institute of Technology, has developed a

new evaluation metric for OPE estimators.

"Risk-return measurement is crucial in ensuring safety in risk-sensitive scenarios such as finance. Inspired by the design principle of the financial risk assessment metric, Sharpe ratio, we developed SharpeRatio@k, which measures both potential risk and return in top-k policy selection," explains Prof. Nakata. The study was presented at the Proceedings of the ICLR 2024 Conference.

SharpeRatio@k treats the top-k policies selected by an OPE estimator as a policy portfolio, similar to financial portfolios, and measures the risk, return and efficiency of the estimator based on the statistics of the portfolio. In this method, a policy portfolio is considered efficient when it contains policies that greatly improve performance (high return) without including poorly performing policies that negatively affect learning in online A/B tests (low risk). This method maximizes return and minimizes risk, thereby identifying the safest and most efficient estimator.

The researchers demonstrated the capabilities of this novel metric through example scenarios and benchmark tests and compared it with existing metrics.

Testing revealed that SharpeRatio@k effectively measures the risk, return and overall efficiency of different estimators under varying online evaluation budgets, while existing metrics fail to do so. Additionally, it also addresses the overestimation and underestimation of policies. Interestingly, they also found that while in some scenarios it aligns with existing metrics, a better value of these metrics does not always result in a better SharpeRatio@k value.

Through these benchmarks, the researchers also suggested several future research directions for OPE estimators, including the need to use

SharpeRatio@k for efficiency assessment of OPE estimators and the need for new estimators and estimator selection methods that account for risk-return tradeoffs. Furthermore, they also implemented their innovative metric in an open-source software for a quick, accurate and insightful evaluation of OPE.

Highlighting the importance of the study, Prof. Nakata concludes, "Our study shows that SharpreRatio@k can identify the appropriate estimator to use in terms of its efficiency under different behavior policies, providing useful insight for a more appropriate estimator evaluation and selection in both research and practice."

Overall, this study enhances policy selection through OPE, paving the way for improved reinforcement learning.

More information: Haruka Kiyohara et al, Towards Assessing and Benchmarking Risk-Return Tradeoff of Off-Policy Evaluation (2024)

Provided by Tokyo Institute of Technology

Citation: Research team develops novel metric for evaluation of risk-return tradeoff in off-policy evaluation (2024, April 24) retrieved 6 May 2024 from https://techxplore.com/news/2024-04-team-metric-tradeoff-policy.html