Efficient Offline Reinforcement Learning Through Dataset Characterization and Reduction

Doctoral Consortium

Enrique Mateos-Melero
Universidad Carlos III de Madrid
Leganés, Spain
enmateos@pa.uc3m.es

ABSTRACT

Offline Reinforcement Learning (RL) relies heavily on the quality of datasets to derive effective policies. The dataset characteristics (such as Trajectory Quality or State-Action Coverage) impact the performance of the learned policy. Typically, these problems are solved by enhancing the algorithms used in the process (such as regularization methods and others). Despite recognizing these characteristics as crucial for determining the quality of a learned policy, what constitutes a "good" dataset remains ambiguous. This thesis explores methodologies for predicting the learning performance of offline RL datasets and optimizing their composition to improve policy outcomes. By representing datasets as images and using Convolutional Neural Networks (CNNs), we predict policy performance and enable efficient dataset reduction using genetic algorithms. Preliminary experiments demonstrate the potential for dataset size reduction while maintaining or enhancing policy quality.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Markov decision processes**; *Neural networks*; Genetic algorithms.

KEYWORDS

Offline Reinforcement Learning; Dataset Quality; Efficient Learning; Performance Predictor

ACM Reference Format:
Enrique Mateos-Melero. 2025. Efficient Offline Reinforcement Learning Through Dataset Characterization and Reduction: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

There has been a significant rise in applying Reinforcement Learning (RL) [15] to domains requiring pre-collected data, such as healthcare, autonomous systems, and industrial processes. Offline RL [10], in particular, has gained attention for its ability to learn effective policies without needing online interactions, thereby avoiding risks associated with exploration in unsafe or costly environments. However, the performance of offline RL is heavily influenced by the quality and composition of the dataset, making the task of evaluating and optimizing datasets crucial for success [13, 14].

Current research often focuses on improving offline RL algorithms but overlooks the dataset's role in determining policy quality. Metrics like Trajectory Quality (TQ) and State-Action Coverage (SACo) [14] provide some insights into dataset suitability but fail to account for some aspects such as state distributions or suboptimal trajectories. These challenges motivate the need for a rapid, scalable method to evaluate dataset quality and optimize its composition.

This thesis introduces a novel approach that represents offline RL datasets as images, leveraging Convolutional Neural Networks (CNNs) [9] to predict policy performance. Additionally, a genetic algorithm [6] framework is proposed to reduce dataset size while improving policy performance. Preliminary experiments demonstrate the feasibility of this approach across a range of RL environments.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 BACKGROUND

Offline RL problems are typically solved by managing the distributional shift [4] from the algorithm perspective. The first algorithm proposed to solve this issue was Batch Constrained Q-Learning (BCQ) [5]. This algorithm suggested restricting the learned policy close to the behavior of the dataset. Some other algorithms implemented slight differences such as BEAR [7] and PLAS [20]. Another approach is to add a regularization term in the training process by learning a conservative function as in Conservative Q-Learning (CQL) [8]. Lastly, algorithms used the estimation of the epistemic uncertainty to vary the restrictions applied to the learning process like Random Ensemble Mixture (REM) [2].

Various methods estimate dataset performance, such as ERI [16], TQ, and SACo [13, 14], which assess data quality. However, these metrics do not accurately predict performance after training, as they only average expected returns from the dataset's episodes and ignore mixtures of policies. Algorithms like LBRAC-v [19] address this by assuming datasets come from different behavior policies and using latent variable models to group trajectories accordingly. While this mitigates the degeneration issue from BRAC-v [17], it still does not estimate overall dataset performance. DVORL [1] improves this by using data valuation to estimate quality for task-specific selection, but it requires a target dataset for the KL divergence metric, making it impractical without one.

3 METHODOLOGY AND RESULTS

In offline RL, dataset quality is estimated by evaluating the performance of a learned policy, denoted as π, in an RL framework with parameters θ. The expected return of a policy can be expressed as:

$$\hat{\mathcal{J}}_{RL}(\mathcal{D} \mid \theta) = \frac{1}{|E|} \sum_{e \in E} \sum_{t=0}^{t_e} \gamma^t r_t^e$$

where E is a set of episodes, t_e is the final timestep of episode e, and r_t^e is the reward at time t. However, this method is time-consuming, prompting the need for a predictive model to estimate performance quickly. The **Performance Prediction Problem (P3)** is defined as $\hat{\mathcal{J}}_M(\mathcal{D} \mid \theta) \approx \mathcal{J}(\pi_{\mathcal{D}} \mid \theta)$ where $\hat{\mathcal{J}}_M$ is a function that approximates the return of the best policy $\pi_{\mathcal{D}}$ learned from $\mathcal{D}$ (the dataset). Since such a function cannot be computed directly, we propose using a neural network. Offline RL datasets are usually composed of numerical values representing states. However, representing each episode with tuples of state values is impractical. Instead, we decide to represent datasets as images, capturing spatial information beneficial for deep neural networks, especially convolutional neural networks (CNNs).

To represent datasets as images, the state dimensionality and environment dynamics must be considered. Different approaches are proposed and evaluated across three domains: Frozen Lake (discrete, bi-dimensional, delayed reward), Mountain Car (continuous, bi-dimensional, delayed reward), and Acrobot (continuous, 6-dimensional, instant reward). The proposed representations are:

- *State Scatter Plots:* States are visualized as scatter plots, with dimensionality reduction techniques like PCA [12] or autoencoders [3] used for high-dimensional spaces. The scatter plots show state distributions or the reduced state spaces.
- *Rendered Images:* State frames are gathered, converted to grayscale, and thresholded to isolate objects from the background. Then the pixel values are summed and normalized to highlight environmental features. This representation captures motion.

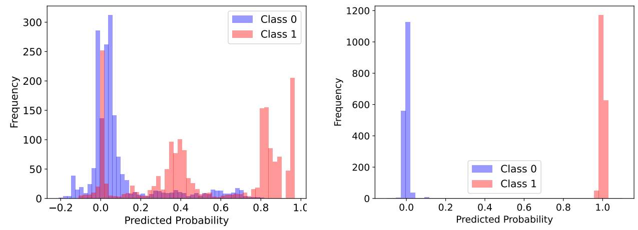
The input to the CNN model is a transformed image of dataset $\mathcal{D}$, generated using either state abstraction or rendered images. The goal is to predict the dataset's expected return. Given the characteristics of each domain, the expected return is considered differently. In Frozen Lake and Mountain Car, we consider the confidence that the model has for a given dataset to be "good". In the case of Acrobot, we predict the actual value of the expected return. Figure 1 shows the results for Frozen Lake which demonstrates that the CNN models can predict the expected return of the datasets with the image representations.

Another problem to tackle is the sub-optimality of the data and the mixture of policies. Several approaches aim to reduce datasets by keeping only useful data for training. Discriminator-Weighted Offline Imitation Learning [18] integrates a discriminator with BC to weight data based on expertise. COIL [11] filters datasets using KL divergence, assuming independent policies for each trajectory.

We frame the problem as a multi-objective optimization task, aiming to minimize dataset size while maximizing policy quality. We do a simplification by combining both objectives into a single function $\psi(\mathcal{D}) = f(\hat{\mathcal{J}}(\mathcal{D} \mid \theta), |\mathcal{D}|)$ where $\hat{\mathcal{J}}(\mathcal{D} \mid \theta)$ is an estimate of the return (the CNN model in our case). The Episode Selection Problem (ESP) seeks to find a reduced subset $\mathcal{D}^* \subseteq \mathcal{D}$ that maximizes ψ, that is $\mathcal{D}^* = arg\,max_{\mathcal{D}' \subseteq \mathcal{D}} \psi(\mathcal{D}')$

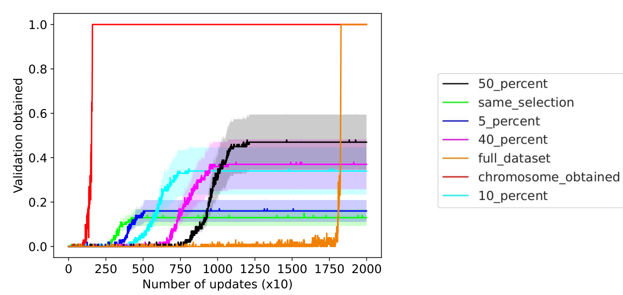
While a brute-force approach could theoretically solve this by evaluating all possible subsets, it is computationally impractical due to the exponential growth of subsets. Instead, we propose using a genetic algorithm (GA) for a suboptimal solution. Each individual in the population represents a subset of episodes from the input dataset. A chromosome is encoded as a binary string, where each bit represents whether an episode is included in the subset. The fitness function is based on the ESP metric ψ.

Formally, if $\mathcal{D} = \{e_1, e_2, \ldots, e_M\}$ is the dataset, with e_i being the i-th episode, each individual $x \in X$ is a binary vector $x = x_1, x_2, \ldots, x_M$, where $x_i \in \{0, 1\}$. The corresponding reduced dataset is mapped by $\rho(x) = \{e_i \in \mathcal{D} \mid x_i = 1\}$. The fitness of an individual x is computed as $\phi(x) = \psi(\rho(x))$.



(a) Frozen Lake Rendered **(b) Frozen Lake Abstraction**

Figure 1: Results from the CNN models using rendered (left column) and state (right column) representations in Frozen Lake. The plot shows the distribution of confidence provided by the model for "good" (red) and "bad" (blue) datasets.



(a) Frozen Lake

Figure 2: Comparison of performance of the genetic algorithm against random and full datasets for Frozen Lake.

Figure 2 shows the results of reducing the dataset to keep only the useful data. We can keep the same policy performance while reducing the number of updates during the training process. Additionally, we compare the results with random selections to prove that the GA obtains better and more consistent results.

4 FUTURE WORK

We intend to test our method in more complex environments with continuous action spaces. Adapting it to other representation forms is also a potential path. Furthermore, exploring new RL applications may lead to developing innovative dataset-reduction techniques and testing their capabilities in other fields like transfer learning.

REFERENCES

[1] Amir Abolfazli, Gregory Palmer, and Daniel Kudenko. 2022. Data Valuation for Offline Reinforcement Learning. *arXiv* 2205.09550 (2022), 9.

[2] Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. 2020. An optimistic perspective on offline reinforcement learning. In *International conference on machine learning*. PMLR, Virtual Conference, 104–114.

[3] Dor Bank, Noam Koenigstein, and Raja Giryes. 2023. *Autoencoders*. Springer International Publishing, 353–374.

[4] Rafael Figueiredo Prudencio, Marcos R. O. A. Maximo, and Esther Luna Colombini. 2024. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE Transactions on Neural Networks and Learning Systems* 35, 8 (2024), 10237–10257.

[5] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*. PMLR, Long Beach, CA, USA, 2052–2062.

[6] John H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. MIT Press, Cambridge, MA, USA.

[7] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems* 32 (2019), 11.

[8] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems. Proceedings of the thirty-fourth Annual Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vol. 33. Curran Associates, Virtual Conference, 1179–1191.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521 (2015), 436–444.

[10] Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *ArXiv* 2005.01643 (2020), 43.

[11] Minghuan Liu, Hanye Zhao, Zhengyu Yang, Jian Shen, Weinan Zhang, Li Zhao, and Tie-Yan Liu. 2021. Curriculum Offline Imitating Learning. In *Advances in Neural Information Processing Systems. Proceedings of the thirty-fifth Annual Conference on Neural Information Processing Systems (NeurIPS 2021)*. Curran Associates, Virtual Conference, 1–12.

[12] Karl Pearson. 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2(11) (1901), 559–572.

[13] Kajetan Schweighofer, Markus Hofmarcher, Marius-Constantin Dinu, Philipp Renz, Angela Bitto-Nemling, Vihang Prakash Patil, and Sepp Hochreiter. 2021. Understanding the Effects of Dataset Characteristics on Offline Reinforcement Learning. In *Deep RL Workshop NeurIPS 2021*. Curran Associates, Virtual Conference, 19.

[14] Kajetan Schweighofer, Andreas Radler, Marius-Constantin Dinu, Markus Hofmarcher, Vihang Patil, Angela Bitto-Nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. 2022. A Dataset Perspective on Offline Reinforcement Learning. In *Proceedings of the 1st Conference on Lifelong Learning Agents*, Vol. 199. PMLR, Montreal, Canada, 470–517.

[15] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press, Cambridge, MA, USA.

[16] Phillip Swazinna, Steffen Udluft, and Thomas Runkler. 2021. Measuring Data Quality for Dataset Selection in Offline Reinforcement Learning. In *Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Orlando, FL, USA, 1–8.

[17] Yueh-Hua Wu, Nontawat Charoenphakdee, Han Bao, Voot Tangkaratt, and Masashi Sugiyama. 2019. Imitation Learning from Imperfect Demonstration. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. PMLR, Long Beach, CA, USA, 6818–6827.

[18] Haoran Xu, Xianyuan Zhan, Honglei Yin, and Huiling Qin. 2022. Discriminator-Weighted Offline Imitation Learning from Suboptimal Demonstrations. In *Proceedings of the 39th International Conference on Machine Learning*, Vol. 162. PMLR, Baltimore, Maryland, USA, 24725–24742.

[19] Guoxi Zhang and Hisashi Kashima. 2023. Behavior estimation from multi-source data for offline reinforcement learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. AAAI Press, Washington, DC, USA, 11201–11209.

[20] Wenxuan Zhou, Sujay Bajracharya, and David Held. 2021. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*. PMLR, Virtual Conference, 1719–1735.