Deep Recurrent Q-Network Approach for Multi Objective Markov Decision Process in Partially Observable Environment

Sharef, Nurfadhlina
UNIVERSITI PUTRA MALAYSIA
UNIVERSITI PUTRA MALAYSIA
SERDANG, SELANGOR, 43400
MYS

08/23/2021
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
23-08-2021	Final	29 Sep 2018 - 28 Mar 2021

4. TITLE AND SUBTITLE
Deep Recurrent Q-Network Approach for Multi Objective Markov Decision Process in Partially Observable Environment

5a. CONTRACT NUMBER
FA2386-18-1-4079

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Nurfadhlina Sharef

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITI PUTRA MALAYSIA
UNIVERSITI PUTRA MALAYSIA
SERDANG, SELANGOR 43400
MYS

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

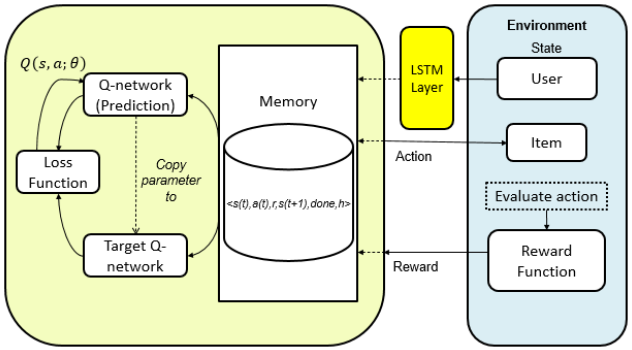
Prediction of relevant items to the users' interest in a recommendation system (RS), is an example of partially observable Markov Decision Process (POMDPs) as user's interests fluctuate over time and the item's satisfaction rating matrix is typically sparse. This problem also requires multi-objectives optimization (MOO) for multi-objectives which are precision, novelty and diversity.
Existing solutions on MOO are based on evolutionary algorithms, which requires combination with rating prediction techniques such as collaborative filtering to fill up the sparse matrix prior to producing recommendation. However, collaborative filtering has limitations when handling cold start or new users.
Most RS merely focus on accuracy of high-rating or trendy items predictions. However, other metrics such as novelty and diversity which are equally essential to generate more quality recommendation have mostly been ignored. The main challenge of considering multiple evaluation metrics is the conflict between the objectives, since to improve either one metrics will hurt the accuracy and vice versa.
Two DRL agents called Deep Q Network for multi-objective recommendation system (DQNMORS) and Deep Q Network with recurrent layer for multi-objective recommendation system (RDQNMORS) are developed to solve the above problems. The conducted evaluation is based on the effect of the features for recommendation (e.g., movieID, movie rating and user information) and optimization parameters (ie.g., weighted average, pareto). The performance of the DRL agents are compared against the benchmarking approaches based on probability multi objective evolutionary algorithms (PMOEA) using a movie recommendation environment.
Results have shown that the DRL approaches, which are the first available DRL approach for MOO in movie recommendation, are better in multi-objective compared to the benchmark. The recurrent layer in the DRL agent is also able to remodel the POMDP as a complete MDP environment, which allows prediction of the sparse rating matrix.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			KRISTOPHER AHLERS
			SAR	0	**19b. TELEPHONE NUMBER** *(Include area code)* 315-227-7000

AFOSR Deliverable Report FA2386-18-1-4079

Title	Deep Recurrent Q-Network Approach for Multi Objective Markov Decision Process in Partially Observable Environment
Duration	September 2018-Jun 2021
Principal Investigator	Nurfadhlina Mohd Sharef, PhD Faculty of Computer Science and Information Technology, Universiti Putra Malaysia nurfadhlina@upm.edu.my
Co-Researchers	Razali Yaakob, PhD Norwati Mustapha, PhD Khairul Azhar Kasmiran, PhD Maslina Zolkepli, PhD Erzam Marlisah, PhD
Output	1. Multi-Objective Deep Reinforcement Learning for Recommendation System, IEEE Access, in press 2. VLearnAI application
Executive Summary	Prediction of relevant items to the users' interest in a recommendation system (RS), is an example of partially observable Markov Decision Process (POMDPs) as user's interests fluctuate over time and the item's satisfaction rating matrix is typically sparse. This problem also requires multi-objectives optimization (MOO) for multi-objectives which are precision, novelty and diversity. Existing solutions on MOO are based on evolutionary algorithms, which requires combination with rating prediction techniques such as collaborative filtering to fill up the sparse matrix prior to producing recommendation. However, collaborative filtering has limitations when handling cold start or new users. Most RS merely focus on accuracy of high-rating or trendy items predictions. However, other metrics such as novelty and diversity which are equally essential to generate more quality recommendation have mostly been ignored. The main challenge of considering multiple evaluation metrics is the conflict between the objectives, since to improve either one metrics will hurt the accuracy and vice versa. Two DRL agents called Deep Q Network for multi-objective recommendation system (DQNMORS) and Deep Q Network with recurrent layer for multi-objective recommendation system (RDQNMORS) are developed to solve the above problems. The conducted evaluation is based on the effect of the features for recommendation (e.g., movieID, movie rating and user information) and optimization parameters (ie.g., weighted average, pareto). The performance of the DRL agents are compared against the benchmarking approaches based on probability multi objective evolutionary algorithms (PMOEA) using a movie recommendation environment. Results have shown that the DRL approaches, which are the first available DRL approach for MOO in movie recommendation, are better in multi-objective compared to the benchmark. The recurrent layer in the DRL agent is also able to remodel the POMDP as a complete MDP environment, which allows prediction of the sparse rating matrix.

Objectives	Methodology	Achievement
To determine the characteristics of partial multi-objective action prediction in partially observable Markov Decision Process (POMDPs). To model the multi-objective decision-making problem in a partially observable environment. To develop a multi-objective deep learning method for partially observable Markov decision processes.	Deep Recurrent Q-Network (DRQN) is a reinforcement learning approach that supports single and multi-objective optimization problems. The difference between this approach with its traditional form (Q-Network) in solving optimization through reinforcement is that instead of calculating the Q-table, Q-value function is being estimated. The challenge in multi-objective optimization compared to the single objective is to achieve Pareto optimization especially when constraints and contrasting objectives exist. Two DRL agent models called Deep Q Network for multi-objective recommendation system (DQNMORS) and Deep Q Network with recurrent layer for multi-objective recommendation system (RDQNMORS) are to address three objectives namely precision, novelty and diversity. Precision, *Pr*, is an essential evaluation metric that relates to measurement on how precise the prediction results. It is indicating the proportion of recommended items from the total user's preferable item list which has a high rating value. $$P_r = \frac{L_u \cap T_u}{L} \quad (1)$$ where *Lu* is the predicted recommendation list that contains items for user *u*, *Lu = [x1, x2, …, xn]*. *Tu* is the actual item in the test set that user, u rated. The high rating item denotes the item which obtained rating 3 or above from that user. *L* is the length of the recommendation list. Novelty, *N* denotes the popularity of the recommended items. It is a measure of the ability to recommend unpopular items to the user and it can be expressed in Eq.2 $$N = \frac{1}{M \cdot L} \sum_{u=1}^{M} \sum_{\alpha \in L_u} \log_2 \left(\frac{M}{N_\alpha}\right) \quad (2)$$ where *M* is the number of total users, *Nα* is the number for the rating of item *α*. The diversity, *DLu* function is the measure for difference between items in the recommendation list. The difference can be described by various topics of items in the recommendation, and the	The proposed models (Figure 1) solve the problem on POMDPs by capturing the sequence of the state (viewed and rated movies) and actions (movie rating) in the MDPs. The Deep Q-Network model could represent the evolving users interests while the LSTM layer captures the sequence of the rating. The proposed approaches are the first deep reinforcement learning model built for multi-objective optimization, and the first solution of its kind for multi-objective RS.  Figure 1: Deep Q-Network with recurrent layer for multi-objective recommendation system An application for video RS based on the DQNMORS (addressing precision and novelty) is deployed, called VLearnAI (Figure 2).

diversity metrics used in this work consisted of three components, which are topic distribution, number of different topics and the distribution of a topic for each item in the recommendation list.

$$D_{L_u} = -\left(\sum_{i \in L_u} \frac{|t_{x_i}|}{|z_{L_u}|} \cdot \log \frac{|t_{x_i}|}{|z_{L_u}|}\right) \cdot Div(L_u) \quad (3)$$

where *Div(Lu)* is the numbers of topics and its distribution in recommendation list *Lu*, the *|t(xi)|* is the amount of topics included in item *xi* and *|z(Lu)|* is the total number of topics in the recommendation list.

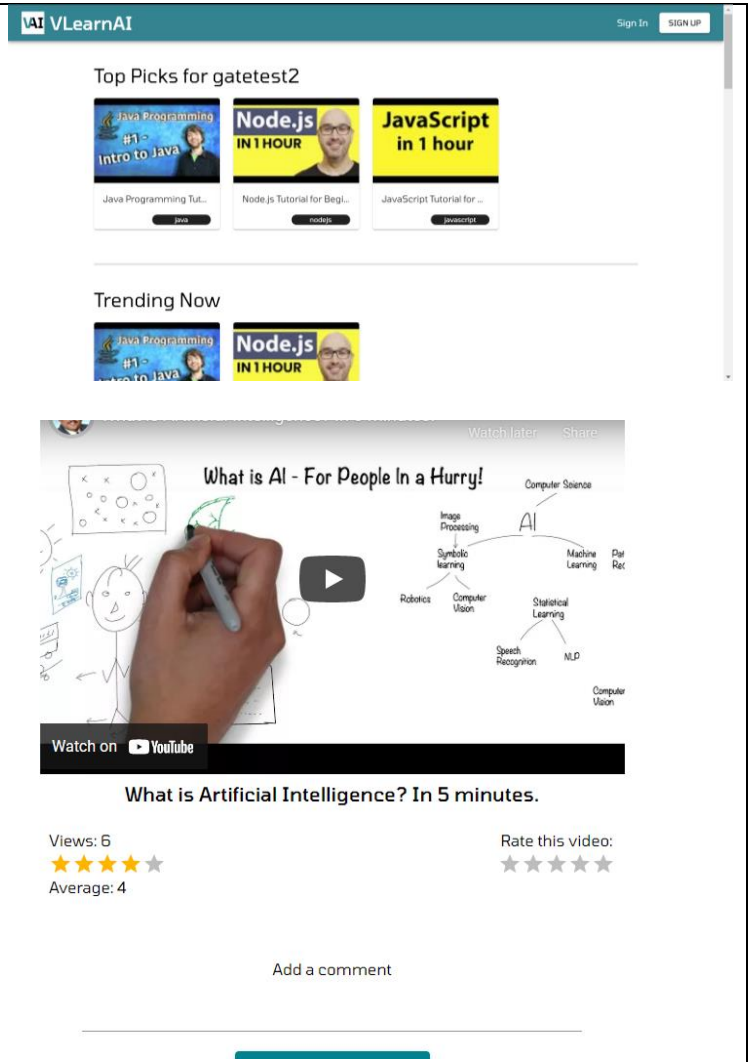


Figure 2: Interfaces from VLearnAI

To improve the performance of multi-objective deep learning methods in the complete Markov decision process.	The Movielens dataset is used for evaluating the performance of the proposed models against the benchmark models called PMOEA [1] which is based on the evolutionary algorithm. Three experiments are devised as follow where parameters setting is in TABLE 1 and 2: a) Experiment 1: to evaluate the performance of pareto optimization method against scalarization MORS involving the features in the models: i. DQNMORS_ps: pareto optimization method in the DQNMORS ii. DQNMORS_*ws*: scalarization method based on simple weighted sum in the DQNMORS b) Experiment 2: to evaluate the effect between the combination of user latent information and movie rating features, against movie rating feature only involving the features in the models: i. DQNMORS_u: combination of user latent information (i.e., age, gender, occupation and zip code) with movie ID as the features in the DQNMORS ii. DQNMORS_m: only use movie ID as the feature in the DQNMORS c) Experiment 3: to evaluate the effect of learning the sequence of the movie rating involving the features in the models: i. RecDQNMORS_ps_m: application of recurrent neural network layer to learn the movie rating sequence. ii. RecDQNMORS_ps_u: application of recurrent neural network layer to learn the user latent.	The experiment settings are as shown in TABLE 3 and the results as shown in TABLE 4 where all models could perform multi-objective optimisation. The analysis of results is as follows: a) Precision In terms of precision, DRL agents achieved lower compared to PMOEA (improvements of PMOEA with a user-based collaborative filtering approach is 59% against the best performing DRL agent, DQNMORS_ps_m_u). b) Novelty In terms of the novelty, both DRL agents except DQNMORS_m are better than all PMOEA (RDQNMORS_ps_u has the best performance and achieved an improvement of 15% against best performing PMOEA). This indicates that the recurrent layer has managed to capture the movie rating sequence, which is the POMDP scenario. Contrarily, without a user latent and recurrent layer, the DQNMORS_m is in a disfavor position overall which indicates that merely utilizing the movieID feature could not support understanding the MDP comprehensively. c) Diversity However, it still demonstrates the ability to optimize multiple objectives. The DQNMORS_ws_m_u shows the best performance in diversity (95% better than the best performing PMOEA model). Meanwhile, RecDQNMORS agent is worse than DQNMORS agent but is still better than PMOEA (68.5% better than the best performing PMOEA model). In conclusion, the DRL agents surpassed all PMOEA approaches in novelty and diversity. The exploration-exploitation nature of DRL agents induce higher potential to explore more variety items, thus,

TABLE I
PARAMETER SETTINGS TESTING RANGE IN TRIAL-AND-ERROR EXPERIMENTS

Parameter	*Range Values*
Learning Rate	$1e^{-2}$ to $1e^{-4}$
Discount Factor	0.1 to 0.9
Min. Epsilon	0.1 to 0.5
Epsilon Decay Rate	0.1 to 1.0
Epoch Number	1 to 60

TABLE 2
OPTIMUM PARAMETER SETTINGS FOR PROPOSED DEEP REINFORCEMENT LEARNING AGENTS

Parameter	DQNMORS_ws_m_u	DQNMORS_ps_m_u	DQNMORS_ps_m	RecDQNMORS_ps_m
Optimization Algorithm	Weighted Sum Method	Pareto Optimal Search	Pareto Optimal Search	Pareto Optimal Search
Learning Rate	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$	$1e^{-3}$
Discount Factor	0.1	0.1	0.1	0.1
Epsilon	3.0	3.0	3.0	3.0
Min. Epsilon	0.5	0.5	0.5	0.5
Epsilon Decay Rate	0.95	0.95	0.95	0.95
Finest Epoch Number	50	10	10	10

increasing the novelty as well as diversity. As a trade-off, precision is affected.

TABLE 3: Experiment settings for evaluation of each model

Exp. 1	Exp. 2	Exp. 3	Proposed methods	Optimization methods		Sequential input handling	Features		
				Pareto method	Scalarize Method	recurrent	movieID, m	user latent, u	movie rating sequence (for recurrent layer)
Yes	No	No	DQNMORS_ws_m_u	No	Yes	No	Yes	Yes	No
Yes	Yes	Yes	DQNMORS_ps_m_u	Yes	No	No	Yes	Yes	No
Yes	Yes	Yes	DQNMORS_ps_m	Yes	No	No	Yes	No	No
No	No	Yes	RecDQNMORS_ps_m	Yes	No	Yes	Yes	No	Yes

TABLE 4: Results of achievement by each model comparing to the benchmark

	PMOEA + ProbS [1]	PMOEA + CF_User [1]	PMOEA + CF_Item [1]	DQNMORS_ws_m_u	DQNMORS_ps_m_u	DQNMORS_ps_m	RecDQNMORS_ps_m
Precision	0.418	**0.499**	0.368	0.184	0.204	0.173	0.126
Novelty	1.956	2.307	3.104	3.131	3.164	2.517	**3.557**
Diversity	2.925	2.832	2.426	**5.695**	5.135	5.186	4.929

References:

[1] L. Cui, P. Ou, X. Fu, Z. Wen, and N. Lu, “A novel multi-objective evolutionary algorithm for recommendation systems,” *J. Parallel Distrib. Comput.*, 2016, doi: 10.1016/j.jpdc.2016.10.014.