Journal of Applied Data Sciences
Vol. 3, No. 3, September 2022, pp. 137-142

ISSN 2723-6471

Policy Optimization Recommendation Algorithm Based on Mapping Network for Behavior Enhancement

Linlin Shan [1,*], Guisong Jiang [2], Shuang Li [2], Shuai Zhao [2], Kunjie Luo [2], Long Zhang [2], Yi Li [3]

[1] School of Fine Arts and Design, Tianjin Normal University, Tianjin, China
[2] School of Computer and Information Engineering, Tianjin Normal University, Tianjin, China
[3] School of Public Administration, Shanxi University of Finance and Economics, Shanxi, China
[1] shanlinlin@tjnu.edu.cnr*
* corresponding author

(Received: April 26, 2022; Revised: June 21, 2022; Accepted: August 21, 2022; Available online: September 30, 2022)

Abstract

The algorithm of policy optimization with learning behavior enhancement based on mapping network technology was proposed, aiming to address the issues of lack and sparsity of learning behavior data and weak generalization ability of the model in AI education. Based on the basic recommendation algorithm and the framework of rein- forcement learning, and model introduces the correlation mapping network to realize the transformation of strong and weak correlation, so as to optimize the input agent policy to improve the performance model of course recommendation. Experiment on MOOC da- tasets show that the proposed algorithm model has a stable improvement compared with the baseline models, and can effectively improve the accuracy of course recommendation.

Keywords: Strong/Weak Correlation, Mapping Network, Policy Optimization, Reinforcement Learning, Course Recommendation

1. Introduction

The rapid development of online education has intensified the overloading of online education resources. So many scholars study course recommendation algorithms to solve this problem. For example deep learn- ing algorithms aims to use machine learning to mine and analyze the learner's historical learning behavior in the limited data [1],so as to predict the learner's learning behavior in the future [2]. However, the model's expression and generalization ability are limited [3] because of data scarcity. But Reinforcement Learning can effective mitigation for the issue. The paper mainly studies course recommendation task of policy optimization based on mapping network under the reinforcement learning (POR_MN).

2. Related Work

In recent years, scholars have found that course recommendation is different from general recommenda- tion algorithms. Due to the lack of relevant data and sparse information, course recommendation model is difficult to be implemented. In the general sense, the basic way to directly use the learner's history learn- ing behaviors to get the similarity. For example, Koren [5] adopted Matrix factorization and Rendle [6] used Bayesian Personalized Rank are all items-based collaborative filtering algorithms, or used the simi- larity between items(FISM) [7]. However, the original design of such a model does not take into account the unique feature of sparse course data, which leads to the mediocre performance of course recommenda- tion and is not universal. Then, He [8] proposed a recommendation model based on deep neural network (NCF), which can alleviate some sparsity problems when applied to course recommendation. Meanwhile, Xia [9] considered the learning cycle and the sequence, proposed a course awareness model, based on course content and considered the prior relationship. But it requires more non-learners' subjective and global external auxiliary data. It is difficult to obtain this data, so the generalization of models is weakly.

In order to improve generalization and make up for the lack of historical behavior data of learners, more scholars use series of dynamic learning behavior data to construct the model. Some scholars proposed Neural Attention Item

Similarity(NAIS)[10],it can calculate the attention coefficient of historical behaviors, to distinguish the different importance of the same learning behavior for different learners. However, just distinguishing different importance does not solve the problem that the low coefficient and weak correlation behaviors will dilute the expression of learners' interest in learning. So Zhang [11] proposed the HRL model, it aims to deal with the noise course, build a subset of learners' historical behavior, and use the reinforcement learning Agent policy for training. Although such methods of optimizing rein- forcement learning Agent policy [12] have some effect, they don't make full use of learners' datas and pay no attention to the strong or weak correlation of their behaviors. Even the attention mechanism [13] does not consider how to handle those behaviors. So the POR_MN was proposed to solve those problems.

3. Model Construction

In the prior representation, the courses are expressed as $C = \{c1,c1,\ldots,cM\}$ the learners are expressed as $S = \{s1,s1,\ldots,sN\}$, the learning behavior matrix of learners and courses is expressed as $\boldsymbol{Y} = \boldsymbol{R}$, If $M * N$ learner has studied the course, it's marked y=1, otherwise 0. Meanwhile $C_1^S, C_2^S \ldots\ldots C_t^S$ as the sequence of s 1,c s 2,…,c s t historical behavior, which from 1 to t moment. The goal is to recommend courses at t+1 moment. Our model is mainly composed of two parts: basic recommendation and agent policy optimization based on mapping network technology. The one part is used to calculate different weight values of history learning behavior and realize the basic algorithm of course recommendation at the time of t +1. Another is the policy input of the optimization based on mapping network. The POR_MN is shown in Figure 1.

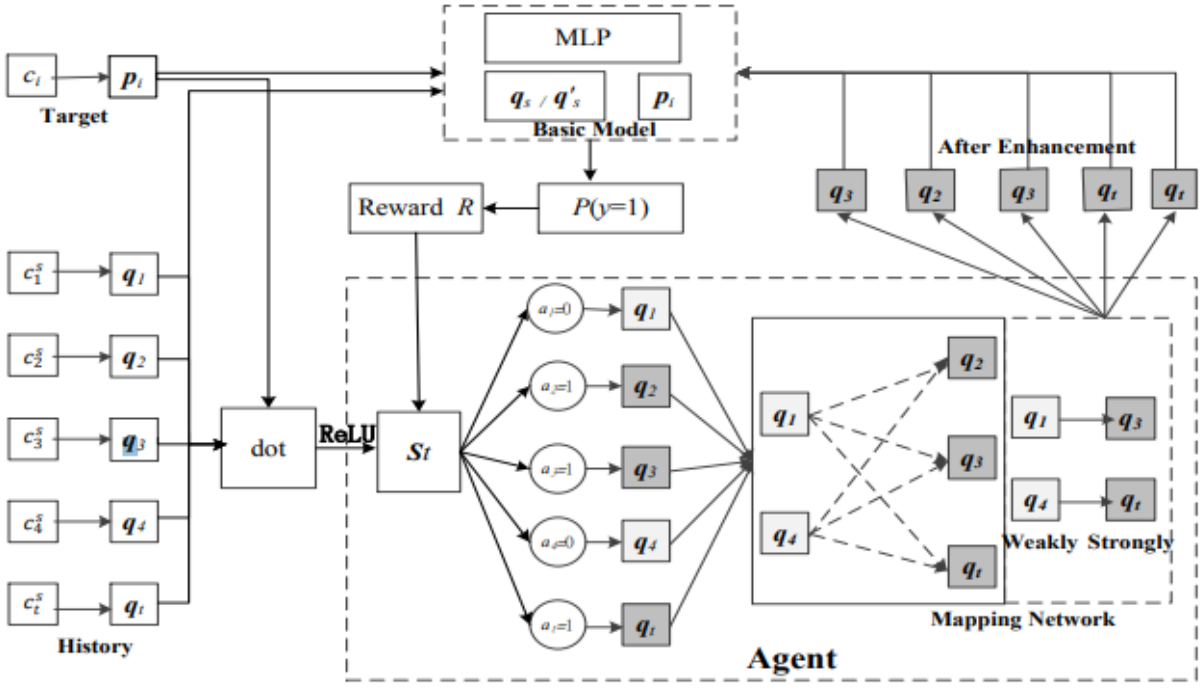

Figure 1. Architecture of POR_MN

3.1. Basic Recommendation

The basic model is based on NAIS[10]. In this part, each course is represented as a vector by real valued low-dimensional embedding, the unregistered course vector P_i as the target. History embedding se- quence $Qs_t = (qs_1, qs_2, \ldots qs_t)$ is obtained in the same way. After input, the first is to calculate the attention coefficient a_{it}^s for each

history behavior by MPL (q_t^s, P_i)The second use the $\sum_{t}^{ts} = 1 a_{it}^{s} q_t$ to calculate the learning interest q_s. Finally uses the $\sigma(q_z^T p_i)$ to calculate the recommendation probability.

3.2. Agent Policy Optimization based on Mapping Network Technology

To further solve the aforementioned problems, the learner's history learning behaviors should be fully utilized and effectively processed. The basic model is embedded into the reinforcement learning frame- work for optimizing agent policy based on mapping network.

In the framework of reinforcement learning, the state S_i defined by it, which calculated by the embed- ding vector of the learner's history learning behavior $qs_1, qs_2, \ldots qs_t$ and target course P_i, then its output characteristic through ReLU (W_1, P_i + b), where W_1 and b are the hyperparameters of the hidden layer.

The action a is defined in the range of [0,1], after calculating the action probability based on the state S_i in the Equation (1). When a=0, it means that the next mapping operation will be performed, St and the behavior is judged to be weakly related to the learner's preference. when a=1, the model determines that it is a strongly correlated behavior, and the mapping operation will not be performed and it will be retained.

$$\pi(\boldsymbol{S}_t, a) = a\,\sigma(\boldsymbol{w}_2 \boldsymbol{S}_t) + (1 - a)(1 - \sigma(\boldsymbol{w}_2 \boldsymbol{S}_t)) \quad (1)$$

Where W_2 is the hyperparameter in the policy function. σ as the activation function to convert the input into a probability. Then the obtained probability value of the action is compared with a given probability threshold γ. If the probability value of the action generated is greater than γ, the action in this state is initially set as a=1, indicating that the corresponding behavior is judged to be strongly correlated and no mapping transformation is required. On the contrary, the action in this state is set as a=0, which means that the weakly correlated behavior is judged and the forward mapping transformation needs to be further performed. Finally, through policy optimization and reward, it is judged whether action a needs to be updated to ensure effective policy adjustment until the optimal plan of the model is adjusted.

To enhance the processing of strong correlation behavior, the mapping network is further used to en- hance the learning behavior data. In this network, the similarity sim_L is calculated between weakly related q_{weak} and strongly related $q_{strongL}$ by using a mapping function. In the mapping network , the strongly related courses with the highest similarity as the mapping results **q** of weakly ' related courses by using 1:1 mapping, according to the result, the enhancement of learning behavior data is realized and an en- hanced learning behavior record $Q_t^{'s}$is constructed. The process is s *t* expressed by the Equation(2)-(4).

$$sim_L = f(\boldsymbol{q}_{weak}^{s}, \boldsymbol{q}_{strongL}^{s}) \quad (2)$$

$$\boldsymbol{q}' = \max\{\boldsymbol{q}_{strongL}^{s} \mid sim_1, sim_2, \ldots, sim_L\} \quad (3)$$

$$\boldsymbol{Q}_t^{'s} = \{\boldsymbol{q}_{strong1}^{s}, \boldsymbol{q}_{strong2}^{s}, \ldots, \boldsymbol{q}_{strongL}^{s}, \ldots, \boldsymbol{q}', \ldots\} \quad (4)$$

Where the f represents using the dot product to calculate the similarity, L is the number of learning behaviors judged as strongly correlated in the network. In the network, there is an identical correlation between strong and weak correlated behaviors through 1:1 mapping. So the results actually use the correlated strong correlation behaviors to express the implicit properties of weak correlation behaviors.

At the same time, the environment sends a reward to the agent, its goal is to maximize the reward. Ac- cording to the signal, the agent policy makes the next choice to illustrate the effect of action a. The reward is the logarithmic

difference between the original learning behavior sequence $Q_t^{'s}$ and the enhanced sequence $Q_t^{'s'}$ of learners, and target course P_i, as shown in Equation (5).

$$R = \log P_{y=1}(\boldsymbol{Q}_t^{'s}, \boldsymbol{p}_i) - \log P_{y=1}(\boldsymbol{Q}_t^{s}, \boldsymbol{p}_i) \tag{5}$$

The above two models construct an augmented learning behavior sequence e $Q_t^{'s}$, then take the s t embed- ding of each of its behaviors as a new input, after calculating the contribution $a_{it}^{'s}$ again using s it MLP, and new representations $q_s^{'}$ of reinforcement learners' learning behavior are still constructed by $\sum_t^{ts} = 1 a_{it}^{'s} q_{it}^{'s}$, after it was input into the model, the final recommendation probability is obtained by using $\sigma\,(q_{it}^{'T} P_i)$

3.3. Model Training

This paper optimizes the Agent policy function by maximizing the expected to get optimal parameter θ. So the gradient descent is used to optimize the training of the model[4,12], as shown in Equation (6).

$$\nabla_\theta = \frac{1}{m}\sum_{m=1}^{M} \nabla_\theta \log \pi_\theta (\boldsymbol{S}_t^m, a_t^m) R \tag{6}$$

In the model training, the learning behavior of learners from 1 to t-1 is taken as the historical learning behavior, and the learning behavior at t is taken as the target behavior to better train the model.

4. Experiment and Analysis

4.1. Datasets and Indicators

The MOOC datasets from Xuetangx.com were selected for verification. The dataset records 458,454 valid course registration behaviors for 1,302 courses generated by 82,532 learners between 2016 and 2018. And the data sparsity of MOOC is as high as 99.57%. Considering the situation and learning cycle, the experiment adopted the time node division method, and divided the data between 2016 and 2018 into training sets, and took the data after 2018 as the testing sets.

In this study, the HR@K and NDCG@K are used, where K are 5 and 10 respectively. To better com- pare the validity and accuracy of the model, FISM[7] of item-based collaborative filtering method, MLP[8] of learning scoring function to learn the user's selection probability of item, NAIS[10] of neural collabora- tive filtering integrating attention mechanism and HRL model[11] of current advanced reinforcement learning framework are selected, the four classical algorithms are compared with the POR_MN model.

4.2. Analysis of experimental results

The setting of parameters is very important, large or small parameters will affect the performance accuracy of the model. In the experiment, we combine previous working experience[11] and practical experience to adjust parameters and obtain results. The experimental results are shown in Table 1.

As shown in Table 1, If the Learning_rate is too high, it may oscillate on both sides of the optimal results, if its too low, the convergence speed and learning ability of the model will be greatly reduced. When Learning_rate=0.001, the model has the best effect. and can see from the table that the best experimental performance was achieved when Weight_size=8. When the dimension of the hidden layer increases gradually, the accuracy and error of the model training will be overfitted due to the complexity of the feature space caused by the over dimension. What's more, when the Batch_size is too small, it will not only increase the training time of the model, but also lead to the instability of the model performance and reduce the generalization ability of the model. The calculation of large gradient Batch_size is more stable, so the model performance is best when Batch_size=256.

To better training for the state and action of the policy function, the experiment also focuses on the comparison of the probability threshold γ of the strength correlation. The results are shown in Table1. when γ=0.25, the performance is the worst, the threshold is too small to precipitation of weak correlation, and the only strong correlation behavior cannot express the forward transition mapping ability of the model. When γ=0.5, the model capability was the best in HR index, which decreased with the increase of the threshold. According to table, γ=0.5 can be selected when the course recommendation task is inclined to recall the hit rate, and γ=0.75 can be selected when the course recommendation task is inclined to recommend the position ranking. In conclusion, this paper considers the average improvement performance of different recommended tasks, and finally chooses γ=0.5 to construct the POR_MN model.

To evaluate the performance of POR_MN, it was compared with the reference data[11] of four baseline models. The experimental results are shown in Table 2.

Table 1. Performance comparison of different parameters(%)

Parameters	Range	HR@5	NDCG@5	HR@10	NDCG@10
Learning Rate	0.01	62.46	45.1	79.89	50.81
	0.001	68.06	49.14	82.11	53.77
	0.0001	64.81	46.74	80.71	51.96
Weight_size	8	68.06	49.14	82.11	53.77
	16	65.41	47.15	81.32	52.38
	32	63.18	45.62	80.13	51.18
Batch_size	64	62.97	45.41	80.16	51.04
	128	64.83	46.69	80.80	51.93
	256	68.06	49.14	82.11	53.77
probability thresholds (γ)	0.25	56.44	43.90	69.01	47.97
	0.5	68.06	49.14	82.11	53.77
	0.65	65.29	50.48	77.98	54.61
	0.75	65.70	52.78	77.28	56.54
	0.95	65.66	52.68	77.30	56.45

Table 2. Performance comparison of different models(%)

Model	HR@5	NDCG@5	HR@10	NDCG@10
FISM	52.73	40.00	65.64	44.98
MLP	52.16	40.39	66.29	44.41
NAIS	56.42	43.73	69.05	47.82
HRL	64.59	45.74	79.68	50.69
POR-MIN	68.06	49.14	82.11	53.77

It can be seen from Table 2, the HR and NDCG of POR_MN model reach 82.11% and 53.77%, respectively. Compared with the FISM and MLP algorithm that recommend directly considering item similarity, POR_MN has the largest improvement on different K values of HR and NDCG indicators. Secondly, in the face of the classic algorithm NAIS model considering the weight of the project, the different Top-K values of HR and NDCG can be improved by more than 11% and 5%. Compared with HRL, the most advanced course recommendation model based on reinforcement learning framework, POR_MN model is improved by 3.47% and 2.43% in HR@5 and HR@10 respectively. It has increased 3.40% in NDCG@5 and 3.08% in NDCG@10. Therefore, we propose the model in the face of learning behavior data. Highly sparse constraints can effectively improve the performance of the recommended curriculum model for learners to recommend more precise course learning resources. At the same time, the generalization ability of the model is improved in the absence of available data fields.

5. Conclusion

For online education, lack of field data and the sparse, POR_MN model is proposed in this paper. Experimental results on MOOC data set show that compared with advanced course recommendation model, POR_MN model can also improve the recommendation performance by 2.43%-3.47%, which is better than other recommendation models. In the future, the introduction of periodic time signals will be considered to further enhance the performance of the course recommendation model in combination with the differences in learning and examination purposes and learning preferences.

References

[1] Qiu, J.Z., Tang, J., Liu, T.X., et al.(2016) “Modeling and Predicting Learning Behavior in Moocs”, Ninth ACM International Conference on Web Search and Data Mining, pp. 93- 102.

[2] Shen, Y.F. (2019) “Personalized Learning Path Recommendation Model Based on Multiple Intelligent Algorithms”, China Educational Technology, (11), pp. 66-72.

[3] Su, Q., Chen, S.Z., Wu, W.M., et al. (2020) “Personalized Recommendation Model Based on Collaborative Filtering Algorithm of Learning Situation”, Data Analysis and Know-ledge Discovery, 4(5), pp. 105-117.

[4] Dong, W.K., Zhang, Z.X., Tan, T.N. (2019) “Attention-aware Sampling Via Deep Reinforcement Learning for Action Recognition”, In the Thirty-Third AAAI Conference on Artificial Intelligence,pp.8247-8254.

[5] Koren, Y., Bell, R. M., Volinsky, C.(2009) “Matrix Factorization Techniques for Recommender Systems”, Computer, 42(8), pp. 30-37.

[6] Rendle, S., Freudentgaler, C., Gantner, Z., et al.(2009) “BPR: Bayesian Personalized Ranking from Implicit Feedback”, In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 452-461.

[7] Kabbur, S., Ning, X., Karpis, G. (2013) “FISM: Factored Item Similarity Models for Top-N Recommender Systems”, In the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 659-667.

[8] He, X.N., Liao, L.Z., Zhang, H.W., et al. (2017) “Neural Collaborative Filtering”, In Proceedings of the 26th International Conference on World Wide Web, pp. 173-182.

[9] Xia, J., Tang, J. (2017) “Guess You Like: Course Recommendation in MOOCs”, In Proceedings of the International Conference on Web Intelligence, pp. 783-789.

[10] He, X.N., He, Z.K., Song, J.K., et al. (2018) “NAIS: Neural Attentive Item Similarity Model for Recommendation”, Transactions on Knowledge and Data Engineering, 30(12), pp. 2354- 2366.

[11] Zhang, J., Hao, B.W., Chen, B., et al. (2019) “Hierarchical Reinforcement Learning for Course Recommendations in MOOCs”, In the Thirty-Third AAAI Conference on Artificial Intelligence, pp. 435-442.

[12] Tang, Y., Wang, W., Huang, S.Q. (2020) “Deep Reinforcement Learning Model in Heuristic Coaching Scenario”, Journal of Systems Engineering, 35(02), pp.145-152.

[13] Zheng, C., Wang, J. (2021) “Collaborative Filtering Recommendation for Joint Attention and Autoencoder”, Computer Engineering and Applications, 57(10), pp.139-145.