International Core Journal of Engineering Volume 7 Issue 12, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

Vending Machine inventory Control Model based on reinforcement learning

Haiyan Hu

School of Shanghai Maritime University, Shanghai 201306, China

Abstract

As a means of retail, demand for vending machines varies according to geographical location and the products sold, so demand is almost unpredictable. In addition, the capacity of vending machines is limited, and replenishment is not as flexible as traditional retail stores, so the replenishment strategy research is more complex, but also has practical significance. Based on the theory of reinforcement learning, consider food vending machine cannot be ignored in the inventory constraints and the cost control of the holding cost, ordering cost and shortage cost factors such as reality, through the design of intensive study quad (environment state observation, agent action, state transition, remuneration), builds a Markov decision process. Then, q-learning algorithm is used to realize the optimal replenishment strategy to find vending machines in the case of unpredictable customer demand. The research goal is to minimize the inventory cost of vending machines and obtain the maximum profit. Finally, the optimal action in different states is obtained through simulation with an example.

Keywords

Reinforcement learning, Markov decision process, Inventory control, Q-Learning.

1. Introduction

With the development of e-commerce in China, vending machine as a mobile business tool, to meet people's fast and convenient shopping needs. With the large increase of vending machines and the diversification of sales products, the problem of inventory management cannot be ignored. Vending machines are constrained by limited machine space, and demand is unpredictable. If stocks are low, the risk of shortages is high. In the case of large inventory, the cost of inventory increases and the rate of product spoilage increases, especially for food vending machines. Therefore, the food vending machine inventory management is an effective way to improve its efficiency, how to make a reasonable and scientific inventory decision is the focus of this paper.

Reinforcement learning problems are often studied by agent/environment interface (Figure 1). Based on this, researchers often further model reinforcement learning as Markov decision process. Reinforcement learning is a specific kind of machine learning problem. Unlike supervised learning, which requires prior knowledge, reinforcement learning learns how to maximize rewards through interaction with the environment[1].Inventory management is a sequential decision making problem. In the face of uncertain demand, reinforcement learning (RL) can be used to solve the inventory management problem of vending machines.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

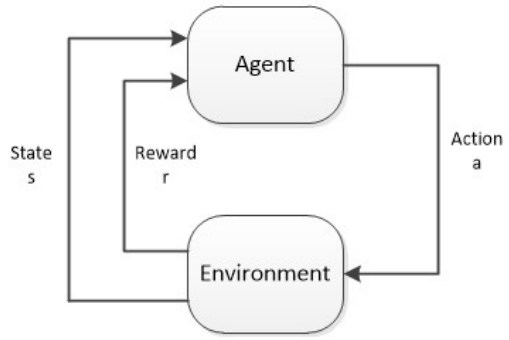


Figure 1. Schematic diagram of reinforcement learning

The uncertainty and diversity of user demand are becoming increasingly obvious and the product life cycle is shortening. In this case, assuming that the demand distribution is known, the traditional methods of inventory management, such as minimizing the total cost for inventory management, or increasing inventory to meet customer demand, are inadequate. In this case, reinforcement learning method does not require prior knowledge and learns how to maximize rewards through interaction with the environment, which is very suitable for inventory management with uncertain demand. Based on the theory of reinforcement learning, consider food vending machines in the inventory constraints and the cost control cannot be neglected, the loss rate, outdated reality factors such as cost, shortage cost and return of the cost, through the design of intensive study quad (environment state observation, agent action, state transition, remuneration), builds a Markov decision process. Then, q-learning algorithm is used to realize the optimal replenishment strategy to find vending machines in the case of unpredictable customer demand. The research goal is to minimize the inventory cost of vending machines and obtain the maximum profit. Finally, the simulation is carried out with an example.

2. Literature review

Inventory control research is not a new problem, however, in the face of changing market demand, inventory management may need some new research direction. Inventory management is a sequential decision problem. Facing uncertain demand, RL can be used to solve the multi-cycle replenishment strategy of retail stores. In the case of non-stationary customer demand, Chengzhi Jiang and Zhaohan Sheng[2] studied the dynamic inventory control problem to meet the target service level in the supply chain of non-stationary customer demand, and proposed a case-based reinforcement learning algorithm. Mohammad Hossein Fazel Zarandi[3] provided a fuzzy reinforcement learning algorithm for supply chain inventory control. In order to optimize the performance of the entire supply chain, Ilaria Giannoccaro and Pierpaolo Pontrandolfo[4] proposed a method to manage inventory decisions in an integrated manner at all stages of the supply chain. Abhilasha Prakash Katariya[5] studied the vMSA cycle consumption and replenishment decision of multi-source components in Dell's supply chain and proposed a heuristic replenishment strategy, which can be used as a reference for suppliers. Ahmet Kara and Ibrahim Dogan[6] applied the reinforcement learning method to specify the ordering strategy of perishable inventory system and studied the perishable product inventory management system under the condition of random demand and definite delivery date with the goal of minimizing the total cost of retailers.

Joint supplementary problem refers to the strategy of purchasing multiple goods from the same supplier and scheduling different goods in different periods. Christian Larsen and Marcel Turkensteen[7] established the inventory level of supplier warehouses and retailers by modeling the vENDOR-managed inventory problem as a joint supplement problem. However, literature[8] studied the joint supplementary problem where demand was a fuzzy variable under the condition of a single supplier, obtained a return function obtained by the system after each action, processed the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

mathematical model through learning algorithm, and finally solved the function to minimize the order cost.

Supply chain management (SCM) provides enterprises with a competitive advantage in the market. Inventory control plays an important role and has been paid attention by many researchers in recent years. To overcome the bullwhip effect, many forms of supply chain coordination, such as vendor managed Inventory (VMI) and continuous replenishment and collaborative forecasting, planning and replenishment (CPFR), have been implemented in supply chain software in recent years. Zheng Sui[9] proposed an approach based on reinforcement learning, which is rooted in The Behrman equation, to determine the replenishment policy of the VMI system for consignment inventory.

We take vending machine as an example to conduct inventory management research. First of all, as a retail way, vending machine is in line with the development of the digital era. In today's fast-paced life, it is difficult for traditional retail stores to keep pace. Vending machines may be the leader in retail. But demand can also be severely affected by its geographical location and the different products it sells, and demand is almost unpredictable. In addition, the capacity of vending machines is limited, and replenishment is not as flexible as traditional retail stores, so the replenishment strategy research is more complex, but also has practical significance.

3. Inventory model based on reinforcement learning

3.1 Reinforcement learning and Markov decision processes

Reinforcement learning is a special kind of machine learning problem. The basic idea is to take actions according to observations in the process of interaction between agents and the environment, and find the optimal strategy by obtaining the maximum reward after the action[10].This chapter uses the most classical mathematical model of reinforcement learning - Markov decision process to model. Markov decision models introduce probability and Markov properties on the basis of states (*S*), actions (*A*) and rewards (*R*). Markov property is simply that, given the current state, the future state is independent of the past state. Thus, the probability of jumping from state $S_t = s$ and action $A_t = a$ to the next state $S_{t+1} = s'$and reward $R_{t+1} = r$ at time t is defined as:

$$P_r\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\} \quad (1)$$

Assuming that a turn ends in T cycles, then T (t<T) return G_t after the moment can be defined as the sum of future rewards, but such return always tends to infinity, so the concept of discount is introduced. Define the return as:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{\tau=0}^{+\infty} \gamma^\tau R_{t+\tau+1} \quad (2)$$

Where the discount factor $\gamma \in [0\ ,\ 1]$. If it is 0, it is greedy, meaning that the value is determined only by the current delay reward. If it is 1, all subsequent state rewards are treated equally as the current reward. Most of the time, we pick a number between zero and one, where the weight of the current delayed reward is greater than the weight of the subsequent reward.

Berman's expectation equation is commonly used to evaluate strategies:

●Use the action value function at time t to represent the state value function at time t:

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a), \quad s \in S \quad (3)$$

●Use the state value function at t+1 to represent the action value function at t:

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s') = \sum_{s',r} p(s',r|s,a)[r + \gamma v_\pi(s')], \quad s \in S, a \in A \quad (4)$$

3.2 Problem description

Vending machines have an unshakable place in the new era of retail. Compared with retail stores, vending machines greatly reduce the cost of sales by eliminating the need for store leasing and manual sales. Moreover, vending machines have fixed prices, and there are no bundling or special offers.

However, vending machines are limited by inventory capacity and cannot offer as many products as retail stores. Even today, with information technology, we are able to quickly detect shortages in

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

vending machines, but vending machines are not replenished as frequently as retail stores and are still vulnerable to loss of stock. Therefore, it is of practical significance to study the inventory control of vending machines, and the purpose of establishing the model is to determine the ordering period and quantity. Markov decision model is established with the goal of maximum return. The demand for vending machines is difficult to determine, and we define the demand as a Poisson distribution with a mean of λ. The demand for each cycle is independent and equally distributed. The demand is reached at the beginning of each cycle. If the inventory directly meets the demand, the order quantity is zero. If not, make replenishment strategy. System activities (figure 2) are as follows:

Step 1: Demand is met;

Step 2: Check stock for expired products and discard;

Step 3: Check whether the remaining inventory meets the demand, Meet demand and update inventory; If not, make order strategy;

Step 4: At this stage, the quantity of inventory has changed.

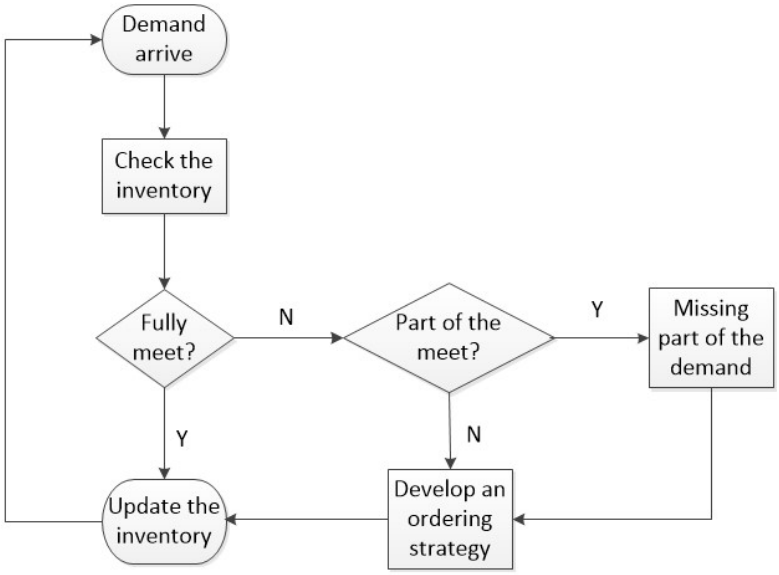


Figure 2. Vending machine system activities in a week

At the end of each cycle, the inventory level will be observed, and the order quantity decision for the next cycle will be made, the goal is to minimize the total expected discount cost for N cycles. Conditional hypothesis and parameter description. This chapter considers a multi-cycle single product ordering model, the goal is to find the optimal ordering strategy to minimize the expected total cost, including ordering cost, storage cost and out-of-stock cost. Assuming that vending machines sell only one product and are on a first-in, first-out basis. Assuming period T, demand follows Poisson distribution with mean value λ. When the demand arrives at the beginning of each cycle, the system judges whether the inventory is satisfied or not. If not, the ordering strategy is formulated. There are the following assumptions:

●Suppliers do not have bundled sales, nor do they have fixed delivery volume;

●Order will produce the corresponding transportation costs, including in the order cost, so frequent order and can not make the maximum benefit;

●There is no advance time for ordering, that is to order;

●Make decisions at the beginning of each cycle. When part of the demand can be met, the unmet demand will disappear, resulting in loss of stock;

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

●Under the boundary condition, at the last moment the retailer's inventory is zero and the inventory is completely consumed;

●The products that arrive at the same time have the same age and are of maximum life at the time of warehousing.

Table 1. Description of model parameters

Symbol	Define
S	System status, representing inventory quantity
A	Action set representing the quantity ordered
K	Maximum inventory capacity
D	The random demand for each period follows a Poisson distribution with a mean value ofλ
C_a	Carrying cost per unit of product
C_s	The selling price per unit of product
C_b	Order cost per unit of product
C_q	Cost due to out of stock
Q_t	Quantity out of stock per cycle
X_t	Quantity sold per cycle
γ	The discount factor, $\gamma \in [0\ ,\ 1]$

3.3 Inventory modeling

The state variable S is defined as the inventory quantity, and the state transition can be expressed as:

$$S_{t+1} = max\ \{\ 0\ ,\ S_t + A_t - D_{t+1}\ \} \tag{5}$$

Action variable A is defined as the order quantity set:

$$A = \{a_1, a_2, \ldots, a_n\} \tag{6}$$

The reward function R is defined as revenue, which consists of sales, holding cost, order cost and out-of-stock cost:

$$R_t = C_s * X_t - (C_a * S_t + C_b * a_t + C_q * Q_t) \tag{7}$$

The sales volume is equal to the demand when the demand is satisfied, and when the demand is not satisfied, the quantity of stock is out of stock, and the surplus demand disappears. Therefore, sales volume X_t and stock out Q_t are defined as follows:

$$X_t = \begin{cases} 0, & S_t = 0 \\ (D_t\text{-}S_t), & 0 < S_t < D_t \\ D_t, & S_t \geq D_t \end{cases}, Q_t = [D_t - S_t]^+ \quad (D_t > S_t) \tag{8}$$

After R is defined, we further define the value function. The action value function is defined in this paper, which represents the expected return of adopting strategy π after action A is taken in state S.

We define the value function as the optimal profit function (that is, the function that needs optimization) :

$$v_\pi(s) = E_\pi(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots | S_t = s) \tag{9}$$

And $V(S_{termination})=0$.

Action probability refers to the probability of taking action A in state S, which is defined as:

$$\pi(a|s) = P(A_t = a|S_t = s) \tag{10}$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

3.4 Experimental design and result analysis

In order to facilitate the simulation, we set the inventory capacity K to 100, and the state S, i.e. the inventory quantity, is discrete into six states{0,20,40,60,80,100}, the action variable A, namely the order quantity, is discretized as{0,20,40,60,80,100}.And set the minimum inventory as 40, that is, when the inventory at the end of the week is less than 40, replenishment must be carried out, and the inventory after replenishment is not less than the minimum inventory. The action set of each state can be expressed as:

$$A(s_0) = \{40, 60,80,100\}, A(s_1) = \{20, 40,60,80\}$$
$$A(s_2) = \{0,20,40,60\}, \quad A(s_3) = \{0,20,40\}$$
$$A(s_4) = \{0,20\}, \quad A(s_5) = \{0\}$$

Take a vending machine that sells bread as an example and set the cost value:

$$C_a = 0.5, C_s = 6.0, C_b = 2.2, C_q = 1.2$$

The demand follows the Poisson distribution with the mean value of λ, and we set λ as 30. In this case, the demand and its probability distribution are shown in Figure 3:

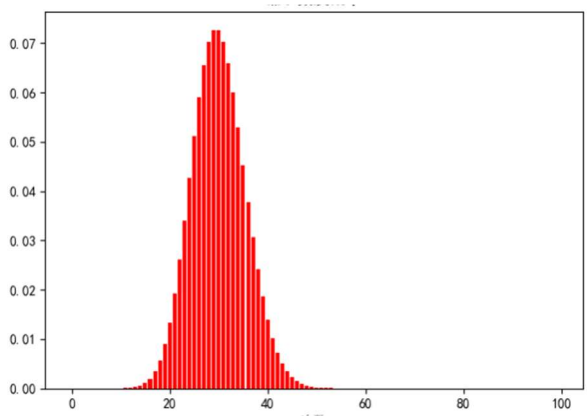


Figure 3. Demand number and probability distribution at λ=30

We use q-learning algorithm for model training to obtain the optimal replenishment strategy, and the specific steps are as follows:

(1) Initialize Q values of all state-action pairs, set gamma parameter and R matrix;

(2) Observe the current state of the systemss_t;

(3) Repeat the following steps:

a Choose an action a;

b Calculate the revenue r brought by this action;

c Observe the new system statess_{t+1};

d Update Q value rule:

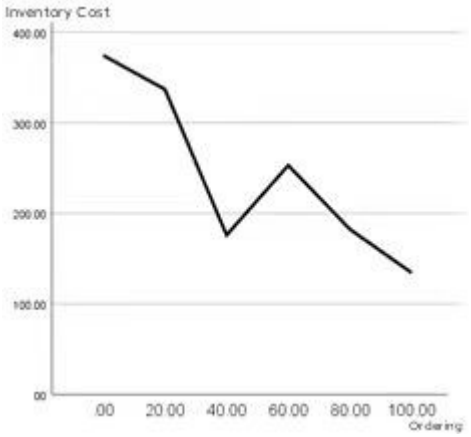
$$Q(s,a) = R(s,a) + \gamma * \max_{\tilde{a}}\{Q(\tilde{s},\tilde{a})\} \tag{11}$$

3.5 Results analysis

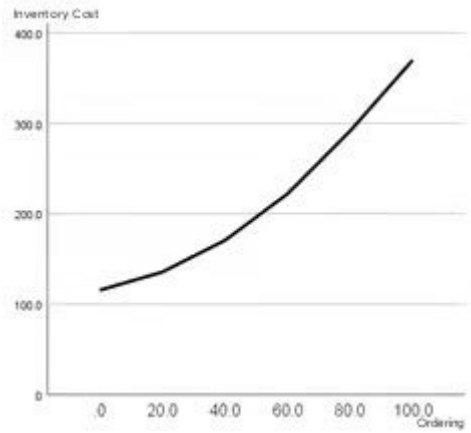
As shown in Figure 4, we first observe the change of out-of-stock cost, which increases and then decreases with the order quantity decreasing to a point first. This is because the uncertainty of demand leads to different random demand arriving in each cycle, so it is not the larger the order quantity, the smaller the out-of-stock cost will be. The carrying cost is different, the larger the order quantity, the higher the carrying cost. This paper assumes out-of-stock costs, but in practice, the ordering and out-of-stock costs are unaffordable when stock is scarce. Therefore, we must find an optimal ordering strategy that minimizes the out-of-stock costs per cycle while minimizing the number of orders. We consider the order fee of each stage into the cost, and it seems that the revenue of each cycle is low.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

However, after considering multiple cycles, the revenue will not be at a low level. Moreover, in the last cycle, when we do not consider the order, the revenue will reach the threshold.



(a)Out of stock cost varies with order quantity



(b)Carrying cost varies with order quantity

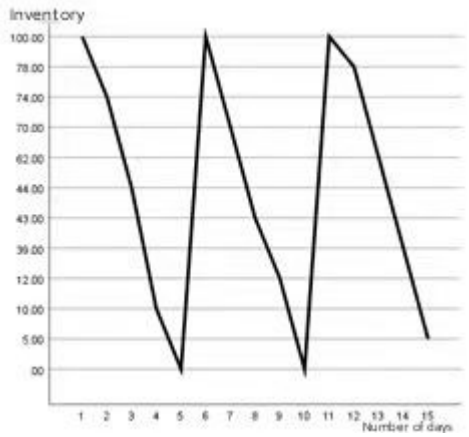
Figure 4. Relationship between order quantity and cost

We take the Poisson distribution of λ=30 to randomly generate demand, and the randomly generated demand set is{29.00 26.00 30.00 34.00 23.00 30.00 27.00 31.00 19.00}.We compared the benefits of choosing different actions in the same state, as shown in Figure 6. We can clearly see that in the same demand situation, when the order quantity is at a higher level, the revenue is lower. This is easy to explain. When the inventory level is reduced, the higher the replenishment level is, the higher the order cost is, and the lower the profit is. In addition, as we observe in the figure, when the demand is closer to the inventory, the benefit is greater. This means that our inventory needs to be as close to demand as possible. We can make a demand forecast according to the demand of the previous cycle and make a reasonable replenishment strategy in the next cycle.

Figure 5. Line chart of demand, inventory, ordering and profits

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

The more times you purchase, the easier it is to meet your needs. However, unit purchase cost is one of the major factors affecting the cost. Once the order cost increases, it will lead to the increase of vending machine cost and the decrease of revenue, thus leading to the shortening of replenishment cycle. So, we need to look at the different order cycle will influence on the cost, we consider the average for three cycles, design different ordering cycle respectively to calculate the return, at this point we assume that each cycle of initial inventory status is full inventory (i.e., s = 100), according to the last day of each cycle inventory replenishment quantity, The demand is still generated randomly by a Poisson distribution of λ=30. Order according to different needs and cycles. See Figure 7 for inventory status. In the figure, we obviously find that when the order cycle is 10 days, the inventory level will be in a low or even zero inventory state for a long time. In this low or even zero inventory state, vending machine revenue will be greatly reduced. replenishment period is 5 days, the replenishment request can be sent immediately whenever the inventory enters 0 state. Although certain out-of-stock loss will be caused, it will not be out of stock for a long time, reducing the out-of-stock loss. It can be seen that the replenishment cycle is 10 angel inventory at a low level for a long time, resulting in a serious loss of stock, which is not reasonable. When the replenishment period is 5 days, there will also be a certain loss of stock shortage. Therefore, we need to set the replenishment period within 5 days.



(a)The ordering cycle is 5 days

(b)The ordering cycle is 10 days

Figure 6. Relationship between inventory and order cycle

Finally, we use q-learning algorithm to find the optimal replenishment strategy by training Q matrix. The Q matrix after training is as follows.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

$$\begin{bmatrix} 80 & 64 & 87.62543954 & 97.29628497 & 94.46185599 & 100 \\ 80 & 0 & 0 & 0 & 0 & 70.38984865 \\ 0 & 0 & 72.14615502 & 81.65188809 & 66.17948326 & 70.38984865 \\ 88.50328696 & 70.05412017 & 83.93792998 & 70.80262957 & 66.17948326 & 70.38984865 \\ 82.72435407 & 78.72251949 & 65.32151047 & 70.80262957 & 66.17948326 & 70.38984865 \\ 87.98731081 & 64 & 65.32151047 & 70.80262957 & 66.17948326 & 70.38984865 \end{bmatrix}$$

We draw the following conclusions:

(1) In the state of 0 inventory, the selection action a_6 is to fill up the vending machine, which can achieve the maximum profit;

(2) When the inventory level is around 20, choose the action a_1 of replenishing 20, which can achieve the maximum profit. The reasons for this are explained below;

(3) When the inventory level is around 40, the maximum profit can be obtained by selecting the action a_3 of replenishing 60 in this state;

(4) When the inventory state is 60, the replenishment action a_0 is generally not selected;

(5) When the inventory status is 80, select the action a_1, replenishment 20 fill;

(6) In full inventory state, choose action a_0 not replenishment.

For conclusion (2), it is because of the low out-of-stock cost that the system would rather produce out-of-stock cost than increase the order cost. But this is not true. In reality, out of stock will lead to consumer loss, its cost is more difficult to estimate. In addition, in order to facilitate the simulation, the area will be divided into large modules, so that the replenishment quantity and cycle can not get accurate results. The influence of out-of-stock cost on replenishment strategy and specific replenishment strategy will be the next research content.

4. Summary and Prospect

As more and more vending machines have been installed, traditional retail stores have been replaced, but vending machines have been unable to restock in the face of unpredictable demand. In this paper, vending machine inventory management as an example of research, trying to find the optimal replenishment cycle and replenishment strategy, the ultimate goal is to make vending machine inventory cost minimum, maximum income. In this paper, we use reinforcement learning theory to construct Markov decision process and design a quad (state variable is inventory state, action variable is order quantity, reward function is revenue) to model inventory and simulate demand as Poisson distribution. By analyzing the impact of different replenishment cycle on revenue, the optimal replenishment cycle is found. Then the q-learning algorithm is used to solve the problem, and the Q-matrix after training is obtained, so as to find the optimal replenishment strategy. Simulation results show that the model can simulate the inventory of vending machines and has practical significance. However, in order to facilitate simulation, inventory status and replenishment quantity are divided into large modules in this study, which makes it difficult for us to obtain accurate replenishment period and replenishment quantity. In the next step of the study, consider further segmentation of status and replenishment.

References

[1] Li Chengao. Machine Reinforcement Learning and Monte Carlo Tree Based Basic Principle and Its Application [J]. Communication world. Vol. 26 (2019) No. 2, p. 212-213.

[2] JIANG C. Z., SHENG Z. H. Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system [J]. Expert Systems with Applications. Vol. 36 (2009) No. 3, p. 6520-6526.

[3] ZARANDI M. H. F., MOOSAVI S. V., ZARINBAL M. A fuzzy reinforcement learning algorithm for inventory control in supply chains [J]. International Journal of Advanced Manufacturing Technology. Vol. 65 (2013) No. 1-4, p. 557-569.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0096

[4] GIANNOCCARO I., PONTRANDOLFO P. Inventory management in supply chains: a reinforcement learning approach [J]. International Journal of Production Economics. Vol. 78 (2002) No. 2, p. 153-161.

[5] KATARIYA A.P., CETINKAYA S., TEKIN E. Cyclic Consumption and Replenishment Decisions for Vendor-Managed Inventory of Multisourced Parts in Dell's Supply Chain [J]. Interfaces. Vol. 44 (2014) No. 3, p. 300-316.

[6] KARA A., DOGAN I. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems [J]. Expert Systems with Applications. Vol. 91 (2018), p. 91:150-158.

[7] Larsen C, Turkensteen M.A vendor managed inventory model using continuous approximations for route length estimates and Markov chain modeling for cost estimates[J].International Journal of Production Economics. Vol. 157 (2014), p. 120-132.

[8] Zhao Shaohang. Reinforcement Learning Algorithm for Supply Chain Joint Supplementary Problem [D]. Harbin University of Science and Technology. Master of engineering , Harbin Institute of Technology, China, 2015.

[9] SUI Z., GOSAVI A., LIN L. A Reinforcement Learning Approach for Inventory Replenishment in Vendor-Managed Inventory Systems With Consignment Inventory [J]. Engineering Management Journal. Vol. 22 (2010), No. 4, p. 44-53.

[10] Ma Pinggan, Xie Wei, Sun Weijie. A Review of Reinforcement Learning [J]. Command Control and Simulation. (2018), No. 6, p. 68-72.