7ZC-07

Logistics System Utilizing Reinforcement Learning to Optimize Shipping Costs for Food Welfare Facilities

– A Temporary Solution in a Trial Environment –

Niko Haapalainen[†], Tomoshi Iiyama[‡], Kota Uematsu[‡], Daisuke Kitakoshi[‡], Masato Suzuki[‡]

Helsinki Metropolia University of Applied Sciences[†]

National Institute of Technology, Tokyo College[‡]

1. Background

This document presents our logistics system project which is being implemented for a food welfare organization called ″Foodbank″ in Japan to optimize their resources in shipping cost expenses. This study presents a temporary, preliminary solution in a trial environment.

As the mentioned Foodbank organization aims to expand its activity to smaller branch food distribution facilities, there is a need to maintain their food stocks balanced cost-efficiently in every facility. We suggest a solution by developing an online software tool, which computes the shortest route between the facilities while taking the deliverable food particles in consideration by utilizing reinforcement learning methods.

Two experiments are to be conducted in this document in order to evaluate the performance of our current algorithm. The results help to consider whether the current setting satisfies the criteria for later development phases.

2. Logistics System to Optimize Shipping Cost for Food Welfare Facilities

A logistics system is a network of organizations, people, activities, information and resources involved in the physical flow of products from supplier to customer [1]. Such model is applied to this project as indicated in Fig. 1, where the Foodbank organization’s personnel, available delivery vehicles, facility locations and real-time stock information elements communicate with each other systematically to achieve the least costly performance in the area of shipping costs. Such setting vacates resources, boosts productivity and systemizes the working flow.

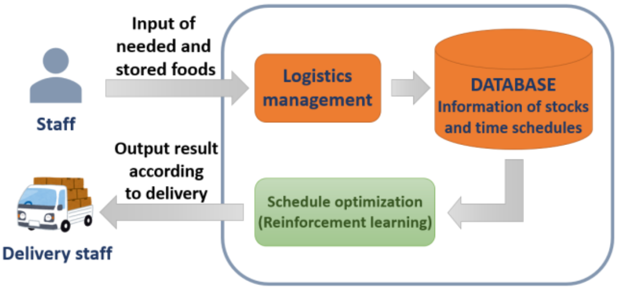
Input of needed and stored foods
Staff
Logistics management
DATABASE
Information of stocks and time schedules
Output result according to delivery
Delivery staff
Schedule optimization (Reinforcement learning)

Fig. 1 - The event flow of our logistics system.

In our setting, the main factor powering up this system is SARSA (state-action-reward-state-action) algorithm from reinforcement learning field of machine learning. A SARSA agent learns its policy by updating its Q-values via the value of its taken actions in its environment in a trial and error method. The Q-values are updated with the following formula:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\Big[R_{t+1} + \gamma Q(S_{t+1},A_{t+1}) - Q(S_t,A_t)\Big] \quad (1),$$

where γ denotes discount rate and α sets the value for learning rate. The algorithm calculation result is deemed to be successful whether the agent has found the optimal solution - or in other words, the shortest route in our environment setting.

3. Experimental Settings

The experimental settings are composed of the environment and machine learning algorithm alongside it.

The environment setting is a simple

two-dimensional XY coordinate plane where the agent is designed to move between user-designated coordinates. These coordinates are regarded as the real-life food facilities' locations. In such environment, the agent learns its behavior by traversing through all coordinates while delivering all food particles required by each facility as the shortest distance. This experiment will be conducted in an environment of four coordinate locations.

The SARSA settings are as follows: Hyperparameters are set as α = 0.1, γ = 0.9 and ϵ = 0.1. Having an ϵ-greedy policy, ϵ can go no higher than the value of 0.99. The maximum episodes are set to 200 and maximum steps to 20. At the end of an episode, a 10-point reward is given if the agent's traversed route is shorter or of equal distance than the previous traversed route. Should the route be longer, a 5-point penalty is given.

4. Experimental Results

It takes approximately three seconds for the agent to learn its optimal policy. The agent could find the shortest route approximately from the 75th episode onwards as indicated in Fig. 2.

In our four-facility-coordinate setting, there were overall five possible routes found and their distances evaluated. The evaluated distance of 20.24 was the most frequent with the event percentage of 75.0% (Table 1). The mentioned distance was the shortest of all other unique distances, which indicates that the optimal policy was found successfully and the algorithm is working properly in this sample run.

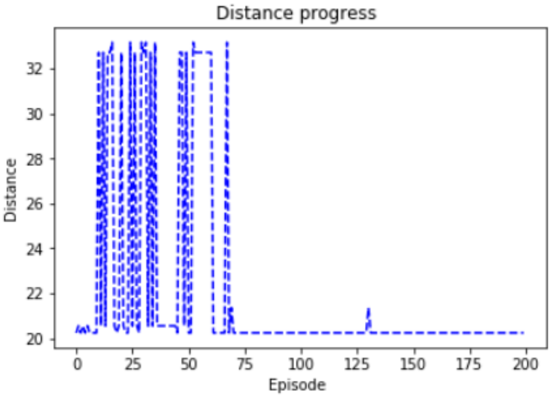

Fig. 2 - The performance of distance progression.

Table 1.
Unique distances found in the environment and their respective frequencies of the same sample run as in Fig.2.

Distance	20.24	20.55	21.36	32.71	33.16
Frequency (%)	*75.0*	*11.0*	*1.0*	*9.5*	*3.5*

To experiment the reliability of the algorithm performance, an experiment of running the program 50 times was conducted. Eventually 44 of 50 program runs were successful, indicating that there is 88% probability for the algorithm to carry out its task of computing the shortest route successfully. The results are satisfying.

5. Conclusion

In this project, a reinforcement learning approach was utilized to study the performance of the agent on finding the shortest route in its environment of coordinates. The experimental results indicate that the algorithm can acquire appropriate policies and the current performance is reasonable.

Occasionally, the agent fails to acquire appropriate policies when increasing the number categories of foods due to the increased number of possible combinations. However, the problem does not affect the overall functionality of the algorithm and can be left to be fixed in later development phases.

With this result, the project can be considered to proceed to the next development phase without concern.

Acknowledgement

We thank the staff and Mr. Kunimoto of Foodbank organization in Hachioji city for cooperation and making this project come true.

References

1. R. Z. Faharani, S. Rezapour, L. Kardar, Logistics Operations and Management, *Logistics Planning and Optimization Problem*, ch. 18.1, p. 371, 2011.

2. R. S. Sutton, A. G. Barto, Reinforcement Learning. An Introduction, 2nd edition. *Sarsa: On-policy TD Control*, ch. 6.4, p. 129-131, 2018.