PAPER

A POMDP-Based Approach to Assortment Optimization Problem for Vending Machine

Gaku NEMOTO[†], ***Nonmember and*** **Kunihiko HIRAISHI**[†a)], ***Member***

SUMMARY Assortment optimization is one of main problems for retailers, and has been widely studied. In this paper, we focus on vending machines, which have many characteristic issues to be considered. We first formulate an assortment optimization problem for vending machines, next propose a model that represents consumer's decision making, and then show a solution method based on partially observable Markov decision process (POMDP). The problem includes incomplete state observation, stochastic consumer behavior and policy decisions that maximize future expected rewards. Using computer simulation, we observe that sales increases compared to that by heuristic methods under the same condition. Moreover, the sales approaches the theoretical upper bound.

key words: *assortment optimization, POMDP, vending machine*

1. Introduction

Appropriate product assortment planning, as well as pricing and inventory management, is an important issue for many retailers. Various approaches are being made to solve problems for each type of business, such as retail stores, convenience stores, supermarkets, and EC sites. Among them, assortment optimization for vending machines (especially beverage vending machines) has several characteristics different from other retail stores. That is, there are many product types with limited available inventory, there is a time lag until the sales data can be obtained, replenishment opportunities are limited, and changes in sales due to the environment is large. These constraints make the problem more complicated.

In this paper, we first present formulation of an assortment optimization problem for vending machines, next propose a model that explains consumer's decision making, and then show a solution method to the problem. We use formulation based on the partially observable Markov decision process (POMDP), a modeling framework for decision making processes where state variables are partially observable. In vending machines, workers, called route men, repeatedly change the assortment of products in order to get better sales. The goal of assortment optimization is to maximize the expected sales by changing the assortment at each replenishment work.

The remainder of the paper is organized as follows. In Sect. 2, overview of the related works is shown. Characteristics of the assortment optimization problem for vending machines are also described. In Sect. 3, the general formulation of the problem is presented. In Sect. 4, the consumer's product selection model is presented together with the theoretical upper bound of the expected sales. By these upper bounds, we can know how the obtained solution is close to the optimal one. The proposed POMDP-based approach to this problem is described in Sect. 5. This is the main contribution of this paper. In Sect. 6 we present a simulation model and its setting for the evaluation, and discuss accuracy and effectiveness of the proposed method in Sect. 7. Section 8 is the conclusion.

2. Related Works

Assortment optimization problem has been widely studied. There are two major themes in literatures on assortment optimization, which deals with static or dynamic substitution mechanisms and models in consumer behavior.

Static substitution assumes that if the initially selected product is out of stock, then the consumer will not purchase another item instead [1]. In contrast, dynamic substitution assumes that if the product is out of stock, then the consumer purchases another item as an alternative [2], [3].

In [4], three models are shown for describing consumer behavior: exogenous demand, locational choice and multinomial logit. The exogenous demand model gives a method for describing consumer behavior from observable sales data of each product, such as Kök and Fisher [5]. The locational choice model was developed by Lancaster [6]. This model introduces multi-dimensional vectors where each dimension corresponds to a product characteristic and consumer's demand. The consumer's selection of products is determined by the proximity of the consumer's ideal vector to the product's vector. The multinomial logit (MNL) model is a random utility model that represents the selection probability of each product as functions of consumer's utility. The basic MNL model was established by McFadden [7]. The MNL model has limitations because it assumes independence of irrelevant alternatives in the selection probabilities of products. To reduce these limitations, the nested MNL model was proposed by Williams [8].

On the other hand, researches on assortment planning for vending machines are not in progress. Instead, consumer behavior for vending machines is being studied. Anupindi et al. [9] propose a model for demand estimation that takes the

Manuscript received March 31, 2023.
Manuscript revised July 20, 2023.
Manuscript publicized September 5, 2023.

[†] The authors are with School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923-1292 Japan.
a) E-mail: hira@jaist.ac.jp

DOI: 10.1587/transfun.2023EAP1036

dynamic substitution into consideration. The reasons why the assortment optimization problem for vending machines has not been well discovered are considered to be (i) demand for solving this problem was small because the assortment is usually decided by route men using their knowledge and experience on sales, and (ii) complexity of the problem. The complexity arises from the following notable characteristics of the problem:

- Sales of products can be observed only when the replenishment is done.
- The replenishment work is done on a regular basis. Therefore, solution to the problem is a decision making process based on past history of observations.
- Nature of customers is not observable and needs to be estimated.

Recently, beverage companies try to introduce information systems that support route men's work. Proposing a method that helps the route men's decision making is the main contribution of this paper.

3. General Formulation

Before discussing the assortment optimization problem, we outline the general formulation. The assortment optimization problem for vending machines is defined by a 6-tuple $AOP = (\boldsymbol{A}, S, G, O, \pi, C)$, where $\boldsymbol{A}$ is the set of assortments, S is the set of states, G is the gain function, O is the observation function, π is the policy, and C is the assortment constraints. Details are described below.

(1) Products and Assortment

Consider n kinds of products $q_i (i = 1, \dots, n)$ and m columns ($m > 0$, normally $n > m$), where columns of a vending machine are containers for stocking products. An assortment is a combination of selecting m products from n kinds of products allowing duplication. Such a combination is represented by a multiset[†]. Let $\boldsymbol{A} = \{\boldsymbol{a}_1, \dots, \boldsymbol{a}_L\}$ denote the set of all assortments, where L is the total number of assortments. The assortment given at time t is denoted by $\boldsymbol{a}(t)$. Note that $\boldsymbol{a}(t)$ takes effect on sales between t and $t + 1$.

We assume that every column has the same capacity, and let cap denote the capacity of each column. Then the number of product q_i in the assortment $\boldsymbol{a}(t)$ is $stk(\boldsymbol{a}(t), q_i) := \#\boldsymbol{a}(t)[q_i] \cdot cap$, where $\#\boldsymbol{a}(t)[q_i]$ is the number of occurrences of q_i in the multiset $\boldsymbol{a}(t)$ and we assume each column is full after replenish work.

(2) State Space

Let $S = \{s_1, \dots, s_v\}$ be the set of states, where each state s_i is a u-dimensional vector and each component of a sate can be a real number, an integer and a discrete value. The states of vending machines consists of environment, weather, background population for the purchase at the vending machine, etc.

The state at time t is denoted by $s(t)$. We define the state transition probability as a function δ: $\mathbb{N} \times S \times S \longrightarrow [0, 1]$, where $\forall t, s_j : \sum_{j'} \delta(t, s_j, s_{j'}) = 1$. It means that the probability that $s(t) = s_j$ and $s(t + 1) = s_{j'}$ is $\delta(t, s_j, s_{j'})$. When the state transition probability depends on time t, it is called time variant, otherwise it is called time invariant. In the time invariant case, δ is defined as $\delta : S \times S \longrightarrow [0, 1]$.

(3) Gain Function

The gain function for assortments is defined as a function $G : S \times \boldsymbol{A} \longrightarrow \mathbb{N}$ that gives the total sales (amount or unit) under a given state and an assortment. $G(s_j, \boldsymbol{a}_l)$ is given by the sum of the sales of all products: $G(s_j, \boldsymbol{a}_l) := \sum_{q_i \in \boldsymbol{a}_l} g_i$, where g_i is the sales of product q_i. The vector $\boldsymbol{g} := [g_1, \dots, g_n]$ is called *the gain vector*. Note that we implicitly assume all products have the same price. How to derive the gain function is explained in the next section.

(4) Observation Function

The observation function is defined as $O : S \longrightarrow W$, where W is some set. The observation at time t is denoted by $o(t) := O(s(t))$. As we have defined, each state s is represented by a u-dimensional vector $s_i := [s_i^1, \cdots, s_i^u]$. In this paper, we assume that the observation function masks some of the substates, e.g., for state $s_i = [s_i^1, s_i^2, s_i^3, s_i^4]$, $O(s_i) = [s_i^1, s_i^4]$ (the function masks the second and the third substates). Here the masked substates imply unobservable substates and the others imply observable ones.

(5) Policy

When $s(k), o(k), \boldsymbol{a}(k), \boldsymbol{g}(k), k = 0, \dots, t-1$ are given, a function that outputs $\boldsymbol{a}(t)$ is called a policy.

(6) Assortment Constraint

The assortment constraint is a set $C \subseteq \boldsymbol{A} \times \boldsymbol{A}$. For any time t, $(\boldsymbol{a}(t), \boldsymbol{a}(t + 1)) \in C$ has to be satisfied. The reason why this constraint arises is that the number of products the route man can exchange at each time is limited. This constraint characterizes the assortment optimization problem for vending machines.

We now define the assortment optimization problem studied in this paper.

Assortment optimization problem: Find a policy that satisfies the assortment constraint and maximizes the total gain during time $t = 0, \dots, T$.

We can also classify the problem by the following characteristics: The state space is known/unknown for the agent (the route man in this case), complete/incomplete observation, gain function is known/unknown, and transition probability is known/unknown. Examples are

- Stores (such as convenience stores, supermarkets):

[†] Multiset: A concept of set that combines the degree of duplication of how many elements are included when the set contains multiple elements of the same value. $\#X[e]$ represents the number of e included in the multiple set X. We denote $e \in X$ if $\#X[e] > 0$.

state is known, complete observation, gain function is known.

- Vending machines: state is known (or unknown), incomplete observation, gain function is known.

4. Product Selection Model

In order to give the gain function, we introduce a consumer's product selection model. Based on MNL model, utility values give the probabilities that a consumer selects one from plural selectable products [10], [11]. Let $P_{q_i,s_j,k}$ denote the probability that consumer C_k tries to purchase product q_i in state s_j. The utility value when consumer C_k tries to purchase product q_i, denoted by $V_{q_i,s_j,k}$, is given by a linear regression model

$$V_{q_i,s_j,k} := \ln \frac{P_{q_i,s_j,k}}{P_{q_1,s_j,k}} = \alpha_i^{j,k} + \sum_l \beta_{i,l}^{j,k} Y_l^j \tag{1}$$

where we assume product q_1 is the reference, $\alpha_i^{j,k}$ is a constant, and $\beta_{i,l}^{j,k}$ is the coefficient of each explanatory variable Y_l^j. Then the probability that consumer C_k selects product q_i in state s_j is given by

$$P_{q_i,s_j,k} = \frac{\exp(V_{q_i,s_j,k})}{\sum_{l=1}^{n} \exp(V_{q_l,s_j,k})} \tag{2}$$

Remark that utility values and the selection probabilities are defined not only for products in the assortment, but also for products not in the assortment.

Using the probability Eq. (2), we give the purchase probability of product q_i by each consumer. Let N be the number of consumers and let X_i denote the stochastic variable representing the number of sales for product q_i without any restriction on the assortment. The purchase probability $\Pr(X_i = r|s_j)$ under state s_j follows Poisson binomial distribution [12]. Poisson binomial distribution is explained as follows. We consider N independent trials each of which has its own success probability. Then Poisson binomial distribution is the discrete probability distribution of the number of successes from the N trials that can be computed recursively by

$$\Pr(X_i = r|s_j) = \begin{cases} \prod_{k=1}^{N} (1 - P_{q_i,s_j,k}) & \text{if } r = 0 \\ \frac{1}{r} \sum_{l=1}^{r} (-1)^{l-1} \Pr(X_i = r - l|s_j) \Upsilon(l) & \text{if } r > 0 \end{cases} \tag{3}$$

where $P_{q_i,s_j,k}$ is that defined by Eq. (2) and

$$\Upsilon(l) = \sum_{k=1}^{N} \left(\frac{P_{q_i,s_j,k}}{1 - P_{q_i,s_j,k}} \right)^l,$$

$$\text{Expected value}: E[X_i|s_j] = \sum_{k=1}^{N} P_{q_i,s_j,k} \tag{4}$$

Next we consider the probability under a given assortment. We assume the static substitution. Since the amount of actual sales g_i is constrained by the assortment, the probability under state s_j and assortment $\boldsymbol{a}_h$ is obtained as follows.

$$\Pr(g_i = r|s_j, \boldsymbol{a}_h) = \begin{cases} 0 & \text{if } r > stk(\boldsymbol{a}_h, q_i) \\ \sum_{l=r}^{N} \Pr(X_i = l|s_j) & \text{if } r = stk(\boldsymbol{a}_h, q_i) \\ \Pr(X_i = r|s_j) & \text{if } r < stk(\boldsymbol{a}_h, q_i) \end{cases} \tag{5}$$

Due to the capacity constraint, all cases $r \leq X_i \leq N$ reduce to $X_i = r$. Also, the expected reward under state s_j and assortment $\boldsymbol{a}_h$ is given by

$$E[G(s_j, \boldsymbol{a}_h)] = \sum_{q_i \in \boldsymbol{a}_h} \min\{stk(\boldsymbol{a}_h, q_i), E[X_i|s_j]\} \tag{6}$$

We can derive the theoretical upper bound on the expected sales. If the agent explicitly knows the state $s(t)$ of the vending machine at time t, the agent can maximize the expected total sales by choosing an appropriate assortment from the set of the entire assortment $\boldsymbol{A}$. We can give the following upper bound of expected sales at each time t, without considering assortment constraint.

$$E_t^{Upper\ bound} = \max_{\boldsymbol{a}(t) \in A} E[G(s(t), \boldsymbol{a}(t-1))] \tag{7}$$

By the assortment constraint, selection of the assortment at time t is constrained by the assortment at time $t - 1$. We consider *the feasible maximum value* of expected sales under the assortment constraint. Let $\boldsymbol{a}_l$, $\boldsymbol{a}_m$ be two assortments and let $C(\boldsymbol{a}_l, \boldsymbol{a}_m)$ denote a Boolean variable such that $C(\boldsymbol{a}_l, \boldsymbol{a}_m) = 1$ if $(\boldsymbol{a}_l, \boldsymbol{a}_m) \in C$ and 0 otherwise. Then the expected sales considering assortment constraint at time t is given by

$$E_t(\boldsymbol{a}(0), \ldots, \boldsymbol{a}(t-1)) = \left(\prod_{i=1}^{t-1} C(\boldsymbol{a}(i-1), \boldsymbol{a}(i)) \right) \cdot E[G(s(t), \boldsymbol{a}(t-1))] \tag{8}$$

and the feasible maximum value at each time t is

$$E_t^{Feasible\ max} = \max_{\boldsymbol{a}(0), \ldots, \boldsymbol{a}(t-1) \in \boldsymbol{A}} E_t(\boldsymbol{a}(0), \ldots, \boldsymbol{a}(t-1)) \tag{9}$$

Clearly, $E_t^{Feasible\ max} \leq E_t^{Upper\ bound}$ holds.

5. Formulation as POMDP

Following Kaelbling et al. [13], we propose a POMDP-based method that select a good assortment policy from a given set of policies. POMDP is a stochastic process that deals with situations where the state can be partially observed, and these observations do not necessarily satisfy Markov process.

5.1 POMDP

POMDP is a model of an agent that synchronously interacts with a world. Given a discrete set Z, let $\Pi(Z)$ denote the set of all discrete probability distributions on Z. Formally, POMDP is defined as a tuple $POMDP = (St, Act, \Delta, Rw, \Omega, Obs)$, where St is the finite set of states, Act is the finite set of actions, $\Delta : St \times Act \longrightarrow \Pi(St)$ is the state transition function, $Rw : St \times Act \longrightarrow \mathbb{R}$ is the reward function, Ω is a finite set of observations, and $Obs : St \times Act \longrightarrow \Pi(\Omega)$ is the observation function. Since the state has to be estimated through the observation function, Kaelbling's method introduces *a belief*. A belief is a variable that represents what the current state is, and it is estimated from the history of observations. At each time step, the agent choose an action to maximize the expected reward depending on the belief. A policy is a description of the behavior of the agent. We adopt POMDP as a framework of the solution method since we focus on the case where the agent can not recognize a part of the current state of the vending machine. But the proposed model is customized from original POMDP to be adapted to the assortment optimization problem of vending machines.

5.2 POMDP Model for AOP

The POMDP model for the assortment optimization problem is described as follows.

(1) State

The set of states in $POMDP$ is given as the set of states in AOP. The state at time t is denoted by $s(t)$.

(2) Action

The agent is the route man. The action given at time t is the assortment $\boldsymbol{a}(t)$ after the exchange.

(3) State Transition Probability

The state transition probability from time t to $t+1$ is denoted by $\delta(t, s(t), s(t+1))$. We assume that assortments do not affect state transitions, and that the state transition probabilites are time invariant. So we denote the probabilites by $\delta(s(t+1)|s(t))$. We define the state transition probability for n time steps, denoted by δ^n, by $\delta^1(s'|s) := \delta(s'|s)$ and

$$\delta^n(s'|s) := \sum_{s'' \in S} \delta^{n-1}(s'|s'')\delta(s''|s) \tag{10}$$

(4) Observation and Reward

The possible observed state $o(t)$ and the sales vector $\boldsymbol{g}(t) = [g_1(t), \cdots, g_n(t)]$ at time t are stochastically given depending on the state $s(t)$ and the assortment $\boldsymbol{a}(t-1)$. Suppose that $\boldsymbol{g}(t) = [r_1, \ldots, r_n]$. Then the probability is given by

$$\begin{aligned}
&O(s(t), \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t)) \\
&\quad := \Pr(o(t), \boldsymbol{g}(t)|s(t), \boldsymbol{a}(t-1)) \\
&\quad = \prod_{i=1}^{n} \Pr(o(t), g_i(t) = r_i|s(t), \boldsymbol{a}(t-1)) \qquad (11) \\
&\quad = \prod_{i=1}^{n} \Pr(g_i(t) = r_i|s(t), \boldsymbol{a}(t-1))
\end{aligned}$$

The last equality follows from the fact that $o(t)$ is uniquely determined from $s(t)$ as described in Sect. 3(4), i.e., the observation masks some substates. $\Pr(g_i(t) = r_i|s(t), \boldsymbol{a}(t-1))$ is computed by Eq. (5).

At time t, the agent obtains the possible observed state $o(t)$ and the sales $g_i(t)$ of each product q_i. The total sales of products is regarded as the reward. Let $rw(t)$ denote the reward at time t. Then $rw(t) := G(s(t), \boldsymbol{a}(t-1)) = \sum_{q_i \in \boldsymbol{a}(t-1)} g_i(t)$.

5.3 Belief and Policy

The belief is represented by a function $b : S \longrightarrow \mathbb{R}$ such that $0 \leq b(s) \leq 1$ and $\sum_{s_j \in S} b(s_j) = 1$. Let b_t denote the belief at time t. For each state $s \in S$, $b_t(s)$ is the strength that the agent believes $s(t) = s$. A policy on assortment exchange is a function that gives the assortment at each time. We assume that the policy π depends only on the latest state $s(t)$, the observed values $o(t)$, and the latest assortment $\boldsymbol{a}(t-1)$. Also, it is assumed that the agent can select one policy from a finite set of policies $\Pi := \{\pi^1, \ldots, \pi^M\}$.

At each time, the policy on assortment exchanges is decided by the current belief. Given a belief b_{t-1} at time $t-1$, the belief that $s(t) = s'$ under the observed state $o(t)$ and the sales $\boldsymbol{g}(t)$ is given by

$$\begin{aligned}
b_t(s') &= \Pr(s'|o(t), \boldsymbol{a}(t-1), \boldsymbol{g}(t), b_{t-1}) \\
&= \frac{\Pr(o(t), \boldsymbol{g}(t)|s', \boldsymbol{a}(t-1), b_{t-1}) \Pr(s'|\boldsymbol{a}(t-1), b_{t-1})}{\Pr(o(t), \boldsymbol{g}(t)|\boldsymbol{a}(t-1), b_{t-1})} \\
&= \frac{\Pr(o(t), \boldsymbol{g}(t)|s', \boldsymbol{a}(t-1))}{\Pr(o(t), \boldsymbol{g}(t)|\boldsymbol{a}(t-1), b_{t-1})} \\
&\quad \times \sum_{s \in S} \Pr(s'|\boldsymbol{a}(t-1), b_{t-1}, s) \Pr(s|\boldsymbol{a}(t-1), b_{t-1}) \\
&= \frac{O(s', \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t))}{\Pr(o(t), \boldsymbol{g}(t)|\boldsymbol{a}(t-1), b_{t-1})} \times \sum_{s \in S} \delta(s'|s) b_{t-1}(s)
\end{aligned} \tag{12}$$

where the denominator $\Pr(o(t), \boldsymbol{g}(t)|\boldsymbol{a}(t-1), b_{t-1})$ can be treated as a normalizing factor.

5.4 Procedure for Determining Policy

We describe the procedure for determining the strategy for assortment exchange at time t.

At time $t = 0$, we assume that $s(0) = s_j$ is equally likely for all possible states $s_j \in S$. In other words, the belief $b_0(s(0))$ is $b_0(s_1) = b_0(s_2) = \cdots = b_0(s_v) = \frac{1}{v}$. At time $t > 0$, the observed value $o(t)$ and the sales $\boldsymbol{g}(t)$ are obtained. Using Eq. (12), we update the belief by

$$b_t(s_j) = O(s_j, \boldsymbol{a}(t-1), o(t), \boldsymbol{g}(t)) \times \sum_{s \in S} \delta(s_j|s) b_{t-1}(s) \tag{13}$$

After the update, b_t is normalized so that $\sum_{s_j \in S} b_t(s_j) = 1$.

The policy on assortment exchanges $\pi_t^{k_0}(k_0 = 1, \ldots, M) \in \Pi$ is decided based on the expected reward obtained in the future. First, we consider the expected value of the reward $rw(t+1)$ at time $t+1$. In the case where $\pi_t^{k_0} : \boldsymbol{a}(t-1) \rightarrow \boldsymbol{a}^{k_0}(t)$ is selected, the expected reward at time $t+1$ is given as follows:

$$E_{\pi_t^{k_0}}[rw(t+1)] =$$
$$\sum_{s' \in S} \left\{ \sum_{s \in S} \delta(s'|s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_0}(t))].$$

Next, we consider the case that $\pi_{t+1}^{k_1}(k_1 = 1, \ldots, M)$ is selected at time $t+1$. The expected value of the reward $rw(t+2)$ is calculated for the each assortment $\boldsymbol{a}^{k_0}(t), \boldsymbol{a}^{k_1}(t+1)$. Note that the assortment at time $t+1$ depends on the assortment at time t because of the assortment constraints.

$$E_{\pi_t^{k_0} \cdot \pi_{t+1}^{k_1}}[rw(t+2)] =$$
$$= \sum_{s' \in S} \left\{ \sum_{s \in S} \delta^2(s'|s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_1}(t+1))].$$

Similarly at time $t + \tau$ ($\tau > 0$), we can obtain the expected value of the reward as follows:

$$E_{\pi_t^{k_0} \cdots \pi_{t+\tau-1}^{k_{\tau-1}}}[rw(t+\tau)] =$$
$$\sum_{s' \in S} \left\{ \sum_{s \in S} \delta^{\tau}(s'|s) b_t(s) \right\} E[G(s', \boldsymbol{a}^{k_{\tau-1}}(t+\tau-1))].$$

Therefore, we can calculate the maximum expected reward in the future when the policy $\pi_t^{k_0}$ is selected at time t. When the policy $\pi_t^{k_0}$ is selected, the total expected reward $E_{t \rightarrow t+\tau}(\pi_t^{k_0})$ is calculated as the sum of them from $t+1$ to $t+\tau$. As usual in POMDP, in order to make recent rewards more effective, we multiply the future expected reward by the discount rate $\gamma (0 < \gamma < 1)$.

$$\begin{aligned} & E_{t \rightarrow t+\tau}(\pi_t^{k_0}) \\ & = E_{\pi_t^{k_0}}[rw(t+1)] + \gamma \max_{\pi_{t+1}^{k_1} \in \Pi} \Big\{ E_{\pi_t^{k_0} \cdot \pi_{t+1}^{k_1}}[rw(t+2)] \\ & + \gamma \max_{\pi_{t+2}^{k_2} \in \Pi} \Big\{ E_{\pi_t^{k_0} \cdot \pi_{t+1}^{k_1} \cdot \pi_{t+2}^{k_2}}[rw(t+3)] + \gamma \max_{\pi_{t+3}^{k_3} \in \Pi} \Big\{ \\ & \cdots + \gamma \max_{\pi_{t+\tau-1}^{k_{\tau-1}} \in \Pi} \Big\{ E_{\pi_t^{k_0} \cdots \pi_{t+\tau-1}^{k_{\tau-1}}}[rw(t+\tau)] \Big\} \cdots \Big\} \Big\} \Big\} \end{aligned} \tag{14}$$

Note that the expected sales in far future is not very important in vending machines, we consider rewards within a finite horizon.

According to the above procedure, we obtain the total expected rewards $E_{t \rightarrow t+\tau}(\pi_t^1), \ldots, E_{t \rightarrow t+\tau}(\pi_t^M)$ by the policies $\pi_t^1, \ldots, \pi_t^M$. Then, the policy on the assortment exchange π_t is decided as one that maximizes the expected reward:

$$\pi_t = \arg \max_{\pi_t^{k_0} \in \Pi} E_{t \rightarrow t+\tau}(\pi_t^{k_0}) \tag{15}$$

6. Computer Simulation

In this section, we show numerical results obtained by computer simulation of the vending machine assortment optimization problem.

6.1 Parameters and Assumptions

We show the parameters for the simulation and some assumptions.

(1) Consumer

We introduce simple assumptions on consumers who are purchasing products at the vending machine. Each consumer has several attributes (gender, age, occupation, etc.), and the preferences for product selection probabilistically depend on these attributes together with the current state. At time t, N consumers try to purchase products at the vending machine. Suppose that the k-th consumer $C_k (k = 1, \ldots, N)$ tries to purchase one of n kinds of products. The products the consumer C_k tries to purchase are determined probabilistically. We assume that the probabilities are determined by the attributes of the consumer C_k and the state of the vending machine $s(t)$. If the product to be purchased is present in the assortment $\boldsymbol{a}(t)$ and the inventory is sufficient, the consumer C_k purchases it. Otherwise (including in case of sold out), the consumer do not purchase, i.e., we assume the static substitution.

In the computer simulation, we assume that the attribute of the consumer is only gender: male or female. Other conditions and attributes are not considered.

(2) Agent

The agent replenishes the vending machine with products and can exchange the assortment at its own discretion. The agent checks sales for each assortment at the next replenishment work. In the simulation, we assume that the agent knows all attributes and parameters including transition probabilities between states. The agent cannot know the entire information on the current state and estimates it from the history of observations as the belief. Based on the belief, the agent selects the next policy.

(3) State of Vending Machine

Originally, the state of vending machines can be considered to have many parameters and factors. The states can be classified into two types: observable and unobservable. In this simulation, we consider three states: location, temperature

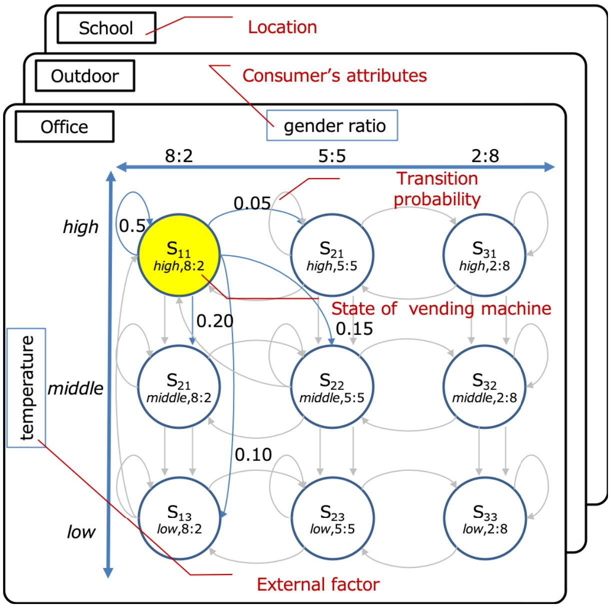

Fig. 1 State of vending machine.

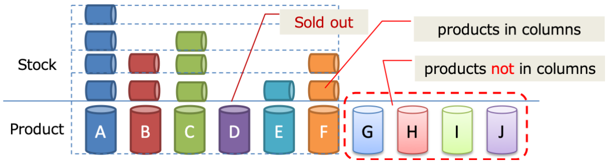

Fig. 2 Products and assortment.

and gender ratio. Location is observable and classified by three types: office, outdoor and school. Temperature is an external factor for vending machines. It is observable and is selected from one of $\{high, middle, low\}$ at each time. The gender ratio is an internal factor for vending machines, and it means the ratio of males and females among consumers who are going to purchase at the vending machine. In the simulation, three patterns are assumed: $\{8:2, 5:5, 2:8\}$. The ratio is selected from one of them at each time. The ratio is unobservable and it cannot be known to the agent. The image of the vending machine states is shown in Fig. 1.

(4) Products and Assortment

All products have the same shape and price, and the number of products that can be replenished in one column is also the same. Products that can be in the assortment are 10 kinds: $\mathsf{A}, \ldots, \mathsf{J}$. The vending machine has 6 columns, and the capacity of each columns is 20 for any kind of products. It is possible to assign the same kind of products to multiple columns. Figure 2 depicts an assortment and stocks.

(5) Selection Probability of Products

In the simulation, the utility value $V_{q_i,s_j,k}$ of product $q_i \in \{\mathsf{A}, \ldots, \mathsf{J}\}$ by the k-th consumer in state s_j is defined as:

$$V_{q_i,s_j,k} = V^0_{q_i} + V^M_{q_i} + \beta^M_{q_i} T_j \quad \text{for male,}$$
$$V_{q_i,s_j,k} = V^0_{q_i} + V^F_{q_i} + \beta^F_{q_i} T_j \quad \text{for female.}$$

Table 1 Parameters of utility value: office.

item	type	COLD /HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	1.0	0.5	0.0	0.5
B	coffee	HOT	0.5	0.5	0.5	-0.5	-0.5
C	café au lait	COLD	0.0	-0.5	0.5	0.0	-0.5
D	green tea	COLD	1.0	1.0	1.0	0.0	0.0
E	tea	COLD	0.5	0.5	1.0	0.0	0.5
F	enegy drink	COLD	-0.5	0.0	-1.0	0.0	0.0
G	soda	COLD	0.5	0.5	0.0	1.0	0.5
H	mineral water	COLD	1.0	0.0	0.0	0.0	0.0
I	tea	HOT	-0.5	-1.0	0.5	-1.0	-0.5
J	sports drink	COLD	0.5	0.5	0.0	0.0	0.5

Table 2 Parameters of utility value: outdoor.

item	type	COLD /HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	0.5	-0.5	0.5	0.5
B	coffee	HOT	0.5	0.5	-1.0	-0.5	-1.0
C	café au lait	COLD	0.0	-1.0	0.0	0.5	1.0
D	green tea	COLD	1.0	0.5	0.5	0.0	0.5
E	tea	COLD	0.5	0.0	1.0	1.0	0.5
F	enegy drink	COLD	-0.5	-0.5	-1.0	0.0	0.5
G	soda	COLD	0.5	0.5	0.0	1.0	1.0
H	mineral water	COLD	1.0	-0.5	0.0	0.5	0.0
I	tea	HOT	-0.5	-1.0	0.5	-0.5	-1.0
J	sports drink	COLD	0.5	1.0	0.5	0.5	1.0

where $V^0_{q_i}$ is the gender-independent constant of utility value, $V^M_{q_i}/V^F_{q_i}$ is the gender-dependent constant, $\beta^M_{q_i}/\beta^F_{q_i}$ is the gender-dependent coefficient, and T_j is the temperature parameter in state s_j, such as $high \rightarrow 1$, $middle \rightarrow 0$, $low \rightarrow -1$.

These parameters are decided by characteristics of locations and products[†]. Each product is classified into types of drink (coffee, green tea, etc.) with attribute COLD or HOT, and the feature of each product is reflected in the values of parameters[††]. The parameter $V^0_{q_i}$ is independent of the location, but $V^M_{q_i}, V^F_{q_i}, \beta^M_{q_i}, \beta^F_{q_i}$ are decided by characteristics of the location. $\beta^M_{q_i}, \beta^F_{q_i}$ are coefficients of temperature's contribution to the utility values. When these values are positive, they indicate that these products become easy to be sold as the temperature rises. The parameter values we adopted for the simulation in this paper are shown in Tables 1–3.

Parameter estimation of the multinominal logit model can be done independently of the assortment optimization. In the simulation, we use artificial values for parameters determined by the following way. We first consider a variety

[†] The values of these parameters are omitted due to space limitations.

[††] For example, the utility value for men is higher for coffee, more COLD products are sold as the temperature rises, etc.

Table 3 Parameters of utility value: school.

item	type	COLD /HOT	V^0	V^M	V^F	β^M	β^F
A	coffee	COLD	1.0	0.0	-0.5	0.0	0.5
B	coffee	HOT	0.5	-1.0	-1.0	-1.0	-1.0
C	café au lait	COLD	0.0	0.0	0.5	-0.5	0.0
D	green tea	COLD	1.0	1.0	1.0	0.0	-0.5
E	tea	COLD	0.5	0.5	1.0	0.5	0.0
F	enegy drink	COLD	-0.5	-0.5	-1.0	0.0	0.0
G	soda	COLD	0.5	0.5	0.0	1.0	0.5
H	mineral water	COLD	1.0	0.0	-0.5	0.0	0.5
I	tea	HOT	-0.5	-1.0	0.5	-0.5	-1.0
J	sports drink	COLD	0.5	1.0	0.5	1.0	0.5

of real products that are for summer/winter, indoor/outdoor, and male/female. Next we assign values to each parameter that seem reasonable from qualitative point of view.

(6) Transition Probability

The transition probabilities $\delta(s(t+1)|s(t))$ are decided by characteristics of locations. However, we assume that a transition of gender ratio and temperature are independent in any location. At outdoor, the transition probability of gender ratio to other states is large so that the variation of the ratio is increased. While at school, the probability of staying in the current state is increased because we consider that the variation is small. The probability in office is adopted an intermediate value. The parameter values for the simulation are shown in Tables 4, 5.

Similarly to parameters in the multinominal logit model, the state transition probabilities should be estimated from empirical data. However, we here assume that the probabilities are already known. In this paper, we aim to show the proposed approach works if all the parameter values are known. If this is not true, then there is no sense to incorporate estimation of unknown factors in the model. Estimating unknown factors during the assort optimization process remains as future work.

(7) Policy

In this simulation, we define the policy set Π includes $M = 8$ policies in Table 6. The assortment constraint is the most strict one that allows all of these policies.

6.2 Models and Evaluation

To evaluate and compare the proposed model, we used the baseline model and the comparative models. As the baseline, we calculated the theoretical upper bound Eq. (8) and the feasible maximum value Eq. (9). We adopted comparative models as heuristic approaches with fixed actions for every time step.

- Fix action = 0: Leave the initial assortment (6 products A, . . . , F) unchanged at all.
- Fix action = 1, 2, 7: Select the policy π^1, π^2, π^7 for every time step. In these cases, it is assumed that the agent considers the gender ratio of the vending machine to be constant at $\{5:5\}$ in the initial state. Since other policies showed lower performance than that by π^1, π^2, π^7, we have picked up these policies in the presentation of graphs and tables.

Table 4 Transition probability of gender ratio.

office		$s(t+1)$		
$\delta(s(t+1)\|s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.60	0.30	0.10
	$\{5:5\}$	0.25	0.50	0.25
	$\{2:8\}$	0.15	0.35	0.50

outdoor		$s(t+1)$		
$\delta(s(t+1)\|s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.45	0.35	0.20
	$\{5:5\}$	0.30	0.40	0.30
	$\{2:8\}$	0.20	0.30	0.50

school		$s(t+1)$		
$\delta(s(t+1)\|s(t))$		$\{8:2\}$	$\{5:5\}$	$\{2:8\}$
$s(t)$	$\{8:2\}$	0.80	0.15	0.05
	$\{5:5\}$	0.15	0.70	0.15
	$\{2:8\}$	0.10	0.15	0.75

Table 5 Transition probability of temperature.

		$s(t+1)$		
$\delta(s(t+1)\|s(t))$		$\{high\}$	$\{middle\}$	$\{low\}$
$s(t)$	$\{high\}$	0.35	0.50	0.15
	$\{middle\}$	0.20	0.60	0.20
	$\{low\}$	0.25	0.45	0.30

Table 6 Policy set Π.

Policy	Detail
π^0	Do nothing
π^1	Exchange one of products in columns which has the lowest utility value for one of products not in columns which has the highest utility value.
π^2	Exchange one of products in columns which has the lowest utility value for one of products not in columns selected randomly.
π^3	Exchange one of products in columns selected randomly for one of products not in columns which has the highest utility value.
π^4	Exchange one of product in columns selected randomly for one of products not in columns selected randomly.
π^5	Add one column for one of products in columns which has the highest utility value, and remove one of that which has the lowest utility value.
π^6	Reduce one column from the multi-column products has the lowest utility value, and add one of products not in columns which has the highest utility value.
π^7	Exchange two products in columns which have the lowest utility values for two products not in columns products which have the highest utility values.

In Sect. 5, we calculate the total expected reward $E_{t\to t+\tau}(\pi_t^{k_0})$ at each time t by Eq. (14), where τ is the length of the future time steps. In the simulation, we set up the following three proposed models at the policy decision Eq. (15):

- Proposed model $\mathcal{M}_1$: $E_{t\to t+1}(\pi_t^{k_0})$ $(\tau = 1)$.

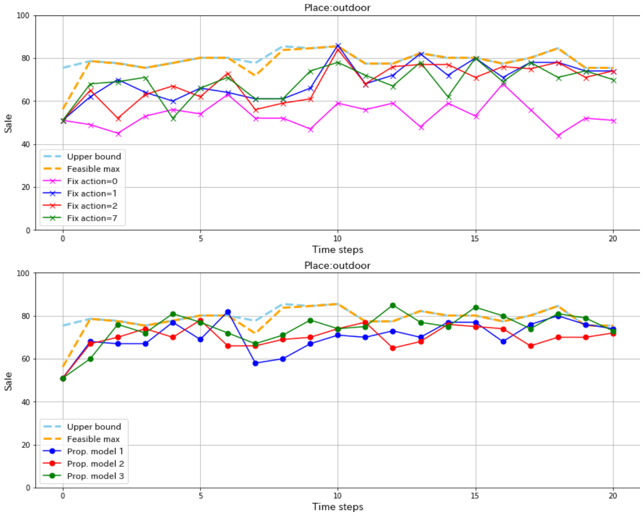

Fig. 3 Example of simulation.

- Proposed model $\mathcal{M}_2$: $E_{t\to t+2}(\pi_t^{k_0})$ $(\tau = 2)$.
- Proposed model $\mathcal{M}_3$: $E_{t\to t+3}(\pi_t^{k_0})$ $(\tau = 3)$.

6.3 Results

For the 2 baselines, 4 comparative models and 3 proposed models, 50 simulations were conducted at each of the 3 location. The length of each simulation is 20 steps, the number of consumers is $N = 100$ and the discount rate is $\gamma = 0.9$ in Eq. (14). The computer environment is as follows: Macbook Pro 15-inch, CPU 2.2 GHz 6cores intel Core i7, RAM 16 GB, Python 3.6.13. The execution time for one simulation was around one minute ($\mathcal{M}_1$), 10 minutes ($\mathcal{M}_2$), and 260 minutes ($\mathcal{M}_3$). Examples of simulation results at outdoor are shown in Fig. 3[†].

Tables 7, 8, and 9 show the summary of "sales" and "sold out" in 50 simulations for each model. Here, the amount of "sold out" cases are counted for the number of consumers who wanted to purchase a product but could not because of sold out in columns.

In these tables, improvement efficiency of assortment for models are evaluated by "Sales(Ave.)/UB", where UB is the theoretical upper bound. This means the ratio of how close the expected sales value is to the upper bound. In all locations, these ratios of "Feasible max" are almost close to 100%. It means that if the agent knows all the state in the future and can make the best exchange of products based on the information, the expected sales that the agent can obtain is almost close to the upper bound.

Table 7 Result summary of office.

Baseline and Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Sales(Ave.) / UB(*)
Upper bound	77.81*	2.45	-	100.0%
Feasible max	77.81	2.45	-	100.0%
Fix action=0	67.28	5.62	1.22	86.5%
Fix action=1	72.34	6.61	1.49	93.0%
Fix action=2	70.53	6.27	1.31	90.6%
Fix action=7	71.21	6.83	1.43	91.5%
$\mathcal{M}_1$	73.55	6.71	1.52	94.5%
$\mathcal{M}_2$	73.49	6.85	1.52	94.4%
$\mathcal{M}_3$	73.43	6.59	1.48	94.4%

Table 8 Result summary of outdoor.

Baseline and Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Sales(Ave.) / UB(*)
Upper bound	79.24*	3.23	-	100.0%
Feasible max	78.94	3.41	-	99.6%
Fix action=0	54.20	5.37	0.99	68.4%
Fix action=1	70.50	8.73	2.69	89.0%
Fix action=2	67.87	8.91	2.58	85.6%
Fix action=7	70.15	8.16	2.90	88.5%
$\mathcal{M}_1$	72.54	6.73	3.18	91.5%
$\mathcal{M}_2$	72.15	6.19	3.34	91.1%
$\mathcal{M}_3$	72.05	6.30	3.11	90.9%

Table 9 Result summary of school.

Baseline and Model	Sales (Ave.)	Sales (Std.)	Sold out (Ave.)	Sales(Ave.) / UB(*)
Upper bound	80.39*	3.19	-	100.0%
Feasible max	80.25	3.25	-	99.8%
Fix action=0	55.30	7.52	2.68	68.8%
Fix action=1	73.63	7.54	5.16	91.6%
Fix action=2	70.82	7.51	4.78	88.1%
Fix action=7	72.53	7.31	5.10	90.2%
$\mathcal{M}_1$	74.40	6.72	4.96	92.5%
$\mathcal{M}_2$	75.50	6.28	5.05	93.9%
$\mathcal{M}_3$	75.51	6.55	4.92	93.9%

In comparative models, the ratio of fix action = 0 is around 65–85%, and that of fix action = 1, 2, 7 are around 90% on each locations. On the other hand, the ratios of the proposed models 1 to 3 are over 90% in all locations, especially at office and school the ratios are 92–94%. Therefore, we can conclude that the proposed models are effective in improving sales compared to the comparative models. Moreover, when state probabilities of state transitions are small like in school, we observe that the performance increases as the future time steps τ in estimation increases. However, the improvement is not very large. For office and outdoor, $\tau = 1$ seems enough.

7. Discussion

If the proposed model works properly, we expect that sales approaches the upper bound. One of possible improvements in the proposed model is to increase τ in Eq. (14), that is, the future time steps for summing up expected rewards. In this paper, we have used $\tau = 1, 2, 3$, but we expect that sales approaches to the upper bound by increasing τ to $4, 5, \ldots$. However, as τ increases, the number of states that must be

[†] In Fig. 3, there are cases where the sales of the proposed models exceed that of the theoretical upper bound at some time steps. This is because the sales of proposed models are calculated by stochastic simulation at each time step while that of the theoretical upper bound is summed up the expected value of sales. Therefore, the proposed models are evaluated by the average values of the total sales in 50 simulations.

calculated increases exponentially. By this reason, we could not try to use larger numbers for τ in the simulation. There are other ways for the improvements, such as increase in the types and patterns of policies.

Next we consider conditions in which the proposed model performs more effectively. The conditions may include the case that the numbers of products and columns are large enough, since when the numbers are small, the effect of future estimation reduces because the assortment reaches the optimal one immediately. When the absence of IoT devices and the current sales data cannot be obtained in real time, it is necessary to estimate the current state based on the limited data and to make accurate plans for stock replenishment and assortment exchanges. In this case, the proposed methods are reasonable and effective. When the environment and the consumer preferences of vending machines frequently change, methods based on demand forecasting from past history on sales can not keep up with the changes. Since the proposed method works adaptively to the current state, it works well even in such cases.

On the other hand, one of less effective cases is that the optimal assortment does not differ significantly on the environment and the consumer preferences. In such a case, the calculation of future expected reward does not necessarily work well.

There is another issue to be studied. Although the state is partially observable in the proposed model, the agent knows detailed information on consumer utility value, and transition probability of vending machine state. The information may be unknown in actual situations and has to be estimated through past history of observations. Incorporating this factor into the model remains as future work.

What we have shown in this paper is that there is a case in which the proposed method outperforms simple policies. Of course, the results will change when the parameter values are changed. From the above discussion, however, we can claim the following properties hold. Compared to simple policies,

- if the diversity of products increases, then the POMDP-based method works well;
- if the volatility of state change increases, then the POMDP-based method works well.

8. Conclusion

We have proposed an assortment optimization model for vending machines that enables the worker to plan an appropriate assortment of products. In model simulation with some assumptions and numerical parameters, we have achieved improvement on sales up to 2-3 points (in percentage ratio to the theoretical upper bound) compared to heuristic methods. The proposed models outperform the heuristic methods under the same conditions. As a result, we have confirmed the effectiveness of estimation for the expected future reward.

There are several remaining works. First, it is necessary to verify the effects of simulation execution under different conditions (the number of products, the number of columns, the number of stocks, the number of consumers, etc.). In addition, in order to bring the model closer to the actual assortment problem, we proceed to formulate to the case that informations that is known by the agent is limited.

References

[1] S.A. Smith and N. Agrawal, “Management of multi-item retail inventory systems with demand substitution,” Operations Research, vol.48, no.1, pp.50–64, 2000.

[2] P. Rusmevichientong, Z.J.M. Shen, and D.B. Shmoys, “Dynamic assortment optimization with a multinomial logit choice model and capacity constraint,” Technical Report, 2008.

[3] V. Gaur and D. Honhon, “Assortment planning and inventory decisions under a locational choice model*,” Technical Report, 2005.

[4] R. Chan, Z. Li, and D. Matsypura, “Assortment optimisation problem: A distribution-free approach,” Omega, vol.95, 102083, June 2019.

[5] A.G. Kök and M.L. Fisher, “Demand estimation and assortment optimization under substitution: Methodology and application,” Operations Research, vol.55, no.6, pp.1001–1021, 2007.

[6] K.J. Lancaster, “A new approach to consumer theory,” The Journal of Political Economy, vol.74, no.2, pp.132–157, 1966.

[7] D.L. McFadden, “Conditional logit analysis of qualitative choice behavior,” Frontiers in Econometrics, vol.8, pp.105–142, 1973.

[8] H.C.W.L. Williams, “On the formation of travel demand models and economic evaluation measures of user benefit,” Environment and Planning A: Economy and Space, vol.9, no.3, pp.285–344, 1977.

[9] R. Anupindi, M. Dada, and S. Gupta, “Estimation of consumer demand with stock-out based substitution: An application to vending machine products,” Marketing Science, vol.17, no.4, pp.406–423, Nov. 1998.

[10] R.D. Luce, Individual Choice Behavior, John Wiley, Oxford, England, 1959.

[11] O. Elshiewy, D. Guhl, and Y. Boztug, “Multinomial logit models in marketing — From fundamentals to state-of-the-art,” Marketing ZFP, vol.39, no.3, pp.32–49, 2017.

[12] X.-H. Chen, A.P. Dempster, and J.S. Liu, “Weighted finite population sampling to maximize entropy,” Biometrika, vol.81, no.3, pp.457–69, 1994.

[13] L.P. Kaelbling, M.L. Littman, and A.R. Cassandra, “Planning and acting in partially observable stochastic domains,” Artificial Intelligence, vol.101, no.1-2, pp.99–134, 1998.

Gaku Nemoto received the B. S. degree in physics from Tohoku University in 2000, the M. S. degree in physics from Chiba University in 2002, and the M. S. degree in information science from Japan Advanced Institute of Science and Technology (JAIST) in 2017. He is currently a Ph.D. student at JAIST. He joined Intage Technosphere Inc., Tokyo, Japan in 2002 and has been engaged in developments of information systems applied forecasting, optimization, and machine learning. His research interests include applications of machine learning and optimization for industry.

Kunihiko Hiraishi received from the Tokyo Institute of Technology the B. E. degree in 1983, the M. E. degree in 1985, and D. E. degree in 1990. He is currently a professor at School of Information Science, Japan Advanced Institute of Science and Technology. His research interests include discrete event systems and formal verification. He is a member of the IEEE, IPSJ, and SICE.