Categorization of material quality using a model-free reinforcement learning algorithm

Categorization of material quality using a model-free reinforcement learning algorithm

Annapoorni Mani[1], Shahriman Abu Bakar[2], Pranesh Krishnan[3], Sazali Yaacob[4]

[1]Ph.D. Research Scholar, British Malaysian Institute Universiti Kuala Lumpur, Gombak, Selangor, Malaysia
[2]Professor Madya, Faculty of Mechanical Engineering Technology, Universiti Malaysia Perlis, Perlis, Malaysia
[3]Talent and Content AI Instructor, Skymind Holdings Berhad, Kuala Lumpur, Malaysia
[4]Professor and Dean, Malaysian Spanish Institute Universiti Kuala Lumpur, Kulim, Kedah, Malaysia

Email: shahriman@unimap.edu.my; annapoorni.pranesh@gmail.com

Abstract. Reinforcement learning is the most preferred algorithms for optimization problems in industrial automation. Model-free reinforcement learning algorithms optimize for rewards without the knowledge of the environmental dynamics and require less computation. Regulating the quality of the raw materials in the inbound inventory can improve the manufacturing process. In this paper, the raw materials arriving at the incoming inspection process are categorized and labeled based on their quality through the path traveled. A model-free temporal difference learning approach is used to predict the acceptance and rejection path of raw materials in the incoming inspection process. The algorithm presented eight routes paths that the raw materials could travel. Four pathways correspond to material acceptance, while the rest lead to material refusal. The materials are annotated using the total scores acquired in the incoming inspection process. The materials traveling on the ideal path (path A) get the highest total score. The rest of the accepted materials in the acceptance path have a 7.37% lower score in path B, whereas path C and path D get 37.28% and 42.44% lower than the ideal approach.

1. Introduction

The objective of reinforcement learning (RL), a branch of artificial intelligence, is to create completely autonomous entities that interact with their surroundings. During the training, they undergo a trial-and-error approach and improve over time to learn the optimal behaviors [1]. There are three approaches to handle an RL problem. Strategies based on value functions, policy search approaches, and a hybrid actor-critic approach uses both value functions and policy search. Algorithms based on value functions estimate the value (expected mean) of being in each state. Unlike value functions, policy search techniques do not require maintaining a value function model and instead search for the best policy directly. RL algorithms are also grouped into model-based and model-free approaches. A model in reinforcement learning is often referred to as the transition dynamic of the environment: state action pair and learning rate. Model-free in RL means that the agent uses just real-world experience to maximize

the predicted reward rather than a model or prior experience. It has no idea what state it will be in after doing a task; all it cares about is the reward associated with the state/state-action.

Recall risks have grown dramatically over the last decade, making product-related risk one of the most severe threats confronting organizations today. According to a recent survey by Allianz Global Corporate & Specialty (AGCS), the automobile sector is the most affected by product recalls, followed by the food and beverage sector, and then IT/electronics. Between 2012 and the first half of 2017, AGCS examined 367 insurance company's product recall claims from 28 countries across 12 industrial sectors for its report [2].AGCS discovered that the most common source of recall claims is a defective product or work, followed by product contamination. A significant incident costs more than US$12 million on average, with the expenses of the most significant occurrences substantially exceeding this figure. Ten incidents account for more than half of all losses. Irrespective of the industries, a product recall is a significant concern that unravels the trust in the product and financial loss to the company. The growing number of product recalls due to poor product quality is a significant concern in the manufacturing industry. The trajectory of the material in the incoming inspection is tracked, thereby estimating the quality of the material.

The rest of this paper is laid out as follows. Section 2.1 expresses the temporal distancing learning approach. Based on the part, incoming inspection data contributes to the weights are estimated from the timestamps data in the primary dataset. Section 2.2 describes the development of the Markov decision process (MDP) model. Section 2.3 details the Q-learning-based temporal difference algorithm to calculate the trajectory. Section 3 specifics the results and discusses the comparison of two optimized models.

2. Model Development

The path of the raw material traveling inside the incoming inspection path is modeled to estimate the quality of the incoming raw materials. The trajectory of the raw material ends in two end states, namely, pack and store and return.

2.1. Incoming inspection case study

Original Equipment Manufacturers (OEM) procure material from outside have agreements with the vendor to supply material or products as per the required dimension[5]&[6]. These agreements, in general, specify the quality of the component both in terms of aesthetics (appearance, color, texture), fit, and function. In any OEM, the initial process is to make the incoming raw materials undergo a set of Quality assurance process known as Incoming inspection (IIP). This process ensures that the material from the vendor meets the expected quality defined by the OEM and accepted by the vendor[7] &[8]. In this case study, seven stations in a typical IIP are considered, namely: Package and store (PI), Visual Inspection (VI), Gauge inspection (GI), Rework (RW1), Rework (RW2), Return to vendor (RT) and Pack & store (PS).

2.2. Finite-state diagram (FSM)

To represent the functions and to describe the possible traversal of the raw materials between the states an FSM. The FSM diagram decomposes the functions of the seven substations of the input inspection.

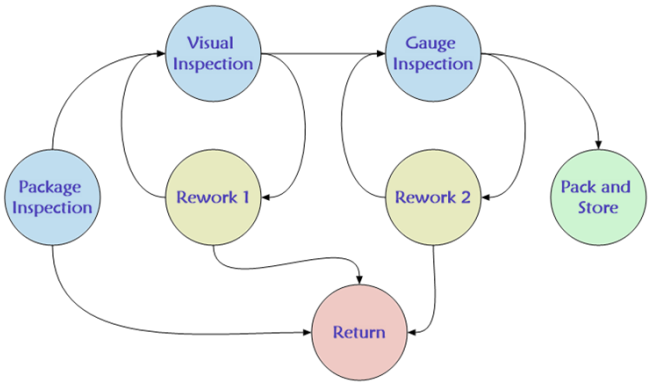

Figure 1. Finite state machine diagram of the seven substations.

Figure 1 depicts the traversal of the raw materials, and Table 1 details the substation and the state transitions.

Table 1. State transition of the incoming inspection of raw material.

	Previous	*Current*	*Input*	*Next*	*Output*	*Description*
PI	-	PI	OK	VI	Conformities	Move to VI
	-	PI	Not OK	RT	Error in packaging	Return to supplier
VI	PI	VI	OK	GI	Conformities	Move to GI
	PI	VI	Not OK	RW1	Minor change	Send to Rework
	RW1	VI	OK	GI	Conformities	Move to GI
	RW1	VI	Not OK	RT	Nonconformities	Return to Supplier
GI	VI	GI	OK	PS	Conformities	Pack and store
	VI	GI	Not OK	RW2	Minor change	Send to Rework
	RW2	GI	OK	PS	Conformities	Pack and Store
	RW2	GI	Not OK	RT	Nonconformities	Return to supplier
RW1	VI	RW1	OK	VI	Minor correction	Move to VI
	VI	RW1	Not OK	RT	Major change	Return to Supplier
RW2	GI	RW2	OK	GI	Minor correction	Move to GI
	GI	RW2	Not OK	RT	Major change	Return to Supplier
PS	GI	PS	OK	-	Conformities	Finish
RT	-	RT	OK	-	Nonconformities	Return to Supplier

2.3. Temporal difference learning

One of the most often used techniques for optimum control is the temporal difference algorithm. There are two primary algorithms in TD learning. 1) State Action Reward State Action (SARSA), and 2) Q-learning. SARSA is an on-policy algorithm, whereas Q-Learning is an off-policy method. The most significant distinction between SARSA and Q-learning is that the highest reward for the following state is not always utilized to update the Q-values. Q-learning is favored over SARSA for the given case study because SARSA learns the safe path, whereas Q–learning learns the optimum path accurately [3-4].

2.3.1. Q – Learning

Q-learning RL algorithm, unlike model-based RL algorithms, does not require any model to handle the problems. There exists a state-action pair in every state. The value of this pair depends on the functionality of the state. In this algorithm, more emphasis is given to the state-action pair rather than the state itself. This RL algorithm is used to figure out how valuable action is in each situation. When

using a model-free approach, the p-value is not used or estimated in any way. During the learning process, Q-learning estimates the value function q(s, a) by interacting with the environment (performing actions and getting rewards). However, it does not know or keep track of the dynamics (i.e., p) of the environment, which is why it is called a model-free method.

2.4. Classification of material quality using the TD method.

This section demonstrates the use of Q-learning to estimate the possible next state. Every raw material that approaches the incoming inspection moves to all the substations for investigation. Depending on the acceptance and rejection at each stage, the trajectory of the material grows. The trajectory is used for grading the material for its quality. The Q-learning-based TD algorithm is employed to achieve this objective.

Q – learning for prediction

$$V(s) = V(s) + \alpha(r + \gamma V(s') - V(s))$$

where V(s)= new estimate
α = learning rate
r = reward
γ = discounted factor
$V(s')$= old estimate

Q- learning – Control or Optimization

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \times maxQ(s',a') - Q(s,a))$$

$Q(s,a)$ = value of current state
α = learning rate
r = reward
γ = discount rate
$Q(s',a')$= value of next state

Figure 2 represents the flowchart of the Q-learning. During the estimation, the value of the following state V' (s) is initialized to zero.

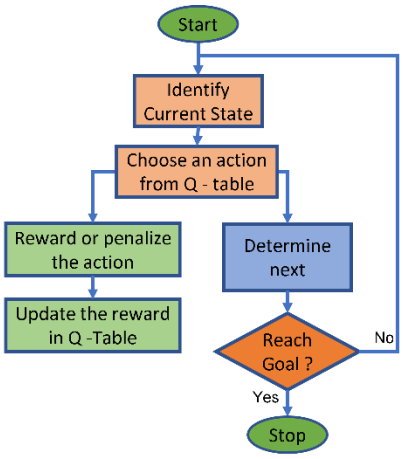

Figure 2. Flowchart of Q-learning algorithm.

During the estimation of the value of the state, the learning rate or step size is responsible for the convergence, and hence the learning rate (α) is chosen to be 0.1 for a smooth convergence based on the guidelines [3]. For every correct decision, the agent gets a reward, whereas, for every undesired action, the agent is penalized. The values for the rewards (r) and the penalization are represented as positive and negative integers, respectively. The discount factor (γ) is a concept that governs the relative relevance of immediate and future rewards. The discount factor and the reward and penalization for each

move are tabulated below in Table 2. The choice of the reward and discount values were chosen arbitrarily.

Table 2. Moves, reward, and discount factors.

Moves	Reward	Discount Factor
PI-VI	5	0.7
Pi-RT	-3	0.1
VI-GI	5	0.8
VI-RW1	-2	0.4
GI-PS	3	0.8
GI-RW2	-4	0.4
RW1-VI	1	0.6
RW1-RT	-2	0.1
RW2-GI	0.5	0.6
RW2-RT	-3	0.1

The numerical values that the agent obtains for completing some action at some state(s) in the environment are referred to as rewards. Based on the agent's activities, the numerical value might be positive or negative. In real life, we are more concerned with maximizing the cumulative reward (all the rewards the agent receives from the environment) than the reward the agent receives from the present condition (also called immediate reward). Returns refer to the overall amount of reward the agent receives from the environment.

3. Results and Discussion

The findings of the study are presented in this section. Q-learning algorithm for two situations: 1) Estimation of the quality index for the Pack and Store path and 2) quality index estimation for the return path. Table 3 demonstrates the four possible paths traveled by the accepted raw material. The substations are referenced as PI for Package Inspection, VI for Visual Inspection, GI for Gauge Inspection, PS for Pack and Store, RW1 and RW2 for Rework 1 and Rework 2, respectively. Figure 3 depicts the temporal graph representation of the acceptance trajectory.

a) **PI → VI → GI → PS**
b) **PI → VI → RWI → VI → GI → PS**
c) **PI → VI → GI → RW2 → GI → PS**
d) **PI → VI → RWI → VI → GI → RW2 → GI → PS**

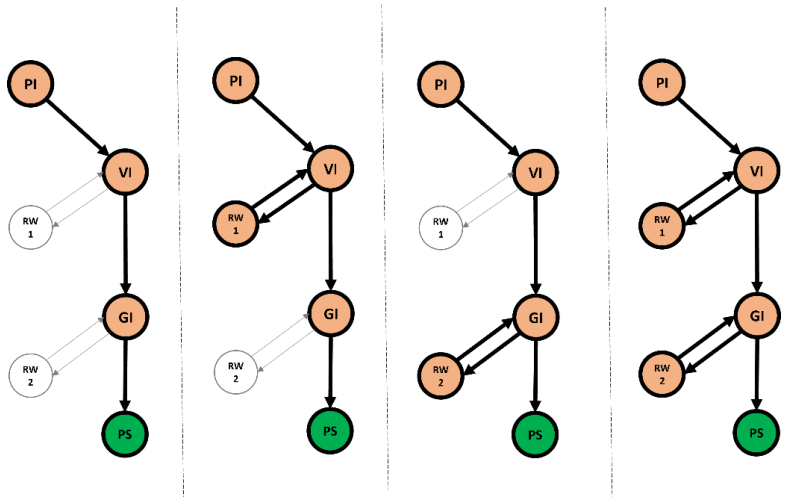

Figure 3. Temporal graph representation of the acceptance trajectory

Table 3. Estimation of the quality index using temporal difference for the Pack and Store path.

State	*Action*	*Next state*	*V(s)*	*Reward*	*State Value*	*Discounted Reward*	*Value of the previous state*
PI	1	VI	0	5	0.8	0.7	0.556
VI	1	GI	0.556	5	0.8	0.8	1.0644
GI	1	PS	1.0644	5	0.6	0.8	1.50596
PI	1	VI	0	5	0.8	0.7	0.556
VI	0	RW1	0.556	-2	0.5	0.4	0.3204
RW1	1	VI	0.3204	1	0.8	0.6	0.43636
VI	1	GI	0.43636	5	0.9	0.8	0.964724
GI	1	PS	0.964724	5	0.8	0.8	1.4322516
PI	1	VI	0	5	0.8	0.7	0.556
VI	1	GI	0.556	5	0.9	0.8	1.0724
GI	0	RW2	1.0724	-4	0.6	0.6	0.60116
RW2	1	GI	0.60116	0.5	0.9	0.6	0.645044
GI	1	PS	0.645044	3	0.8	0.8	0.9445396
PI	1	VI	0	5	0.8	0.7	0.556
VI	0	RW1	0.556	-2	0.5	0.4	0.3204
RW1	1	VI	0.3204	1	0.8	0.8	0.45236
VI	1	GI	0.45236	5	0.9	0.8	0.979124
GI	0	RW2	0.979124	-4	0.6	0.4	0.5052116
RW2	1	GI	0.5052116	0.5	0.9	0.6	0.55869044
GI	1	PS	0.55869044	3	0.8	0.8	0.866821396

Consider the results shown in Table 3. The total score obtained for acceptance path A is **1.50596**. This score is labeled for every raw material which follows this path. Similarly, for acceptance path B, the score is reduced to **1.4322516** as a penalty (negative reward) for undergoing a rework for visual correction. In acceptance path C, the score further reduces to **0.9445356** for a dimensional correction. The difference between acceptance paths B and C is due to the significance of the next state. Technically, the visual correction is less significant than dimension correction. When the raw material undergoes both the visual and dimensional correction, the score is now **0.866821396**. The score for the raw materials traversing through the ideal path (path A) is 1.50596. The score for the raw materials following path B is **7.37%** lower than the ideal path. Path C and D are **37.28%** and **42.44%,** respectively lower than the ideal path. Therefore, any irregularities in the dimension pave the way for more penalization. Every raw material now gets a label after the incoming inspection process signifying the trajectory and thus the quality of the accepted material. Figure 4 showcases the temporal graph representations of the rejection trajectory.

Table 3 lists the four possible paths traversed by the raw materials to reach the Return (RT).

e) **PI → RT**
f) **PI → VI → RWI → RT**
g) **PI → VI → GI → RW2 → RT**
h) **PI → VI → RWI → VI → GI → RW2 → RT**

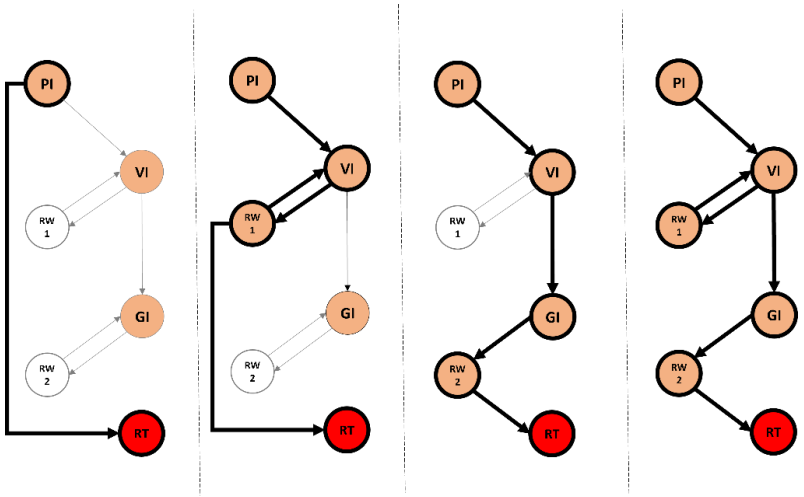

Figure 4. Temporal graph representation of the rejection trajectory

Table 4. Estimation of the quality index using temporal difference for the Return-path.

State	*Action*	*Next state*	*V(s)*	*Learning rate*	*Reward*	*State Value*	*Discounted Reward*	*Value of the previous state*
PI → RT								
PI	1	RT	0	0.1	-3	0.1	0.1	-0.299
PI → VI → RWI → RT								
PI	1	VI	0	0.1	5	0.8	0.7	0.556
VI	0	RW1	0.556	0.1	-2	0.5	0.4	0.3204
RW1	0	RT	0.3204	0.1	-2	0.1	0.1	0.08936
PI → VI → GI → RW2 → GI → RT								
PI	1	VI	0	0.1	5	0.8	0.7	0.556
VI	1	GI	0.556	0.1	5	0.9	0.8	1.0724
GI	0	RW2	1.0724	0.1	-4	0.6	0.4	0.58916
RW2	1	RT	0.58916	0.1	-3	0.1	0.1	0.231244
PI → VI → RWI → VI → GI → RW2 → GI → RT								
PI	1	VI	0	0.1	5	0.8	0.7	0.556
VI	0	RW1	0.556	0.1	-2	0.5	0.4	0.3204
RW1	1	VI	0.3204	0.1	1	0.8	0.6	0.43636
VI	1	GI	0.43636	0.1	5	0.9	0.8	0.964724
GI	0	RW2	0.964724	0.1	-4	0.6	0.4	0.4922516
RW2	1	RT	0.4922516	0.1	-3	0.1	0.1	0.14402644

The results for the rejection path are shown in Table 4. Packaging inspection has an essential role to play in the incoming inspection process. Irrespective of the material quality, the package is judged for adherence to the packaging standards. Upon failure, the raw materials are directly rejected and sent to the supplier with a total score of **-0.299**. In path F, the subsequent rejection stage can occur at Rework station 1 (RW1) due to visual nonconformities with a total rejection score of **0.08936**. Fit and function inspection using a gauge plays a vital role in checking dimensional conformity. The raw materials failing at this stage get the lower total score of **0.231244.** If a raw material fails in both visual and gauge inspection is awarded the least total score of **0.14402644.** The supplier usually scraps these materials.

4. Conclusion

In this paper, an incoming inspection problem is considered as a reinforcement learning task. A Temporal difference learning approach predicts the acceptance and rejection path of raw materials in the incoming inspection process. The algorithm presented eight possible paths that the raw materials could travel. Four trajectories contribute to material acceptance, whereas the remaining paths lead to material rejection. The materials are labeled using the total scores obtained in the incoming inspection process. The materials traveling on the ideal path (path A) get the highest total score. The rest of the accepted materials have a 7.37% lower score in path B, whereas path C and path D get 37.28% and 42.44% lower from the ideal path.

Acknowledgments

The research has been carried out under the Malaysian Technical University Network (MTUN) Research Grant by Ministry of Higher Education of Malaysia (MOHE) under a grant number of (9028-00005) & (9002-00089) with the research collaboration with thanks to Center of Excellence Automotive & Motorsport and Faculty of Mechanical Engineering Technology, Universiti Malaysia Perlis (Malaysia) for their productive discussions and input to the research.

References

[1] Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A brief survey of deep reinforcement learning. arXiv preprint arXiv:1708.05866.
[2] Allianz Global Corporate & Specialty. Product recall: Managing the impact of the new risk landscape. Munich; 2017.
[3] Ravichandiran, S. (2018). Hands-on reinforcement learning with Python: master reinforcement and deep reinforcement learning using OpenAI gym and TensorFlow. Packt Publishing Ltd.
[4] Nazia Habib, (2019). Hands-On Q-Learning with Python: Practical Q-learning with OpenAI Gym, Keras, and TensorFlow. Packt Publishing Ltd.
[5] Pogorelov, V., & Ismayilov, R. I. O. (2017). Mathematical modelling of the process of quality control of construction products. In *MATEC Web of Conferences* (Vol. 106, p. 08009). EDP Sciences.
[6] Pyzdek, T. (1991). A Simulation of Receiving Inspection. Quality Engineering, 4(1), 9-19.
[7] Er, M., Astuti, H. M., & Pramitasari, D. (2015). Modeling and analysis of incoming raw materials business process: a process mining approach. International Journal of Computer and Communication Engineering, 4(3), 196.
[8] Mahendrawathi, E. R., Astuti, H. M., & Wardhani, I. R. K. (2015, June). Material movement analysis for warehouse business process improvement with process mining: a case study. In Asia-Pacific Conference on Business Process Management (pp. 115-127). Springer, Cham.