การจัดการวัตถุดิบคงคลังของร้านอาหาร โดยใช้ Proximal Policy Optimization

ณัฐวัฒน์ เอกธรรมนิตย์*[1] และ วรพล พงษ์เพ็ชร[2]
สถาบันบัณฑิตพัฒนบริหารศาสตร์ 148 ถนนเสรีไทย แขวงคลองจั่น เขตบางกะปิ กรุงเทพมหานคร 10240

Received: 3 February 2021; Revised: 22 April 2021; Accepted: 27 April 2021

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ในการจัดทำขึ้นเพื่อศึกษาหลักการ Proximal Policy Optimization ของกระบวนการ Reinforcement Learning ในการสร้างแบบจำลองการพยากรณ์จำนวนการสั่งวัตถุดิบของร้านอาหาร เนื่องมาจากปัญหาการสั่งวัตถุดิบของร้านอาหารในแต่ละวัน เกิดความคลาดเคลื่อนจากจำนวนการใช้วัตถุดิบจริงสูง วัตถุดิบที่เหลือจึงกลายเป็นขยะอาหาร ทำให้เกิดการหมักและก่อตัวเป็นก๊าซมีเทนลอยขึ้นไปทำลายโอโซนในชั้นบรรยากาศ ซึ่งเป็นสาเหตุหลักที่ก่อให้เกิดปัญหาภาวะเรือนกระจก โดยแบ่งการศึกษาออกเป็น 2 รูปแบบหลักๆ คือ 1. แบบจำลองแบบหนึ่งแอตทริบิวต์ 2. แบบจำลองแบบหลายแอตทริบิวต์ และชุดข้อมูลจำลองที่ใช้มาจากการจำลองโดยอาศัยหลักการจำลองข้อมูลแบบแจกแจงปกติ ส่วนการประเมินประสิทธิภาพของแบบจำลองจะใช้เครื่องมือ 3 อย่าง คือ F-Statistics ,R-Square และRMSE ในการศึกษาครั้งนี้แต่ละแบบจำลองจะมีการเรียนรู้ทั้งหมด 12 ล้านไทม์สเต็ป ผลจากการศึกษาในงานวิจัยนี้ พบว่า ในการเรียนรู้ของแบบจำลองทั้งหมด 12 ล้านไทม์สเต็ป แบบจำลองแบบหลายแอตทริบิวต์ จะเกิดการลู่เข้าหาค่าที่เหมาะสมของแบบจำลอง ได้เร็วกว่าแบบจำลองแบบหนึ่งแอตทริบิวต์ และประสิทธิภาพของค่าความถูกต้องในการพยากรณ์จำนวนการสั่งวัตถุดิบ อยู่ที่ร้อยละ 82 ของค่าการพยากรณ์ทั้งหมด ซึ่งจำนวนค่าที่ใช้ในการทดสอบทั้งหมด คือ 1,000 ค่า จากงานวิจัยนี้ทำให้ทราบถึงแนวทางในการนำหลักการ Proximal Policy Optimization ของกระบวนการ Reinforcement Learning ไปสร้างแบบจำลองการพยากรณ์จำนวนการสั่งวัตถุดิบ ให้สามารถพยากรณ์จำนวนวัตถุดิบที่จะสั่ง ได้ใกล้เคียงกับจำนวนวัตถุดิบที่ใช้จริงมากที่สุด และสามารถลดจำนวนขยะอาหารให้มีจำนวนเหลือน้อยลง

คำสำคัญ: การเรียนแบบเสริมกำลัง, การเรียนรู้ของเครื่อง, วัตถุดิบคงคลัง

* Corresponding author. E-mail: tsr.nutthawat@gmail.com
[1] นักศึกษาปริญญาโท สาขาวิชาการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์
[2] อาจารย์ สาขาวิชาการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

Proximal Policy Optimization on Casual Restaurant Raw Material Stock

Nutthawat Ekthammanit*[1] and Worapol Pongpech[2]
National Institute of Development Administration (NIDA)
148 Serithai Road, Klong-Chan, Bangkapi, Bangkok THAILAND 10240

Received: 3 February 2021; Revised: 22 April 2021; Accepted: 27 April 2021

Abstract

This research focuses on the Proximal Policy Optimization Algorithm of Reinforcement Learning to make a forecasting model of raw material stock in restaurants. Due to the restaurant's daily raw material stock ordered. There is a high deviation from the number of raw material stock used. The rest of the raw material stock become food waste. This causes fermentation and the formation of methane gas to rise to destroy ozone in the atmosphere. Which is the main cause of the greenhouse effect. This research investigated a One-Attribute Model and a Multi-Attribute Model. The dataset used in this research is synthetic data that use the normal distribution theory to make it. The model's performance was assessed using

F-statistics, R-Square, and RMSE. We trained each model trained 12 million timesteps. The result showed that the Multi-Attribute Model would converge to the value optimization faster than the One-Attribute Model. We found that both models' accuracy is about 82 percent of the number of the test set where the number of the test set is 1,000. From this research, we can learn how to apply the Proximal Policy Optimization Algorithm of Reinforcement Learning make a forecasting model of raw material stock. To be able to forecast the number of raw material stock. As close as possible to the number of raw material stock used and it can reduce the number of food waste.

Keywords: reinforcement learning, machine learning, raw material stock

* Corresponding author. E-mail: tsr.nutthawat@gmail.com
1 Master's degree in Graduate School of Applied Statistics (GSAS), National Institute of Development Administration (NIDA)
2 Lecturer in Graduate School of Applied Statistics (GSAS), National Institute of Development Administration (NIDA)

1. บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันหลายประเทศได้ประสบปัญหาขยะที่เกิดจากอาหารขึ้นเป็นจำนวนมาก โดยจากสถิติในประเทศนิวซีแลนด์ ขยะอาหารที่เกิดจากร้านอาหารมีมากถึง 24,372 ตันต่อปี [10] โดยขยะอาหารเหล่านี้เป็นส่วนหนึ่งของต้นเหตุที่ทำให้เกิดปัญหาใหญ่ๆบนโลก เช่น ปัญหาภาวะเรือนกระจก ที่ขยะอาหารที่ถูกฝังกลบก่อให้เกิดก๊าซมีเทนไปทำลายชั้นโอโซนของโลก ทำให้อุณหภูมิของโลกจึงเพิ่มสูงขึ้น และปัญหาการขาดแคลนอาหาร ซึ่งจากงานศึกษาของ Food and Agriculture Organization of the United Nations ขยะอาหารคิดเป็น1 ใน3 ของจำนวนอาหารทั้งหมดผลิตได้ คิดเป็นน้ำหนักของขยะอาหารที่ถูกทิ้งไปเป็นจำนวน 1,300 ล้านตัน ทำให้มีผู้ที่ขาดแคลนอาหารจำนวน 820 ล้านคน [1]

โดยงานวิจัยของ Hu. Chen, and McCain [4] ในปี2004 ได้วิจัยเกี่ยวกับการวางแผนและจัดการในการพยากรณ์ระยะสั้นของร้านอาหารในคาสิโน โดยในงานวิจัยได้ใช้เครื่องมือทางสถิติและอนุกรมเวลาในการสร้างโมเดล คือ Naïve model, Single Moving Average model, Double Moving Average model, Single Exponential Smoothing model, Holt-Winters method และ Regression model ซึ่งผลลัพธ์ที่สามารถพยากรณ์ได้ค่าความถูกต้องมากที่สุด จะเป็นของ Double Moving Average model โดยในส่วนสุดท้ายของงานวิจัย Hu. Chen, and McCain [4] ได้แนะนำถึงการค้นหาหลักการสร้างโมเดลพยากรณ์รูปแบบใหม่ มาประยุกต์รวมกับหลักการพยากรณ์รูปแบบเดิม เพื่อให้มีประสิทธิภาพในการพยากรณ์ที่ดีขึ้น

ปี 2019 งานวิจัยของ Takashi Tanizaki et al. [8] ได้นำหลักการ Machine Learning แบบ Supervised Learning และหลักทางสถิติ มาใช้ในการพยากรณ์จำนวนความต้องสินค้าของร้านอาหาร โดยได้ทำทั้งหมด 4 วิธี คือ Bayesian Linear Regression, Boosted Decision Tree Regression, Decision Forest Regression และ Stepwise ซึ่งผลลัพธ์วิธี Bayesian Linear Regression, Decision Forest Regression และ Stepwise มีค่าอัตราการพยากรณ์ที่ไม่แตกต่างกัน แต่จะมีวิธี Boosted Decision Tree Regression ที่จะได้ค่าอัตราการพยากรณ์ที่ต่ำเล็กน้อย

ต่อมาในปี 2020 Takashi Tanizaki et al. [9] ได้ทำงานวิจัยเกี่ยวกับการจัดการร้านอาหาร โดยเน้นการพยากรณ์จากความต้องการของลูกค้า ที่อาศัย Machine Learning แบบ Supervised Learning มาสร้างโมเดลการพยากรณ์ โดยในงานวิจัยได้ใช้กระบวนการ Random Forest และ Random Forest Regression ซึ่งผลลัพธ์ที่ได้จากการพยากรณ์จากสินค้าคงคลังมีค่าอัตราการพยากรณ์อยู่ที่ 71% ซึ่ง Takashi Tanizaki กล่าวว่าเป็นค่าความถูกต้องที่ไม่สูง

จากงานวิจัยที่ได้กล่าวมาข้างต้น จึงนำมาสู่แรงจูงใจในการเลือกใช้กระบวนการ Reinforcement Learning ในการสร้างโมเดลการพยากรณ์ในงานวิจัยนี้ เนื่องจากการสร้างโมเดลแบบ Machine Learning สามารถแบ่งย่อยได้ออกเป็น 3 กระบวนการ คือ Supervised Learning, Unsupervised Learning และ Reinforcement Leaning โดยในงานวิจัยของ Takashi Tanizaki et al. [8], [9] ทั้งสองฉบับได้เลือกใช้กระบวนการสร้างโมเดลแบบ Supervised Learning ซึ่งผลลัพธ์ของความแม่นยำในการพยากรณ์ยังไม่สูง ส่วนกระบวนการแบบ Unsupervised Learning จะไม่เหมาะกับการพยากรณ์จำนวนการสั่งวัตถุดิบ จึงเหลือกระบวนการแบบ Reinforcement Learning ที่มีจุดเด่นในของการนำหลักการของ Markov Chain มาใช้ในการกำหนดการตัดสินใจในแต่ละครั้งโดยอาศัยค่าความน่าจะเป็น [7]

โดยในปี 2020 งานวิจัยของ Geevers [2] ได้นำกระบวนการ Deep Reinforcement Learning (DRL) มาประยุกต์ใช้กับการจัดการสินค้าคงคลัง โดยใช้หลักการของ Q-Learning ซึ่งพบกับข้อจำกัดของตาราง Q-Table ที่ทำหน้าที่จัดเก็บค่า Q-Value ที่เมื่อมีค่าเกิน 60 ล้านเซลล์จะไม่สามารถเก็บเพิ่มได้อีก Geevers [2] จึงได้เปลี่ยนมาใช้หลักการของ Proximal Policy Optimization ซึ่งให้ผลลัพธ์ที่ออกมาจากการนำไปทดสอบกับบอร์ดเกม Beer game, Divergent และ CBC โดยเปรียบเทียบค่า Total costs ระหว่างค่าที่ได้จาก Proximal Policy Optimization

กับค่าที่ได้จากการ Benchmark ซึ่ง Proximal Policy Optimization ให้ค่า Total costs ที่น้อยกว่าการ Benchmark 2 เกมส์ คือ Beer game และ Divergent

1.2 ขอบเขตของการศึกษา

งานวิจัยนี้จัดทำขึ้นเพื่อสร้างโมเดลการพยากรณ์จำนวนการสั่งวัตถุดิบที่ต้องใช้ในแต่ละวัน โดยใช้ Proximal Policy Optimization Algorithm ของกระบวนการ Reinforcement Learning

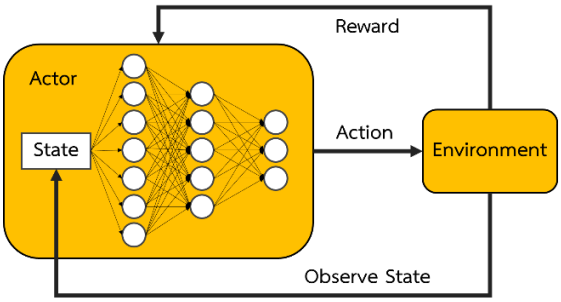
2. วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎี Reinforcement Learning

Reinforcement Learning คือ ศาสตร์หนึ่งของการทำ Machine Learning โดยจะใช้โรบอท (Agent) มาช่วยในการพิจารณาตัดสินใจ ว่าในสภาพแวดล้อม ณ ขณะนั้น ควรจะต้องตัดสินใจทำสิ่งใดที่ให้ได้ผลลัพธ์ที่ดีที่สุด โดยการเรียนรู้ของโรบอทจะเรียนรู้จากการลองผิดลองถูก และจากการลองผิดลองถูกนี้จะมีการให้คะแนน (Reward) เพื่อเป็นตัวใช้ในการบ่งบอกให้แก่โรบอทว่าควรจะต้องปรับปรุงการตัดสินใจไปในทิศทางใด เพื่อให้ได้ผลลัพธ์ที่ดีที่สุด [3]

องค์ประกอบของระบบ Reinforcement Learning จะแบ่งออกเป็น 5 ส่วน ดังนี้

1) Actor คือ โรบอทที่มีหน้าที่ตัดสินใจว่าจะทำอะไรในสถานการณ์ที่เกิดขึ้น ณ ขณะนั้น โดยการตัดสินใจในแต่ละครั้งจะอ้างอิงจากประสบการณ์การลองผิดลองถูกในสถานการณ์ที่เคยเกิดขึ้นมาก่อนหน้า แล้วได้ผลลัพธ์หรือคะแนน (Reward) ออกมาดี
2) Action คือ การกระทำของโรบอททั้งหมดที่เป็นไปได้ เช่น งานวิจัยครั้งนี้การจะทำให้การสั่งวัตถุดิบออกมาได้ค่าที่เหมาะสม จะมีการกระทำที่เป็นไปได้ทั้งหมด 2 อย่าง คือ เพิ่มจำนวนสั่งวัตถุดิบและลดจำนวนการสั่งวัตถุดิบ
3) Environment คือ สภาพแวดล้อมที่กำหนดให้โรบอทเรียนรู้และต้องทำการตัดสินใจว่าจะทำอย่างไรในสถานการณ์ ณ ขณะนั้น เพื่อให้ได้ค่า Reward ที่สูงที่สุด
4) State คือ ประสบการณ์การเรียนรู้ของโรบอทว่าถ้าเกิดสถานการณ์แบบนี้ควรจะต้องตัดสินใจไปในทิศทางอย่างไร เช่น ตัวอย่าง Hidden Markov Model จะประกอบด้วยสถานการณ์ที่แดดออก และสถานการณ์ที่ฝนตก โดยจะมีความน่าจะเป็นที่จะเกิดสถานการณ์ต่างๆ
5) Reward คือ คะแนนที่ระบบประเมินผลให้กับการกระทำครั้งนั้นๆที่โรบอทได้ตัดสินใจทำลงไป ค่าคะแนนรางวัลนี้จะเป็นสิ่งบ่งบอกให้โรบอททราบว่าการกระทำกับสภาพแวดล้อมแบบที่เกิดขึ้นนั้น ควรจะกระทำร่วมกันหรือไม่ โดยโรบอทต้องตัดสินใจกระทำในสิ่งที่ได้ค่าคะแนนรางวัลออกมาสูงที่สุด



รูปที่ 1 Reinforcement Learning Diagram

2.2 ทฤษฎี Proximal Policy Optimization (PPO)

Proximal Policy Optimization [6] เป็นหลักการที่ได้รวมจุดเด่นของ Advantage Actor Critic (A2C) ในเรื่องของการประมวลผลแบบ Multiple Workers ทำให้สามารถประมวลผลได้รวดเร็ว และจุดเด่นของ Trust Region Policy Optimization (TRPO) ที่มีการกำหนดขอบเขตของการอัพเดต Policy ใหม่ เพื่อป้องกันการเกิดการ Divergent ออกจากค่า Optimization และอีกจุดเด่นของกระบวนการ Proximal Policy Optimization คือ การอัพเดต Policy ของ Stochastic Gradient Ascent

แบบ Multiple Epochs ซึ่งทำให้โอกาสที่ข้อมูล Outlier จะส่งผลกระทบกับโมเดลเกิดขึ้นได้ยากยิ่งขึ้น

โดยกระบวนการ Proximal Policy Optimization จะมีส่วนที่เพิ่มเติมขึ้นมาจากระบบ Reinforcement Learning ทั่วไป คือตัว Critic ที่จะทำหน้าที่เป็นผู้ช่วยของ Actor ในการประเมินผลประสิทธิภาพของ Action ที่ Actor ได้กระทำต่อ Environment ด้วยค่า Advantage Function ดังสมการที่ 1 ที่จะเป็นการหาค่าผลต่างระหว่างค่า Discounted Rewards (G_t) ที่ได้จากการคำนวณค่า Reward ที่ t แบบ Multiple Epochs ดังสมการที่ 2 โดยที่ค่า k คือ ลำดับที่ของ Epoch และ ค่า γ คือค่า Discount Factor ซึ่งในงานวิจัยนี้ได้ใช้ค่า 0.99 ตามงานวิจัย Proximal Policy Optimization [6] กับค่า Value Function ที่จะได้จากการคาดคะเนของ Value Function Neural Network ซึ่งถ้าค่าของ Advantage Function ออกมาเป็นบวกแสดงว่า Action ที่กระทำต่อ State นั้นเป็น Action ที่ดีต่อ State นั้น และจะมีการปรับค่าความน่าจะเป็นที่จะใช้ Action ดังกล่าวเมื่อเกิด State นั้นเพิ่มขึ้น ในทางตรงกันข้าม ถ้าค่าของ Advantage Function ออกมาเป็นลบ จะมีการลดความน่าเป็นที่จะใช้ Action ดังกล่าวเมื่อเกิด Stateนั้น

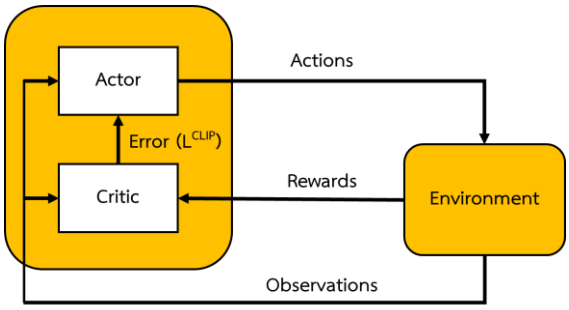
$$\hat{A}_t = G_t - \text{Value Function} \quad (1)$$

$$G_t = \sum_{k=0}^{x} \gamma^k R_{t+k+1} \quad (2)$$

ต่อมาจะเป็นการคำนวณค่า Objective Function (L^{CLIP}) ที่ใช้ในการอัพเดต Policy ด้วยวิธีการ Clip ที่จะช่วยในเรื่องการจำกัดกรอบของโมเดลในการอัพเดต Policy ไม่ให้เกิดกรณีการอัพเดตค่าที่มีการกระโดดมาเกินไป เพื่อป้องกันผลกระทบที่จะเกิดจากการเจอข้อมูลที่เป็น Outlier โดย Objective Function ดังในสมการที่ 3 จะเป็นการเปรียบเทียบค่า Minimum ระหว่างค่าอัตราส่วนของค่า Policy ใหม่ ($\pi_\theta(a_t|s_t)$) ต่อค่า Policy เดิม ($\pi_{\theta_{old}}(a_t|s_t)$) และอัตราส่วนของค่า Policy ใหม่ กับ Policy เดิม แบบที่ Clip ค่าให้อยู่ระหว่าง 1-ε กับ 1+ε โดยที่ ε กำหนดค่าตามงานวิจัย Proximal Policy Optimization [6] ให้มีค่าอยู่ที่ 0.2 ซึ่งเป็นค่าที่ให้ประสิทธิภาพออกมาสูงที่สุด โดยเมื่อได้ค่า Minimum ของอัตราส่วนระหว่างค่า Policy ใหม่กับ Policy เก่า จะนำค่าดังกล่าวไปคูณกับค่า Advantage Function ที่ได้คำนวณมาก่อนหน้า หลังจากนั้นนำค่า L^{CLIP} ที่ได้ไปอัพเดต Gradient Ascent ของระบบใหม่

$$L^{CLIP}(\theta) = \hat{E}[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t] \quad (3)$$

เมื่อ $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$



รูปที่ 2 Proximal Policy Optimization Diagram

2.3 ทฤษฎีการพยากรณ์ (Forecasting)

การพยากรณ์ คือ การทำนายผลลัพธ์หรือสิ่งที่จะเกิดในอนาคต โดยอาศัยการพิจารณาจากเหตุการณ์ที่เคยเกิดขึ้นมาในอดีต ซึ่งจะพิจารณาจากปัจจัยรอบข้างที่ส่งผลกระทบต่อค่าที่จะพยากรณ์ แล้วนำค่าที่ได้มาคำนวณเพื่อพิจารณาหาลักษณะการเกิด แล้วจดจำใช้ในการพยากรณ์ครั้งต่อไป [5], [11]

การพยากรณ์สามารถแบ่งได้เป็น 2 ประเภท ดังนี้

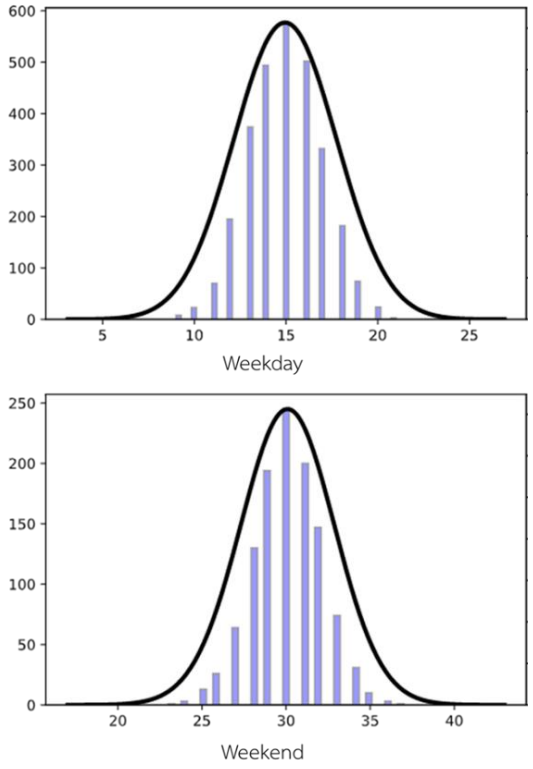
1) การพยากรณ์เชิงคุณภาพ คือ การพยากรณ์แบบใช้ความรู้และประสบการณ์ที่ผ่านมาในอดีตมาเป็นเกณฑ์ในการตัดสินใจ
2) การพยากรณ์เชิงปริมาณ คือ การพยากรณ์ที่ใช้การคำนวณทางคณิตศาสตร์ หรือหลักทาง

สถิติมาช่วยเป็นเกณฑ์ในการประกอบตัดสินใจร่วมกับข้อมูลในอดีตที่ผ่านมา สามารถแบ่งได้เป็น แบบอนุกรมเวลา คือ การพยากรณ์ที่มีการนำตัวแปรของเวลา มาเป็นสิ่งที่ใช้ในการคำนวณทางคณิตศาสตร์ หรือใช้ประกอบกับหลักทางสถิติ

3. วิธีการวิจัย

3.1 ชุดข้อมูล

ชุดข้อมูลที่ใช้ในงานวิจัยครั้งนี้จะเป็นชุดข้อมูลจำลอง (Synthetic Data) ที่อาศัยหลักการจำลองข้อมูลแบบแจกแจงปกติ (Normal Distribution) เนื่องมาจากชุดข้อมูลจริงที่ได้มาจากเครื่อง POS ในแต่ละสาขาของบริษัท เอ เป็นชุดข้อมูลที่มีความแปรปรวนที่สูงและไม่มีแบบฟอร์มการกระจายตัวของข้อมูลที่แน่นอน โดยผู้วิจัยได้ทำการคัดกรองข้อมูลที่เป็น Outlier และข้อมูลส่วนที่ไม่เกี่ยวข้องออก แต่ชุดข้อมูลยังคงมีความแปรปรวนที่สูงอยู่ ผู้วิจัยจึงตัดสินใจใช้ชุดข้อมูลจำลอง โดยใช้เลือกใช้ไลบรารี่ของ Numpy ในส่วนคำสั่งของการ Random แบบ Normal Distribution มาเป็นคำสั่งที่ใช้ในการจำลอง โดยในชุดข้อมูลจำลองจะมีการแยกพฤติกรรมการใช้วัตถุดิบเบื้องต้นเป็นแบบความต้องการวัตถุดิบในวันธรรมดา และความต้องการใช้วัตถุดิบในวันหยุด ดังรูปที่ 3 โดยค่าเฉลี่ยที่ใช้ในการจำลองชุดข้อมูล นำมาจากค่าเฉลี่ยของชุดข้อมูลจริงที่ได้จากเครื่อง POS และกำหนดค่าส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 2



รูปที่ 3 Data Distribution Diagram

โดยชุดข้อมูลจำลองจะแบ่งย่อยตามโมเดลที่จะทำการเรียนรู้ต่อออกเป็น 2 ประเภท คือ แบบ One-Attribute ที่จะประกอบด้วย Attribute ของการใช้วัตถุดิบในแต่ละวันเพียง Attribute เดียว และแบบ Multi-Attribute ที่จะประกอบด้วย Attribute 4 ตัว คือ Attribute ของการใช้วัตถุดิบในแต่ละวัน ที่เป็นข้อมูลชุดเดียวกับแบบ One-Attribute ต่อมาจะเป็น Attribute ของความชื้นสัมพัทธ์ในบรรยากาศกับค่าฝุ่นละออง PM2.5 ที่อาจส่งผลต่อการเก็บรักษาวัตถุดิบ และสุดท้ายจะเป็น Attribute ของอัตราการเกิดฝนที่อาจส่งผลต่อการเข้าใช้บริการที่สาขาของลูกค้า

3.2 เครื่องมือที่ใช้ในการวิจัย

งานวิจัยนี้จะใช้ชุดข้อมูลมาประกอบกับการใช้เทคนิค Reinforcement Learning เพื่อให้โรบอทเรียนรู้ในการหาแนวทางการตัดสินใจที่ให้ได้ค่าการสั่งวัตถุดิบที่สอดคล้องกับการขายอาหารในแต่ละสภาพแวดล้อมมากที่สุด โดยงานวิจัยนี้จะใช้ Proximal Policy Optimization เป็นอัลกอริทึมในการเรียนรู้ของ Reinforcement Learning

3.3 กระบวนการวิจัย

1) การทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องกับการทำระบบ Reinforcement Learning เช่น องค์ประกอบสำคัญของระบบ ระบบการทำงานของระบบ รูปแบบของหลักการทำงานแบบต่างๆของระบบ Reinforcement Learning เพื่อให้ได้รูปแบบของการทำงานที่เหมาะสมกับปัญหาของงานวิจัย
2) การสร้างชุดข้อมูลจำลอง (Synthetic Data) ขึ้นมาเพื่อใช้ในการป้อนให้แบบจำลองโมเดลใช้ในการเรียนรู้
3) การสร้างระบบ Reinforcement Learning เช่น การสร้าง Environment ของระบบที่จะให้แบบจำลองใช้ในการเรียนรู้
4) การดำเนินการให้โรบอทในระบบ Reinforcement Learning เรียนรู้จากชุดข้อมูลที่กำหนดให้
5) การประเมินผลระบบ Reinforcement Learning จากชุดข้อมูลอีกชุดที่ถูกแบ่งไว้

4. ผลการวิจัย

งานวิจัยครั้งนี้ใช้การประมวลผลการเรียนรู้โมเดลภายใต้ระบบ Google Cloud Platform โดยใช้เครื่องแบบ n2-standard-4 มีหน่วยประมวลผลกลางจำนวน 4 cores หน่วยความจำหลัก(RAM) 16GB ระบบปฏิบัติการ Debian 9

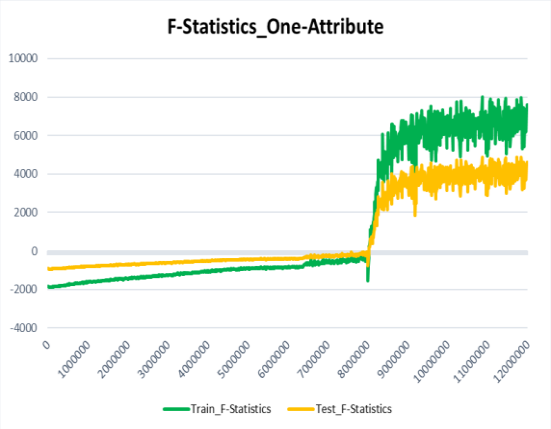
ในงานวิจัยนี้จะทำการเรียนรู้โมเดล 2 รูปแบบ จำนวนรูปแบบละ 12 ล้าน timesteps โดยรูปแบบแรกเป็นโมเดลแบบ One-Attribute ใช้เวลาในการเรียนรู้ทั้งหมด 11 ชั่วโมง ส่วนรูปแบบที่สองเป็นโมเดลแบบ Multi-Attribute ใช้เวลาในการเรียนรู้ทั้งหมด 12 ชั่วโมง โดยทั้งสองรูปแบบจะป้อนข้อมูลแบบ Time-Series ย้อนหลัง 7 วัน ทุก Attribute ของแต่ละโมเดล เพื่อใช้ในการพยากรณ์ค่าวัตถุดิบ

โดยที่โมเดลแบบ One-Attribute จะป้อนเฉพาะค่าวัตถุดิบที่ใช้จริงย้อนหลัง 7 วัน เพื่อใช้เป็นปัจจัยให้อัลกอริทึม Proximal Policy Optimization ใช้ในการเรียนรู้และสร้างเป็นโมเดลออกมา ส่วนโมเดลแบบ Multi-Attribute จะมีการเพิ่ม Attribute อีก 3 ตัวคือ Humidity, PM2.5, Rainy ที่มีความสัมพันธ์กับ Attribute ของค่าวัตถุดิบที่ใช้ แล้วเรียนรู้ด้วยกระบวนการของอัลกอริทึม Proximal Policy Optimization เหมือนดังโมเดลแบบ One-Attribute

ซึ่งผลลัพธ์ที่ได้ผู้วิจัยจะใช้เกณฑ์ในการวัดผลในงานวิจัยครั้งนี้จำนวน 3 รูปแบบ ดังนี้

1) การวัดผลด้วยค่า F-Statistics เป็นการวัดผลในด้านของความแม่นยำของโมเดล โดยจะวัดจากค่าของความคลาดเคลื่อนที่โมเดลทำการพยากรณ์ออกมา อยู่ภายใต้ค่านัยสำคัญที่กำหนดไว้
2) การวัดผลด้วยค่า R-Square เป็นการวัดผลในด้านของการที่ค่าตัวแปรที่ป้อนให้กับโมเดลสามารถอธิบายค่าที่พยากรณ์ออกมาได้ เป็นจำนวนกี่เปอร์เซนต์
3) การวัดผลด้วยค่า RMSE เป็นการวัดผลในด้านของค่าความผิดพลาดโดยเฉลี่ย จากการโมเดลที่ได้ทำการพยากรณ์ค่าออกมา

4.1 One-Attribute Model

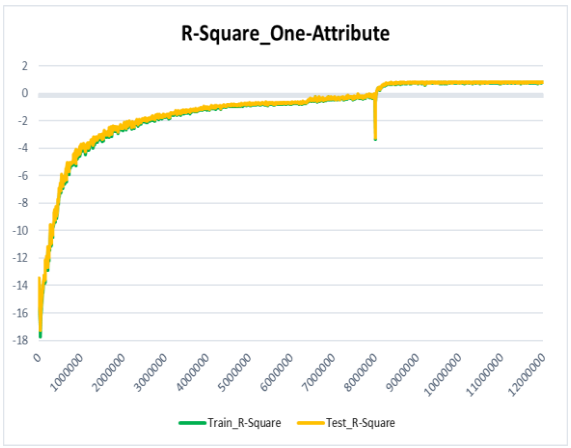


รูปที่ 4 ค่า F-Statistics ของโมเดลแบบ One-Attribute

ในรูปที่ 4 เป็นกราฟการวัดผลจากค่า F-Statistics ของโมเดลแบบ One-Attribute ที่มีการวัดผลจากการเรียนรู้ของโมเดลทุกๆ 10,000 timesteps โดยในงานวิจัยนี้จะกำหนดค่าความเชื่อมั่นที่ 95% (α = 0.05) และในแต่ละครั้งจะทดสอบโมเดลด้วยข้อมูลจำนวนทั้งหมด 1,000 records ซึ่งเมื่อทำการเปิดตารางค่า F จะได้ค่า $F_{0.05,1,998} \approx 3.85$

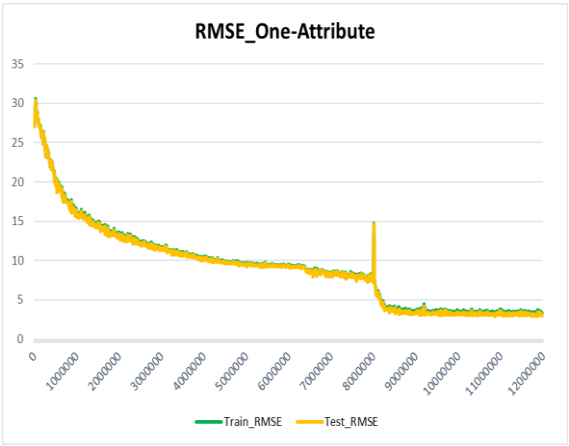
นั่นหมายความว่าเมื่อใดที่ค่า $F > F_{0.05,1,998} \approx 3.85$ ในโมเดลนั้นค่าความผิดพลาดส่วนใหญ่ สามารถอธิบายได้ด้วยค่าตัวแปรที่ป้อนเข้าไป

โดยจากรูปที่ 4 โมเดลที่มีค่า $F > F_{0.05,1,998} \approx 3.85$ คือโมเดล Timestep ที่ 8,040,000 และหลังจากนั้นจะมีค่าที่ค่อยๆสูงขึ้นไปเรื่อยๆ



รูปที่ 5 ค่า R-Square ของโมเดลแบบ One-Attribute

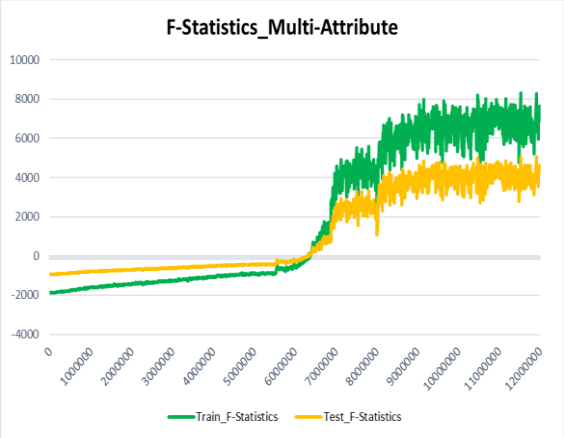
ในรูปที่ 5 เป็นกราฟการวัดผลจากค่า R-Square ของโมเดลแบบ One-Attribute ซึ่งค่า R-Square จะมีค่าอยู่ระหว่าง 0 – 1 โดยจากรูปที่ 5 ในช่วงแรกของกราฟค่า R-Square จะมีค่าต่ำกว่า 0 ซึ่งหมายความว่าค่าตัวแปรที่ป้อนให้กับโมเดล ยังไม่สามารถอธิบายค่าที่พยากรณ์ออกมาได้ จนมาถึงในช่วง Timestep ที่ 8,040,000 ค่า R-Square มีค่ามากกว่า 0 และมีค่าสูงขึ้นเรื่อยๆจน R-Square มีค่าเท่ากับ 0.82 ใน Timestep ที่ 12 ล้าน นั่นหมายความว่าค่าตัวแปรที่ป้อนให้กับโมเดล สามารถอธิบายค่าที่พยากรณ์ออกมาได้ 82%



รูปที่ 6 ค่า RMSE ของโมเดลแบบ One-Attribute

ในรูปที่ 6 เป็นกราฟการวัดผลจากค่า RMSE ของโมเดลแบบ One-Attribute โดย RMSE เป็นการวัดค่าความผิดพลาดโดยเฉลี่ยของโมเดล จากรูปที่ 6 ค่า RMSE มีค่าลดลงเรื่อยๆจนมาถึง Timestep ที่ 8,170,000 ค่า RMSE มีค่าต่ำกว่า 5 และมีค่าลดลงเรื่อยๆจนใน Timestep ที่ 12 ล้าน มีค่าเท่ากับ 3 ซึ่งหมายความว่า ค่าความผิดพลาดโดยเฉลี่ยของค่าการพยากรณ์ที่โมเดลได้พยากรณ์ออกมาอยู่ที่ ±3 ชิ้น

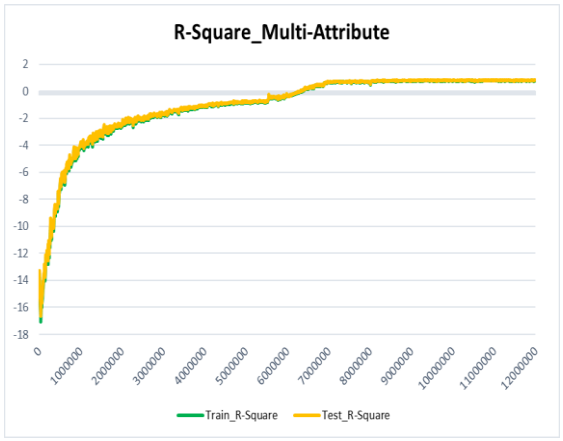
4.2 Multi-Attribute Model



รูปที่ 7 ค่า F-Statistics ของโมเดลแบบ Multi-Attribute

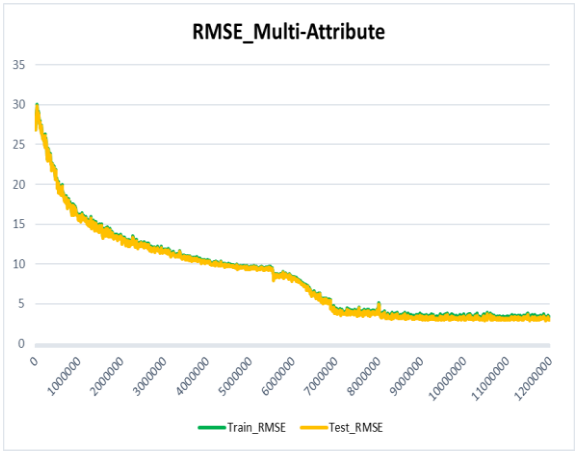
ในรูปที่ 7 เป็นกราฟการวัดผลจากค่า F-Statistics ของโมเดลแบบ Multi-Attribute ที่มีการวัดผลจากการเรียนรู้ของโมเดลทุกๆ 10,000 timesteps กำหนดค่าความเชื่อมั่นที่ 95% (α = 0.05) และทำการทดสอบโมเดลด้วยข้อมูลจำนวนทั้งหมด 1,000 records ซึ่งเมื่อทำการเปิดตารางค่า F จะได้ค่า $F_{0.05,1,998} \approx 3.85$ เช่นเดียวกับโมเดลแบบ One-Attribute

แต่จะแตกต่างกันที่โมเดลแบบ Multi-Attribute จะได้ค่า $F > F_{0.05,1,998} \approx 3.85$ ที่ Timestep ที่ 6,360,000 ซึ่งไวกว่าโมเดลแบบ One-Attribute



รูปที่ 8 ค่า R-Square ของโมเดลแบบ Multi-Attribute

ในรูปที่ 8 เป็นกราฟการวัดผลจากค่า R-Square ของโมเดลแบบ Multi-Attribute ซึ่งค่า R-Square จะมีค่าอยู่ระหว่าง 0 – 1 โดยในช่วงแรกของกราฟจะมีค่า R-Square ที่ต่ำกว่า 0 เหมือนโมเดลแบบ One-Attribute จนมาถึง Timestep ที่ 6,360,000 ค่า R-Square ที่ได้จะมากกว่า 0 และใน Timestep ที่ 12 ล้าน มีค่า R-Square เท่ากับ 0.82 หรือค่าตัวแปรที่ป้อนให้กับโมเดล สามารถอธิบายค่าที่พยากรณ์ออกมาได้ 82%



รูปที่ 9 ค่า RMSE ของโมเดลแบบ Multi-Attribute

ในรูปที่ 9 เป็นกราฟการวัดผลจากค่า RMSE ของโมเดลแบบ Multi-Attribute โดย RMSE เป็นการวัดค่าความผิดพลาดโดยเฉลี่ยของโมเดล จากรูปที่ 9 ค่า RMSE มีค่าลดลงเรื่อยๆจนมาถึง Timestep ที่ 6,890,000 RMSE มีค่าต่ำกว่า 5 และใน Timestep ที่ 12 ล้าน RMSE มีค่าเท่ากับ 3 หรือค่าความผิดพลาดโดยเฉลี่ยของค่าการพยากรณ์ของโมเดลอยู่ที่ ±3 ชิ้น

5. สรุป

งานวิจัยครั้งนี้มีวัตถุประสงค์ที่จะศึกษา Proximal Policy Optimization Algorithm ของกระบวนการ Reinforcement Learning เพื่อสร้างโมเดลการพยากรณ์จำนวนการสั่งวัตถุดิบของร้านอาหาร โดยได้สร้างชุดข้อมูลจำลอง (Synthetic Data) ที่อาศัยหลักการจำลองข้อมูลแบบแจกแจงปกติ (Normal Distribution) ขึ้นมา และสร้าง Environment ของระบบ Reinforcement Learning จากนั้นป้อนชุดข้อมูลจำลองให้กับระบบเรียนรู้ในการสร้างโมเดล

โดยการศึกษาในครั้งนี้แบ่งการศึกษาออกเป็นโมเดลแบบ One-Attribute และโมเดลแบบ Multi-Attribute และทำการเรียนรู้ทั้งหมด 12,000,000 timesteps พบว่าโมเดลทั้งสองใน Timestep ที่ 12 ล้าน ให้ผลลัพธ์ที่ใกล้เคียงกัน จากการประเมินผลด้วย 3 เครื่องมือ คือ ค่า F-Statistics,

ค่า R-Square และค่า RMSE ด้วยชุดข้อมูลจำลองที่แยกจากชุดข้อมูลที่ใช้เรียนรู้จำนวน 1,000 records แต่ทั้งสองโมเดลจะแตกต่างกันที่ โมเดลแบบ Multi-Attribute จะเกิดการลู่เข้า (Convergence) หาค่าที่เหมาะสมของโมเดลได้เร็วกว่าการเรียนรู้ของโมเดลแบบ One-Attribute และในส่วนของค่าความถูกต้องในการพยากรณ์จำนวนการสั่งวัตถุดิบ อยู่ที่ร้อยละ 82 ของค่าการพยากรณ์ทั้งหมด ซึ่งจำนวนค่าที่ใช้ในการทดสอบทั้งหมด คือ 1,000 ค่า

โดยงานวิจัยนี้จัดทำขึ้นเพื่อเป็นแนวทางในการประยุกต์ใช้หลักการทาง Machine Learning ที่เป็นแบบ Reinforcement Learning ที่มีจุดเด่นในเรื่องของการใช้ค่าความน่าจะเป็นแบบ Markov Decision มาช่วยในการตัดสินใจกับการจัดการวัตถุดิบคงคลังในร้านอาหาร ว่าในวันต่อมาควรสั่งวัตถุดิบเพิ่มขึ้นหรือสั่งลดลงเป็นจำนวนเท่าไรจากวันดังกล่าว เพื่อให้สามารถลดจำนวนขยะอาหารที่อาจเกิดขึ้นลงได้

6. เอกสารอ้างอิง

[1] FAO, “The State of Food and Agriculture 2019. Moving Forward on Food Loss and Waste Reduction,” [Online]. Available: http://www.fao.org/3/ca6030en/ca6030en.pdf. [Accessed 8 March 2020].

[2] K. Geevers, “Deep Reinforcement Learning in Inventory Management,” M.S. thesis, University of Twente, Netherlands, 2020.

[3] A. a. R. Hill et al., “Stable Baselines,” *GitHub repository.*, 2018.

[4] C. Hu, M. Chen and S.-L. C. McCain, “Forecasting in Short-Term Planning and Management for a Casino Buffet Restaurant,” *Journal of Travel & Tourism Marketing.,* vol. 16, pp. 79-98, 2004.

[5] S. Makridakis, S. Wheelright and R. Hyndman, “Manual of Forecasting: Methods and Applications,” 2000.

[6] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Proximal Policy Optimization Algorithms,” CoRR, vol. abs/1707.06347, 2017.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford and O. Klimov, “Reinforcement Learning: An Introduction: A Bradford Book,” 2018.

[8] T. Tanizaki, T. Hoshino, T. Shimmura and T. Takenaka, “Demand forecasting in restaurants using machine learning and statistical analysis,” *Procedia CIRP.*, vol. 79, pp. 679-683, 2019.

[9] T. Tanizaki, T. Hoshino, T. Shimmura and T. Takenaka, “Restaurants store management based on demand forecasting,” *Procedia CIRP.*, vol. 88, pp. 580-583, 2020.

[10] WasteMINZ, “Food Waste in The Cafe & Restaurant Sector in New Zealand,” [Online]. Available: http://www.wasteminz.org.nz/wp-content/uploads/2018/10/New-Zealand-cafe-and-resturant-food-waste-WasteMINZ-2018.pdf. [Accessed 8 March 2020].

[11] เจย์. ไฮเซอร์, การจัดการการผลิตและการปฏิบัติการ = Operations management, พิมพ์ครั้งที่ 4., กรุงเทพฯ: เพียร์สันเอ็ดดูเคชั่นอินโดไชน่า, 2551.