บทที่ 2

ปริทัศน์วรรณกรรม งานวิจัยที่เกี่ยวข้อง

2.1 กล่าวนำ

เนื้อหาในบทนี้กล่าวถึงวิธีหาเส้นทางการเชื่อมต่อที่รองรับคุณภาพการบริการสำหรับเส้นทาง (QoS routing) ในเครือข่ายเคลื่อนที่แบบแอดฮอค (mobile ad hoc network หรือ MANET) ที่มีรูปร่างเครือข่ายแบบพลวัติ รายงานวิจัยฉบับนี้นำกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ (reinforcement Learning หรือ RL) ประยุกต์ใช้กับเครือข่ายเคลื่อนที่แบบแอดฮอคที่มีรูปร่างเครือข่ายพลวัติ กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ไม่ต้องการรูปแบบการคำนวณทางคณิตศาสตร์ที่ชัดเจน วิธีนี้มีกระบวนการตัดสินใจที่ดีซึ่งได้จากการเรียนรู้แบบลองผิดลองถูก (trial and error) ซึ่งผู้เรียนจะเรียนรู้ผลได้รับจากการกระทำซึ่งได้รับจากสิ่งแวดล้อม แล้วนำไปปรับปรุงนโยบายจนบรรลุเป้าหมายที่ต้องการ

กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์จะกำหนดปัญหาให้เป็นกระบวนการตัดสินใจแบบมาร์คอฟ (Markov decision process หรือ MDP) ด้วยการระบุปัญหาว่า ระบบที่มีสิ่งแวดล้อมแบบพลวัติจะสามารถเรียนรู้นโยบายในการเลือกการกระทำที่ดีที่สุดเพื่อให้บรรลุเป้าหมายได้อย่างไร ในกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์นั้นจะมองปัญหาในรูปแบบของผู้เรียนสามารถเรียนรู้พฤติกรรมที่ดีจากการลองผิดลองถูกซึ่งเป็นการเลือก การกระทำ และสังเกตผลจากการกระทำที่ส่งผลต่อสิ่งแวดล้อมแบบพลวัติ รายงานวิจัยฉบับนี้จะประยุกต์ใช้กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ที่แบ่งการเรียนรู้ออกเป็นเอพพิโซด ด้วยวิธีการที่เรียกว่า ออนโพลิซี มอนติ คาร์โล (On-policy Monte Carlo หรือ ONMC) [11] โดยฟังก์ชั่นค่าการกระทำจะถูกประมาณค่า และนโยบายจะถูกปรับปรุงหลังจากการเรียนรู้ในแต่ละเอพพิโซด ภายใต้สมมุติฐานที่กล่าวมา วิธีออนโพลิซี มอนติ คาร์โล จะสามารถหาผลเฉลยที่ลู่เข้าสู่

นโยบายที่ดีที่สุดได้และได้ค่าฟังก์ชั่นที่ดีที่สุดจากการเรียนรู้ในแต่ละเอพพิโซดโดยไม่จำเป็นต้องรู้ข้อมูลของสิ่งแวดล้อมแบบพลวัติที่ชัดเจน

ดังนั้นรายงานฉบับนี้จะนำเสนอวิธีONMC ประยุกต์ใช้ในปัญหาการค้นพบเส้นทางในเครือข่ายเคลื่อนที่แบบแอดฮอค หัวข้อถัดไปจะนำเสนอทฤษฎีพื้นฐานเกี่ยวกับกระบวนการตัดสินใจแบบมาร์คอฟ และกล่าวแนะนำกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ในหัวข้อที่ 2.3 ในหัวข้อ2.4 จะกล่าวแนะนำวิธี ONMC และกล่าวสรุปเนื้อหาในบทนี้ในหัวข้อสุดท้าย

2.2 พื้นฐานทฤษฎีการตัดสินใจแบบมาร์คอฟ

2.2.1 คุณสมบัติมาร์คอฟ

คุณสมบัติมาร์คอฟกล่าวว่าทุกสิ่งที่เกิดขึ้นในระยะยาวเป็นผลสืบเนื่องมาจากสถานะปัจจุบัน ดังนั้นความน่าจะเป็นของสถานะถัดไป ณ เวลา $k+1$ สามารถนิยามได้โดยใช้เงื่อนไขอย่างง่ายภายใต้ข้อมูลที่ทราบจากสถานะปัจจุบัน ณ เวลา k ดังนี้

$$\Pr\{s_{k+1} = s' | s_k = s\} = \Pr\{s_{k+1} = s' | s_k = s, s_{k-1} = s, ..., s_0 = s\}. \quad (2.1)$$

โดยรายงานฉบับนี้จะประยุกต์ใช้กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ด้วยการกำหนดปัญหาให้มีคุณสมบัติสอดคล้องกับคุณสมบัติมาร์คอฟ สถานะของผู้เรียนอธิบายถึงข้อมูลของสิ่งแวดล้อมซึ่งเป็นประโยชน์ต่อการตัดสินใจ ถ้าสถานะของผู้เรียนมีคุณสมบัติมาร์คอฟจะทำให้การตอบสนองของสิ่งแวดล้อมที่เวลา $k+1$ ขึ้นอยู่กับผลที่เกิดจากสถานะปัจจุบันที่เวลา k ซึ่งเรียกสถานะเช่นนี้ว่า สถานะมาร์คอฟ (Markov state)

2.2.2 กระบวนการตัดสินใจแบบมาร์คอฟ

กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ที่มีสิ่งแวดล้อมที่สอดคล้องกับคุณสมบัติมาร์คอฟ ถูกเรียกว่า กระบวนการตัดสินใจแบบมาร์คอฟ (Markov decision process หรือ MDP) สมมุติให้เวลาปัจจุบันคือ ช่วงเวลา k ซึ่งมีสถานะจากสิ่งแวดล้อม s ผู้เรียนจะเลือกการกระทำ a ผลที่ได้รับจากการเลือกการกระทำ a ณ สถานะ s คือการตอบสนองจากสิ่งแวดล้อมทำให้ได้สถานะใหม่เป็น s' ดังนั้นความน่าจะเป็นในการเกิดสถานะใหม่ที่เป็นไปได้ คือ

$$P_{ss'}^{a} = \Pr\{s_{k+1} = s' | s_k = s, a_k = a\}. \quad (2.2)$$

สมการนี้ถูกเรียกว่า ความน่าจะเป็นในการส่งสถานะ(transition probabilities) สมมุติให้สถานะและการกระทำ ณ เวลาปัจจุบันคือ s_k และ a_k และสถานะใหม่ที่เกิดขึ้นคือ s_{k+1} ส่งผลให้ได้รับ

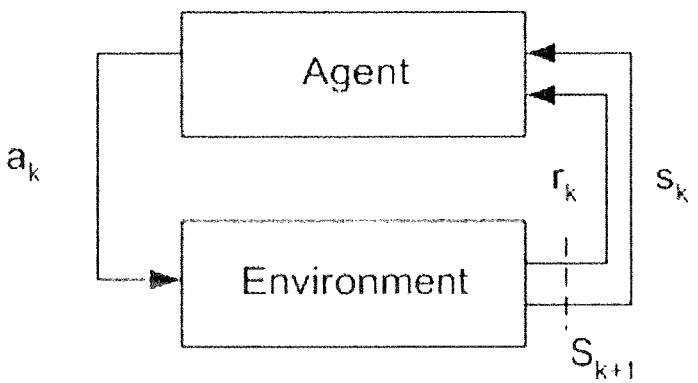
ค่าตอบแทน g_k ซึ่งเป็นค่าที่ได้รับจากการกระทำของผู้เรียน ค่าคาดหวัง (the expected value) ของผลตอบแทนถัดไปคือ

$$G_{ss'}^{a} = E\{g_k \mid s_k = s, a_k = a, s_{k+1} = s'\} \tag{2.3}$$

เมื่อผู้เรียนได้รับผลตอบแทนจากสมการนี้แล้ว ผู้เรียนจะสามารถเรียนรู้เพื่อหาการกระทำที่ดีและปรับปรุงกระบวนการตัดสินใจเพื่อให้ได้รับผลตอบแทนสูงสุดยะระยาว

2.3 กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์

กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์เป็นวิธีการคำนวณด้วยการกำหนดให้ระบบที่มีสิ่งแวดล้อมแบบพลวัติสามารถเรียนรู้เพื่อเลือกการกระทำที่ส่งผลให้ผู้เรียนบรรลุเป้าหมายที่วางไว้ [12] ผู้เรียน (The learner) จะมีกลไกการเรียนรู้โดยจะไม่สามารถระบุอย่างชัดเจนว่าการกระทำไหนควรเลือก แต่จะค้นหาการกระทำที่เหมาะสมจากการกระทำที่ให้ผลตอบแทนมากที่สุดซึ่งได้มาจากการลองผิดลองถูกภายใต้ขอบเขตของสิ่งแวดล้อมที่ศึกษา ผู้เรียนจำเป็นต้องใช้ประโยชน์จากประสบการณ์ที่ได้จากการทดลองเลือกการะกระทำและผลที่ได้รับจากการกระทำนั้น การเปลี่ยนแปลงสถานะไปจนถึงผลตอบแทนที่ได้รับเพื่อใช้ในการปรับปรุงการเลือกการกระทำที่ดีที่สุดให้กับตัวเอง นอกจากนี้กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์เป็นการเรียนรู้แบบออนไลน์



รูปที่ 2.1 แผนผังการกระทำโต้ตอบระหว่างผู้เรียนและสิ่งแวดล้อมในกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์

รูปที่ 2.1 แสดงให้เห็นถึงแนวทางการเรียนรู้ในกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์เพื่อหาผลเฉลยจากปัญหาที่ซับซ้อนด้วยการเรียนรู้จากการกระทำโต้ตอบระหว่างผู้เรียนและสิ่งแวดล้อม

ซ้ำไปซ้ำมา เมื่อเอเจนท์หมายถึงผู้เรียน (learner) หรือผู้ตัดสินใจ (decision maker) ทุกสิ่งนอกเหนือเอเจนท์ถูกกำหนดให้เป็นสิ่งแวดล้อม โดยทั่วไป *การกระทำ*ใช้นิยามถึงการตัดสินใจของผู้เรียน ในขณะที่สถานะหมายถึงข้อมูลที่ทราบจากสิ่งแวดล้อมซึ่งใช้ประกอบการตัดสินใจของผู้เรียน

เป้าหมายหลักของผู้เรียนคือการหานโยบาย π ซึ่งเป็นการจับคู่ระหว่างสถานะและการกระทำที่ให้ผลตอบแทนระยะยาวสูงสุด รูปแบบมาตรฐานของกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ คือ ผู้เรียนจะพยายามเลือก *การกระทำ* จากชุดของการกระทำที่เป็นไปได้ทั้งหมด ณ สภาพแวดล้อมปัจจุบัน จากนั้น การกระทำที่ถูกเลือก จะส่งผลให้สภาพแวดล้อมมีการเปลี่ยนแปลงและผู้เรียนก็จะได้*ผลตอบแทน* ซึ่งขึ้นอยู่กับว่าการกระทำดังกล่าวส่งผลให้สภาพแวดล้อมเปลี่ยนไปในทิศทางใด กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ประกอบด้วยปัจจัยพื้นฐานสามอย่างคือ สิ่งแวดล้อม, ฟังก์ชั่นรีอินฟอร์สเมนท์ และฟังก์ชั่นมูลค่า (value function)

1) สิ่งแวดล้อม

ในระบบการเรียนรู้แบบรีอินฟอร์สเมนท์จะเรียนรู้การจับคู่จากสถานะไปยังการกระทำด้วยการสุ่มทดลองการกระทำโต้ตอบ (interactions) กับสิ่งแวดล้อมแบบพลวัติสิ่งแวดล้อมเหล่านี้จะถูกเฝ้าสังเกตจากกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ด้วยการสังเกตจากค่าที่อ่านได้จากอุปกรณ์เซนเซอร์, สัญลักษณ์ หรือ จากสถานการณ์ที่ผิดปกติ ถ้ากระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์สามารถสังเกตรายละเอียดต่างๆของสิ่งแวดล้อมที่สนใจได้ดีเยี่ยมจะส่งผลให้กระบวนการนี้สามารถเลือกการกระทำที่เหมาะสมกับสถานะจริงที่เกิดขึ้นได้ แนวคิดนี้จึงเป็นแนวคิดพื้นฐานที่ดีที่สุดของกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ อย่างไรก็ตามกระบวนการนี้ยังต้องอาศัยปัจจัยที่จำเป็นต่อการเรียนรู้เพื่อการจับคู่ที่เหมาะสมตามทฤษฎีดังกล่าว

2) ฟังก์ชั่นรีอินฟอร์สเมนท์

จากที่กล่าวมาข้างต้น ระบบที่มีการเรียนรู้แบบรีอินฟอร์สเมนท์ซึ่งทำการจับคู่จากสถานะไปยังการกระทำด้วยการทดลองสุ่มการกระทำโต้ตอบกับสิ่งแวดล้อม โดยเป้าหมายของกระบวนการเรียนรู้นี้คือ การใช้แนวคิดของฟังก์ชั่นรีอินฟอร์สเมนท์ซึ่งเป็นฟังก์ชั่นจริงของการเสริมกำลังในอนาคต (future reinforcements) ของผู้เรียนเพื่อค้นหาการกระทำที่ให้ผลตอบแทนระยะยาวสูงสุด โดยหลังจากการเลือก *การกระทำ* ที่ *สถานะปัจจุบัน* แล้วผู้เรียนจะได้รับ ผลตอบแทน (reward) ในรูปแบบของปริมาณเชิงตัวเลข ผู้เรียนจะเรียนรู้การกระทำที่ให้ผลตอบแทนระยะยาวสูงสุด

3) ฟังก์ชั่นมูลค่า

ฟังก์ชั่นมูลค่าคือการจับคู่จาก *สถานะ* (state) ไปยัง *มูลค่าของสถานะ* (state values) สมมุติให้นโยบาย π ใช้ในการกำหนด การกระทำที่ควรเลือกในแต่ละสถานะ มูลค่าของสถานะ $V^{\pi}(s)$ ถูกนิยามด้วยผลรวมของค่าคาดหวังที่ผู้เรียนจะได้รับเมื่ออยู่ในสถานะ s

$$V^{\pi}(s) = E_{\pi}\left\{\sum_{n=1}^{\infty} g_{t+n} | s_t = s\right\} \quad (2.4)$$

ดังนั้นนโยบายที่ดีที่สุด V^{*} จะถูกจับคู่จาก สถานะ ไปยัง การกระทำ ที่ให้ผลตอบแทนสูงสุดซึ่งเริ่มต้นจากสถานะแรกและทำการเลือกการกระทำจนกระทั่งสิ้นสุดสถานะสุดท้ายจึงได้ว่า

$$V^{*} = \max_{\pi}\left\{V^{\pi}(s)\right\} \quad (2.5)$$

โดยทั่วไปการเลือกการกระทำในแต่ละช่วงเวลามักถูกคาดหวังให้เป็นการกระทำที่ให้ผลตอบแทนสูงที่สุดในระยะยาว

2.3.1 วิธีมอนติ คาร์โล

วิธีมอนติ คาร์โลเป็นวิธีที่ใช้หาคำตอบของปัญหาในกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์บนพื้นฐานของการเรียนรู้จากมูลค่าเฉลี่ยสะท้อนกลับ วิธีมอนติ คาร์โลต้องการเพียงประสบการณ์ในการเรียนรู้เพื่อหาคำตอบเท่านั้น เช่น ลำดับของสถานะตัวอย่าง, การกระทำ และผลตอบแทนที่ได้รับจากกระบวนการเรียนรู้แบบออนไลน์ กระบวนการเรียนรู้จากประสบการณ์ตรงแบบออนไลน์กำลังเป็นที่สนใจเนื่องจากวิธีนี้ไม่จำเป็นต้องใช้ข้อมูลจากสิ่งแวดล้อมแบบพลวัติแต่ยังสามารถตอบโจทย์ของปัญหาโดยการได้มาซึ่งพฤติกรรมที่ดีที่สุดที่ควรปฏิบัติ รายงานฉบับนี้ประยุกต์วิธีมอนติ คาร์โลกับปัญหาโดยแบ่งการเรียนรู้เป็นเอพพิโซด โดยการสมมุติให้ประสบการณ์ในการเรียนรู้ถูกแบ่งออกเป็นเอพพิโซด แต่ละเอพพิโซดจะจบลงเมื่อมีมีการเปลี่ยนแปลงของมูลค่าที่ประมาณและนโยบายการเลือกพฤติกรรมเท่านั้น ดังนั้นวิธีมอนติ คาร์โลจึงเป็นการเรียนรู้แบบเอพพิโซดต่อเอพพิโซด

พิจารณาวิธีมอนติ คาร์โลสำหรับการเรียนรู้ *ฟังก์ชั่นมูลค่าสถานะ*(the state-value function) สมมุติให้นโยบาย $\pi : S \rightarrow A$ โดยมี *มูลค่าของสถานะ* เป็นค่าคาดคะเนผลตอบแทนย้อนกลับ (the expected return) หรืออาจกล่าวได้อีกนัยว่า เป็นค่าคาดคะเนผลตอบแทนในอนาคตแบบลดทอน (the expected cumulative future discounted reward) ของสถานะนั้น [12] วิธีการดั้งเดิมที่ใช้ประมาณค่าฟังก์ชั่นมูลค่าสถานะคือค่าผลเฉลี่ยย้อนกลับ (สมการที่ 2.4) ซึ่งได้รับหลังการพบสถานะนั้น โดยที่ค่าตอบแทนผลเฉลี่ยนี้ควรลู่เข้าสู่ค่า *ค่าคาดคะเน* (the expected value) ปัญหาการ

ประเมินนโยบายสำหรับเลือกพฤติกรรมคือ การประมาณค่า $Q^{\pi}(s,a)$ ซึ่งเป็น *ค่าคาดคะเนย้อนกลับ* หลังจากการเลือกการกระทำ a ที่สถานะ s แล้วได้นโยบาย π

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{n=1}^{\infty} g_{t+n} \middle| s_t = s, a_t = a\right\}. \tag{2.6}$$

ค่าการพบสถานะแรกของวิธีมอนติคาร์โล (the first-visit Monte Carlo method) ได้จากการเฉลี่ย *ค่าย้อนกลับ* จากการพบสถานะแรกที่เป็นผลจากเกิดจากการการกระทำที่ถูกเลือกดังนี้

$$Q^{\pi}(s,a) = \frac{c(s,a,1)}{1}. \tag{2.7}$$

เมื่อ *c(s, a, 1)* คือ ค่าย้อนกลับ หลังจากพบคู่สถานะ-การกระทำแรก *(s,a)*

ดังนั้นค่าการพบสถานะถัดไปที่เหลือทั้งหมดของวิธีมอนติคาร์โล (the every-visit Monte Carlo method) จึงได้จากการประมาณค่าจากคู่สถานะ-การกระทำนั้น ซึ่งเป็นการเฉลี่ยค่าย้อนกลับจากการพบสถานะที่เป็นผลจากเกิดจากการการกระทำที่ถูกเลือกดังนี้

$$Q^{\pi}(s,a) = \frac{\sum_{k=1}^{n} c(s,a,k)}{n(s,a)}. \tag{2.8}$$

เมื่อ *c(s, a, k)* คือ ค่าย้อนกลับ หลังจากพบคู่สถานะ-การกระทำ *(s, a)* และ *n(s, a)* คือจำนวนครั้งในการพบคู่สถานะ-การกระทำ *(s, a)*

วิธีคำนวณค่าผลตอบแทนทั้งสองวิธีที่กล่าวมาจะทำให้ค่าย้อนกลับลู่เข้าสู่ค่าคาดคะเนจริงได้ถ้าจำนวนครั้งของการพบคู่สถานะ-การกระทำแต่ละคู่เป็นอนันต์กระบวนการนี้ถูกเรียกว่า การประเมินนโยบายภายใต้นโยบายคงที่ π ในแต่ละเอพพิโซด ค่าผลตอบแทน จะถูกสังเกตเพื่อนำไปประเมินและปรับปรุงนโยบายเมื่อทุกสถานะจะต้องถูกพบครบทุกสถานะในการเรียนรู้แต่ละเอพพิโซด การปรับปรุงนโยบายเป็นกระบวนการที่ประกอบด้วย นโยบายใหม่ ซึ่งถูกปรับปรุงมาจาก นโยบายเดิมด้วยการใช้ค่ากรีดี (หรือ ε-greedy) นโยบายกรีดี (the greedy policy) จะเลือกการกระทำที่ดีที่สุดจากการคาดการณ์ค่าประมาณมูลค่าการกระทำปัจจุบัน (the current action-value estimates) สำหรับนโยบายอีกรีดี (the ε-greedy policy) ผู้เรียนจะประพฤติตัวอย่างละโมบด้วยการการกระทำที่ดีที่สุดจากการคาดการณ์ค่าประมาณมูลค่าการกระทำปัจจุบันเป็นส่วนใหญ่ แต่จะมีช่วงเวลาขณะหนึ่งด้วย

ค่าความน่าจะเป็นน้อยๆที่นโยบายอีกรีดีจะเลือกการกระทำจากการสุ่มค่าประมาณมูลค่าการกระทำเหตุที่กระทำเช่นนี้เนื่องมาจากการเลือกการกระทำการการประมาณมูลค่าเพื่อให้ได้การกระทำที่ดีที่สุดอาจยังไม่เพียงพอถ้ายังไม่มีการเข้าพบทุกสถานะที่เป็นไปได้ทั้งหมด ดังนั้นแล้วทุกคู่สถานะ-การกระทำที่ยังไม่เคยถูกพบก็จะไม่มีมูลค่าผลตอบแทน ซึ่งคู่สถานะ-การกระทำที่ไม่เคยถูกสำรวจนี้อาจมีมูลค่าย้อนกลับที่ดีกว่าคู่สถานะ-การกระทำอื่นก็ได้ แต่ด้วยการใช้นโยบายอีกรีดีทำให้คู่สถานะ-การกระทำที่ยังไม่เคยถูกพบได้มีโอกาสถูกสำรวจมากขึ้นดังนั้นผู้เรียนจึงจำเป็นอย่างยิ่งในการประมานมูลค่าของการกระทำที่เป็นไปได้ทั้งหมดในแต่ละสถานะเพื่อที่จะได้ข้อมูลประกอบการตัดสินใจที่ครอบคลุมซึ่งจะทำให้ได้พฤติกรรมที่ถูกต้อง ดังนั้นด้วยการใช้นโยบายอีกรีดีจะช่วยให้ผู้เรียนสามารถสำรวจการกระทำที่เป็นไปได้ทั้งหมดในแต่ละสถานะ

ระหว่างการเรียนรู้การประเมินนโยบายจากหลายๆเอพพิโซดด้วยวิธีการประมาณค่าฟังก์ชั่นมูลค่าการกระทำ(the action-value function)และฟังก์ชั่นมูลค่าคาดคะเน (the expected value function) สมมุติให้ผู้เรียนมีการเฝ้าสังเกตการเปลี่ยนแปลงในเอพพิโซดแบบอนันต์และแต่ละเอพพิโซดจะเริ่มต้นจากการสำรวจ สุดท้ายสมมุติให้ทุกๆคู้สถานะ-การกระทำในแต่ละเอพพิโซดมีความน่าจะเป็นที่ไม่เป็นศูนย์ ภายใต้สมมุติฐานเหล่านี้จะทำให้วิธีมอนติ คาร์โลสามารถคำนวณค่า Q^{π_k} ได้โดยเริ่มจากการสุ่มเลือกนโยบาย π_k หลังจากจบแต่ละเอพพิโซดผู้เรียนจะต้องสังเกตมูลค่าย้อนกลับที่ถูกใช้ไปในการประเมินนโยบาย และนโยบายจะถูกปรับปรุงที่ทุกๆสถานะที่ถูกสำรวจของแต่ละเอพพิโซด การปรับปรุงนโยบายทำได้ด้วยการใช้นโยบายกรีดีคาดประมาณฟังก์ชั่นมูลค่าการกระทำปัจจุบันฟังก์ชั่นมูลค่าการกระทำ $Q^{\pi}(s,a)$ ใดๆภายใต้นโยบาย π ซึ่งสัมพันธ์นโยบายกรีดีกล่าวคือ แต่ละสถานะ s ในชุดของสถานะ $(s \in S)$ จะเลือกการกระทำที่คาดการณ์ได้ด้วยฟังก์ชันมูลค่าการกระทำสูงสุด (ซึ่งอ้างถึง มูลค่าQ หรือ Q-value)

$$\pi(s) = \arg\max_{a} \{Q(s,a)\} \tag{2.9}$$

การปรับปรุงนโยบายจะถูกสร้างขึ้นจากนโยบายใหม่ π_{k+1} ที่ได้รับในแต่ละครั้งด้วยนโยบายกรีดีที่พิจารณาจาก Q^{π_k} นโยบายจะถูกปรับปรุงจาก π_k และ π_{k+1} สำหรับทุกชุดของสถานะ $(s \in S)$,

$$\begin{aligned} Q^{\pi_k}\left(s,\pi_{k+1}(s)\right) &= Q^{\pi_k}\left(s,\arg\max_{a}\left\{Q^{\pi_k}(s,a)\right\}\right) \\ &= \max_{a}\left\{Q^{\pi_k}(s,a)\right\} \\ &\geq Q^{\pi_k}\left(s,\pi_k(s)\right). \end{aligned} \tag{2.10}$$

ด้วยความสัมพันธ์ข้างต้นทำให้มั่นใจได้ว่า แต่ละนโยบาย π_{k+1} จะมีค่าที่ดีกว่า π_k หรืออย่างน้อยก็เท่ากัน ในกรณีที่มีนโยบายที่ดีที่สุดสองทาง นอกจากนี้วิธีการดังกล่าวทำให้เชื่อมั่นได้อีกว่าจะสามารถลู่เข้าสู่นโยบายที่ดีที่สุดได้ด้วยมูลค่าฟังก์ชั่นสูงที่สุด

2.4 วิธีออนโพลิซีมอนติ คาร์โล

รายงานฉบับนี้นำเสนอกระบวนการเลือกเส้นทางที่ใช้พลังงานอย่างมีประสิทธิภาพสำหรับเครือข่ายเคลื่อนที่แบบแอดฮอค (MANET) ด้วยการประยุกต์ใช้กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ที่แบ่งการเรียนรู้เป็นเอพพิโซดด้วยวิธีการที่เรียกว่า ออนโพลิซี มอนติ คาร์โล (the on-policy Monte Carlo หรือ ONMC) วิธีการนี้เน้นการเรียนรู้แบบเป็นเอพพิโซดสำหรับประเมินว่าสถานะหรือการกระทำใดที่เหมาะสมในการดำเนินงานระยะยาว ฟังก์ชั่นดังกล่าวเรียกว่า ฟังก์ชั่นมูลค่าการกระทำ ซึ่งเป็นฟังก์ชั่นคู่สถานะ-การกระทำที่ใช้กำหนดปริมาณเฉลี่ยของผลตอบแทนที่ผู้เรียนใช้คาดคะเนเพื่อให้ได้ผลตอบแทนสูงสุดในระยะยาวจากผลตอบแทนเฉลี่ยย้อนกลับที่ได้รับจากคู่สถานะ-การกระทำ

วิธีออนโพลิซีจะพยายามประเมินหรือปรับปรุงนโยบายที่เกิดขึ้นในปัจจุบันเพื่อใช้ในการตัดสินใจเลือกการกระทำ วิธีทั่วไปพยายามทำให้มั่นใจว่าทุกการกระทำได้ถูกเลือกอย่างต่อเนื่องจนเป็นอนันต์แล้ว กำหนดให้ S เป็นชุดของสถานะที่เป็นไปได้ทั้งหมด และ A เป็นชุดของการกระทำที่เป็นไปได้ทั้งหมด สมมุติให้การกระทำที่ถูกเลือกในเอพพิโซด t ถูกควบคุมโดยนโยบาย π_t เมื่อ $\pi_t : S \to A$ กำหนดให้ฟังก์ชั่นมูลค่าการกระทำที่ (s,a) โดย $Q^{\pi_t}(s,a)$ คือผลตอบแทนที่ถูกคาดหวังว่าจะได้รับจาก (s,a) และนโยบาย π_t จะถูกปฏิบัติตามอีกครั้งในภายหลัง กำหนดให้นโยบายเริ่มต้นเป็น π_0 และ $Q^{\pi_t}(s,a)$ เริ่มต้นที่จุดเริ่มต้นในแต่ละเอพพิโซด สำหรับเอพพิโซด t ใดๆ จะเลือกการกระทำที่เป็นไปได้จากสถานะนั้นตามนโยบาย π_t เมื่อเอพพิโซด t จบลงค่า $Q^{\pi_t}(s,a)$ จะถูกอัพเดทตามสมการดังนี้

$$Q^{\pi_t}(s,a) = Q^{\pi_{t-1}}(s,a) + \frac{1}{t}\left[\sum_{n=\tau_t(s,a)}^{N_{t-1}} g(s_n,a_n) - Q^{\pi_{t-1}}(s_n,a_n)\right]. \qquad (2.11)$$

เมื่อ N_t คือช่วงเวลาหรือจำนวนครั้งของขั้นเวลา (time step) ในเอพพิโซด t และ $\tau_t(s,a)$ เมื่อ $0 \le \tau_t(s,a) \le N_t$ คือ ขั้นเวลาเมื่อเกิดการสำรวจคู่สถานะ-การกระทำ (s,a) ในเอพพิโซด t และ $g(s,a)$ ผลตอบแทนที่ได้รับจากการเลือก การกระทำ a ที่สถานะ s โดยที่เทอมของผลรวมคือค่าผลตอบแทนสะสมที่เกิดจากการสำรวจคู่สถานะ-การกระทำ (s,a) แรก

นโยบายใหม่สำหรับเอพพิโซดถัดไป π_{t+1} ถูกปรับปรุงมาจากนโยบายเดิม π_t ซึ่งใช้นโยบาย อีกรีดี (the ε-greedy policy) ในกระบวนการปรับปรุงดังนี้

$$\pi_{t+1}(s)=\begin{cases} a^* & \text{with probability } 1-\varepsilon+\dfrac{\varepsilon}{|\mathrm{A}|} \\ a\in A-\{a^*\} & \text{with probability } \dfrac{\varepsilon}{|\mathrm{A}|}, \end{cases} \tag{2.12}$$

เมื่อ a^* คือ นโยบายกรีดีที่ถูกพบโดย $a^*=\arg\max_{a\in A}\{Q^{\pi_t}(s,a)\}$, $\varepsilon\in[0,1]$ และ $|A|$ คือขนาดของชุดการกระทำ ภายใต้เงื่อนไขดังกล่าวจะทำให้นโยบาย ε-greedy ซึ่งได้จาก Q^{π} ถูกการันตีว่าเหมาะสมกว่าหรือเทียบเท่า π

2.5 สรุป

เนื้อหาบทนี้กล่าวถึงแนวคิดของกระบวนการตัดสินใจแบบมาร์คอฟ การแนะนำแนวคิดของกระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์เพื่อหาผลเฉลยของปัญหาที่มีการตัดสินใจแบบมาร์คอฟ กรอบงานของกระบวนการตัดสินใจแบบมาร์คอฟถูกนำมาใช้ในการกำหนดปัญหาการเลือกเส้นทางในเครือข่ายเคลื่อนที่แบบแอดฮอค สำหรับปัญหาการเลือกเส้นทางในรายงานเล่มนี้ได้แบ่งลักษณะงานออกเป็นเอพพิโซดเมื่อเอพพิโซดหนึ่งๆเริ่มต้นจากการที่โหนดต้นทางค้นหาเส้นทางไปยังโหนดปลายทาง เอพพิโซดสิ้นสุดเมื่ออย่างน้อยมีหนึ่งเส้นทางที่ถูกพบ ด้วยเหตุนี้กระบวนการเรียนรู้แบบรีอินฟอร์สเมนท์ที่แบ่งการเรียนรู้ออกเป็นเอพพิโซดด้วยวิธีการที่เรียกว่า วิธีออนโพลิซี มอนติ คาร์โลจึงถูกอธิบายในบทนี้ สำหรับบทถัดไปจะเสนอถึงการกำหนดปัญหาการเลือกเส้นทางที่รองรับคุณภาพการบริการในเครือข่ายเคลื่อนที่แบบแอดฮอค ด้วยวิธีออนโพลิซี มอนติ คาร์ พร้อมทั้งเปรียบเทียบประสิทธิภาพของวิธีออนโพลิซี มอนติ คาร์กับวิธีเลือกเส้นทางที่มีอยู่เดิมเช่น [10], [11]