การจัดเส้นทางบนพื้นฐานของการเรียนรู้แบบเสริมแรงด้วยเร็พพิวเทชันของเส้นทางในโครงข่ายเซนเซอร์ไร้สายเพื่อการประยุกต์ใช้ในการเตือนเหตุอุทกภัย

นางสาวณัฐธิดา ขาวสะอาด

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557

REINFORCEMENT LEARNING-BASED ROUTING WITH PATH REPUTATION IN WIRELESS SENSOR NETWORK FOR FLOOD WARNING APPLICATION

Miss. Nuttida Khawsaard

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Electrical Engineering
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2014

หัวข้อวิทยานิพนธ์	การจัดเส้นทางบนพื้นฐานของการเรียนรู้แบบเสริมแรงด้วยเร็พพิวเทชันของเส้นทางในโครงข่ายเซนเซอร์ไร้สายเพื่อการประยุกต์ใช้ในการเตือนเหตุอุทกภัย
โดย	นางสาวณัฐธิดา ขาวสะอาด
สาขาวิชา	วิศวกรรมไฟฟ้า
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.ชัยเชษฐ์ สายวิจิตร

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.บัณฑิต เอื้ออาภรณ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร.วาทิต เบญจพลกุล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.ชัยเชษฐ์ สายวิจิตร)

.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เชาวน์ดิศ อัศวกุล)

.....กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.ภูมิพัฒ แสงอุดมเลิศ)

ณัฐธิดา ขาวสะอาด : การจัดเส้นทางบนพื้นฐานของการเรียนรู้แบบเสริมแรงด้วยเร็พพิวเทชันของเส้นทางในโครงข่ายเซนเซอร์ไร้สายเพื่อการประยุกต์ใช้ในการเตือนเหตุอุทกภัย (Reinforcement Learning-Based Routing with Path Reputation in Wireless Sensor Network for Flood Warning Application)
อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร. ชัยเชษฐ์ สายวิจิตร, 71 หน้า.

วิทยานิพนธ์นี้มีจุดมุ่งหมายเพื่อพัฒนากรอบการวิเคราะห์เชิงคณิตศาสตร์เพื่อใช้ในการพัฒนาการเรียนรู้แบบอัตโนมัติเพื่อช่วยในกระบวนการตัดสินใจเลือกเส้นทางในโครงข่ายเซนเซอร์ไร้สายโดยประยุกต์ใช้ในระบบเตือนภัยอุทกภัย ซึ่งกระบวนการเรียนรู้ที่นำมาใช้คือการเรียนรู้แบบเสริมแรง (Reinforcement learning: RL) โดยเส้นทางที่ดีที่สุดจะถูกหาผ่านจากนโยบายซึ่งประเมินจากค่าของแอคชั่นแวลู่ฟังก์ชันโดยทดสอบผ่านโครงข่ายขนาดเล็กที่ระบบมีการเปลี่ยนแปลงแบบคงที่และโครงข่ายขนาดใหญ่แบบที่ระบบมีการเปลี่ยนแปลงแบบไม่คงที่

การประยุกต์ใช้วิธีการแบบมอนติคาร์โลในระบบเตือนภัยอุทกภัย โดยมีการนำมาใช้เพื่อช่วยในการตัดสินใจเลือกเส้นทางที่ดีที่สุดในโครงข่ายขนาดเล็กที่มีการควบคุมเงื่อนไขบังคับ (constraints) ทั้งหมดสามตัวแปรได้แก่ระดับพลังงาน อายุการใช้งาน และค่าเร็พพิวเทชันของเส้นทาง โดยค่าตัวแปรเงื่อนไขทั้งสามตัวได้ถูกนำมาพิจารณาเป็นเงื่อนไขสำคัญในโครงข่ายขนาดเล็ก โดยค่าเร็พพิวเทชันของเส้นทางเป็นตัวแปรสำคัญเพื่อทำให้การตัดสินใจเลือกเส้นทางนั้นสามารถหลีกเลี่ยงโอกาสของโนดที่จะเป็นจุดปัญหา (single point of failure) ของระบบได้ ผลการทดลองที่ได้พบว่าการนำวิธีการมอนติคาร์โลมาประยุกต์ใช้ในโครงข่ายขนาดเล็กสามารถทำงานได้อย่างมีประสิทธิภาพและทำการพิจารณาในโครงข่ายขนาดใหญ่เป็นลำดับถัดไป

การทดสอบในระบบโครงข่ายขนาดใหญ่ได้มีการพัฒนาแนวคิดด้วยการเพิ่มขีดความสามารถของวิธีการมอนติคาร์โลโดยการปรับเปลี่ยนตัวแปรสถานะจากมอนติคาร์โลอย่างง่าย ที่ไม่ได้นำข้อมูลที่เกี่ยวข้องกับระบบมารวมเข้ากับตัวแปรสถานะให้นำเข้ามาพิจารณาร่วมเพื่อปรับปรุงนโยบายของการเลือกเส้นทางที่จะเปลี่ยนไป โดยการทดสอบได้นำเข้าสู่การพิจารณาเรื่องของผลกระทบของฟังก์ชันผลรางวัลต่อการตัดสินใจของมอนติคาร์โลในสถานการณ์ต่างๆ และการกระจายภาระงานที่เหมาะสม (optimal load balancing) โดยพบว่าสำหรับตัวแปรสถานะที่มีค่าพลังงานปัจจุบันและค่าเร็พพิวเทชัน และการใช้ฟังก์ชันผลรางวัลที่มีสามตัวแปร ที่มีการถ่วงน้ำหนักอย่างเท่าเทียม เช่นเดียวกับในโครงข่ายขนาดเล็กเป็นหนึ่งในคำตอบที่เป็นไปได้และในการทดลองลำดับสุดท้ายจะเป็นการพิจารณาเปรียบเทียบความสามารถในการรักษาการเชื่อมต่อของโครงข่ายได้โดยทำการเปรียบเทียบกับวิธีการที่มีอยู่เดิมซึ่งพบว่าความสำคัญของการเพิ่มโอกาสเลือกตัวกระทำใหม่โดยเปิดโอกาสให้ระบบเข้าถึงทุกสภาวะของสถานะจะเป็นวิธีที่เหมาะสมที่สุดในระบบที่มีการเปลี่ยนแปลงสูง เมื่อพิจารณาเปรียบเทียบกับวิธีอื่นจะพบว่าความสามารถในการรักษาการเชื่อมต่อของโครงข่ายนั้นดีขึ้นถึง 20% เมื่อเทียบกับวิธีมาตรฐานจากการเลือกเส้นทางที่สั้นที่สุด สำหรับการพิจารณาในด้านของความซับซ้อนของระบบเมื่อเทียบกับวิธีการที่นำเสนอและข้อจำกัดในการนำวิธีการมอนติคาร์โลไปใช้รวมถึงข้อเสนอของงานวิจัยในอนาคตได้ถูกแสดงในวิทยานิพนธ์ฉบับนี้

ภาควิชา วิศวกรรมไฟฟ้า ..... **ลายมือชื่อนิสิต** .....
สาขาวิชา วิศวกรรมไฟฟ้า ..... **ลายมือชื่ออ.ที่ปรึกษาหลัก** .....
ปีการศึกษา 2557 .....

5470192021 : MAJOR ELECTRICAL ENGINEERING
KEYWORDS: WIRELESS SENSOR NETWORKS/ LOAD BALANCING/ ENERGRY/ LIFTIME/ PATH REPUTATION/ MONTE-CARLO/ FLOODING APPLICATION.

NUTTIDA KHAWSA-ARD : REINFORCEMENT LEARNING-BASED ROUTING WITH PATH REPUTATION IN WIRELESS SENSOR NETWORK FOR FLOOD WARNING APPLICATION.
ADVISOR: ASST. PROF. CHAIYACHET SAIVICHIT, Ph.D., 71 pp.

This dissertation aims at developing a mathematical framework by using an automated learning to help in the decision making in a route selection process of wireless sensor network in flood warning application. The automated learning is the reinforcement learning (RL) where the best possible route will be determined from the policy (set of actions) obtained from the action-value function by considering on the small-scale network scenario with a static topology and the large-scale network scenario with a dynamic topology.

The application of the monte Carlo algorithm in flood warning application has been used to determine the best possible route selection with three constraints which are remaining energy, remaining lifetime and path reputation. These three constraints have been considered as a reward function in small-scale network scenario. By using path reputation, this technique can be alleviated the single point of failure of the system. The preliminary result confirms that the monte Carlo algorithm performs well and effective. Therefore, the investigation on a large-scale network has been considered.

The experimental settings for a large-scale network begin with the consolidation of Monte Carlo algorithm by changing state variable from simple to the advance. Previously, the state variable has not incorporated the environmental status but it will be included in a large-scale network. The main idea is to extend the proposed algorithm towards the actual environment scenarios. Thus, the effect of reward function to the solution of monte Carlo algorithm has been observed as well as the optimal load balancing. Only remaining energy and its reputation are required for being state variables. From the investigations, the reward function with three components (remaining energy, remaining lifetime and the path reputation value) performs well. Note that there are many possible weighted functions can be used but the results in terms of total remaining energy from all sensor nodes are slightly different. Finally, the weight has been chosen equally. The performance measurement has been done in terms of the ability to maintain the network connectivity by balancing and sharing the traffic load to alternative routes. The benchmarking methods are shortest path, max-min method, uniform random, monte Carlo with non-intelligence and monte Carlo with intelligence. The results show that the proposed method can guarantee the link connectivity time longer than the worst case method up to 20% (the shortest path). Finally, the computational complexity is considered as well as the possibility work towards future.

Department : Electrical Engineering	**Student's Signature** .....
Field of Study : Electrical Engineering	**Advisor's Signature** .....
Academic Year : 2014	

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่ง จากอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.ชัยเชษฐ์ สายวิจิตร ซึ่งได้ให้ความรู้และคำแนะนำอันมีค่ายิ่งต่อนิสิต รวมทั้งได้มอบหมายงานที่เป็นประโยชน์ที่ทำให้นิสิตแนวความคิดและเติมเต็มความรู้นอกเหนือความรู้ที่ได้รับจากการทำวิทยานิพนธ์ ตลอดจนเมตตาและใส่ใจต่อนิสิตเสมอมา จึงใคร่ของกราบขอบพระคุณมา ณ ที่นี้ ขอขอบพระคุณ รศ.ดร.วาทิต เบญจพลกุล ประธานกรรมการสอบวิทยานิพนธ์ ผศ.ดร.เชาวน์ดิศ อัศวกุลและ รศ.ดร.ภูมิพัฒ แสงอุดมเลิศ กรรมการสอบวิทยานิพนธ์ที่ได้สละเวลาตรวจสอบและให้คำแนะนำเพื่อให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้นและขอขอบพระคุณคณาจารย์ทุกท่านในสาขาวิชาไฟฟ้าสื่อสารที่ได้ประสิทธิประสาทความรู้อันเป็นพื้นฐานในการศึกษาและทำวิทยานิพนธ์นี้

นอกจากนี้วิทยานิพนธ์ฉบับนี้ได้รับทุนสนับสนุนจากโครงการขับเคลื่อนการวิจัย กองทุนรัชดาภิเษกสมโภช (Special Task Force for Activating Research (STAR) ภายใต้กลุ่มวิจัยโครงข่ายไร้สายและอินเทอร์เน็ตอนาคต (Wireless Network and Future Internet Research Group) จุฬาลงกรณ์มหาวิทยาลัย

ขอขอบคุณเพื่อนๆ รุ่นพี่และรุ่นน้องในห้องปฏิบัติการวิจัยโทรคมนาคมที่ให้กำลังใจและคำปรึกษาโดยเฉพาะอย่างยิ่งกลุ่มวิจัยโครงข่ายไร้สายและอินเทอร์เน็ตอนาคต ซึ่งดูแลโดย ผศ.ดร.เชาวน์ดิศ อัศวกุล ผศ.ดร.ชัยเชษฐ์ สายวิจิตร และผศ.ดร.กุลธิดา โรจน์วิบูลย์ชัย ที่จัดกิจกรรมเพื่อส่งเสริมการเรียนรู้และการทำงานให้มีประสิทธิภาพที่ดียิ่งขึ้น ทำให้งานวิทยานิพนธ์นี้สำเร็จได้อย่างสะดวกราบรื่น ขอขอบคุณ ดร.ปิติพงศ์ ชาญโลหะ และนายธีรพล ศิลาวรรณ์ สำหรับคำแนะนำในการแก้ปัญหาและข้อคิดเห็น อันเป็นประโยชน์ต่องานวิทยานิพนธ์นี้ด้วยดีเสมอมา

ขอขอบคุณ โครงการศิษย์ก้นกุฏิของภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนค่าเล่าเรียนและค่าใช้จ่ายรายเดือนตลอดการศึกษา

ขอขอบคุณห้องปฏิบัติการวิจัยโทรคมนาคม ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สำหรับทรัพยากรต่างๆในการศึกษาค้นคว้าและวิจัย

สุดท้ายนี้กราบขอบพระคุณบิดามารดาและครอบครัว ซึ่งได้ให้การสนับสนุนและเป็นกำลังใจให้แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษาระดับมหาบัณฑิต

สารบัญ

บทที่ หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

คำอธิบายสัญลักษณ์และคำย่อ

MANET	=	Mobile ad hoc network
WSN	=	Wireless sensor network
MP	=	Markov property
MDP	=	Markov decision process
RL	=	Reinforcement learning
MC	=	Monte Carlo
ONMC	=	On-policy Monte Carlo
s	=	State
a	=	Action
r	=	Reward
$\mathbf{S}$	=	State vector
$\mathcal{S}$	=	State space
A	=	Action space
π	=	Policy
ε	=	Greedy policy
V^{π}	=	State-value function
Q^{π}	=	Action-value function
α_i	=	Remaining energy of the sensor node i
β_j	=	Remaining energy of the actor node j
η	=	Time-invariant fraction of energy leaked
χ	=	Total energy consumption for one communication pairs
$\tau_{i,i'}$	=	Transmitting energy consumption
$\mu_{i'',i}$	=	Receiving energy consumption
$\mathcal{R}$	=	Reputation value

บทที่ 1

บทนำ

ระบบการสื่อสารแบบไร้สายในปัจจุบันได้รับความสนใจเป็นอย่างมากทั้งในด้านการศึกษาและในด้านอุตสาหกรรม เป็นผลให้เกิดความก้าวหน้าด้านเทคโนโลยีการให้บริการด้านการสื่อสารที่มีความสามารถและหลากหลายเพื่อตอบสนองความต้องการในการสื่อสารที่แตกต่างกัน ขณะเดียวกันอุปกรณ์สำหรับการติดต่อสื่อสารนั้นก็ได้รับการออกแบบเพื่อพัฒนาประสิทธิภาพ ความคุ้มค่า ความปลอดภัย และความคล่องตัวในการใช้งาน ดังนั้นการสื่อสารแบบไร้สายเป็นเทคโนโลยีหนึ่งที่ได้ความนิยมอย่างมาก อย่างไรก็ตามการนำเทคโนโลยีนี้มาประยุกต์ใช้ก็ยังมีข้อจำกัดในการใช้งานเมื่อเกิดสภาวการณ์ในการส่งข้อมูลที่ต้องการความรวดเร็วและต่อเนื่อง เทคโนโลยีการสื่อสารไร้สายจึงได้รับการพัฒนารูปแบบการเชื่อมต่อ อาทิการเชื่อมต่อแบบกลุ่มโครงสร้าง (infrastructure) ลักษณะการเชื่อมต่อที่มีอุปกรณ์กระจายสัญญาณ (access point) ที่เพียงพอต่อความต้องการเข้าใช้ ซึ่งหากปริมาณความต้องการเข้าใช้งานมากเกินความสามารถรองรับการให้บริการ จะมีผลกระทบต่อคุณภาพการให้บริการโครงข่ายลดลง ผู้ให้บริการโครงข่ายจึงจำเป็นต้องขยายโครงข่ายเพื่อตอบสนองความต้องการที่เพิ่มขึ้นของผู้ใช้งาน เพื่อรองรับประสิทธิภาพการให้บริการจึงเกิดการเชื่อมต่อแบบกลุ่มเฉพาะกิจหรือแอดฮอก (ad-hoc) กลายเป็นทางเลือกใหม่เพื่อตอบโจทย์ในด้านความคุ้มค่า เมื่อเทียบงบประมาณการลงทุนและความซับซ้อนที่ต่ำในการติดตั้งโครงข่ายพื้นฐาน

โครงข่ายแอดฮอกเป็นโครงข่ายที่มีการติดต่อสื่อสารโดยไม่มีการอำนวยความสะดวกจากส่วนกลาง เป็นลักษณะการเชื่อมต่อหนึ่งในโครงข่ายไร้สายที่มีการส่งแพ็กเกตข้อมูลโดยตรงระหว่างโนดตัวรับ-ส่ง ภายในระยะการส่งข้อมูล (transmission range) แต่เนื่องจากการสื่อสารแบบโครงข่ายแอดฮอกเป็นการติดต่อสื่อสารแบบไม่มีการอำนวยความสะดวกจากส่วนกลาง (infrastructureless) เป็นการทำงานแบบกระจายศูนย์ (distributed) ในตัวของโครงข่ายเอง ทำให้การติดต่อแบบไร้สายเพื่อส่งแพ็กเกตข้อมูลระหว่างกันของโนดที่มีการเคลื่อนที่อย่างอิสระมีความไม่แน่นอนและไม่สามารถทำนายได้ล่วงหน้า รูปแบบการเชื่อมต่อของโนดหรือโครงรูปของโครงข่ายมีการเปลี่ยนแปลงอยู่ตลอดเวลา อีกทั้งการสื่อสารแอดฮอกเป็นการสื่อสารแบบหลายช่วงเชื่อมต่อ (multi-hop communication) ซึ่งหากโนดต้นทางต้องการส่งแพ็กเกตข้อมูลไปยังโนดปลายทางที่อยู่ในตำแหน่งไกลกว่าระยะการส่งข้อมูล การสื่อสารดังกล่าวจำเป็นต้องอาศัยโนดระหว่างทาง (intermediate node) ทำหน้าที่ถ่ายทอดข้อมูลดังกล่าวไปยังโนดปลายทาง ด้วยเหตุนี้ระบบโครงข่ายแอดฮอกจึงสามารถขยายระยะการเชื่อมต่อระหว่างโนดต้นทางและโนดปลายทางไกลกว่าโครงข่ายที่ต้องมีสถานีฐาน

1.1 ความเป็นมาและความสำคัญของปัญหา

ช่วงหลายปีที่ผ่านมาเซนเซอร์ไร้สายถูกนำมาพัฒนาความสามารถชิ้นส่วนอุปกรณ์ เป็นผลให้ปัจจุบันโครงข่ายเซนเซอร์ไร้สายจึงกลายมาเป็นรูปแบบหนึ่งของการสื่อสารผ่านโครงข่ายแอดฮอกด้วยเทคโนโลยีที่ร่วมกันสื่อสารผ่านเซนเซอร์ ซึ่งเป็นอุปกรณ์ขนาดเล็กกะทัดรัด เคลื่อนย้ายสะดวก นอกจากนี้ยังใช้พลังงานต่ำและมีอายุการใช้งานยาวนาน ภายในประกอบไปด้วยหน่วยประมวลผลและหน่วยรับ-ส่งข้อมูลภายในโครงข่ายเซนเซอร์ไร้สาย ทำให้โครงข่ายเซนเซอร์ไร้สายสามารถนำไปประยุกต์ใช้ในหลายแขนงสายงาน เช่น ด้านการแพทย์, ด้านอุตสาหกรรม, ด้านการเกษตรกรรม และ

ด้านความมั่นคง เป็นต้น

ตัวอย่างการนำเทคโนโลยีโครงข่ายเซนเซอร์ไร้สายมาประยุกต์ใช้ เช่น การประยุกต์ใช้ในเหตุการณ์ภัยพิบัติทางธรรมชาติ โครงข่ายเซนเซอร์ไร้สายสามารถรวบรวมและจัดการข้อมูล แสดงผลการทำงาน การตรวจรับส่งข้อมูลผ่านโครงข่าย เพื่อใช้ประโยชน์ในคาดการณ์ การติดตามเฝ้าระวัง แจ้งเตือนสถานการณ์เพื่อพยายามป้องกันผลกระทบหรือลดความสูญเสียต่อชีวิตและทรัพย์สิน ด้วยเหตุนี้งานวิจัยเล่มนี้จึงมองจากผลกระทบของอุทกภัยของประเทศไทยในปี พ.ศ. 2554 ที่ผ่านมาและปัญหาของระบบพยากรณ์น้ำและเตือนภัยอุทกภัย ที่ไม่สามารถรับส่งข้อมูลอุตุอุทกวิทยาที่ทันต่อสถานการณ์ จึงจำเป็นต้องปรับปรุงเสถียรภาพและความน่าเชื่อถือของระบบการตรวจวัดและการรับส่งข้อมูลอย่างต่อเนื่อง นอกจากนี้เนื่องจากการนำโครงข่ายเซนเซอร์ไร้สายมาประยุกต์ใช้ในสถานการณ์การแจ้งเตือนอุทกภัยนั้น โนดในโครงข่ายจำนวนมากต้องการติดต่อสื่อสารระหว่างกันโดยปราศจากสถานีฐาน แต่เนื่องจากสถานการณ์ การเข้าเปลี่ยนอุปกรณ์หรือการเปลี่ยนแบตเตอรี่ของโนดเป็นเรื่องที่ทำได้ลำบาก ข้อจำกัดทางกายภาพหรือภาวการณ์เช่นนี้ส่งผลให้การใช้พลังงานอย่างมีประสิทธิภาพรวมไปถึง การยืดอายุการใช้งานโครงข่าย ยังคงเป็นประเด็นสำคัญต่อเสถียรภาพของระบบโครงข่าย

เพื่อปรับปรุงเสถียรภาพ (stability) และความน่าเชื่อถือ (reliability) โครงข่ายเซนเซอร์ไร้สายบนแม่น้ำจำต้องเพิ่มความสามารถในการจัดสรรเส้นทาง (routing) ทำหน้าที่เสมือนเฝ้าตรวจสถานะโครงข่ายเมื่ออุปกรณ์บางตัวในโครงข่ายขัดข้องหรือได้รับผลกระทบจากข้อจำกัด เช่น พลังงาน อายุการใช้งาน รวมไปถึงสภาพแวดล้อมภายนอก โครงข่ายนั้นจึงควรมีความสามารถหาเส้นทางใหม่เพื่อคงสมรรถภาพเมื่อเกิดการเปลี่ยนแปลงของโครงรูปของโครงข่าย ฉะนั้นเพื่อพัฒนาประสิทธิภาพของโครงข่าย การออกแบบโพรโทคอลการจัดสรรเส้นทาง (routing protocol) สำหรับโครงข่ายแบบเซนเซอร์ไร้สายจึงประเด็นหลักอย่างหนึ่งและยังคงเป็นประเด็นที่นักวิจัยให้ความสนใจในการพัฒนาอย่างต่อเนื่อง โพรโทคอลการจัดหาเส้นทางโดยทั่วไปสามารถแบ่งออกเป็น 2 ประเภทหลัก คือ ประเภทที่หนึ่งเป็นวิธีการจัดเส้นทางแบบเตรียมเส้นทางไว้ล่วงหน้าหรือโปรแอคทีฟ (proactive หรือ table-driven) และประเภทที่สองเป็นการจัดเส้นทางเมื่อต้องการส่งข้อมูลหรือรีแอคทีฟ (reactive หรือ on-demand) โพรโทคอลทั้งสองประเภทนี้แตกต่างกันที่รูปแบบการจัดสรรเส้นทางและกระบวนการในการปรับปรุงข้อมูลเส้นทาง โดยที่โพรโทคอลการจัดหาเส้นทางแบบเตรียมเส้นทางไว้ล่วงหน้าจะให้แต่ละโนดเก็บข้อมูลเส้นทางการเชื่อมโยงกับโนดอื่นๆทั้งหมดของโครงข่ายในตารางการจัดเส้นทาง ทำให้สามารถส่งข้อมูลได้ทันทีเมื่อต้องส่งแพ็กเกตข้อมูล โดยพยายามที่จะรักษาการเชื่อมโยง (connectivity) และอัพเดตข้อมูลเส้นทางในทุกโนด จะเหมาะสมกับสภาพโครงสร้างของโครงข่ายที่มีการเปลี่ยนแปลงเล็กน้อย ตัวอย่างโพรโทคอลที่อยู่ในประเภทนี้คือ destination-sequenced distance vector (DSDV) [1] และ optimized link-state routing (OLSR) [2] ส่วนการจัดเส้นทางประเภทที่สองเมื่อต้องการส่งข้อมูล จะเป็นการค้นหาเส้นทางการส่งข้อมูลเมื่อโนดมีความต้องการจะส่งข้อมูลใช้การสื่อสารบนความกว้างแถบความถี่ (bandwidth) ได้อย่างมีประสิทธิภาพ ซึ่งเหมาะสมกับสภาพโครงสร้างของโครงข่ายไร้สายที่มีการเปลี่ยนแปลงบ่อย ตัวอย่างโพรโทคอลที่อยู่ในประเภทนี้คือ ad hoc on-demand distance vector (AODV) [3] และ dynamic source routing (DSR) [4]

กระบวนการจัดสรรเส้นทางของโครงข่ายเซนเซอร์ไร้สายสามารถแบ่งได้เป็นสามขั้นตอน ประกอบด้วย การสร้างเส้นทางโครงข่าย (route construction) การตัดสินใจเลือกเส้นทางภายในโครงข่าย (route selection) และการซ่อมบำรุงเส้นทาง (route maintenance) ในขั้นตอนของการสร้างเส้นทางโครงข่าย จะเริ่มตั้งแต่การพิจารณาการค้นหาเส้นทางโดยมีการร้องขอเส้นทางและการตอบกลับของแพ็จเกจร้องขอ โดยมีหลายเส้นทางของการเชื่อมต่อของอุปกรณ์ภายในโครงข่ายเพื่อจะใช้ในการตัดสินใจว่าเส้นทางใดสามารถใช้งานได้และเส้นทางใดเหมาะสม ในขั้นตอนของการตัดสินใจเลือก

เส้นทางภายในโครงข่าย จะเป็นช่วงการพิจารณาการตัดสินใจว่าควรเลือกเส้นทางใดในกรณีที่มีหลายเส้นทางที่สามารถใช้งานได้พร้อมกัน โดยการตัดสินใจจะพิจารณาจากค่าชี้วัดหลายอย่าง เช่น การใช้พลังงาน สถานะของแบตเตอรี่ของอุปกรณ์ในขณะนั้น อายุการใช้งานของอุปกรณ์ หรือความน่าเชื่อถือในการส่งข้อมูลของอุปกรณ์นั้น ในขั้นตอนการซ่อมบำรุงเส้นทาง เป็นการพิจารณาในสภาวะที่อุปกรณ์เสียหายมีความต้องการในการซ่อมบำรุงตัวอุปกรณ์นั้น

งานวิจัยนี้จะเลือกพิจารณาการจัดสรรเส้นทางแบบรีแอคทีฟซึ่งพิจารณากระบวนการตัดสินใจเลือกเส้นทางของโครงข่ายเซนเซอร์ไร้สาย โดยคำนึงถึงพลังงานที่เหลืออยู่ อายุการใช้งานรวมไปถึงความน่าเชื่อถือของการรับส่งข้อมูลของอุปกรณ์ เป็นปัจจัยสำคัญในการตัดสินใจเพื่อเลือกเส้นทางที่ดีที่สุด งานวิจัยที่ผ่านมาได้มีการนำเสนอเกี่ยวกับอัลกอริทึมต่างๆ ที่ใช้ในการตัดสินใจเลือกเส้นทาง 2 ลักษณะ คือ ลักษณะวิธีการที่มีการตัดสินใจแบบไม่ฉลาด (non-intelligent) และวิธีการแบบฉลาดหรือวิธีการที่ระบบสามารถพัฒนาและปรับปรุงตน (intelligent) ซึ่งลักษณะวิธีการที่มีการตัดสินใจแบบไม่ฉลาดหรือวิธีการที่ระบบไม่สามารถพัฒนาและปรับปรุงตน จะมีการการตัดสินใจโดยการใช้กฎเกณฑ์ (rule based) ส่วนใหญ่จะใช้กับลักษณะรูปแบบที่มีความแน่นอนและชัดเจน ยกตัวอย่าง เช่น การตัดสินแบบสุ่ม (randomized) [5] การเลือกเส้นทางที่สั้น (shortest path) โดยมีจุดประสงค์หลักเพื่อประหยัดพลังงานในการส่งข้อมูลหลายๆฮอพ [6] [7] หรือสามารถรักษาอายุการใช้งานของแต่ละโนดได้ [8] การเลือกเส้นทางแบบละโมบ (greedy) ภายใต้เงื่อนไขบังคับ เช่น พลังงานที่เหลืออยู่ของเซนเซอร์โนด [7] และการเลือกแบบตรรกศาสตร์คลุมเครือหรือฟัซซี่ลอจิก (fuzzy logic) [9] โดยแบ่งการตัดสินใจออกเป็นส่วนๆและมีการตีความในรูป If-Then คือไม่ถูกก็ผิดเพียงสองสถานะ ซึ่งจะมีดีกรีความถูกผิดขึ้นอยู่กับเหตุการณ์ที่พิจารณาในโครงข่ายเซนเซอร์ไร้สาย ซึ่งจะเห็นได้ว่าลักษณะของวิธีการตัดสินใจแบบไม่ฉลาดจะไม่มีกระบวนการปรับปรุงโครงสร้างของกฎและตัวแปรในระบบ ดังนั้นจึงไม่สามารถแก้ปัญหาระบบที่มีการเปลี่ยนแปลงตลอดเวลาหรือฉับพลันและการทำงานของระบบที่มีการตัดสินใจในลักษณะเช่นนี้จะทำงานไม่ได้เต็มประสิทธิภาพเนื่องจากมีเงื่อนไขจำกัดของตนเอง (hard policy) ด้วยเหตุนี้จึงมีนักวิจัยได้นำเสนอวิธีการสินใจแบบฉลาดเพื่อนำมาแก้ปัญหาข้อจำกัดที่เกิดขึ้น โดยอาศัยกระบวนการปรับปรุงกฎการตัดสินใจ ยกตัวอย่างเช่น การเลือกเส้นทางแบบอาณาจักรมด (ant colony) [10] จะมีการตัดสินใจเลือกเส้นทางโดยการจำลองพฤติกรรมของมด เคลื่อนที่ผ่านปริภูมิพารามิเตอร์ซึ่งในที่นี้คือพลังงานของแต่ละโนดและจะมีการเลือกใช้ค่าน้ำหนักของฟิโรโมนและฮิวริสติคในการป้อนกลับที่เป็นประโยชน์ (positive feedback) สำหรับการนำไปสู่คำตอบที่ดีที่สุด การเลือกเส้นทางแบบโครงข่ายประสาท (neural network) [11] หลักการของงานวิจัยดังกล่าวจะมีการจำลองทำงานเสมือนสมองสิ่งมีชีวิตสำหรับการตัดสินใจ ซึ่งจะมีความสามารถในการปรับตัวและประมวลผลจากแบบขนานจากอินพุตไปยังเอาต์พุต โดยขั้นตอนจะมีการปรับเปลี่ยนค่าของน้ำหนักประสาทหรือเงื่อนไขในการฝึกฝนของโครงข่ายคือพลังงานสูญเสียของเซนเซอร์โนด และใช้วิธีการแก้ปัญหาแบบโปรแกรมเชิงเส้น (linear programming) การเลือกเส้นทางแบบเชิงพลวัต [12] งานวิจัยดังกล่าวนำเอาวิธีการแบบเชิงพลวัตมาใช้ในการแก้ปัญหาแบบไม่เป็นโพลิโนเมียลหรือเอ็นพีคอมพรีท (NP-complete) สำหรับการเลือกกลุ่มของเส้นทางที่ดีที่สุด แต่อย่างไรก็ตามกลุ่มของวิธีการสินใจแบบฉลาดแบบดังกล่าว ยังคงมีข้อจำกัดในด้าน การใช้การคำนวณสูง มีความซับซ้อนของกระบวนการตัดสินใจ ต้องมีขนาดความจุในการเก็บตัวแปรที่ใหญ่ การนำไปประยุกต์ใช้งานได้ยากในกรณีขนาดของระบบใหญ่มาก และไม่สามารถแก้ปัญหาของระบบที่มีการเปลี่ยนแปลงเชิงเวลาสูง

ดังนั้นงานวิจัยนี้จึงเลือกวิธีการเรียนรู้แบบเสริมแรง (reinforcement learning) โดยมีการแก้ปัญหาแบบมอนติคาร์โล (Monte Carlo) มาใช้ในการลดความซับซ้อนของระบบ ซึ่งวิธีการดังกล่าวเป็นสาขาหนึ่งของการเรียนรู้ของเครื่อง (machine learning) [13] ประเภทการเรียนรู้แบบไม่มี

ผู้สอน (unsupervised learning) เป็นวิธีการเรียนรู้จากสภาพแวดล้อมของระบบด้วยตนเองและค้นหาคุณลักษณะของตนเองจากสภาพแวดล้อมดังกล่าว โดยกระบวนการทั้งหมดเกี่ยวข้องกับความสัมพันธ์ระหว่าง สภาพแวดล้อม (environment) และตัวกระทำการตัดสินใจของระบบ (agent) โดยตัวกระทำจะเลือกการกระทำ (action) ใดๆ จากชุดการกระทำที่เป็นไปได้ทั้งหมดในสถานะ (state) ปัจจุบัน เป็นผลให้สภาพแวดล้อมเปลี่ยนไปและผู้เรียนจะได้รับรางวัล (reward) ซึ่งขึ้นอยู่กับการกระทำดังกล่าวมีผลให้สภาพแวดล้อมเปลี่ยนแปลงไปในทางใด โดยจะแตกต่างจากประเภทการเรียนรู้แบบมีผู้ฝึกสอน (supervised learning) เป็นวิธีการเรียนรู้จากรูปแบบของอินพุตระบบหรือสร้างผลลัพธ์ที่ต้องการให้ได้ตามตัวอย่างที่ได้รับ ซึ่งเมื่อมีการเปลี่ยนแปลงเชิงแบบจำลองของระบบแล้วไม่ตรงกับรูปแบบของอินพุตที่เคยให้มา จะไม่สามารถหาผลของคำตอบได้

งานวิจัยที่ผ่านมาสำหรับการนำวิธีการมอนคาร์โลมาประยุกต์ใช้สำหรับการตัดสินใจเลือกเส้นทาง งานวิจัย [14] ได้มีการประยุกต์ใช้วิธีการมอนติคาร์โลในการตัดสินใจเลือกเส้นทางเพื่อหลีกเลี่ยงโนดประสงค์ร้าย (malicious node) สำหรับโครงข่ายเคลื่อนที่แบบแอดฮอกที่เป็นแบบไม่คงที่ในสภาวะปิด โดยในงานวิจัยถัดมา [15] เป็นการพัฒนาต่อยอดของวิธีการมอนติคาร์โลบนโครงข่ายชนิดเดียวกัน โดยทำการพิจารณาการเลือกเส้นทางเพื่อรักษาสมดุลของการใช้พลังงานและอายุการใช้งาน จากนั้นได้มีงานวิจัย [16] – [19] ที่ได้ทำการพิจารณาสภาวะของระบบที่มีพฤติกรรมของสภาพแวดล้อมที่ไม่คงที่และแก้ปัญหาด้วยวิธีการแบบปรับตัว (adaptive) มาใช้

ตารางที่ 1.1: ตารางเปรียบเทียบงานวิจัย

Comparison criterion	Nurmi 2007 [16]	Naruephiphat and Usaha 2008 [15]	Chettibi and Chikhi 2014 [19]	Our Proposition
Objective function	Maximum-lifetime routing	Balancing objectives of Maximum-lifetime routing and minimizing total transmission power	Maximum-lifetime routing	Balancing objectives of Maximum-lifetime routing, minimizing total transmission power and path reputation value
Other design considerations	The network contains selfish nodes	None	None	The network contains malicious nodes
RL algorithm	Stochastic gradient descent	First Visit On-Policy Monte Carlo	Q-Learning SARSA	Every Visit On-Policy Monte Carlo
State definition	Energy	Energy and lifetime	Lifetime	Energy and path reputation
The learned action	Next-hop for routing	Routing path	RREQs forwarding rate	Routing path

ตารางที่ 1.1 แสดงถึงการเปรียบเทียบงานวิจัยที่อยู่ในกลุ่มงานใกล้เคียงกันโดยสามารถอธิบายได้ดังนี้

สำหรับในด้านของวัตถุประสงค์ของงานวิจัยทั้ง 4 งานจะพบว่า ในงานวิจัย [16] และ [19] ได้เลือกวัตถุประสงค์ของงานวิจัยด้วยการเลือกค่าของอายุการใช้งานที่เหลืออยู่สูงที่สุดในการสร้างเส้นทาง แต่ในงานวิจัย [15] ได้มีการพิจารณาเพิ่มเติมจากอายุการใช้งานที่เหลือที่มากที่สุดแล้วยังได้พยายามเลือกเส้นทางที่ใช้พลังงานต่ำที่สุดอีกด้วย ส่วนในมุมมองของสภาวะโนดที่ผิดปรกตินั้นในงานวิจัย [16] ได้ทำการพิจารณาโนดที่มีปัญหาที่อาจจะไม่ส่งต่อข้อมูลซึ่งจะทำให้ระบบเกิดปัญหาได้ การนำกระบวนการแบบมอนติคาร์โลมาใช้งานนั้น การกำหนดค่าของตัวแปรสถานะเพื่อให้ตัวตัดสินใจเป็นตัวเลือกและพิจารณาผลการเลือกนั้นจากฟังก์ชันผลรางวัล หากค่าของตัวแปรเหล่านี้มี

ความสัมพันธ์กันอย่างเหมาะสม ย่อมจะทำให้ระบบเรียนรู้ได้เร็วและทำงานได้อย่างมีประสิทธิภาพ โดยในงานวิจัย [16] ใช้ระดับพลังงานเป็นตัวแปรสถานะ งานวิจัย [15] ได้เลือกพลังงานและอายุการใช้งาน งานวิจัย [19] ได้เลือกระดับอายุการใช้งาน ซึ่งจะมีความแตกต่างกันไปตามแต่ระบบที่พิจารณา เช่นเดียวกับค่าของตัวตัดสินใจที่มีความแตกต่างกันไปตามแต่ละระบบ

ดังนั้นในงานวิจัยที่นำเสนอนี้ได้พิจารณาการนำเอาระบบที่ใช้ระดับพลังงานและค่าของเร็พพิวเทชันของเส้นทาง ซึ่งการพิจารณาในเรื่องของระดับพลังงานกับอายุการใช้งานนั้นมีความใกล้เคียงกัน แต่อย่างไรก็ตามอายุการใช้งานยังเป็นเรื่องสำคัญ จึงได้นำอายุการใช้งานมาพิจารณาในฟังก์ชันผลรางวัลแทน โดยที่การนำค่าเร็พพิวเทชันของเส้นทางมาใช้นั้นจะครอบคลุมการพิจารณาโนดที่ประสงค์ร้ายได้อีกด้วยและสำหรับการเลือกใช้วิธีการแบบมอนติคาร์โลนั้นก็จะแตกต่างกับงานวิจัย [15] ที่นำการพิจารณาแบบวิธีการเอฟวรี-วิสซิทมอนติคาร์โลมาใช้ซึ่งจะสามารถรับประกันเรื่องของผลรางวัลตอบแทนเฉลี่ยระยะยาวได้ดีกว่าแบบวิธีการเฟิร์สท-วิสซิทมอนติคาร์โลซึ่งอัลกอริธึมที่เลือกใช้ในงานวิจัยนี้จะเป็นมอนติคาร์โลที่ใช้วิธีการแบบฮิวริสทิค (heuristic) เพื่อแก้ปัญหากับสภาพแวดล้อมในระบบเปิดแบบไม่คงที่ ซึ่งมอนติคาร์โลแบบปรกติจะไม่ได้นำมาพิจารณาในสภาวะในรูปแบบดังกล่าว

กล่าวโดยสรุป ในวิทยานิพนธ์ฉบับนี้ได้ประยุกต์เทคนิคการเรียนรู้แบบเสริมแรงที่แบ่งการเรียนรู้ออกเป็นเอพพิโซด (episode) ด้วยวิธีการที่เรียกว่า ออนโพลิซี มอนติ คาร์โล (On-policy Monte Carlo หรือ ONMC) เพื่อหาผลคำตอบของกระบวนการตัดสินใจเลือกเส้นทางเพื่อใช้ในระบบการเตือนเหตุอุทกภัย โดยจะพยายามรักษาสมดุลของฟังก์ชันวัตถุประสงค์คือพลังงาน อายุการใช้งานและค่าเร็พพิวเทชันของเส้นทาง

1.2 วัตถุประสงค์ของงานวิทยานิพนธ์

นำเสนอกรอบความคิดเชิงคณิตศาสตร์สำหรับวิธีการมอนติคาร์โล (monte carlo) มาประยุกต์ใช้ในกระบวนการตัดสินใจเลือกเส้นทางที่ดีที่สุด โดยคำนึงถึงการแบ่งกระจายภาระงาน (load balancing) ที่ควบคุมการใช้พลังงาน อายุการใช้งานและหลีกเลี่ยงโนดที่มีโอกาสที่จะมีปัญหาบนโครงข่ายเซนเซอร์ไร้สายโดยประยุกต์ใช้ในแบบจำลองการเตือนอุทกภัย

1.3 ขอบเขตวิทยานิพนธ์

1. ศึกษาวิธีการตัดสินใจเลือกเส้นทางเพื่อลดการใช้พลังงาน เพิ่มอายุการใช้งานของโครงข่ายเซนเซอร์ไร้สายในระบบเตือนภัยอุทกภัยในระบบที่มีสภาวะแวดล้อมคงที่และไม่คงที่
2. พัฒนากรอบความคิดเชิงคณิตศาสตร์ด้วยการประยุกต์ใช้อัลกอริธึมมอนติคาร์โลและฟังก์ชันผลรางวัลที่ประกอบด้วยพลังงาน อายุการใช้งานและเร็พพิวเทชันของเส้นทางเพื่อหานโยบายที่ดีที่สุดสำหรับการตัดสินใจเลือกเส้นทางในระบบเตือนภัยอุทกภัย
3. เปรียบเทียบสมรรถนะเชิงระบบ ทั้งผลดีและผลเสีย ของการประยุกต์ใช้มอนติคาร์โล ด้วยวิธีการเร็พพิวเทชันของเส้นทางกับวิธีอื่น เช่น การเลือกเส้นทางแบบวิธีสั้นที่สุด การเลือกเส้นทางแบบสุ่ม และการเลือกเส้นทางแบบค่าสูงสุดของค่าต่ำสุดพลังงาน (max-min) เป็นต้น
4. เปรียบเทียบสมรรถนะเชิงระบบ ในทอพอโลยีแบบคงที่ของเซนเซอร์โนด, ทอพอโลยีแบบไม่คงที่ของเซนเซอร์โนด และทอพอโลยีแบบไม่คงที่ที่มีความผิดปกติของเซนเซอร์โนด
5. พัฒนาโปรแกรม MATLAB® เพื่อใช้ในการประเมินผลวิธีการที่นำเสนอ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ข้อสรุปอันเป็นประโยชน์เกี่ยวกับขั้นตอนวิธีการการแก้ปัญหามอนติ คาร์โลสามารถหลีกเลี่ยงโนดที่ผิดปกติในโครงข่ายเซนเซอร์ไร้สาย
2. ได้ข้อสรุปอันเป็นประโยชน์เกี่ยวกับกระบวนการในการตัดสินใจที่สามารถเพิ่มประสิทธิภาพและสมรรถภาพในด้านการใช้พลังงาน อายุการใช้งานรวมถึงการได้เส้นทางที่มีความปลอดภัยโดยใช้วิธีการมอนติ คาร์โล ซึ่งสามารถเลือกเส้นทางที่ใกล้เคียงเส้นทางที่ดีที่สุดในโครงข่ายเซนเซอร์ไร้สายในสภาวะทอพอโลยีแบบไม่คงที่และมีความผิดปกติของเซนเซอร์โนด
3. สามารถนำผลการศึกษาที่ได้รับจากงานวิจัยไปประยุกต์ใช้งานกับสถานการณ์เตือนภัยพิบัติที่มีพื้นที่การพิจารณาขนาดใหญ่และสภาพแวดล้อมแบบไม่คงที่ได้อย่างมีประสิทธิภาพ

1.5 ประมวลวิทยานิพนธ์

บทที่ 1 บทนำ: กล่าวถึงลักษณะทั่วไปของโครงข่ายเซนเซอร์ไร้สาย การนำไปประยุกต์ใช้กับสถานการณ์เตือนเหตุอุทกภัย ความสำคัญของการควบคุมพลังงาน อายุการใช้งานรวมไปถึงความน่าเชื่อถือของอุปกรณ์ในโครงข่ายและการแก้ไขปัญหาที่เกิดขึ้นจากงานวิจัยในอดีต รวมไปถึงการเลือกใช้แบบจำลองที่เหมาะสมต่อการปรับปรุงสมรรถนะของโครงข่าย

บทที่ 2 หลักการและทฤษฎี: กล่าวถึงหลักการของการพิจารณาโครงข่ายแบบมาร์คอฟ หลักการของการเรียนรู้แบบเสริมแรงและวิธีการมอนติคาร์โล โดยมีคุณสมบัติและองค์ประกอบการนำไปใช้ในการแก้ปัญหา

บทที่ 3 การประยุกต์ใช้งานวิธีการมอนติคาร์โลด้วยการพิจารณาค่าเร็พพิวเทชันของเส้นทางในตำแหน่งคงที่: กล่าวถึงการนำวิธีการมอนติคาร์โลมาประยุกต์ใช้ในกระบวนการตัดสินใจเลือกเส้นทางในโครงข่ายเซนเซอร์ไร้สาย โดยมีการทดสอบสมมุติฐานอย่างง่ายในการนำวิธีการดังกล่าวมาใช้เพื่อลดการใช้พลังงาน อายุการใช้งานและความน่าเชื่อถือของโครงข่ายภายใต้การพิจารณาสภาวะแวดล้อมแบบคงที่และไม่ซับซ้อน เปรียบเทียบกับวิธีการตัดสินใจเลือกเส้นทางพื้นฐานอื่น โดยจะแสดงผลของการทดสอบของงานวิจัยเบื้องต้น เช่น พลังงานที่เหลืออยู่และอายุการใช้งานที่เหลืออยู่โดยรวมของเซนเซอร์โนดของแต่ละเหตุการณ์

บทที่ 4 การประยุกต์ใช้งานวิธีการมอนติคาร์โลด้วยการพิจารณาค่าเร็พพิวเทชันของเส้นทางในตำแหน่งไม่คงที่: กล่าวถึงการนำวิธีการมอนติคาร์โลมาประยุกต์ใช้ในกระบวนการตัดสินใจเลือกเส้นทางในโครงข่ายเซนเซอร์ไร้สายมาประยุกต์ใช้ในสภาวะแวดล้อมจริงที่มีลักษณะของโครงข่ายแบบไม่คงที่ โดยนำกรอบความคิดทางคณิตศาสตร์ทดสอบสมรรถนะของระบบและนำมาเปรียบเทียบกับการเลือกเส้นทางแบบอื่นๆ เช่น การเลือกเส้นทางที่สั้นที่สุด การเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน การเลือกเส้นทางแบบสุ่มเอกรูป

บทที่ 5 บทสรุปและข้อเสนอแนะ: สรุปงานวิจัยทั้งหมดในวิทยานิพนธ์ฉบับนี้และเสนอแนวทางในการพัฒนางานวิจัยในอนาคต

บทที่ 2

หลักการและทฤษฎี

วิทยานิพนธ์เล่มนี้นำเสนอการศึกษาการตัดสินใจเลือกเส้นทางในโปรโตคอลการจัดสรรเส้นทางที่มีการพิจารณาด้านพลังงาน อายุการใช้งานและค่าเร็พพิวเทชันของเส้นทางของโครงข่ายเซนเซอร์ไร้สาย ซึ่งทอพอโลยีจำลองจากลักษณะทางกายภาพของแม่น้ำโดยมีทอพอโลยีที่ไม่คงที่ อุปกรณ์ภายในโครงข่ายหรือเซนเซอร์โนดสามารถเคลื่อนที่ได้ดังนั้นเส้นเชื่อมโยงหรือลิงค์จะถูกสร้างเมื่อโนดอยู่ในระยะการส่งซึ่งกันและกันและจะขาดหายเมื่อหลุดออกจากระยะการส่ง ซึ่งโนดจะไม่สามารถส่งข้อมูลได้เมื่อพลังงานที่เหลืออยู่ไม่เพียงพอต่อการรับ-ส่งข้อมูล ดังนั้นเพื่อให้ได้สมรรถนะของโครงข่ายที่ดีการจัดสรรเส้นทางจึงควรมีกระบวนการตัดสินใจเลือกเส้นทางที่ดีโดยคำนึงถึงปัจจัยด้านพลังงาน อายุการใช้งานและค่าเร็พพิวเทชันของโนดหรืออุปกรณ์ในโครงข่าย

การนำเอาวิธีมอนติคาร์โลมาประยุกต์ใช้ ในการตัดสินใจเลือกเส้นทางของโครงข่ายโดยทำการศึกษากับโครงข่ายคงที่และโครงข่ายไม่คงที่ โดยเนื้อหาในบทนี้เป็นการแนะนำทฤษฎีพื้นฐานของวิธีและเทคนิคต่างๆ ที่กล่าวไว้ข้างต้นโดยประกอบไปด้วย หัวข้อ 2.1 กล่าวถึงคุณสมบัติแบบมาร์คอฟซึ่งเป็นคุณสมบัติที่ใช้พิจารณาการเปลี่ยนสถานะของสิ่งแวดล้อม เช่นสถานะของระบบโครงข่าย หัวข้อ 2.2 ทฤษฎีวิธีการเรียนรู้แบบเสริมแรง โดยมีรายละเอียดและองค์ประกอบและการนำมาใช้ในการแก้ปัญหา หัวข้อ 2.3 ทฤษฎีวิธีมอนติคาร์โลจะกล่าวถึงองค์ประกอบ กระบวนการตัดสินใจรวมไปถึงการปรับปรุงนโยบาย ส่วนสุดท้าย หัวข้อ 2.4 เป็นส่วนของบทสรุป

2.1 คุณสมบัติแบบมาร์คอฟและกระบวนการตัดสินใจแบบมาร์คอฟ

ระดับพลังงานที่เหลืออยู่หรือการเกิดความผิดปกติในโครงข่ายส่งผลให้เกิดสถานะของการพิจารณาโครงข่ายที่เปลี่ยนไป เช่นกรณีที่สถานะของโครงข่ายเซนเซอร์ไร้สายที่โนดทุกตัวสามารถรับ-ส่งข้อมูลเป็นปกติเมื่อเวลาผ่านไปโนดในโครงข่ายบางตัวไม่สามารถรับ-ส่งข้อมูลได้ ลักษณะเช่นนี้จะเป็นการเปลี่ยนของสถานะจะเป็นการเปลี่ยนจากสถานะก่อนหน้านี้เพียงสถานะเดียวเท่านั้นโดยไม่มีการพิจารณาผลที่ได้รับจากสถานะอดีตที่เคยผ่านมา ดังนั้นพฤติกรรมของการเปลี่ยนสถานะของระบบโครงข่าย จึงจัดได้ว่ามีคุณสมบัติของการเป็นมาร์คอฟ (Markov property)

2.1.1 คุณสมบัติแบบมาร์คอฟ

การพิจารณาลักษณะการเปลี่ยนสถานะของระบบ โดยที่สถานะถัดไปจะขึ้นอยู่กับสถานะก่อนหน้านั้นหรือสถานะปัจจุบันเพียงสถานะเดียว แต่ไม่ขึ้นกับสถานะในอดีตที่ผ่านมา โดยจะเรียกการพิจารณาระบบดังกล่าวว่ามีคุณสมบัติแบบมาร์คอฟ (Markov Property หรือ MP) ซึ่งการวิเคราะห์จะเป็นลักษณะลำดับของสถานะที่เชื่อมโยงกันด้วยความน่าจะเป็นของการเปลี่ยนสถานะ (State Transition Probability) ดังนั้นถ้าให้ $\{s_t\}$ คือการพิจารณาเซตของความน่าจะเป็นของการเปลี่ยนสถานะ โดย s_t แทนสถานะ ณ เวลา t เมื่อ s' เป็นสถานะถัดไปดังสมการ (2.1)

$$Pr\{s_{t+1} = s' \mid s_t = s\} = Pr\{s_{t+1} = s' \mid s_t = s, s_{t-1} = s, \ldots, s_0 = s\} \tag{2.1}$$

2.1.2 กระบวนการตัดสินใจแบบมาร์คอฟ

เมื่อระบบนั้นมีการเปลี่ยนแปลงสถานะที่มีคุณสมบัติแบบมาร์คอฟ โดยที่มีการตัดสินใจเลือกการกระทำแล้วการกระทำนั้นส่งผลต่อระบบ ทำให้สถานะของระบบนั้นเปลี่ยนแปลงไปจากสถานะก่อนหน้านั้นเป็นสถานะใหม่แต่ไม่ขึ้นกับสถานะในอดีตที่ผ่านมา จะเรียกระบบที่มีการเลือกการกระทำเช่นนี้ว่ามีคุณสมบัติของกระบวนการตัดสินใจแบบมาร์คอฟ (Markov Decision Process หรือ MDP) ซึ่งหากสถานะและการกระทำของระบบที่ทำการพิจารณามีขอบเขตที่ชัดเจน จะเรียกกระบวนการตัดสินใจเช่นนี้ว่า กระบวนการตัดสินใจแบบมาร์คอฟที่มีขอบเขตจำกัด (finite Markov Decision Process หรือ finite MDP) [13]

ถ้าให้ $P^a_{ss'}$ เป็นความน่าจะเป็นที่จะเกิดการเปลี่ยนสถานะของระบบที่เคยเป็นอยู่จากสถานะ s จะเปลี่ยนสถานะเป็น s' เมื่อเลือกการกระทำ a ดังสมการ (2.2)

$$P^a_{ss'} = Pr\{s_{t+1} = s' \mid s_t = s, a_t = a\} \tag{2.2}$$

การเปลี่ยนไปยังสถานะถัดไป s_{t+1} ของระบบจะขึ้นอยู่เพียง สถานะในปัจจุบัน s_t และการกระทำ a_t ที่กระทำเมื่ออยู่ในสถานะปัจจุบันเท่านั้น ดังนั้น R ผลรางวัลระยะยาวสูงสุดที่ระบบคาดว่าจะได้รับเมื่อมีการตัดสินใจกระทำจะสามารถเขียนได้ดังสมการ (2.3)

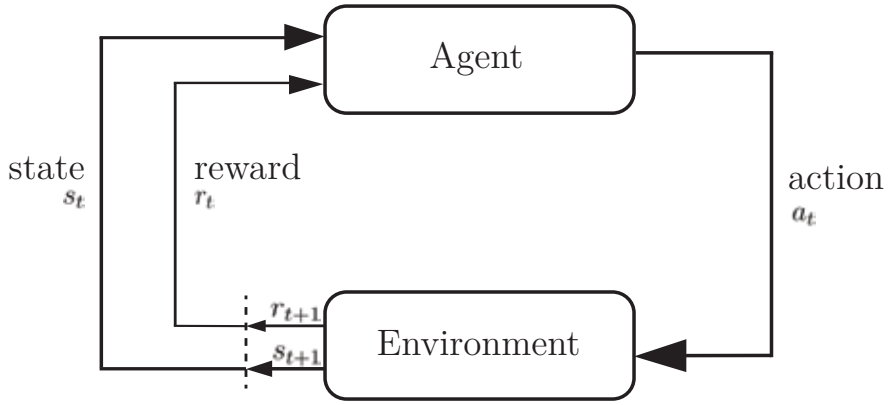
$$R^a_{ss'} = E\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\} \tag{2.3}$$

ซึ่งสรุปได้ว่าระบบที่มีคุณสมบัติมาร์คอฟจะเป็นระบบที่ทราบถึงสถานะถัดไป s' เมื่อรู้สถานะ s และการกระทำ a ในปัจจุบัน คุณสมบัติเช่นนี้จึงเป็นลักษณะของระบบที่มีการเรียนรู้ของเครื่อง (Machine Learning) ที่ต้องการ

2.2 การเรียนรู้แบบเสริมแรง (Reinforcement Learning)

การเรียนรู้แบบเสริมแรงหรืออินฟอร์สเมนท์เลิร์นนิ่ง (Reinforcement Learning หรือ RL) [13] เป็นอีกสาขาย่อยหนึ่งของการเรียนรู้ของเครื่อง โดยจะศึกษากระบวนการเรียนรู้เพื่อให้ระบบสามารถตัดสินใจเชิงตรรกะและแก้ปัญหาอย่างอัตโนมัติด้วยข้อมูลที่มีอยู่ในลักษณะความสัมพันธ์ของ สถานะ (state) การกระทำ (action) และผลรางวัลที่ได้รับ (reward) โดยที่การเรียนรู้จะเป็นการเรียนรู้จากกระบวนการการตัดสินเลือกกระทำจากสภาวะแวดล้อมหนึ่งไปยังสภาวะหนึ่ง เพื่อให้ได้ผลของการกระทำที่ดีที่สุด ดังกระบวนการเรียนรู้รูปที่ 2.1 ซึ่งเริ่มจากระบบทราบถึงสถานะปัจจุบันที่ชัดเจนโดยตัวกระทำการตัดสินใจของระบบ (agent) จะทำการตัดสินใจเลือกกระทำการอย่างใดอย่างหนึ่งจากชุดการกระทำที่เป็นไปได้ทั้งหมด จากนั้นตัวกระทำการตัดสินใจจะรับทราบถึงการเปลี่ยนไปของสภาพแวดล้อม (environment) และผลรางวัลที่ได้รับกลับมา ระบบจะมีลักษณะการวนซ้ำเช่นนี้และได้รับผลรางวัลกลับมาอยู่โดยตลอด จนกระทั่งตัวกระทำการตัดสินใจเกิดการเรียนรู้ว่าเมื่ออยู่ในสถานะใดควรเลือกการกระทำแบบใด เพื่อให้ได้ผลรางวัลระยะยาวสูงสุด (long-term expected reward)

เพื่อความสะดวก กำหนดให้ $s_t = s$, $a_t = a$ และ $r_t = r$ โดยเมื่อ s สถานะของระบบ $\mathcal{S}$ ชุดสถานะทั้งหมดที่เป็นไปได้ โดย $s_t \in \mathcal{S}$ และ a จะทำการตัดสินใจเลือกกระทำอย่างใดอย่างหนึ่งจาก $A(s_t)$ ชุดการกระทำที่เป็นไปได้ทั้งหมดในแต่ละสถานะ โดย $a_t \in A(s_t)$ และ r ผลรางวัลที่ได้รับจากการกระทำ $\mathcal{R}$ ผลรางวัลระยะยาวสูงสุดที่ระบบคาดว่าจะได้รับเมื่อมีการตัดสินใจกระทำโดย $r_{t+1} \in R$ เมื่อ t คือเวลาแต่ละขั้น โดย $t = 0, 1, 2, 3, \ldots$



รูปที่ 2.1: กระบวนการเรียนรู้ของเครื่อง

State s_t	คือ สถานะของระบบ ณ เวลา t
Action a_t	คือ การตัดสินใจเลือกการกระทำ ณ เวลา t
Reward r_t	คือ ผลรางวัลที่ได้รับจากการกระทำ ณ เวลา t
State s_{t+1}	คือ สถานะของระบบ ณ เวลา $t+1$
Reward r_{t+1}	คือ ผลรางวัลที่ได้รับจากการกระทำ ณ เวลา $t+1$

2.2.1 องค์ประกอบของการเรียนรู้ของเครื่อง

โดยทั่วไปแล้ววิธีการเรียนรู้แบบเสริมแรงจะประกอบด้วย 4 องค์ประกอบหลัก [13] คือ

2.2.1.1 นโยบาย (policy)

ตัวกำหนดแนวทางของการเรียนรู้และการตัดสินใจในการเลือกการกระทำในลำดับถัดไป เพื่อให้ได้ผลรางวัลเฉลี่ยในระยะยาวที่สูงสุด โดยกำหนดให้นโยบายแทนด้วย π ซึ่งได้มาจากค่าสูงสุดของผลเฉลี่ยรางวัลสะสมเพื่อนำไปใช้เป็นฟังก์ชันในการเลือกการกระทำดังสมการ (2.4)

$$\pi(s) = \arg\max_{a} Q(s,a) \tag{2.4}$$

เมื่อ $\pi(s)$ คือนโยบายที่ใช้ในการเลือกการกระทำในแต่ละสถานะ s และ $Q(s,a)$ คือค่าผลเฉลี่ยรางวัลสะสมที่ได้จากการกระทำ a ในสถานะ s

2.2.1.2 ฟังก์ชันผลรางวัล (reward function)

ฟังก์ชันที่ใช้คำนวณผลรางวัลในระยะยาวของระบบที่ได้ดำเนินอยู่ในขณะนั้น โดยผลรางวัลนี้จะเกิดจากการตัดสินใจในการเลือกการกระทำเมื่อระบบอยู่ในสถานะต่างๆ โดยหากเป็นการตัดสินใจที่ถูกต้องจะได้ผลรางวัลในระดับที่สูง แต่หากเป็นการตัดสินใจที่ถูกต้องน้อยกว่าหรือเป็นการตัดสินใจที่ผิดจะได้รับผลรางวัลในระดับที่ต่ำลดหลั่นกันลงมา ซึ่งผลรางวัลนี้จะเป็นตัวแสดงถึงความสามารถของตัวตัดสินใจว่าสามารถทำการตัดสินใจได้ดีหรือไม่ในช่วงเวลาขณะนั้น จุดมุ่งหมายหลักของการแก้

ปัญหาด้วยการเรียนรู้แบบเสริมแรงคือ ต้องการการเรียนรู้เพื่อให้ได้นโยบายในการตัดสินใจที่ส่งผลให้ได้ผลรางวัลในระยะยาวที่สูงที่สุด โดยลักษณะการคำนวณผลรางวัลจะแบ่งออกเป็น 2 ลักษณะ คือ

ภารกิจเป็นตอน (episodic tasks) คือ การพิจารณาผลรางวัลที่ทราบจุดสิ้นสุดของการทำงานเป็นฉากหรือรอบ โดยจะรวมผลรางวัลที่เกิดขึ้นภายหลังช่วงเวลา t จนถึงเวลาสิ้นสุดของการทำงาน T จะสามารถเขียนได้ดังสมการ (2.5)

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots + r_T \tag{2.5}$$

เมื่อ R_t คือผลรวมของผลรางวัลช่วงตั้งแต่เวลา t จนถึงเวลา T และ r_t คือผลรางวัลที่เกิดขึ้น ณ เวลา t

ภารกิจต่อเนื่อง (continuing tasks) คือ การพิจารณาผลรางวัลที่ไม่ทราบจุดสิ้นสุดของการทำงานหรือแบบต่อเนื่องที่ไม่มีขีดจำกัดทางเวลา (infinite horizon) โดย $T = \infty$ ดังนั้นจากสมการ (2.5) จึงจำเป็นต้องมีพารามิเตอร์ควบคุมผลรางวัลไม่ให้เกิดการลู่ออก จึงเขียนใหม่ได้ดังสมการ (2.6)

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2.6}$$

เมื่อ γ อัตราการลดทอน (discount rate) โดยที่ $\gamma \in [0, 1)$

อัตราการลดทอนจะกำหนดความสำคัญของผลตอบแทนในอนาคต ผลรางวัลที่ได้รับ k ครั้ง โดยจะสามารถเขียนอัตราการลดทอนของผลรางวัลที่ได้รับ ณ ขณะนั้นได้ว่า γ^{k-1} เมื่ออัตราการลดทอนมีค่าเป็น 0 จะส่งผลให้ตัวตัดสินใจของระบบพิจารณาเฉพาะผลรางวัล ณ ปัจจุบันเท่านั้น และเมื่ออัตราการลดทอนเข้าใกล้ค่า 1 จะส่งผลให้เกิดการรวมของค่าผลรางวัล ณ เวลาต่างเท่ากับ $\{r_k\}$ ซึ่งจะทำให้ตัวตัดสินใจของระบบสามารถพิจารณาผลรางวัลระยะยาวยิ่งขึ้น แต่ถ้าอัตราการลดทอนมีค่าเป็น 1 จะไม่สามารถพิจารณาถึงจุดสิ้นสุดของสถานะได้

2.2.1.3 ฟังก์ชันมูลค่า หรือ แวลู่ฟังก์ชัน (value function)

คือฟังก์ชันที่ใช้คำนวณหาผลรางวัลในระยะยาวที่คาดว่าจะได้รับหากเลือกการกระทำนั้นๆ ภายใต้นโยบายเดียวกัน ในการจำลองการกระทำซึ่งตัวกระทำการตัดสินใจ จะใช้ผลรางวัลนี้ในการตัดสินใจเลือกการกระทำในลำดับถัดไป โดยถ้าให้นโยบาย π เป็นนโยบายที่ทำการเชื่อมโยงกับสถานะ s ฟังก์ชันมูลค่าจะดำเนินไปภายใต้นโยบาย π นี้ ดังสมการ (2.7)

$$V^{\pi}(s) = E_{\pi}\{R_t | s_t = s\} = E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right\} \tag{2.7}$$

ดังนั้นแวลู่ฟังก์ชัน $V^{\pi}(s)$ ที่สถานะ s ภายใต้นโยบาย π จะได้จากค่าคาดหวังเมื่อตัวกระทำการตัดสินใจกระทำตามนโยบาย $E_{\pi}\{\}$ และค่าอัตราการลดทอน γ ต้องอยู่ในช่วงระหว่าง $0 \leq \gamma < 1$ ทั้งหมดที่กล่าวมานี้จึงสามารถเรียกฟังก์ชันมูลค่า V^{π} ได้ว่าสเตท-แวลู่ฟังก์ชันและค่าผลรางวัลเฉลี่ยสะสมที่ได้จากแต่ละการกระทำ a ในสถานะ s นั้นๆ ภายใต้นโยบายจึงเรียกได้ว่าแอคชั่น-แวลู่ฟังก์ชันหรือคิวฟังก์ชันดังสมการ (2.8)

$$Q^{\pi}(s, a) = E_{\pi}\{R_t | s_t = s, a_t = a\} = E_{\pi}\left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a \right\} \tag{2.8}$$

การนำการเรียนรู้ของเครื่องมาประยุกต์ใช้แก้ปัญหานั้น จะเป็นการหานโยบายที่จะก่อให้เกิดผลรางวัลสะสมระยะยาวสูงที่สุด โดยจะพยายามทำการปรับปรุงนโยบายนี้ให้ดีขึ้นอยู่โดยตลอด ดังนั้นนโยบายใหม่จึงดีกว่าหรือเท่ากับนโยบายเดิม ซึ่งจะเรียกนโยบายที่ส่งผลถึงการดำเนินไปของระบบที่ดีที่สุดเท่าที่สภาวะแวดล้อมแบบนั้นควรจะเป็นได้ว่า นโยบายที่เหมาะสมที่สุด (optimal policy) และเมื่อให้ระบบดำเนินไปตามนโยบายนี้ จะได้สเตต-แวลู่ฟังก์ชันที่เป็นสเตต-แวลู่ฟังก์ชันที่เหมาะสมที่สุด (optimal state-value function) ซึ่งแทนด้วย V^* ดังสมการ (2.9)

$$V^*(s) = \max_{\pi} V^{\pi}(s) \tag{2.9}$$

ในแนวทางเดียวกันเมื่อระบบดำเนินไปตามนโยบายที่เหมาะสมนี้ จะได้แอคชั่น-แวลู่ฟังก์ชันที่เป็นแอคชั่น-แวลู่ฟังก์ชันที่เหมาะสมที่สุด (optimal action-value function) ซึ่งแทนด้วย Q^* ดังสมการ (2.10)

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a) \tag{2.10}$$

สำหรับในทุกๆ คู่ (s,a) ของสถานะและการกระทำ (state-action pair) จะได้ความสัมพันธ์ของแอคชั่น-แวลู่ฟังก์ชันที่เหมาะสมที่สุด Q^* และสเตต-แวลูฟังก์ชั่นที่เหมาะสมที่สุด V^* ดังสมการ (2.11)

$$Q^*(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a\} \tag{2.11}$$

2.2.1.4 แบบจำลองของสภาวะแวดล้อม (model of environment)

แบบจำลองของสิ่งแวดล้อมที่จะนำการเรียนรู้แบบเสริมแรงเข้าไปประยุกต์ใช้ ซึ่งต้องมีความสามารถที่จะแสดงพฤติกรรมได้เหมือนกับสภาวะแวดล้อมจริงที่จะนำไปประยุกต์ใช้ ตัวอย่างเช่น แบบจำลองโครงข่ายเซนเซอร์ไร้สายกรณีที่สถานะของโครงข่ายที่อุปกรณ์ทุกอุปกรณ์ทำงานเป็นปกติ เมื่อเวลาผ่านไปปรากฏว่ามีอุปกรณ์ขัดข้องเกิดขึ้น เช่น การสื่อสารขัดข้อง ซึ่งในวิทยานิพนธ์เล่มนี้จะกล่าวถึงการนำวิธีการเรียนรู้แบบเสริมแรงมาประยุกต์ใช้กับแบบจำลองดังกล่าวนี้ในบทที่ 3

2.3 วิธีการมอนติคาร์โล (Monte Carlo Method)

วิธีการมอนติคาร์โล (Monte Carlo Method หรือ MC) นี้เป็นวิธีการหนึ่งของการแก้ปัญหาการเรียนรู้แบบเสริมแรง ซึ่งในวิทยานิพนธ์เล่มนี้นำวิธีการมอนติคาร์โลประยุกต์ใช้เนื่องจากลักษณะของสิ่งที่วิธีการมอนติคาร์โลต้องการในการพิจารณา ที่ไม่ต้องการลักษณะของรูปแบบทางคณิตศาสตร์ที่ชัดเจน ลักษณะการปรับปรุงของนโยบายและความซับซ้อนของวิธีการมอนติคาร์โลที่ไม่มากนัก เมื่อเทียบกับวิธีการแก้ปัญหาแบบอื่นๆในการเรียนรู้ของเครื่อง โดยวิธีการมอนติคาร์โลเป็นวิธีการเรียนรู้โดยพิจารณาจากค่าเฉลี่ยของผลรางวัลที่ได้รับจากการจำลองแบบโดยมีการทำงานเป็นฉากหรือรอบเรียกว่าเอพพิโซด (episode) โดยจะทำการแบ่งประสบการณ์ของการเรียนรู้ออกเป็นหลายๆ เอพพิโซดและในทุกๆ ครั้งของการสิ้นสุดเอพพิโซดจะแสดงถึงการเลือกการกระทำที่สมควรที่จะกระทำ และในแต่ละเอพพิโซดจะนำค่าเฉลี่ยของผลรางวัลจากเอพพิโซดนั้นมาปรับปรุงนโยบายหรือจะกล่าวได้ว่านโยบายถูกปรับปรุงในทุกๆ เอพพิโซด (episode-by-episode) ไม่ใช่การปรับปรุงนโยบายในลักษณะเป็นขั้น (step-by-step) [13] จนกระทั้งระบบได้รับนโยบายที่ดีที่สุดในระยะยาว

2.3.1 การประมาณค่าแอคชั่น-แวลู่ของมอนติคาร์โล (Monte Carlo Estimation of Action Values)

วิธีการมอนติคาร์โลจะเรียนรู้จากแอคชั่น-แวลู่ฟังก์ชันภายใต้นโยบาย $\pi : s \rightarrow a$ ซึ่งนโยบายดังกล่าวจะถูกเรียกใหม่ว่า อัตราการลดทอนของผลรางวัลสะสมที่คาดว่าจะได้รับ (expected cumulative discounted reward) หรือ แอคชั่น-แวลู่ของคู่สถานะและการกระทำ $Q^{\pi}(s,a)$ โดยวิธีการประมาณค่าแอคชั่น-แวลู่ฟังก์ชันจะประเมินจากประสบการณ์ที่เคยทำมา จึงหมายถึงการนำค่าเฉลี่ยผลรางวัลหรือในวิธีการมอนติคาร์โลจะเรียกว่าผลตอบแทน (return) มาพิจารณาภายหลังการเข้าสู่สถานะนั้น ซึ่งจะแบ่งวิธีการประมาณค่าแอคชั่น-แวลู่ฟังก์ชันออกเป็นสองวิธีดังนี้

วิธีการเอฟวรี-วิสซิทมอนติคาร์โล (every-visit MC method) คือการคำนวณผลรางวัลสะสมที่คาดว่าจะได้รับ จากการเฉลี่ยของผลตอบแทนทุกครั้ง ที่ระบบอยู่ในสถานะ s และเลือกการกระทำ a ของแต่ละเอพพิโซด ดังสมการ (2.12)

$$Q^{\pi}(s,a) = \frac{\sum_{k=1}^{n(s,a)} r(s,a,k)}{n(s,a)} \tag{2.12}$$

เมื่อ $r(s,a,k)$ คือจำนวนครั้งที่เกิดขึ้นของผลตอบแทน เมื่อระบบอยู่ในสถานะ s และเลือกการกระทำ a และ $n(s,a)$ คือจำนวนครั้งที่ระบบอยู่ในสถานะ s และเลือกการกระทำ a

วิธีการเฟิร์สท-วิสซิทมอนติคาร์โล (first-visit MC method) คือวิธีการคำนวณผลรางวัลสะสมที่คาดว่าจะได้รับ จากการเฉลี่ยค่าของผลตอบแทนเพียงครั้งแรกเท่านั้น ที่ระบบอยู่ในสถานะ s และเลือกการกระทำ a ของแต่ละเอพพิโซด หรือจะกล่าวได้ว่า $r(s,a,k)$ ของวิธีการเฟิร์สท-วิสซิทจะกำหนดค่า $k = 1$ ดังสมการ (2.13)

$$Q^{\pi}(s,a) = \frac{r(s,a,1)}{1} \tag{2.13}$$

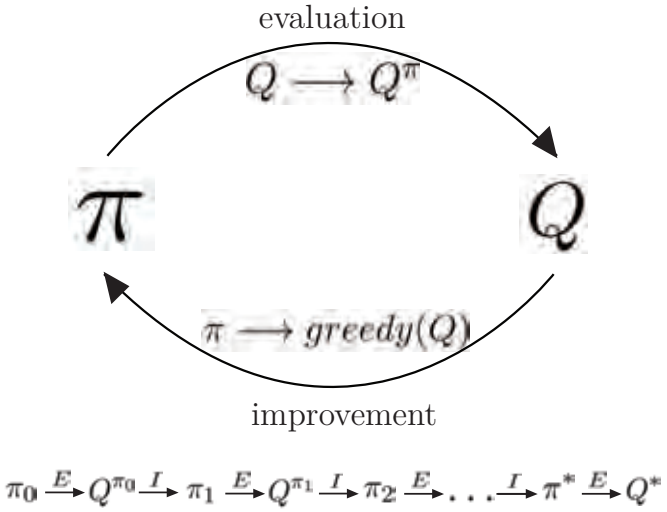
เมื่อ $r(s,a,1)$ คือผลตอบแทนครั้งแรกที่ได้รับ เมื่อระบบอยู่ในสถานะ s และเลือกการกระทำ a

จากวิธีการประมาณแอคชั่น-แวลู่ทั้งหมดที่กล่าวมา ถ้า π เป็นนโยบายที่เลือกแบบคงตัวเสมอ ฉะนั้นถ้ามีการเลือกการกระทำภายใต้นโยบายเพียงอย่างเดียวจะส่งผลกระทบให้มีเพียงแค่การกระทำเดียวเท่านั้นในแต่ละสถานะที่จะถูกเลือกหรือจะกล่าวได้ว่าระบบไม่มีการปรับปรุงนโยบายการเลือกกระทำ ซึ่งนัยสำคัญที่แท้จริงของการเรียนรู้ของมอนติคาร์โลนั้นคือการทำให้เกิดการเรียนรู้ของแอคชั่น-แวลู่และมีการปรับปรุงนโยบายอย่างสม่ำเสมอ ทำให้ระบบมีโอกาสในการตัดสินเลือกกระทำทุกๆการกระทำและนโยบายนั้นต้องสามารถเปลี่ยนแปลงได้ จึงมีการแก้ปัญหาดังกล่าวโดยการกำหนดค่าความน่าจะเป็นในการเลือกคู่สถานะและการกระทำให้มีค่าความน่าจะเป็นไม่เท่ากับ 0 เพื่อให้ระบบมีโอกาสในการเลือกคู่สถานะ-การกระทำใหม่เพิ่มขึ้นแม้ระบบไม่เคยทำการเลือกก่อนหน้านี้

2.3.2 การเข้าสู่ภาวะที่เหมาะสมที่สุดของมอนติคาร์โล (Convergence Optimality for Monte Carlo)

ลักษณะพื้นฐานของวิธีการมอนติคาร์โลคือการเรียนรู้จากประสบการณ์ที่เคยกระทำซ้ำๆ (iteration) โดยจะเริ่มการทำงานของเอพพิโซดแรกด้วยการใช้นโยบาย π_0 ระบบจะดำเนินไปเพื่อค้นหา

การกระทำที่เหมาะสมที่สุด และเมื่อได้การกระทำที่เหมาะสมที่สุดแล้วจะเป็นการสิ้นสุดเอพพิโซด และจะเก็บผลรางวัลที่เกิดขึ้น ตั้งแต่เริ่มต้นเอพพิโซดจนจบเอพพิโซด จากนั้นเก็บผลรางวัลนี้ในรูปของแอคชั่น-แวลู่ฟังก์ชันภายใต้นโยบาย π_0 แทนด้วย Q^{π_0} และนำค่าเฉลี่ยของผลรางวัลจากเอพพิโซดนั้นมาปรับปรุงนโยบาย ซึ่งมีผลให้นโยบายถูกปรับปรุงในทุกๆ เอพพิโซด จากนั้นระบบจะเริ่มเอพพิโซดใหม่และมีการทำงานเช่นนี้ต่อๆ ไปจนกระทั่งได้นโยบายที่เหมาะสมแทนด้วย π^* ซึ่งเป็นจุดประสงค์ของวิธีมอนติคาร์โล



รูปที่ 2.2: ลักษณะของการปรับปรุงนโยบาย

ดังรูปที่ 2.2 จะเห็นได้ว่าประสบการณ์เรียนรู้ที่ระบบจะได้รับจะได้จากช่วงที่ระบบมีการประเมินนโยบาย ซึ่งหากได้ค่าแอคชั่น-แวลู่ที่เหมาะสมก็จะส่งผลให้ระบบเข้าสู่สภาวะการลู่เข้า (converge) ดังนั้นทุกๆเหตุการณ์ (event) ในเอพพิโซดใดๆ ระบบจะทำการสำรวจทุกๆคู่ของสถานะและการกระทำและทำเช่นนี้ต่อไปเรื่อยๆ จากนั้นจึงมีการปรับปรุงนโยบายจากนโยบายแบบละโมบ (ε-greedy policy) ดังนั้นการกระทำที่จะถูกเลือกนั้นจะได้จากค่าสูงสุดของค่าแอคชั่น-แวลู่ฟังก์ชั้น ดังสมการ (2.14)

$$\pi(s) = \arg\max_a Q(s,a) \tag{2.14}$$

ในทำนองเดียวกันจะสามารถเขียนค่าแอคชั่น-แวลู่ฟังก์ชันภายใต้การปรับปรุงจากนโยบายแบบละโมบ (Greedy) ได้ดังสมการ (2.15) ซึ่งการเรียนรู้ดังกล่าวจะพยายามทำการปรับปรุงนโยบายให้ดีขึ้นอยู่โดยตลอด ดังนั้นนโยบายใหม่ π_{k+1} ควรดีกว่าหรือเท่ากับนโยบายเดิม π_k เพื่อค่านโยบายดังกล่าวเข้าสู่สภาวะที่เหมาะสมที่สุดและเมื่อระบบดำเนินไปตามนโยบายนี้จะส่งผลให้ได้แวลู่ฟังก์ชันที่เหมาะสมที่สุด

$$\begin{aligned} Q^{\pi_k}(s,\pi_{k+1}(s)) &= Q^{\pi_k}(s,\arg\max_a Q^{\pi_k}(s,a)) \\ &= \max_a Q^{\pi_k}(s,a) \\ &\geq Q^{\pi_k}(s,\pi_k(s)) \\ &\geq V^{\pi_k}(s) \end{aligned} \tag{2.15}$$

2.3.3 แบบออนโพลิซีมอนติคาร์โล (On-policy Monte Carlo)

วิธีการมอนติคาร์โลจะมี 2 วิธีในการหาผลลัพธ์ คือ วิธีการออนโพลิมอนติคาร์โล (On-policy Monte Carlo) และวิธีการออฟโพลิซีมอนติคาร์โล (Off-policy Monte Carlo) โดยในวิธีการออนโพลิมอนติคาร์โลจะเป็นการเรียนรู้จากการดำเนินการของนโยบายในปัจจุบัน ซึ่งตัวตัดสินใจในการเลือกกระทำจะพยายามสำรวจในทุกคู่สถานะและการกระทำเสมอ จนกว่าจะได้นโยบายที่เหมาะสมที่สุดจากการสำรวจนั้น และในส่วนของวิธีการออฟโพลิซีมอนติคาร์โล ตัวตัดสินใจในการเลือกการกระทำก็จะพยายามสำรวจในทุกคู่สถานะและการกระทำเช่นกัน แต่ลักษณะการเรียนรู้ของระบบอาจไม่ได้มาจากค่านโยบาย และนอกจากนั้นมีเพียงวิธีการออนโพลิมอนติคาร์โลเท่านั้นที่มีการพิสูจน์ว่านโยบายสามารถเข้าสู่สภาวะที่เหมาะสมได้ในเชิงคณิตศาสตร์

วิธีการออนโพลิมอนติคาร์โลจะใช้วิธีการเรียนรู้โดยพิจารณาค่าเฉลี่ยผลรางวัลที่ได้รับจากการจำลองแบบโดยการทำงานเป็นแอพพิโซดต่อแอพพิโซด ซึ่งจะมีการประเมินและปรับปรุงนโยบายอย่างสม่ำเสมอจนกระทั้งได้ค่านโยบายและค่าแอคชั่น-แวลู่ฟังก์ชันที่เหมาะสมที่สุด เพื่อประกันได้ว่าการกระทำที่เกิดขึ้นจะถูกเลือกจากตัวตัดสินใจภายใต้เงื่อนไขของนโยบายดังกล่าว จึงเขียนสมการได้ว่า เซตของสถานะทั้งหมดที่พิจารณา $\mathcal{S}$ และการกระทำทั้งหมดที่เป็นไปได้คือ A โดยพิจารณาคู่สถานะและการกระทำ (s,a) ที่มีการวิสซิทในแต่ละเอพพิโซด เมื่อ $s \in \mathcal{S}$ และ $a \in A$ เริ่มการทำงานของนโยบายในเอพพิโซดแรก π_0 ดังนั้นในแต่ละเอพพิโซดจะมีเหตุการณ์แต่ละเหตุการณ์ในการเลือกการกระทำและผลจากการกระทำเหล่านั้นจะถูกนำมาสร้างนโยบาย π_t ที่จุดสิ้นสุดของเอพพิโซด t ฉะนั้นจะประมาณค่าแอคชั่น-แวลู่ฟังก์ชันที่ถูกอัพเดทได้ดังสมการ (2.16) [20]

$$Q^{\pi_t}(s,a) = Q^{\pi_{t-1}}(s,a) + \frac{1}{t}\left[\sum_{n=\tau_t(s,a)}^{N_{t-1}} r(s_n,a_n) - Q^{\pi_{t-1}}(s,a)\right] \tag{2.16}$$

เมื่อ N_t จำนวนเหตุการณ์หรือจำนวนครั้งในแต่ละเอพพิโซด โดย $\tau_t(s,a)$ คือจำนวนครั้งที่เกิดคู่สถานะและการกระทำครั้งแรก (s,a) และผลรางวัล $r(s,a)$ ที่ได้รับจากการกระทำณ สถานะต่างๆ ถ้าให้เหตุการณ์ที่เกิดขึ้น $\{s_0,a_0,r(s_0,a_0),\ldots,s_n,a_n,r(s_n,a_n)\}$

พิจารณาในเทอมของผลรวมของผลรางวัลที่เกิดจากคู่สถานะและการกระทำ (s,a) ในเอพพิโซด t จะได้การกระทำที่อาจจะถูกเลือก a^* หากใช้วิธีการสำหรับนโยบายแบบละโมบดังสมการ (2.17) [20]

$$a^* = \arg\max_a \{Q^{\pi_1}(s,a)\} \tag{2.17}$$

ในทำนองเดียวกันจะสามารถปรับปรุงนโยบายจากนโยบายแบบละโมบ (ε-greedy policy) โดย $\varepsilon \in [0,1]$ ได้ดังสมการ (2.18) [20]

$$\pi_{t+1}(s) = \begin{cases} a^* & \text{with probability } 1-\varepsilon+\frac{\varepsilon}{|A|} \\ a \in A - a^* & \text{with probability } \frac{\varepsilon}{|A|} \end{cases} \tag{2.18}$$

นโยบายละโมบ ดังกล่าวจะมีการเลือกการกระทำ 2 ลักษณะคือการเลือกการกระทำในลักษณะเดิมซ้ำๆด้วยความน่าจะเป็น $1-\varepsilon$ เรียกว่า เอคซพลอยท (exploit) และจะมีการเลือกการกระทำใหม่จะมีความน่าจะเป็นที่เหลืออยู่ ε เรียกว่า เอคซพลอร์ (explore) ดังนั้นจากสมการ (2.18) เมื่อระบบอยู่ในสถานะ s จะมีการเลือกการกระทำ a^* ด้วยความน่าจะเป็น $1-\varepsilon+\frac{\varepsilon}{|A|}$ และเลือกการกระทำ

อื่นๆ $a \in A - a^*$ ด้วยความน่าจะเป็น $\frac{\varepsilon}{|A|}$ เมื่อ $|A|$ คือขนาดของปริภูมิการกระทำ (action space) ซึ่งทั้งหมดที่กล่าวมาจะสรุปเป็นขั้นตอนการทำงานของวิธีแบบออนโพลิซีมอนติคาร์โลดังตารางที่ 2.1

การนิยามตัวแปรขั้นตอนการทำงานของวิธีออนโพลิซีมอนติคาร์โล

s	สถานะของระบบ (state)
a	การกระทำที่เลือกกระทำ (action)
$Q(s,a)$	แอคชั่น-แวลู่ฟังก์ชันที่ใช้ในการเก็บผลรางวัลที่เกิดขึ้นในทุกๆ คู่สถานะของระบบและการกระทำ
$Returns(s,a)$	ใช้ในการเก็บผลตอบแทนที่เกิดขึ้น ณ ทุกๆ คู่สถานะของระบบและการกระทำในหนึ่งเอพพิโซดซึ่งในการเริ่มต้นจะถูกตั้งค่าเป็น 0
π	เป็นนโยบายซึ่งใช้แบบละโมบ ε คือใช้การเลือกการกระทำที่มีค่า $Q(s,a)$ สูงสุดด้วยความน่าจะเป็นเท่ากับ $1 - \varepsilon + \frac{\varepsilon}{\|A(s)\|}$ และเลือกการกระทำอื่นๆ ด้วยความน่าจะเป็นเท่ากับ $\frac{\varepsilon}{\|A(s)\|}$
$\|A(s)\|$	การกระทำทั้งหมดที่เป็นไปได้จากสถานะของระบบ
R	เป็นผลรางวัลที่ได้จากผลตอบแทนทั้งหมด จากการไปพบคู่ของสถานะของระบบและการกระทำตั้งแต่การวิสซิทเป็นครั้งแรก (first visit) และทำการเพิ่มเข้าสู่ $Returns(s,a)$
a^*	เป็นการกระทำที่ก่อให้เกิดผลรางวัลสูงสุดที่ได้มาจาก $Q(s,a)$
$\pi(s,a)$	เป็นนโยบายที่ใช้ในการเลือกการกระทำเมื่อระบบอยู่ในสถานะใดๆ s โดยจะเลือกการกระทำ a^* ด้วยความน่าจะเป็น $1 - \varepsilon + \frac{\varepsilon}{\|A(s)\|}$ และเลือกการกระทำอื่นๆ ด้วยความน่าจะเป็น $\frac{\varepsilon}{\|A(s)\|}$
ε	กรีดี้ ของนโยบาย โดยมีค่ามากกว่าหรือเท่ากับ 0 แต่น้อยกว่าหรือเท่ากับ 1

ตารางที่ 2.1: ลำดับขั้นตอนการทำงานของวิธีออนโพลิซีมอนติคาร์โล [13]

1. Initialisation, for all $s \in S$, $a \in A(s)$:
2. $Q(s,a) \leftarrow$ arbitrary
3. $Returns(s,a) \leftarrow$ empty list
4. $\pi \leftarrow$ an arbitrary $\epsilon-$soft policy
5. Repeat forever:
6. (a) Generate an episode using π
7. (b) For each pair s,a appearing in the episode:
8. $R \leftarrow$ return follow the first occurrence of s,a
9. Append R to $Returns(s,a)$
10. $Q(s,a) \leftarrow$ average$(Returns(s,a))$
11. For each s in the episode:
12. $a^* \leftarrow argmax_a Q(s,a)$
13. For all $a \in A(s)$:
14. $\pi(s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{\lvert A(s)\rvert} & \text{if } a = a* \\ \frac{\varepsilon}{\lvert A(s)\rvert} & \text{if } a \neq a* \end{cases}$

2.4 สรุป

บทนี้กล่าวถึงภาพรวมของทอพอโลยีที่พิจารณาคุณลักษณะแบบมาร์คอฟ กระบวนการตัดสินใจแบบมาร์คอฟซึ่งเป็นกระบวนลักษณะหนึ่งภายใต้พื้นฐานการเรียนรู้ นอกจากนั้นยังมีการนำเสนอแนวคิดการเรียนรู้แบบเสริมแรงและวิธีการมอนติคาร์โล สำหรับการแก้ปัญหาในลักษณะแบบมาร์คอฟซึ่งเคยมีการนำมาพิจารณาในลักษณะปัญหาของการจัดสรรเส้นทางในโครงข่ายแอดฮอก สำหรับปัญหาของการจัดสรรเส้นทางในงานวิจัยนี้จะมีการพิจารณาแบบการทำงานเป็นฉากหรือรอบและสามารถรู้ถึงจุดสิ้นสุดการทำงาน โดยแต่ละรอบของการทำงานจะเริ่มจากโนดต้นทางค้นหาเส้นทางไปยังโนดปลายทาง เมื่อสิ้นสุดรอบของการทำงานจะมีหนึ่งเส้นทางที่ถูกเลือก ด้วยเหตุนี้วิธีการมอนติคาร์โลที่ได้นำเสนอจึงมีกระบวนการทำงานบนพื้นฐานการจำลองแบบเป็นรอบ ซึ่งถูกเรียกว่าออนโพลิซีมอนคิคาร์โล หลังจากบทนี้ในบทที่ 3 จะนำเสนอการใช้วิธีมอนติคาร์โล เพื่อทำการสมมติฐานเบื้องต้นในการเลือกเส้นทางสำหรับสภาวะแวดล้อมแบบคงที่ และบทที่ 4 ทำการสมมุติฐานการเลือกเส้นทางสำหรับสภาวะแวดล้อมแบบไม่คงที่

บทที่ 3

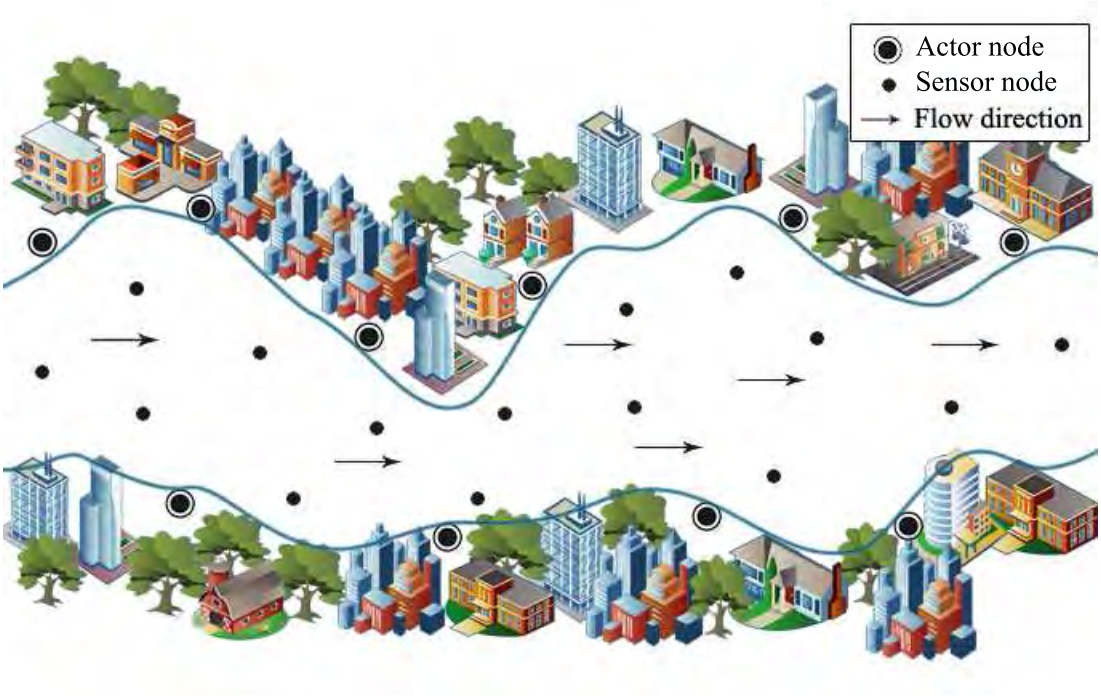
การประยุกต์ใช้งานวิธีการมอนติคาร์โลด้วยการพิจารณาค่าเร็พพิวเทชันของเส้นทางในตำแหน่งคงที่

ในบทนี้จะนำเสนอการตัดสินใจเลือกเส้นทางที่ดีที่สุดภายใต้วัตถุประสงค์ของการพิจารณาการลดการใช้พลังงาน ยืดอายุการใช้งานและค่าเร็พพิวเทชันของโนดภายในโครงข่ายเซนเซอร์ไร้สาย ด้วยวิธีการมอนติคาร์โล โดยการพิจารณาเพื่อการนำไปประยุกต์ใช้ในเหตุการณ์อุทกภัย

เนื้อหาของบทนี้จะประกอบด้วย หัวข้อ 3.1 สมมติฐานของแบบจำลองโดยจะกล่าวถึงโครงรูปของโครงข่ายที่ใช้ในการสมมติฐานแบบจำลอง การกำหนดพารามิเตอร์ต่างๆ หัวข้อ 3.2 แบบจำลองพลังงานจะกล่าวถึงวิธีการคำนวณพลังงานเพื่อเป็นพารามิเตอร์ตัวหนึ่งที่นำมาตัดสินใจเลือกเส้นทาง หัวข้อ 3.3 วิธีการเร็พพิวเทชัน เป็นส่วนของการพิจารณาความน่าเชื่อถือของเซนเซอร์โนดในระบบและวิธีการคำนวณพารามิเตอร์ดังกล่าว หัวข้อ 3.4 การนิยามปัญหาโดยจะกล่าวถึงการจำลองแบบในรูปของวิธีการออนโพลิซีมอนติคาร์โล การกำหนดพารามิเตอร์ที่ใช้ประกอบในการเรียนรู้ เช่น สถานะของโครงข่าย การกระทำที่เกิดขึ้นในระบบ และการกำหนดและพิจารณาผลรางวัล ส่วนสุดท้ายหัวข้อ 3.5 จะเป็นส่วนของบทสรุป

3.1 โครงข่ายที่พิจารณา

แบบจำลองจะจำลองจากลักษณะภูมิศาสตร์ของแม่น้ำเจ้าพระยาดังรูปที่ 3.1 โดยกำหนดการวางตำแหน่งของโนดให้มีลักษณะการกระจายตัวแบบสุ่ม (random) ตลอดช่วงแม่น้ำที่พิจารณา โดยโนดที่กระจายตัวอยู่นั้นสามารถแบ่งได้ 2 ประเภท คือ โนดที่ลอยติดทุ่นอย่างคงที่บนผิวน้ำ ทำหน้าที่เก็บรวบรวมข้อมูลและสามารถถ่ายทอดแพ็กเกตข้อมูลนั้นเรียกว่า เซนเซอร์โนด (sensor node) ในขณะเดียวกันโนดที่รับการติดตั้งประจำที่ในพื้นที่ที่สามารถเข้าถึงเฉพาะบริเวณริมฝั่งแม่น้ำรวมไปถึงสถานีฐานซึ่งเป็นระบบโทรมาตรเดิมที่มีอยู่จริง ทำหน้าที่ส่งผ่านแพ็กเกตข้อมูลที่ได้รับจากเซนเซอร์โนดไปยังโครงข่ายบรอดแบนด์หรือโนดปลายทาง เรียกว่า แอคเตอร์โนด (actor node) ซึ่งในที่นี้กำหนดให้ค่าตัวแปรของเซนเซอร์โนดของ i โดยที่ $i = 1, 2, 3, \ldots, I$ และแอคเตอร์โนดเป็น j โดยที่ $j = 1, 2, 3, \ldots, J$ ลักษณะส่งแพ็กเกตข้อมูลเป็นแบบทิศทางเดียวโดยส่งจากโนดที่อยู่ต้นทางไปจนถึงปลายทาง ทุกๆโนดจะมีการส่งสัญญาณแบบรอบทิศทาง (omni-directional) ที่มีระยะการส่งสัญญาณสูงสุดของเซนเซอร์โนด τ^s และ แอคเตอร์โนด τ^a ตามลำดับ กำหนดให้ระยะการส่งสัญญาณสูงสุดแอคเตอร์โนดมีระยะการส่งที่สูงกว่าเซนเซอร์โนด ดังนั้น $\tau^s \leq \tau^a$ แต่ละโนดสามารถเชื่อมต่อกันได้และมีการแลกเปลี่ยนข้อมูลโดยอาศัยข้อมูลจากการเก็บรวบรวมเองหรือเป็นข้อมูลตำแหน่งที่ได้รับจาก GPS



รูปที่ 3.1: แบบจำลองลักษณะภูมิศาสตร์ของแม่น้ำเจ้าพระยา

3.2 ปัจจัยที่พิจารณา

3.2.1 แบบจำลองของพลังงาน (Energy Model)

การสูญเสียพลังงานอันเนื่องมาจากการรับ-ส่งแพ็กเกตข้อมูล กำหนดให้มีความต้องการในการส่งแพ็กเกตข้อมูลเป็นลักษณะปัวซอง (Poisson) โดยที่การส่งข้อมูลแต่ละครั้งจากโนดต้นทางไปยังโนดปลายทางจะเรียกว่า เหตุการณ์ (event) เพื่อความสะดวกจะขอนิยามเวลา t ณ ที่นี้หมายถึงเหตุการณ์ใดๆ โดยการคำนวณการสูญเสียพลังงานจะแบ่งออกเป็น 2 ลักษณะตามประเภทของโนดนั้นๆ กำหนดให้แบตเตอรี่ของเซนเซอร์โนดไม่สามารถชาร์จพลังงานได้ใหม่ ในขณะที่แบตเตอรี่ของแอคเตอร์โนดมีความสามารถชาร์จพลังงานได้ จึงเขียนได้ว่า $\mathcal{E}^s = \{[\alpha_i(t)]\}$ คือเซตของพลังงานที่เหลืออยู่ของเซนเซอร์โนด i ณ เวลา t ใดๆ และ $\mathcal{E}^a = \{[\beta_j(t)]\}$ คือเซตของพลังงานที่เหลืออยู่ของแอคเตอร์โนด j ณ เวลา t ใดๆ โดย $i = 1, 2, \ldots, I$ และ $j = 1, 2, \ldots, J$ ตามลำดับ

พลังงานที่เหลืออยู่ของเซนเซอร์โนด ณ เวลาถัดไป เมื่อมีการตัดสินใจเลือกเส้นทาง $i \in l(t)$ จะคำนวณได้จากการพลังงานที่เหลืออยู่ของเซนเซอร์โนด ณ เวลาปัจจุบัน เมื่อเกิดการรั่วไหลของพลังงานในเซนเซอร์โนดลบด้วยพลังงานที่เกิดการสูญเสียสำหรับการรับส่ง หรือแลกเปลี่ยนข้อมูล แต่หากเส้นทางนั้นไม่ถูกเลือก $i \notin l(t)$ จะไม่มีการคำนวณในส่วนพลังงานที่สูญเสียจากการแลกเปลี่ยนข้อมูลดังสมการ (3.1)

$$\alpha_i\left(t+1\right) = \begin{cases} \alpha_i(t)\eta & , \quad i \notin l(t) \\ \alpha_i(t)\eta - \chi(t) & , \quad i \in l(t) \end{cases} \tag{3.1}$$

พลังงานที่เหลืออยู่ของแอคเตอร์โนด ณ เวลาถัดไป จะมีการพิจารณาคล้ายกับแบบจำลองพลังงานของเซนเซอร์โนด แต่เนื่องจากในงานนี้แอคเตอร์โนดทำหน้าที่เป็นโนดปลายทางและมีความสามารถชาร์จพลังงานได้เมื่อพลังงานไม่เพียงพอสำหรับการแลกเปลี่ยนข้อมูล จึงทำให้ในส่วนที่เป็นพลังงานที่จะเกิดการรั่วไหลไม่ถูกนำมาพิจารณาดังสมการที่ (3.2)

$$\beta_j(t+1) = \begin{cases} \beta_j(t) & , \quad j \notin l(t) \\ \beta_j(t) - \chi(t) & , \quad j \in l(t) \end{cases} \tag{3.2}$$

เมื่อ $\alpha_i(t)$ คือ พลังงานที่เหลืออยู่ของเซนเซอร์โนด i ณ เวลา t
$\beta_j(t)$ คือ พลังงานที่เหลืออยู่ของแอคเตอร์โนด j ณ เวลา t
$\alpha_i(t+1)$ คือ พลังงานที่เหลืออยู่ของเซนเซอร์โนด i ณ เวลา $t+1$
$\beta_j(t+1)$ คือ พลังงานที่เหลืออยู่ของแอคเตอร์โนด j ณ เวลา $t+1$
$\chi(t)$ คือ พลังงานที่เกิดการสูญเสียสำหรับการรับส่งหรือแลกเปลี่ยนข้อมูล ณ เวลา t
$l(t)$ คือ เส้นทางที่ถูกเลือก ณ เวลา t
η คือ สัดส่วนของพลังงานที่จะเกิดการรั่วไหล ณ เวลา t ใดๆ

พลังงานที่เกิดการสูญเสียสำหรับการแลกเปลี่ยนข้อมูล $\chi(t)$ สามารถพิจารณาได้จากรูปการแบบการสื่อสารสัญญาณวิทยุโดยจะได้จากพลังงานที่สูญเสียในการส่งจากเซนเซอร์โนด i ไปยังเซนเซอร์โนดที่ตำแหน่ง downstream i' ที่เวลา t แทนด้วยตัวแปร $\tau_{i,i'}(t)$ และพลังงานที่สูญเสียในการรับข้อมูลจากโนดปลายทาง i'' มายังโนดต้นทาง i แทนด้วยตัวแปร $\mu_{i'',i}(t)$ ดังสมการที่ (3.3)

$$\chi(t) = \tau_{i,i'}(t) + \mu_{i'',i}(t) \tag{3.3}$$

สำหรับพลังงานที่สูญเสียในฝั่งส่งจะคำนวณได้จากแลกเปลี่ยนข้อมูล b ไบต์ ระหว่างหนึ่งคู่โนดโดยมีระยะการส่ง d เมตรซึ่งถูกกำหนดโดยค่าเลขชี้กำลังการสูญเสียตามระยะทาง (path loss exponent) σ ดังสมการ (3.4) [21]

$$\tau_{i,i'}(t) = \begin{cases} 0 & , \quad i = \text{destination} \\ e_d b d^{\sigma} + e_t b & , \quad \text{otherwise} \end{cases} \tag{3.4}$$

โดยพลังงานที่สูญเสียเมื่อมีการส่ง e_t แพ็กเกตข้อมูลมีค่าเท่ากับ $50\ nJ/bit$ และค่าความคาดเคลื่อนของการส่งสัญญาณที่สูญเสียของเซนเซอร์โนด e_d ต่อระยะทางหนึ่งเมตร เมื่อพิจารณาในพื้นที่ว่าง (free space) มีค่าเท่ากับ $100\ pJ/bit \times m^{\sigma}$ ซึ่งค่าการลดทอนของสัญญาณ σ มีค่าเท่ากับ 2 และคำนวณพลังงานที่สูญเสียในการรับจากโนด i'' ไปยังอัปสตรีมโนด (upstream node) i ดังสมการที่ (3.5)

$$\mu_{i'',i}(t) = \begin{cases} 0 & , \quad i = \text{source} \\ e_l b & , \quad \text{otherwise} \end{cases} \tag{3.5}$$

โดย e_l คือพลังงานที่สูญเสียเมื่อมีการรับแพ็กเกตข้อมูลมีค่าเท่ากับ $50\ nJ/bit$ ซึ่งในงานนี้ในส่วนของ $e_l b$ มีการสูญเสียพลังงานในการรับที่มีค่าน้อยมากดังนั้นจึงไม่ส่งผลต่อการสูญเสียพลังงานในส่วนของการส่งแพ็กเกต ดังนั้นจึงสามารถเขียนได้ดังสมการ (3.6)

$$\chi(t) = \tau_{i,i'}(t) + \mu_{i'',i}(t) = e_d b d^{\sigma} \tag{3.6}$$

ดังนั้นจึงกำหนดให้ปริมาณพลังงานที่เหลืออยู่ในเซนเซอร์โนด i จะขึ้นจำนวนแพ็กเกตที่ส่งรวมถึงขนาดของแพ็กเกตที่ใช้ในการส่งตลอดเส้นทาง

3.2.2 แบบจำลองของอายุการใช้งานของโนด (Node Lifetime)

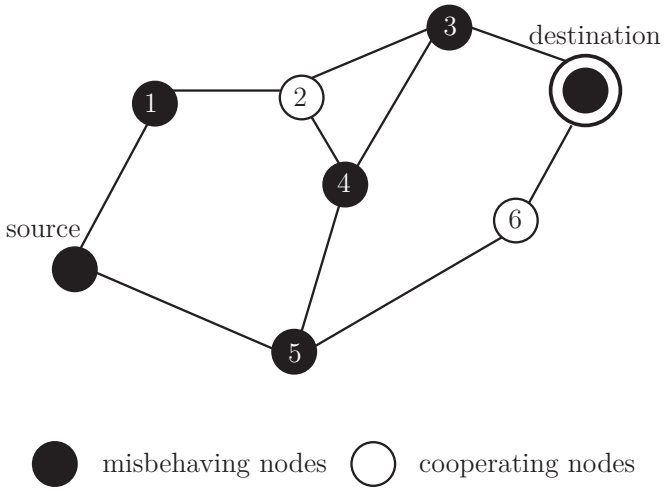
สำหรับอายุการใช้งานของเซนเซอร์โนดและแอคเตอร์โนดจะทำการพิจารณาในลักษณะเดียวกัน คืออายุการใช้งานของเซนเซอร์โนดและแอคเตอร์โนด ณ เวลาถัดไป $t+1$ จะขึ้นอยู่กับอายุการใช้งานปัจจุบัน t หักลบค่าคงที่ค่าหนึ่ง เมื่อเซนเซอร์โนดหรือแอคเตอร์โนดนั้นๆอยู่บนเส้นทางที่ถูกเลือก

$$\ell_i\,(t+1) = \begin{cases} \ell_i(t) & ;\ \ i \notin l(t) \\ \ell_i(t) - c & ;\ \ i \in l(t) \end{cases}$$
$$\ell_j\,(t+1) = \begin{cases} \ell_j(t) & ;\ \ j \notin l(t) \\ \ell_j(t) - c & ;\ \ j \in l(t) \end{cases} \tag{3.7}$$

เมื่อ $\ell_i(t)$ คือ อายุการใช้งานของเซนเซอร์โนด i ณ เวลา t
$\ell_j(t)$ คือ อายุการใช้งานของแอคเตอร์โนด j ณ เวลา t
c คือ ค่าคงที่

3.2.3 วิธีการค่าเร็พพิวเทชัน (Reputation Method)

วิธีการค่าความมีชื่อเสียงหรือเร็พพิวเทชันเป็นเทคนิควิธีการที่มีขึ้นเพื่อหาความเหมาะสมสำหรับการระบุความน่าเชื่อถือของโนดในการส่งแพ็กเกตข้อมูลสำหรับโครงข่ายเฉพาะกิจ โดยวิธีค่าเร็พพิวเทชันในที่นี้สามารถใช้ในการตรวจสอบโนดที่ทำงานผิดพลาดโดยพิจารณาอัตราการทำงานของโนดนั้นๆเมื่อมีการจัดสรรเส้นทางและมีการรับ-ส่งแพ็กเกตข้อมูลเกิดขึ้นภายในโครงข่าย ซึ่งวิธีนี้เป็นการตรวจสอบสถานะและให้คะแนนโนดอื่นๆข้างเคียงด้วยการพิจารณาพฤติกรรมจากความเห็นและประสบการณ์จากพฤติกรรมของโนดข้างเคียง ดังนั้นให้ความเห็นหรือประสบการณ์เกี่ยวกับโนดอื่นๆเรียกว่าค่าเร็พพิวเทชันหรือค่าความมีชื่อเสียง เป้าหมายของเร็พพิวเทชันนั้นคือค่าความสามารถหรือลักษณะของพฤติกรรมของโนด พฤติกรรมของโนดชนิดใดที่ควรเอาเข้ามาใช้และโนดใดควรหลีกเลี่ยง ซึ่งค่าเร็พพิวเทชันที่เกิดขึ้นจึงสามารถนำมาตัดสินใจและตรวจสอบในทุกๆประเภทของโนด ตราบใดที่โนดนั้นสามารถตรวจสอบได้ดังรูปที่ 3.2 โดยกำหนดให้โนดที่ให้ความร่วมมือ (cooperating node)



รูปที่ 3.2: ลักษณะของโนดที่ผิดปกติในโครงข่ายเซนเซอร์

แทนด้วยโนดสีขาว และ โนดที่มีพฤติกรรมไม่ปรกติหรือมีความเสี่ยงที่จะเกิดปัญหา (misbehaving node) แทนด้วยโนดสีดำ

ในงานวิจัย [22] ค่าเร็พพิวเทชันของแต่ละโนดจะได้มาจากจำนวนแพ็กเกตที่โนดสามารถส่งไปได้ก่อนหน้าของโนดนั้นๆ โดยโนดต้นทางค้นหาเส้นทางที่เป็นไปได้เพื่อไปยังโนดปลายทางโดยโพรโทคอลการจัดสรรเส้นทาง โนดต้นทางจะเริ่มส่งแพ็กเกตไปยังโนดข้างเคียงด้วยค่าเร็พพิวเทชันที่สูงที่สุด ซึ่งโนดนี้จะทำการส่งแพ็กเกตดังกล่าวไปยังฮอพข้างหน้าด้วยค่าเร็พพิวเทชันที่สูงสุดและจะมีการทำซ้ำไปเรื่อยๆ จนกระทั่งโนดปลายทางได้รับแพ็กเกตนั้น จากนั้นจะมีการปรับปรุงค่าเร็พพิวเทชันภายในตารางการหาเส้นทาง routing table) ซึ่งจะมีเกณฑ์ของค่าเร็พพิวเทชัน (reputation threshold) ที่จะเป็นการบอกถึงค่าต่ำสุดของเร็พพิวเทชันที่ยอมรับได้ ด้วยการคาดเดาจากความน่าจะในการส่งแพ็กเกตผ่านไปยังฮอพถัดไปในเส้นทางนั้นๆ ถ้าฮอพถัดไปนั้นไม่มีค่าเร็พพิวเทชันตามที่ต้องการฮอพดังกล่าวจะไม่ได้รับการส่งแพ็กเกตผ่าน ซึ่งเกณฑ์ของค่าเร็พพิวเทชันจะทำการคัดสรรเฉพาะโนดที่เป็นไปได้ในการส่งผ่านแพ็กเกต

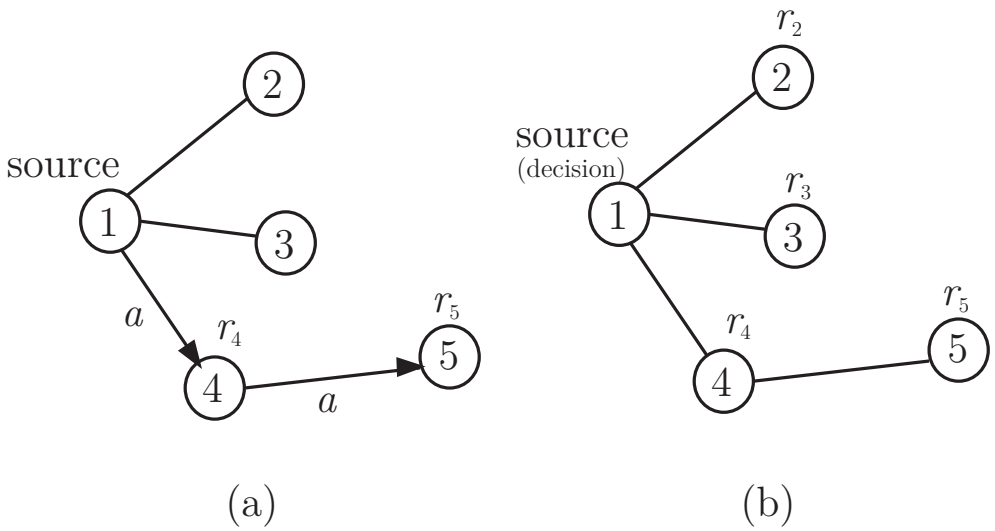
ตัวอย่างเมื่อมีการพิจารณาโครงข่ายที่ประกอบด้วยโนด A โนด B และโนด C ถ้าโนด A ต้องการส่งแพ็กเกตไปยังโนด C ซึ่งโนด A จะต้องค้นหาเส้นทางที่จะสามารถส่งแพ็กเกตไปยังโนด C ในที่นี้คือต้องผ่านโนด B จากนั้นเมื่อส่งแพ็กเกตผ่านไปยังโนด B และโนด C ตอบรับการได้รับแพ็กเกตดังกล่าว โนด A จึงจะให้ความเห็นและแนะนำผ่าน Rec_{AB} ได้ว่าโนด B เป็นโนดที่สามารถส่งแพ็กเกตได้อย่างถูกต้องและให้ค่าเร็พพิวเทชัน +1 จะได้ว่า $Rec_{AB} = +1$

โดยค่าเร็พพิวเทชัน (reputation value) ที่กล่าวถึงข้างต้นจะคำนวณจากความเห็นและแนะนำที่ได้รับจากโนดหนึ่ง ตัวอย่างเช่นโนด B ได้รับแพ็กเกต 100 แพ็กเกตและทำการส่งแพ็กเกตต่อไปยังโนดถัดไป 90 แพ็กเกตแต่เกิดการดร็อป 10 แพ็กเกต ดังนั้นโนดที่ทำการส่งแพ็กเกตให้โนด B จะถูกให้ความเห็นจากโนด A Rec_{AB} ด้วยค่า +1 หรือ -1 ซึ่งขึ้นอยู่กับเกณฑ์ของค่าเร็พพิวเทชัน จากข้างต้นค่าเร็พพิวเทชันของโนด B ที่ได้รับคือ $R_B = (100-10)/100 = 90/100 = 0.9$ ซึ่งลักษณะดังกล่าวจะสามารถเขียนค่าเร็พพิวเทชันของแต่ละโนดทั้งเซนเซอร์โนดและแอคเตอร์โนดได้ดังสมการ (3.8)

$$r_{i,l}\,(t) = \frac{\phi_{i,l}(t) - \varphi_{i,l}(t)}{\phi_{i,l}(t)}, \tag{3.8}$$

เมื่อ $r_{i,l}(t)$ คือ ค่าเร็พพิวเทชันของแต่ละโนดในเส้นทางนั้นๆ l ณ เวลา t
$\phi_{i,l}(t)$ คือ จำนวนการแพ็กเกตข้อมูลที่มีการส่งผ่านโนด
ในเส้นทาง l ณ เวลา t
$\varphi_{i,l}(t)$ คือ คือจำนวนการแพ็กเกตข้อมูลที่สูญหายหรือเกิดการดรอป
ในเส้นทาง l ณ เวลา t

การนำวิธีเร็พพิวเทชันไปใช้ในการประกอบการเลือกเส้นทางนั้น เป็นกระบวนการที่ง่ายโดยอาศัยจากข้อมูลเดิมได้รับถึงขีดความสามารถของโนดแต่ละโนดในการส่งข้อมูล ซึ่งการตัดสินใจในการเลือกโนดข้างเคียงนั้นจะเกิดขึ้นและพยายามทำการสร้างเส้นทางขึ้นมาจากเพียงข้อมูลเพียง 1 ฮอพ จึงทำให้โอกาสที่จะเกิดปัญหาอันเนื่องมาจากการเลือกเส้นทางที่ผิดพลาดจากโนดตัวกลาง ตัวใดตัวหนึ่งที่ไม่ทราบถึงข้อมูลของโนดถัดไปในฮอพที่ $h+1$ ดังนั้นในงานวิจัยนี้จึงได้นำเสนอค่าเร็พพิวเทชันที่พิจารณาตลอดเส้นทาง (path reputation) ดังรูปที่ 3.3 (a) การตัดสินใจเลือกเส้นทางด้วยการใช้ค่าเร็พพิวเทชันโดยโนดที่มีความต้องการในการส่งข้อมูล จะทำการตัดสินใจเลือกโนดข้างเคียงจากตารางการจัดสรรเส้นทางที่บรรจุค่าเร็พพิวเทชันของโนดข้างเคียง โดยเส้นทางจะถูกสร้างตัวขึ้นและดำเนินต่อไปเรื่อยๆจนถึงโนดปลายทาง ในทางตรงกันข้ามดังรูปที่ 3.3 (b) โนดที่มีความต้องการในการส่งข้อมูล จะทำการรวบรวมข้อมูลจากโนดรอบข้างไปจนถึงโนดปลายทาง และ ทำการรวบรวมข้อมูลเหล่านั้นกลับมายังโนดต้นทาง หลังจากนั้นโนดต้นทางจะทำการคำนวณค่าเร็พพิวเทชันของเส้นทางและเลือกเส้นทางที่ดีที่สุด ซึ่งการทำวิธีนี้สามารถลดโอกาสที่จะเจอโนดที่มีปัญหาระหว่างการส่งข้อมูลได้ ซึ่งจะมีประสิทธิภาพมากกว่าวิธีแรกในรูปที่ 3.3 (a)



รูปที่ 3.3: (a) การตัดสินใจเลือกเส้นทางด้วยการใช้ค่าเร็พพิวเทชัน [22] (b) การตัดสินใจเลือกเส้นทางด้วยการใช้ค่าเร็พพิวเทชันของเส้นทาง

ดังนั้นค่าเร็พพิวเทชันที่พิจารณาตลอดเส้นทาง $\rho_l(t)$ เกิดจากการคูณของค่าเร็พพิวเทชันประจำเซนเซอร์โนด i บนเส้นทาง l ดังสมการ (3.9)

$$\rho_l(t) = \prod_{\forall i \in l} r_{i,l}(t). \tag{3.9}$$

โดยค่าเร็พพิวเทชันของแต่ละโนดที่ได้รับจะเสมือนการแสดงตัวตนหรือบ่งชี้ในโครงข่าย ซึ่งถ้ามีโนดใดโนดหนึ่งประสงค์ร้ายหรือไม่ให้ความร่วมมือต่อโครงข่าย ค่าเร็พพิวเทชันของโนดนั้นๆจะถูกลดลงอย่างรวดเร็วและโนดดังกล่าวจะเสมือนถูกตัดออกจากโครงข่ายแม้โนดนั้นยังคงอยู่

3.3 การนิยามปัญหาในรูปออนโพลิซีมอนติคาร์โล

หัวข้อนี้จะกล่าวถึงการนิยามปัญหาบนพื้นฐานของวิธีการออนโพลิซีมอนติคาร์โลที่มีการอธิบายไว้ในบทก่อนหน้านี้ โดยวิธีดังกล่าวจะมีการใช้การจำลองแบบการทำงานเป็นรอบเพื่อพิจารณาว่าการตัดสินใจเลือกการกระทำเช่นไร โดยการตัดสินใจนั้นจะมีการปรับปรุงและเรียนรู้แบบรอบต่อรอบ (episode-by-episode) โดยอาศัยการประเมินค่าจากการประเมินแอคชั่น-แวลู่ฟังก์ชัน (หัวข้อ 2.3.1) เพื่อใช้ในการปรับปรุงนโยบายการตัดสินใจ (หัวข้อ 2.3.2) ผ่านนโยบายแบบ ϵ-กรีดี้ โดยในแต่ละรอบของการเรียนรู้มีเป้าหมายเพื่อหาค่าของผลรางวัลสูงสุดในระยะยาว (long term average reward)

งานวิจัยฉบับนี้นำวิธีออนโพลิซีมอนติคาร์โลมาประยุกต์ใช้เพื่อหาผลคำตอบ เนื่องจากลักษณะของพฤติกรรมการค้นหาเส้นทางภายในโครงข่ายแอดฮอกและการพิจารณาหาผลตอบแบบที่มีขีดจำกัดของการพิจารณา (finite horizon) ซึ่งแต่ละเอพพิโซดจะเริ่มจากโนดต้นทางเริ่มต้นค้นหาเส้นทางเพื่อจะไปสู่โนดปลายทาง ซึ่งในวิทยานิพนธ์นี้จะพิจารณาเหตุการณ์ (event) ของการส่งข้อมูลที่มากกว่า 1 ครั้ง เพื่อใช้ในการจำลองสถานการณ์ที่เกิดขึ้นจริงในระบบเตือนภัยอุทกภัย ดังนั้นในแต่ละเอพพิโซด อาจจะเกิดเหตุการณ์ที่มากกว่า 1 ครั้ง โดยตัวอย่างของเหตุการณ์เป็นสถานการณ์ของการที่เซนเซอร์โนดต้องการที่จะรายงานข้อมูลที่วัดได้เข้าสู่แอคเตอร์โนดจะสิ้นสุดเหตุการณ์ใดๆ ก็ต่อเมื่อมีการเชื่อมต่อและส่งข้อมูลเสร็จสิ้น โดยยกตัวอย่างเช่นนำค่าของระดับแบตเตอรี่ที่เหลืออยู่และค่าเร็พพิวเทชันที่ได้รับจากแต่ละโนดข้างเคียงมาพิจารณา เพื่อเป็นข้อมูลของสถานะของระบบสำหรับการตัดสินใจเลือกเส้นทาง ซึ่งแต่ละโนดต้นทางจะทำหน้าที่เสมือนตัวแทนที่ตัดสินใจในการเลือกเส้นทางบนสถานะ ณ ปัจจุบัน โดยสมมติว่าแต่ละโนดมีการเคลื่อนที่ที่เป็นอิสระ (ตำแหน่ง ทิศทาง และความเร็ว) ซึ่งจะเห็นได้ว่าลักษณะของทอพอโลยีขึ้นอยู่กับลักษณะทอพอโลยีปัจจุบัน แต่ไม่ขึ้นกับลักษณะของทอพอโลยีที่เคยเป็นมา ดังนั้นจึงกล่าวได้ว่าสถานะ ณ เวลาถัดไป (ระดับแบตเตอรี่ที่เหลืออยู่และค่าเร็พพิวเทชันของแต่ละโนด) ขึ้นอยู่กับสถานะปัจจุบันเท่านั้น โดยไม่ขึ้นกับสถานะในอดีตที่เคยมีมา ดังนั้นพฤติกรรมของระบบโครงข่ายจึงมีคุณสมบัติของการเป็นมาร์คอฟ

ทอพอโลยีที่ใช้ในการจำลองแบบทั้งหมดสองแบบด้วยกันคือ แบบแรกคือวิธีการโปรแอคทีฟเน็ตเวิร์คมาเนจเมนท์แบบออนโพลิซีมอนติคาร์โล แบบที่สองคือวิธีการรีแอคทีฟเน็ตเวิร์คมาเนจเมนท์แบบออนโพลิซีมอนติคาร์โล โดยทั้งสองแบบที่กล่าวมานั้นมีการจำลองแบบโดยทอพอโลยีในระบบมีการไม่เคลื่อนที่ กำหนดให้ลิงค์ทำงานเป็นปกติตลอดเวลา ไม่มีการพิจารณาถึงสภาวะที่โนดบางโนดไม่สามารถทำงานได้หรือในกรณีที่โครงข่ายมีการค้นหาเส้นทางใหม่เกิดขึ้น นอกจากนั้นจะมีการพิจารณาการจำลองแบบที่สนใจเหตุการณ์ที่เปลี่ยนแปลงไป (discrete event simulation) ดังนั้นการเลื่อนไปของเวลาที่ใช้ในแบบจำลองจะขึ้นอยู่กับการเกิดเหตุการณ์ ในที่นี้คือเหตุการณ์ที่โนดต้นทางต้องการส่งข้อมูลไปยังโนดปลายทาง ดังนั้นจึงสามารถกำหนดค่าพารามิเตอร์ต่างๆ ได้ดังนี้

3.3.1 กำหนดนิยามสถานะของโครงข่าย (state)

สถานะของสิ่งแวดล้อมซึ่งในที่นี้คือสถานะของการแจ้งเตือนหรือความต้องการในการส่งข้อมูลที่เกิดขึ้นในระบบโครงข่าย ซึ่งโดยปกติแล้วโนดทุกๆโนดจะไม่ทำการแจ้งเตือนสถานะในกรณีที่ไม่เกิดการเปลี่ยนแปลงของสภาพแวดล้อม ซึ่งในที่นี้หมายถึงการเปลี่ยนแปลงของความเร็ว ค่า PH และระดับน้ำ แต่หากสภาพแวดล้อมที่กล่าวมานั้นเกิดการเปลี่ยนแปลงจะถือได้ว่าสิ่งแวดล้อมนั้นอยู่ในสถานะที่ไม่ปกติ ดังนั้นสถานะที่เป็นไปได้ทั้งหมดของโครงข่ายที่มีทอพอโลยีดังรูปที่ 3.4, 3.8 กำหนดให้ $s_i(t)$ เป็นตัวแปรสถานะ (state space) ของเซนเซอร์โนด i ที่เวลา t โดยที่

$$\mathcal{S} = \{\mathbf{s}(t) = [s_i(t)], \forall i\}, i = 1, 2, 3, ..., I \tag{3.10}$$

โดย $\mathcal{S}$ แทนด้วยปริภูมิสถานะของระบบและ $\mathbf{s(t)}$ แทนด้วยเวคเตอร์สถานะของระบบ ซึ่งหากเซนเซอร์โนด i ใดๆไม่มีความต้องการที่จะส่งข้อมูลที่เวลา t ดังนั้นค่าตัวแปรสถานะของเซนเซอร์โนด i จะมีค่าเป็น $s_i(t) = 0$ แต่ในทางตรงกันข้าม หากเซนเซอร์โนด i สามารถตรวจจับการเปลี่ยนแปลงของแม่น้ำและมีความต้องการที่จะส่งข้อมูลไปยังสถานีฐานหรือแอคเตอร์โนด ตัวแปรสถานะของเซนเซอร์โนดจะเป็น $s_i(t) = 1$ โดยในงานวิจัยนี้กำหนดให้มีเพียงเหตุการณ์เดียวหรือการแจ้งเตือนเดียวเท่านั้นที่จะเกิดขึ้นในระหว่างช่วงเวลา t ที่พิจารณาซึ่งจะทำให้ $\sum_{i \in I} s_i(t) = 1$

3.3.2 กำหนดนิยามการกระทำของโครงข่าย (action)

ตัวแปรสถานะที่เวลา t ใดๆ กำหนดให้ i^* คือเซนเซอร์โนดที่ทำการแจ้งเตือนเพื่อขอส่งข้อมูล โดยกำหนดให้ปริภูมิการกระทำ (action space) A โดยที่ A หมายถึงชุดของทุกการกระทำที่เป็นไปได้ที่จะสามารถตัดสินใจเลือกกระทำ ซึ่งการกระทำ $a(s_{i^*}(t)) \in A$ คือทุกการกระทำที่เป็นไปได้ของโนดที่ทำการแจ้งเตือน i^* ดังนั้นหากเกิดการแจ้งเตือนหรือมีโนดที่ต้องการส่งข้อมูลเกิดขึ้น ตัวตัดสินใจ (agent) จะดำเนินการเลือกเส้นทางที่ดีที่สุดจากเส้นทางทั้งหมดที่มีในปริภูมิการกระทำผ่านข้อมูลพื้นฐานของประสบการณ์ของตัวตัดสินใจ (agent's experiences) ซึ่งจะตัดสินใจเลือกโดยนโยบายแบบละโมบ (สมการ 2.18) ซึ่งถ้าการทำงานของชั้น MAC ทำงานได้อย่างสมบูรณ์ ดังนั้นค่าของเวคเตอร์เส้นทางทั้งหมดที่เป็นไปได้หากเซนเซอร์โนด i^* ทำการแจ้งเตือนเพื่อส่งข้อมูล $\mathbf{l}_{i^*}(t)$ จะเขียนได้เป็น

$$a(s_{i^*}(t)) = \mathbf{l}_{i^*}(t) \tag{3.11}$$

โดย $\mathbf{l}_{i^*}(t) = [l_{i^*,1}(t), l_{i^*,2}(t), l_{i^*,3}(t), ..., l_{i^*,n}(t)] = [l_{i^*,n}(t)]$ โดยที่ n คือจำนวนของเส้นทางทั้งหมดที่เกิดขึ้นเมื่อชั้น MAC ทำงานได้อย่างสมบูรณ์ และ โนด i^* เป็นโนดที่ต้องการจะส่งข้อมูล

3.3.3 กำหนดผลรางวัลของโครงข่าย (reward)

วิธีการมอนติคาร์โลใช้ผลรางวัลที่ได้เป็นการสะท้อนการดำเนินการของการตัดสินใจเลือกกระทำที่เวลา t เมื่อระบบอยู่ในสถานะ s_i และ เมื่อตัดสินใจกระทำ a_i จะให้ผลรางวัลทันที $f(s_i(t), a_i(t))$

ซึ่งผลรางวัลดังกล่าวจะดีหรือไม่ขึ้นอยู่กับฟังก์ชันที่มีการพิจารณาพลังงานที่เหลืออยู่ของเซนเซอร์โนด $\overline{\mathcal{E}}$ อายุการใช้งาน $\bar{\ell}$ และค่าเร็พพิวเทชันของเส้นทาง $\bar{R}_l$ ฉะนั้นพลังงานที่เหลือ อายุการใช้งานของแต่ละโนดเหลืออยู่และค่าเร็พพิวเทชันของเส้นทางจะแสดงถึงความสามารถในการตัดสินใจ ณ ขณะนั้น มีค่ามากจะถือว่าการตัดสินใจเลือกเส้นทางนั้นๆ ถูกต้อง

$$f(s_{i^*}(t), a(s_{i^*}(t))) = w_{\mathcal{E}}\overline{\mathcal{E}}_l(t) + w_{\ell}\bar{\ell}_l(t) + w_R\bar{R}_l(t) \tag{3.12}$$

สมการที่ (3.12) แสดงถึงค่าผลรางวัลที่ได้รับเมื่อโนด i^* มีความต้องการที่จะส่งข้อมูล

ในบทนี้การทดสอบการจำลองแบบที่เกิดขึ้นเป็นเพียงการทดสอบเบื้องต้นเพื่อทดสอบความสามารถของการตัดสินใจเลือกอย่างง่าย ภายหลังจากบทนี้จึงจะมีการทดสอบระบบประสิทธิภาพของผลรางวัลที่นำเสนอไปใช้ในวิธีการมอนติคาร์โลไปประยุกต์ใช้กับระบบเตือนอุทกภัย ฉะนั้นเบื้องต้นจึงกำหนด $f(s_i(t), a_i(t)) = \{0, 1\}$ โดยมีเงื่อนไขบังคับ (constraints) เป็นตัวบังคับในการเลือกเส้นทางภายใต้ข้อจำกัดที่วางไว้ หลังจากนั้นจะมีการเฉลี่ยของผลรางวัลโดยการใช้วิธีการเอพพิโซดิก-ท๊าสค์ ซึ่งเป็นวิธีการประเมินหาคำตอบและปรับปรุงคำตอบแบบทีละเอพพิโซด ซึ่งเขียนแมททิกฟังก์ชันของผลรางวัลได้ว่า

$$Q\left(s_i\left(t\right), a\left(s_i\left(t\right)\right)\right) = E\left[f(s_i(t), a(s_i(t))\right] \tag{3.13}$$

โดย $E\left[.\right]$ คือค่าคาดหวังผลรางวัลเฉลี่ยในแต่ละสถานะ s_i และการกระทำ a_i ที่เวลา t

3.3.4 กำหนดการอัพเดตของระบบ

เมื่อเหตุการณ์ของการส่งข้อมูลที่เวลา t เสร็จสิ้นลง การอัพเดตค่าพารามิเตอร์ต่างๆของระบบที่เวลาถัดไป $t+1$ จะมีการอัพเดตจากสมการ 3.1–3.8 โดยค่าระดับพลังงานที่เหลืออยู่ของแต่ละเซนเซอร์โนด i ที่เวลาถัดไปเมื่อเซนเซอร์โนดบนเส้นทาง $l \in L$ ถูกเลือกจะพิจารณาจาก

$$\alpha_i(t+1) \geq u_{i,th} + \alpha_i(t)\eta - \chi\left(t\right), i \in l \tag{3.14}$$

โดย $u_{i,th}$ คือระดับพลังงานขีดแบ่งต่ำที่สุด (minimum energy threshold) ที่ยอมรับได้ของแต่ละเซนเซอร์โนด i ก่อนที่โนดจะไม่สามารถทำงานได้อีก (โนดตาย) และสำหรับค่าที่พลังงานที่เหลืออยู่ของแต่ละแอคเตอร์โนด j จะพิจารณาจาก

$$\beta_j(t+1) \geq u_{j,th} + \beta_j(t) - \chi\left(t\right), j \in l \tag{3.15}$$

โดย $u_{j,th}$ คือระดับพลังงานขีดแบ่งต่ำที่สุด (minimum energy threshold) ที่ยอมรับได้ของแต่ละแอคเตอร์โนด j ก่อนที่จะถูกเติมพลังงาน ในส่วนการอัพเดตอายุการใช้งานของแต่ละโนดระบบจะมีการอัพเดตจากสมการ (3.7) โดยจะต้องมีอายุการใช้งานเหลืออยู่ในแต่ละโนดที่เวลาถัดไป

$$\begin{aligned} \ell_i\left(t+1\right) > 0, i \in l \\ \ell_j\left(t+1\right) > 0, j \in l \end{aligned} \tag{3.16}$$

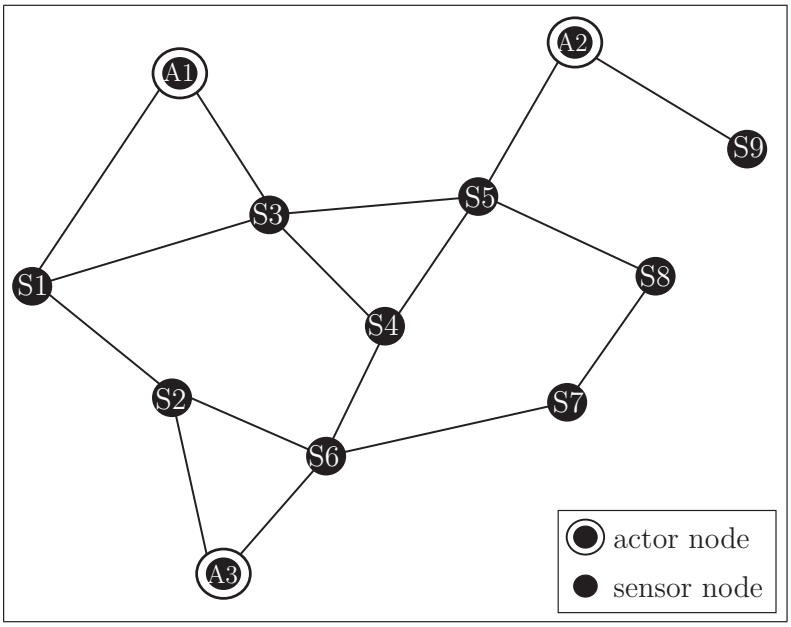
สุดท้ายจะมีการอัพเดตของค่าเร็พพิวเทชันประจำโนดด้วยสมการ (3.8) ซึ่ง $r(t+1) \leftarrow r(t)$

3.4 ผลการทดสอบเบื้องต้นในรูปออนโพลิซีมอนติคาร์โลในสภาวะคงที่

ในหัวข้อนี้จะกล่าวถึงการใช้แบบจำลองอย่างง่ายเบื้องต้นโดยมีวัตถุประสงค์ เพื่อพิจารณาถึงความเป็นไปได้ในการประยุกต์ใช้วิธีการมอนติคาร์โลที่มีการพิจารณาค่าเร็พพิวเทชันของเส้นทางโดยนำมาประยุกต์ใช้กับสถานการณ์การเตือนอุทกภัย ซึ่งแบ่งช่วงการทดสอบออกเป็นสองส่วนคือ ส่วนแรกใช้ในการทดสอบวิธีการด้วยการจัดสรรเส้นทางแบบโปรแอคทีฟบนโครงข่ายขนาดเล็กที่มีเส้นทางจำกัด และ ส่วนที่สองใช้การทดสอบวิธีการด้วยการจัดสรรเส้นทางแบบรีแอคทีฟบนโครงข่ายขนาดเล็กที่มีรูปร่างแตกต่างไปจากโครงข่ายของโปรแอคทีฟ โดยภาพของแบบจำลองนั้นถูกนำมาพิจารณาผ่านโปรแกรม MATLAB® และผลที่ได้จากบทที่ 3 จะนำไปสู่การวิเคราะห์ระบบในบทที่ 4 ต่อไป

การทดสอบวิธีการมอนติคาร์โลทั้งสองส่วนนั้น พิจารณาในโครงข่ายขนาดเล็กโดยมีความแตกต่างในรูปแบบของลักษณะโครงรูปโครงข่ายหรือทอพอโลยี (topology) และทั้งนี้ เพื่อให้เห็นถึงขีดความสามารถในการประยุกต์ใช้งานของวิธีการมอนติคาร์โลในลักษณะงานที่แตกต่างกันและด้วยข้อจำกัดของโปรแกรม MATLAB® จึงทำให้การพิจารณาในเชิงแพ็กเกตหรือในระดับชั้นการพิจารณาระดับที่ 2 (layer 2) นั้น ไม่ถูกนำมาพิจารณาร่วมในงานวิจัยนี้ และจากผลการทดลองทั้งสองระบบพบว่าเพื่อทำให้ระบบมีความสมบูรณ์และสามารถเข้าใจพฤติกรรมการเปลี่ยนแปลงที่เกิดขึ้นได้อย่างชัดเจน จึงไม่มีความจำเป็นต้องทำการเปรียบเทียบวิธีการจัดสรรเส้นทางทั้งสองแบบ แต่จำเป็นต้องพิจารณาโดยการปรับเปลี่ยนปริภูมิสถานะ รวมถึงฟังก์ชันผลรางวัลใหม่ทั้งหมดดังแสดงในบทที่ 4

3.4.1 การจัดสรรเส้นทางแบบโปรแอคทีฟบนโครงข่ายขนาดเล็กที่มีเส้นทางจำกัด [23]

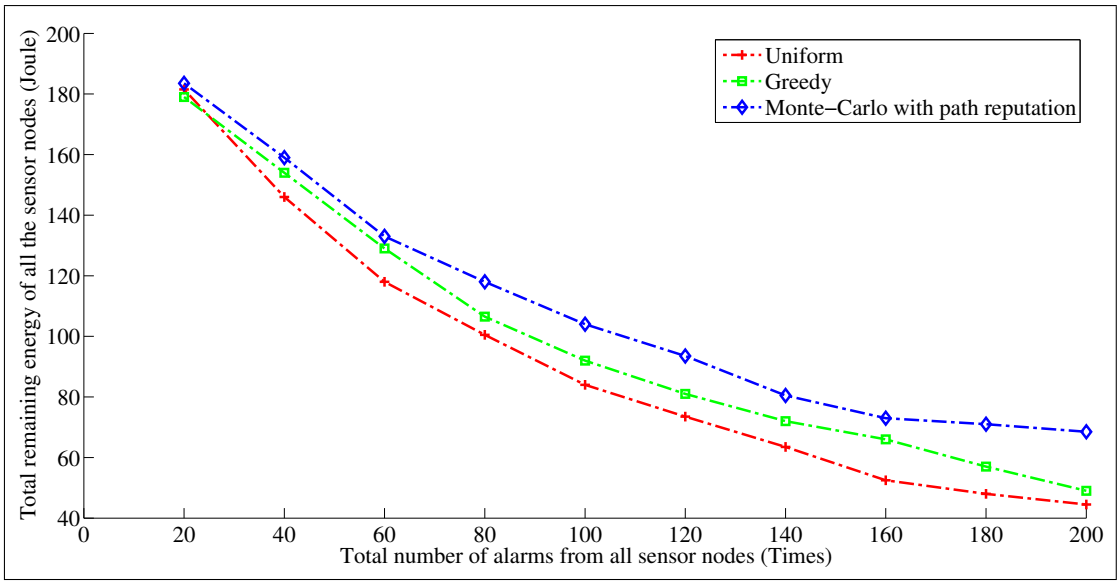


รูปที่ 3.4: ทอพอโลยีขนาดเล็กที่ใช้ในการทดสอบด้วยการเชื่อมต่อแบบโปรแอคทีฟ

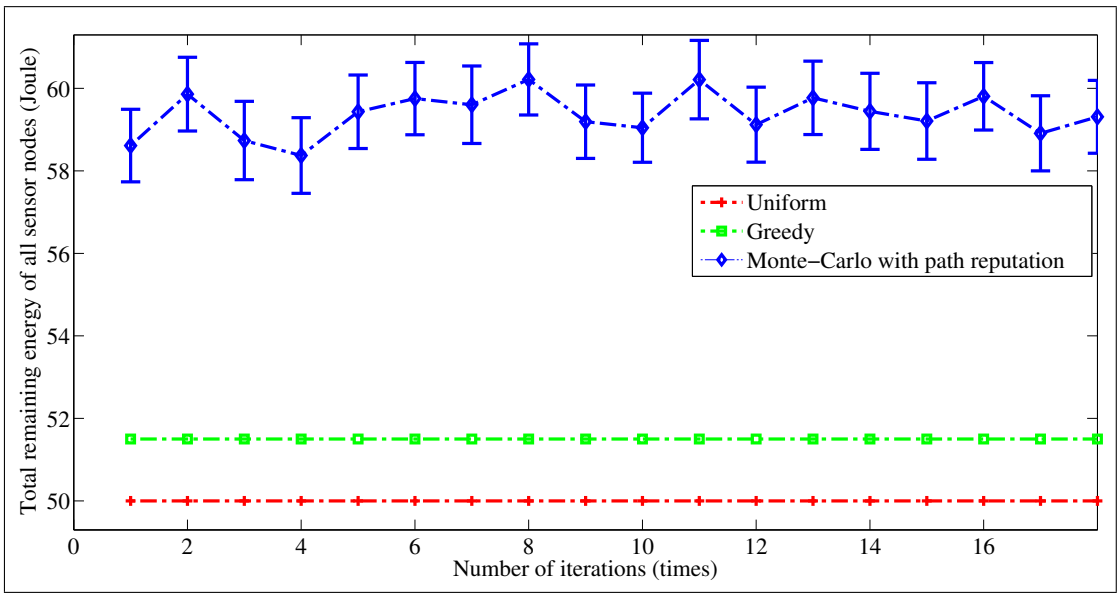
กำหนดให้แบบจำลองที่พิจารณาประกอบด้วยเซนเซอร์โนดจำนวน 9 โนด และแอคเตอร์โนดจำนวน 3 โนด ดังรูปที่ 3.4 โดยการวางโนดแบบสุ่มกระจายตัวบนพื้นที่พิจารณา 1000×1000 เมตร

ระยะการส่งของเซนเซอร์โนด 200 เมตรและแอคเตอร์โนด 200 เมตร และโครงข่ายในระบบกำหนดให้โนดไม่มีการเคลื่อนที่ สภาวะแวดล้อมในระบบคงที่ กำหนดให้มีการเชื่อมต่อของลิงค์อยู่เสมอบนพื้นฐานการค้นหาเส้นทางในลักษณะวิธีโปรแอคทีฟ พลังงานเริ่มต้นของแต่ละเซนเซอร์โนดมีค่า 15 จูล และของแต่ละแอคเตอร์โนดมีค่า 75 จูล งานวิจัยนี้มุ่งเน้นการพิจารณาพลังงานการรับส่งแพ็กเกตข้อมูล ไม่มีการคำนวณพลังงานระหว่างการค้นหาเส้นทาง ซึ่งหนึ่งคู่ระยะการส่งจะเกิดการสูญเสียพลังงาน 0.5 จูล โดยแต่ละเหตุการณ์จะมีการส่งแพ็กเกตข้อมูล 100 แพ็กเกตด้วยขนาดของข้อมูล 3.2 กิโลไบต์ การทดลองมีการจำลองเหตุการณ์ 200 เหตุการณ์ในแต่ละรอบการวนซ้ำ โดยระบบที่พิจารณาจะมีการวนซ้ำ 400 รอบ ภายใน 18 เอพพิโซดเพื่อรับประกันว่าอยู่ในช่วงความเชื่อมั่น 95% ซึ่งหมายความว่าใน 1 เอพพิโซด จะมีการทดสอบวนซ้ำ 400 ครั้ง เพื่อลดผลกระทบจากการสุ่มข้อมูลของการเกิดเหตุการณ์

บนพื้นฐานการนิยามข้างต้นที่ได้กล่าวมา ผลการทดลองที่ได้ดังรูปที่ 3.5 ทำการเปรียบเทียบวิธีการตัดสินใจเลือกเส้นทางแบบเอกรูป (uniform) วิธีการเลือกทางแบบละโมบ (การเลือกเส้นทางจะพิจารณาจากพลังงานรวมที่เหลือสูงที่สุดในระบบเท่านั้น) กับวิธีการมอนติคาร์โลที่มีการพิจารณาสามฟังก์ชันวัตถุประสงค์ จะเห็นถึงประสิทธิภาพของการตัดสินใจเลือกเส้นทางที่สามารถรักษาพลังงานที่เหลืออยู่โดยรวมของเซนเซอร์โนดสูงกว่าการเลือกแบบเอกรูปและการเลือกแบบละโมบ นอกจากนั้นเพื่อรับประกันช่วงความเชื่อมั่นของระบบว่าระบบเข้าสู่สภาวะเสถียรแล้ว จึงทำการทดสอบเพื่อให้ได้มาซึ่งความเชื่อมั่น 95% ของการพิจารณาพลังงานที่เหลืออยู่โดยเฉลี่ยของรอบการวนซ้ำเปรียบเทียบกับสองวิธีการที่ได้กล่าวมาดังรูปที่ 3.6 เพื่อยืนยันการทำงานว่าวิธีการมอนติคาร์โลที่นำเสนอนั้น สามารถนำไปใช้งานได้

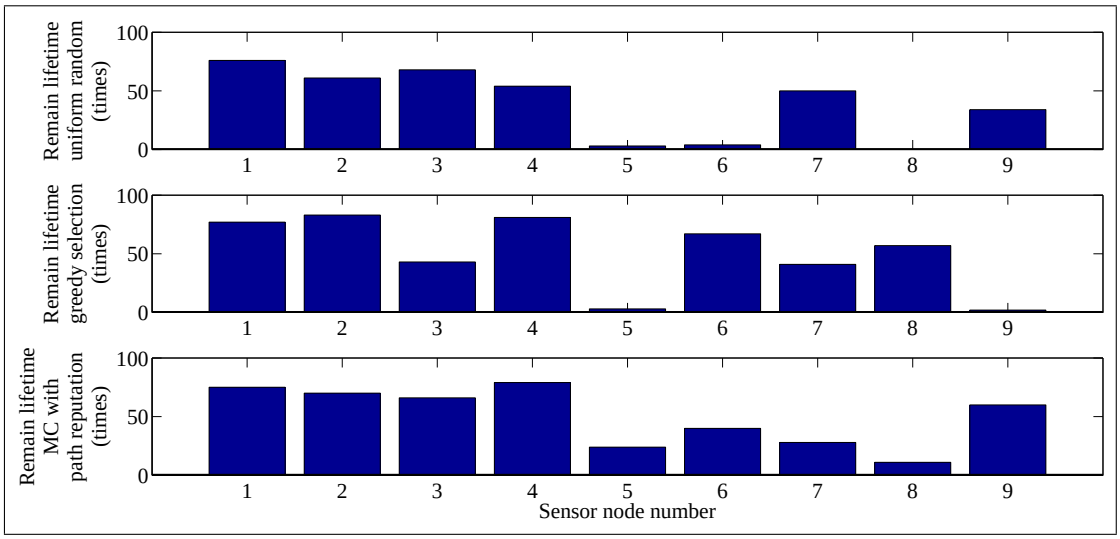


รูปที่ 3.5: พลังงานโดยรวมที่เหลืออยู่ของเซนเซอร์โนดในแต่ละเหตุการณ์



รูปที่ 3.6: การพิจารณาช่วงความเชื่อมั่น 95%

โดยรูปที่ 3.5 นั้นจะพิจารณาตามลำดับของเหตุการณ์ที่เกิดการเตือนของเซนเซอร์โนดในระบบ จะเห็นได้ว่าพลังงานจะค่อยๆลดลงเรื่อยๆ ตามเหตุการณ์ที่เพิ่มขึ้น แต่มอนติคาร์โลตจะลดช้าที่สุดเมื่อเทียบกับวิธีอื่นๆ โดยหากพิจารณา ณ จุดที่ เป็นค่าแบ่งขีด (threshold value) พบว่าวิธีการอื่นนั้นระดับพลังงานไม่สามารถที่จะส่งต่อได้ หรืออีกนัยหนึ่งคือระบบขาดการเชื่อมต่อเนื่องจากโนดบางโนดมีปัญหา และ รูปที่ 3.6 นั้นจะเป็นการนำพลังงานทั้งหมดมาเฉลี่ย ทำการทดสอบในกรณีที่แตกต่างกันทั้งหมด 18 เอพิโซด จะพบว่าวิธีการแบบมอนติคาร์โลจะสามารถรักษาระดับให้พลังงานที่เหลืออยู่ในระบบนั้นนานกว่าวิธีการเลือกแบบละโมบ 15.68% และสูงกว่าวิธีแบบเอกรูป 18%



รูปที่ 3.7: แสดงอายุการใช้งานที่เหลืออยู่ของแต่ละเซนเซอร์โนด

การทดลองสุดท้ายในหัวข้อนี้คือการวัดค่าเฉลี่ยอายุการใช้งานของแต่ละเซนเซอร์โนดที่พิจารณา

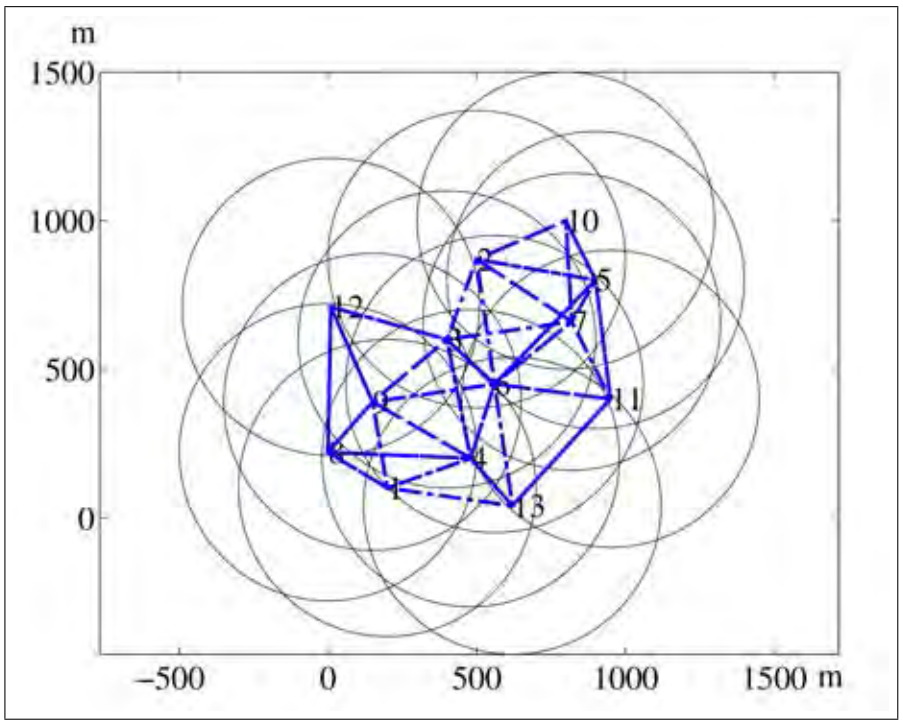
จากช่วงความเชื่อมั่น 95% ดังรูปที่ 3.7 โดยผลที่ได้แสดงถึงอายุการใช้งานเฉลี่ยของวิธีการที่ได้นำเสนอสามารถมีอายุการใช้งานที่ยาวนานกว่า 10% เมื่อเทียบวิธีแบบเอกรูป และ แบบละโมบ ซึ่งทั้งการเลือกแบบเอกรูปและการเลือกแบบละโมบไม่สามารถรับประกันด้านอายุการใช้งานในระยะยาวของระบบ ดังจะเห็นได้จากอายุการใช้งานของโนดที่ 5 6 8 จากวิธีเอกรูป หรือ โนด 5 9 จากวิธีแบบละโมบ นั้น ค่าของอายุการใช้งานต่ำมากจนไม่สามารถใช้งานได้แต่หากพิจารณาในวิธีของมอนติคาร์โลแล้ว จะพบว่าอายุการใช้งานของโนด ยังมีการเฉลี่ยออกไปยังเส้นทางอื่น ซึ่งจะทำให้ภาพรวมของระบบในด้านของอายุการใช้งานนั้นดีกว่าวิธีอื่น

3.4.2 การจัดสรรเส้นทางแบบรีแอคทีฟบนโครงข่ายขนาดเล็กที่มีเส้นทางจำกัด [24]

จากในหัวข้อย่อยที่ 3.4.1 ได้ทำการทดสอบวิธีการมอนติคาร์โลอย่างง่ายด้วยการพยายามยืดอายุการใช้งานของเซนเซอร์โนดและรวมไปถึงการลดการใช้พลังงานของเซนเซอร์โนดด้วย และพบว่าวิธีการที่นำเสนอสามารถใช้งานได้และทำงานได้ดีกว่าสองวิธีที่เป็นการเลือกแบบสุ่มอย่างคงตัว (uniform random) และรวมถึงการเลือกแบบละโมบที่มีการพิจารณาจากพลังงานที่เหลืออย่างเดียวซึ่งทั้งสองวิธีที่กล่าวมานั้นจะไม่สามารถควบคุมในส่วนของอายุการใช้งานได้

สำหรับในส่วนของหัวข้อนี้จะทำการทดสอบวิธีการมอนติคาร์โลอีกครั้ง ในรูปแบบโครงข่ายที่มีรูปร่างของทอพอโลยีซึ่งแตกต่างไปจากในหัวข้อย่อยที่ 3.4.1 โดยการทดสอบนี้เลือกใช้ทอพอโลยีใหม่เพื่อใช้ทดสอบระบบในสภาวะของระบบที่แตกต่างกันของทอพอโลยีและพิจารณาแนวโน้มการนำไปใช้จริงที่จะเกิดขึ้น ถึงแม้ว่าการออกแบบการทดสอบโดยใช้ทอพอโลยีคงตัวในรูปแบบเช่นเดียวกับในหัวข้อย่อยที่ 3.4.1 นั้นสามารถกระทำได้ในหัวข้อนี้ แต่ในหัวข้อนี้มีวัตถุประสงค์เพื่อทดสอบความสามารถในการทำงานของมอนติคาร์โล ในระบบที่แตกต่างกันออกไป เพื่อนำไปใช้ในบทที่ 4 จะเป็นการทดลองที่มีข้อมูลจริงมากขึ้น และ ซับซ้อนมากขึ้นตามลำดับ

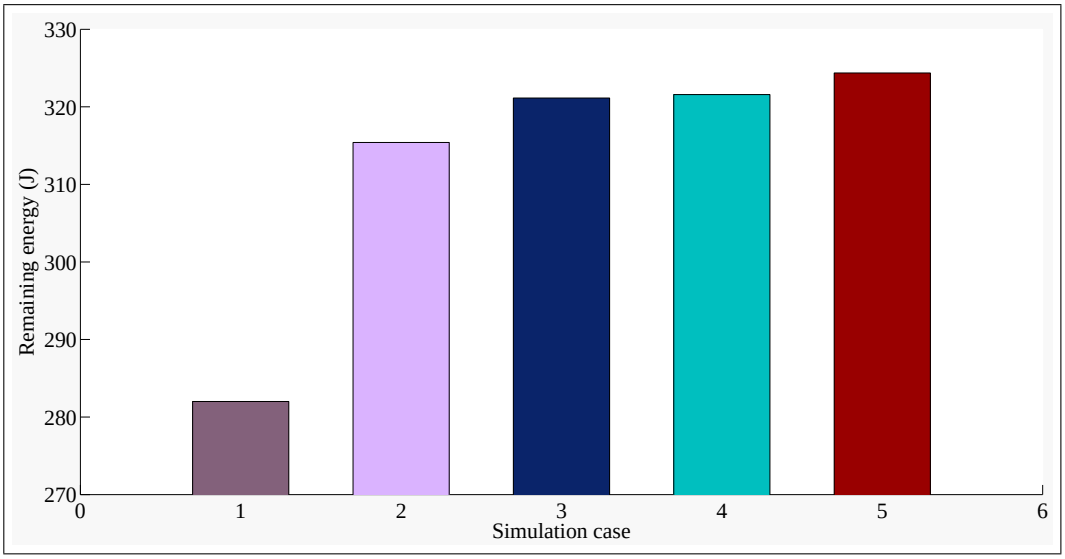
ในการทดลองนี้จะพิจารณาระบบที่ประกอบไปด้วย 13 โนดที่มีระยะการส่งเท่ากับ 500 เมตร และแต่ละโนดมีการวางตัวอย่างสุ่มและกระจายตัวตลอดในช่วงระยะ 1 ตารางกิโลเมตร ดังแสดงในรูปที่ 3.8 โดยกำหนดให้การส่งข้อมูลในแต่ละคู่โนดมีค่าเท่ากับ 100 แพ็กเกต(ค่าจริงของการส่งข้อมูลจะถูกใช้ในบทที่ 4) และพลังงานในการส่งข้อมูลเริ่มต้นที่ 100 จูล เท่ากันโดยระดับพลังงานจะลดลงตามสมการที่ 3.3 สำหรับแอคเตอร์โนดนั้นนิยามให้เป็นโนดที่ใช้ในการรับข้อมูลและสามารถชาร์จพลังงานเข้าไปใหม่ได้ในกรณีที่เกิดพลังงานต่ำกว่าระดับพลังงานที่ตั้งไว้เพื่อที่จะรักษาความมีเสถียรภาพของระบบให้ได้นานที่สุด เนื่องจากแอคเตอร์โนดนั้นมีหน้าที่รวบรวมข้อมูลที่รายงานมาจากเซนเซอร์โนดไม่ว่าการรายงานนั้นจะมาจากวิธีการใดก็ตาม การจำลองสถานการณ์จะเริ่มต้นจากเซนเซอร์โนด i^* มีความต้องการที่จะส่งข้อมูลไปยังแอคเตอร์โนด j ใดๆ ตัวเซนเซอร์โนดจะสร้างเส้นทางที่เกิดขึ้นด้วยกระบวนการรีแอคทีฟ ตัวอย่างเช่นวิธีการแบบ AODV โดยที่งานวิจัยนี้กำหนดให้การทำงานในชั้น MAC สามารถทำงานได้อย่างสมบูรณ์ดังนั้น การสร้างเส้นทางจะถูกสร้างขึ้น และหน้าที่ของตัวตัดสินใจ (agent) นั้นจะทำการเลือกเส้นทางที่ดีที่สุดผ่านค่า ϵ-กรีดี้ ที่ในงานวิจัยนี้เลือกจากการทำการทดสอบแบบลองผิดลองถูก (trial-and-error) และเลือกใช้ค่าเท่ากับ 0.8 โดยกำหนดให้ มีเหตุการณ์ทั้งหมด 200 ครั้ง เกิดการทำซ้ำ 100 เอพิโซด และมีการวนซ้ำ 20 ครั้งเพื่อพิจารณาที่ความเชื่อมั่น 95% โดยค่าต่ำสุดของพลังงานและอายุการใช้งานของเซนเซอร์โนดที่ยอมรับได้เท่ากับ 50%



รูปที่ 3.8: ทอพอโลยีขนาดเล็กที่ใช้ในการทดสอบด้วยการเชื่อมต่อแบบรีแอคทีฟ

ในการทดลองอย่างง่ายนี้ได้มีการกำหนดค่าของตัวแปรสถานะ (state variable) ของวิธีการมอนติคาร์โลไว้โดยใช้เป็นเพียงการแจ้งเตือนของเซนเซอร์โนดเท่านั้น เพียงเพื่อใช้ในการประเมินความเป็นไปได้ในการนำวิธีการมอนติคาร์โลมาใช้ต่อในบทถัดไป การทดลองจะมีการแบ่งการพิจารณาสถานการณ์ออกเป็น 5 กรณี และจำลองระบบผ่านโปรแกรม MATLAB® ด้วยการใช้คอมพิวเตอร์แบบพกพารุ่น Intel® core(TM) i7-2360QM มีหน่วยประมวลผล 2.0 GHz และมีหน่วยความจำ 4 GB โดยมีเงื่อนไขที่พิจารณาคือพลังงานต่ำกว่า 50% จากพลังงานเริ่มต้นหรืออายุการใช้งานลดลงจากเริ่มต้น 50% อย่างใดอย่างหนึ่ง โดยกำหนดให้ค่าของค่าเร็พพิวเทชันของแต่ละโนดมีค่าอยู่ในช่วง $0.5 - 1$ โดยในการทดสอบนี้จะวางค่าของเร็พพิวเทชันให้อยู่ในค่าที่จำกัดเนื่องจากต้องการจำกัดปัจจัยเงื่อนไขที่เป็นผลกระทบอื่นๆ และการทดสอบที่เหลือจะถูกพิจารณาในบทที่ 4 สำหรับสถานการณ์ที่พิจารณามีดังต่อไปนี้

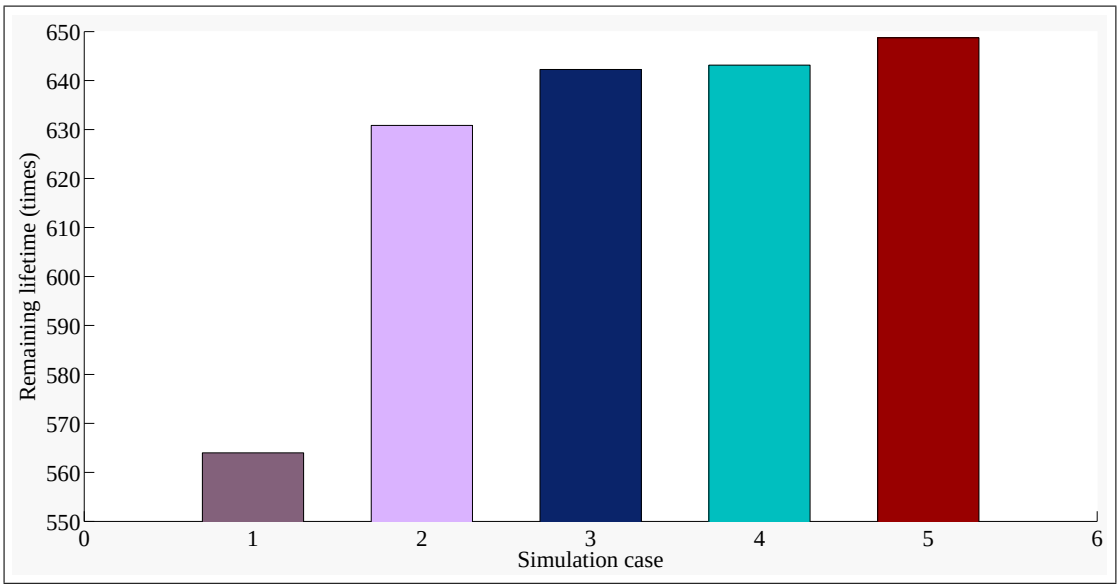
1. การจัดสรรเส้นทางรีแอคทีฟทั่วไปโดยไม่มีการพิจารณาเงื่อนไขใดๆ
2. การจัดสรรเส้นทางที่มีการตัดสินใจเลือกเส้นทางที่คำนึงถึงพลังงานที่เหลืออยู่มากที่สุด
3. การจัดสรรเส้นทางที่มีการตัดสินใจเลือกเส้นทางที่คำนึงถึงพลังงานที่เหลืออยู่มากที่สุดและอายุการใช้งานเหลืออยู่มากที่สุด
4. การจัดสรรเส้นทางที่มีการตัดสินใจเลือกเส้นทางที่คำนึงถึงพลังงานที่เหลืออยู่มากที่สุด อายุการใช้งานเหลืออยู่มากที่สุดและค่าเร็พพิวเทชัน
5. การจัดสรรเส้นทางที่มีการตัดสินใจเลือกเส้นทางที่คำนึงถึงพลังงานที่เหลืออยู่มากที่สุด อายุการใช้งานเหลืออยู่มากที่สุดและค่าเร็พพิวเทชันของเส้นทาง



รูปที่ 3.9: แสดงพลังงานที่เหลืออยู่ของแต่ละการจัดสรรเส้นทาง 5 กรณี

ตารางที่ 3.1: เปรียบเทียบพลังงานที่เหลืออยู่เฉลี่ยของระบบของการจัดสรรเส้นทาง 5 กรณี

	Type 1	Type 2	Type 3	Type 4	Type 5
Remaining energy (J)	282.00	315.41	321.13	321.58	324.37
% improvement		11.85	13.88	14.04	15.03



รูปที่ 3.10: แสดงอายุการใช้งานที่เหลืออยู่ของแต่ละการจัดสรรเส้นทาง 5 กรณี

ตารางที่ 3.2: เปรียบเทียบอายุการใช้งานที่เหลืออยู่เฉลี่ยของระบบของการจัดสรรเส้นทาง 5 กรณี

	Type 1	Type 2	Type 3	Type 4	Type 5
Remaining lifetime (times)	564.00	630.83	642.27	643.16	648.74
% improvement		11.85	13.63	14.04	15.03

จากตารางที่ 3.1 และตารางที่ 3.2 แสดงค่าเฉลี่ยของพลังงานที่เหลืออยู่และค่าเฉลี่ยอายุการใช้งานที่เหลืออยู่ของ 5 กรณีที่ได้กล่าวมาข้างต้น เมื่อพิจารณาจาก 3.1 ในกรณีที่ 1 คือการจัดสรรเส้นทางแบบรีแอคทีฟนั้นเมื่อไม่มีการพิจารณาเงื่อนไขใดๆ จะไม่สามารถยืดอายุการใช้งานโดยรวมของระบบได้เนื่องจาก การจัดสรรเส้นทางและการเลือกนั้นเป็นกระบวนการเลือกแบบสุ่ม (random) โดยไม่มีเงื่อนไข จึงทำให้สภาวะโดยรวมของระบบนั้นแย่ที่สุดหากเทียบกับกรณีอื่น การพิจารณาระบบนี้สามารถแยกพิจารณาแยกทั้งในด้านพลังงานและอายุการใช้งานแบบแยกตามตัวของเซนเซอร์โนดได้ ซึ่งจะพบว่ามีคุณลักษณะเช่นเดียวกับ การพิจารณาค่าของอายุการใช้งานในรูปที่ 3.7 สำหรับในกรณีที่ 2 จะเป็นการวางเงื่อนไขเพิ่มขึ้นโดยการเลือกเส้นทางหลังจากที่ระบบได้สร้างเส้นทางแล้วโดยเงื่อนไขที่เพิ่มขึ้นคือการพิจารณาเส้นทางที่มีโนดที่เหลือพลังงานรวมในระบบสูงที่สุด ซึ่งจะดีกว่าวิธีแรกสุดถึง 11.85% และในกรณีที่ 3 เป็นการวางเงื่อนไขเพิ่มขึ้นโดยพิจารณานอกจากระดับพลังงานในเส้นทางแล้วยังจะมีการพิจารณาอายุการใช้งานที่เหลืออยู่อีกด้วย ซึ่งก็จะเป็นการเพิ่มความซับซ้อนขึ้นมาอีก 1 ระดับ โดยทั้งสามกรณีที่ได้นำเสนอนั้น ไม่ได้นำกระบวนการแบบมอนติคาร์โลมาพิจารณา และสำหรับในกรณีที่ 4 และ กรณีที่ 5 จะเป็นกรณีที่นำกระบวนการแบบมอนติคาร์โลมาพิจารณาโดยเพิ่มเงื่อนไขของค่าเร็พพิวเทชันและค่าเร็พพิวเทชันของเส้นทางเข้าไปพิจารณาด้วย โดยหากเทียบกระบวนการอนติคาร์โล ทั้งใน กรณีที่ 4 และกรณีที่ 5 ความแตกต่างอย่างมีนัยสำคัญนั้น ไม่สามารถนำมาแสดงได้อย่างชัดเจนเนื่องจากระบบที่พิจารณานั้นมีความคงที่ จึงทำให้ความแตกต่างของค่าเร็พพิวเทชัน และค่าเร็พพิวเทชันของเส้นทางนั้นมีความแตกต่างกันน้อยมาก ทั้งนี้เหตุปัจจัยอีกประการหนึ่งคือการตั้งค่าของเร็พพิวเทชันให้อยู่ในช่วงจำกัดอีกด้วย

จะเห็นได้ว่าการจัดสรรเส้นทาง ในกรณีที่ 5 ที่มีการตัดสินใจเลือกเส้นทางที่คำนึงถึงพลังงานที่เหลืออยู่มากที่สุด อายุการใช้งานที่เหลืออยู่มากที่สุดและค่าเร็พพิวเทชันของเส้นทางที่นำเสนอเมื่อเปรียบเทียบกับการจัดสรรเส้นทางแบบอื่นๆ จะเห็นว่าการใช้ค่าสูงสุดของผลรางวัลที่ได้จะได้มาจากค่า ϵ-กรีดี้ ซึ่งค่าดังกล่าวจะเลือกการกระทำที่ดีที่สุดจากค่า ϵ ดังนั้นถ้าพลังงานที่เหลืออยู่ของเส้นทางใดๆต่ำกว่าค่าที่กำหนดไว้ จะมีการเลือกเส้นทางอื่นแทนที่โดยเลือกจากค่าพลังงานที่เหลืออยู่ที่มีค่าสูงกว่า แต่อย่างไรก็ตามหากเกิดเหตุการณ์ที่โนดบางโนดมีค่าเร็พพิวเทชันที่มีปัญหาการใช้งานแบบกรณีที่ 4 จะไม่สามารถทำงานได้อย่างมีประสิทธิภาพเนื่องจากไม่สามารถเห็นข้อมูลของโนดตลอดเส้นทางได้ ดังนั้นการนำกรณีที่ 5 ไปใช้งานจะเหมาะสมมากกว่าในสภาวะที่ระบบมีการเปลี่ยนแปลงสูงมากกว่าระบบคงที่และเมื่อพิจารณาจากตารางที่ 3.2 จะเห็นได้ว่าวิธีการที่นำเสนอยังคงรักษาอายุการใช้งานได้ดีเมื่อเทียบกับวิธีการอื่นๆ

3.4.3 การคำนวณความซับซ้อนของระบบ

ในหัวข้อย่อยนี้จะนำเสนอการวิเคราะห์การคำนวณความซับซ้อนของระบบโดยมีการเปรียบเทียบกับเทคนิคการหาคำตอบ 5 กรณี การคำนวณความซับซ้อนดังกล่าวสามารถแบ่งออกเป็น 2 ประเภท

คือการพิจารณาความซับซ้อนทางเวลาหรือการพิจาณาความซับซ้อนเชิงขนาด จากผลการประเมินความซับซ้อนเชิงเวลาพบว่าการคำนวณทั้ง 5 วิธีนั้นไม่มีความแตกต่างนัยสำคัญเชิงเวลาที่อยู่ในระดับมิลลิวินาที ซึ่งไม่ส่งผลกระทบต่อการคำนวณในการใช้งานจริง แต่ในบทนี้พิจารณาระบบที่มีขนาดเล็กเมื่อระบบถูกเพิ่มขนาดการพิจารณาขึ้นในบทถัดไป ความซับซ้อนของระบบจึงต้องถูกนำมาพิจารณาอีกครั้ง ดังนั้นเมื่อพิจารณาความซับซ้อนเชิงขนาดดังตารางที่ 3.3 จะเห็นได้ว่าขนาดของความซับซ้อนของการหาผลคำตอบสุดท้ายที่ได้จะมีขนาดที่ใหญ่ขึ้นเรื่อยๆตามขนาดของระบบ

ตารางที่ 3.3: คำนวณความซับซ้อนของการจัดสรรเส้นทาง 5 กรณี

Protocol	Computational Complexity (Space)
Type 1	$O(1)$
Type 2	$O(a)$
Type 3	$O(a)$
Type 4	$O(\|\mathcal{S}\| \times \|A\|)$
Type 5	$O(\|\mathcal{S}\| \times \|A\|)$

ดังแสดงในตารางที่ 3.3 จะเห็นว่าความซับซ้อนของระบบเมื่อพิจารณากรณีที่ 1 2 และ 3 จะมีปริมาณที่ใช้ในการเก็บค่าเท่ากันและมีค่าต่ำที่สุดเพราะว่าทั้ง 3 วิธีนี้เป็นวิธีการเลือกเส้นทางโดยไม่ต้องใช้ความฉลาดในการปรับปรุงตนในการเลือกเส้นทางและดังนั้นความซับซ้อนของระบบจะมีค่าเท่ากับ $O(a)$ โดย a แทนด้วยค่าคงที่โดยความซับซ้อนของระบบที่ 2 และ 3 นั้นมีค่าเท่ากับ

$$\begin{aligned} O(a) + O(a) &= 2 \times O(a) \\ 2 \times O(a) &= O(a) \end{aligned}$$

โดยตัวอย่างการคำนวณที่แสดงข้างต้นคือการพิจารณาการค้นหาในกรณีที่ 3 ที่มีการทำการค้นหาคำตอบ 2 ครั้งโดยมีการวางเงื่อนไข แต่อย่างไรก็ตาม ในท้ายที่สุด ค่าของการกระทำซ้ำเชิงเส้นในท้ายที่สุดค่าของ $O(a)$ ที่มีอิทธิพลเท่ากันจะถูกรวมเข้าเป็นฟังก์ชันเดียวกัน ตามทฤษฎีความซับซ้อนของระบบ (computational complexity theory) อย่างไรก็ตามการเพิ่มการพิจารณาให้ระบบมีความสามารถในการปรับปรุงตนดังกรณีที่ 3 และ 4 การพิจารณาความซับซ้อนจะมีค่าเพิ่มขึ้นตามขนาดของปริภูมิสถานะ $\mathcal{S}$ และขนาดของปริภูมิการกระทำ $\mathcal{A}$ ดังนั้นจะเห็นได้อย่างชัดเจนถึงประมาณการคำนวณที่แตกต่างกันในการใช้ระบบที่มีความฉลาดและความสามารถในการปรับปรุงตนที่เพิ่มขึ้นเป็นแบบไม่เชิงเส้น (non-linear increasing) ซึ่งจำเป็นที่จะต้องมีการคำนวณเพิ่มขึ้นมากเมื่อเทียบกับ 3 วิธีแรกที่นำเสนอวิธีแบบไม่มีความฉลาด ดังนั้นหากโครงข่ายมีขนาดที่เพิ่มมากขึ้นหรือใหญ่ขึ้นไม่ว่าจะเป็นปริภูมิสถานะหรือปริภูมิการกระทำ การทำการแบ่งนับหรือควอนไตเซชัน (Quantization) จะเป็นแนวทางหนึ่งที่สามารถนำเข้ามาประยุกต์ใช้งานเพื่อแก้ปัญหาของการเพิ่มขึ้นอย่างมหาศาลของปริภูมิสถานะ (state explosion) ของระบบซึ่งวิธีการทำการแบ่งนับเป็นวิธีที่นิยมใช้หากระบบที่พิจารณานั้นมีขนาดการเพิ่มขึ้นอย่างมหาศาลของปริภูมิสถานะหรืออีกนัยหนึ่งคือความซับซ้อนและความล่าช้าทางการคำนวณเพิ่มขึ้นอย่างมีนัยสำคัญและส่งผลกระทบรุนแรงต่อการหาคำตอบที่ดีที่สุดของระบบ

3.5 สรุป

จากจุดเริ่มต้นของแนวทางการพิจารณาการประยุกต์ใช้วิธีมอนติคาร์โลในสถานการณ์การเตือนอุทกภัยนั้น ได้มีการนำเสนอกรอบความคิดเชิงคณิตศาสตร์อย่างง่ายเพื่อใช้ในการพิจารณาถึงความเป็นไปได้ของการนำวิธีการมอนติคาร์โลนี้ เพื่อไปใช้ในสถานการณ์การเตือนอุทกภัย โดยมีการแบ่งช่วงของการทดสอบออกเป็นสองชุดได้แก่ การใช้วิธีการจัดสรรเส้นทางแบบโปรแอคทีฟและรีแอคทีฟตามลำดับ

ทั้งสองชุดการทดลองนั้นมีวัตถุประสงค์เพื่อใช้ในการประเมินเบื้องต้นของการนำเอาวิธีการมอนติคาร์โลไปใช้ ดังนั้นการเปรียบเทียบในส่วนของสมรรถนะเชิงระบบของทั้งสองวิธีนั้นจะไม่ถูกนำมาพิจารณาในบทนี้ และยิ่งไปกว่านั้นทั้งสองระบบของการทดสอบได้อาศัยทอพอโลยีที่มีรูปร่างลักษณะแตกต่างกันอีกด้วยและสำหรับในผลการทดลองของทั้งสองวิธีก็ได้พิจารณาให้เห็นถึงการวิเคราะห์ที่แตกต่างกัน ประกอบด้วยการพิจารณาความสามารถในการควบคุมระดับพลังงานรวมของระบบ การหาค่าของช่วงความเชื่อมั่น (confidence interval) เพื่อดูภาพเฉลี่ยของระดับพลังงานรวมของทุกโนดในระบบ การแจกแจงค่าของอายุการใช้งานที่เหลือของแต่ละโนดในระบบ การเปรียบเทียบประสิทธิผลเชิงสัมพัทธ์แบบร้อยละ (relative percentage of performance comparison) ของระดับพลังงานและอายุการใช้งาน ในท้ายที่สุดได้นำเสนอการวิเคราะห์เรื่องของความซับซ้อนของระบบ

ข้อสรุปที่ได้จากทอพอโลยีและกระบวนการเลือกเส้นทางที่แตกต่างกันนั้น พบว่าการเลือกใช้กรอบความคิดเชิงคณิตศาสตร์อย่างง่ายของวิธีการมอนติคาร์โลที่ได้แสดงไว้ในบทนี้นั้น สามารถทำได้ดีแต่หากจะนำวิธีการมอนติคาร์โลไปใช้ในทางปฏิบัติแล้วจำเป็นที่จะต้องปรับปรุงค่าของปริภูมิสถานะ ตัวแปรสถานะ รวมถึงฟังก์ชันผลรางวัล โดยการปรับปรุงปริภูมิสถานะนี้จะเพิ่มข้อมูลให้กับค่าของแอคชั่น-แวลู่ฟังก์ชัน เพื่อทำให้ความสามารถในการตัดสินใจของวิธีการมอนติคาร์โลนั้นดีขึ้น และนอกจากนั้นในส่วนของการใช้ฟังก์ชันผลรางวัลในบทนี้ ถึงแม้ว่าจะสามารถควบคุมพารามิเตอร์ได้ถึงสามตัว ทั้งการพยายามทำให้ระบบประหยัดพลังงาน ยืดอายุการใช้งานของเซนเซอร์โนดและการเลือกเส้นทางจากค่าเร็พพิวเทชันของเส้นทาง แต่ในบทนี้ ทั้งสามพารามิเตอร์ได้ถูกนำมาผูกกับฟังก์ชันผลรางวัลด้วยการแปรเปลี่ยนเป็นเงื่อนไขบังคับ (constraint) เพื่อกำหนดเงื่อนไขในการเลือกเส้นทาง ดังนั้นเพื่อให้ระบบทำงานได้อย่างเต็มที่ด้วยขีดความสามารถของวิธีการที่นำเสนอเองนั้น ในบทถัดไปจะมีการเปลี่ยนแปลงฟังก์ชันผลรางวัลโดยการนำเงื่อนไขบังคับปัจจุบันสร้างเป็นฟังก์ชันผลรางวัลใหม่ดังแสดงในบทถัดไป

บทที่ 4

การประยุกต์ใช้งานวิธีการมอนติคาร์โลด้วยการพิจารณาค่าเร็พพิวเทชันของเส้นทางในตำแหน่งไม่คงที่

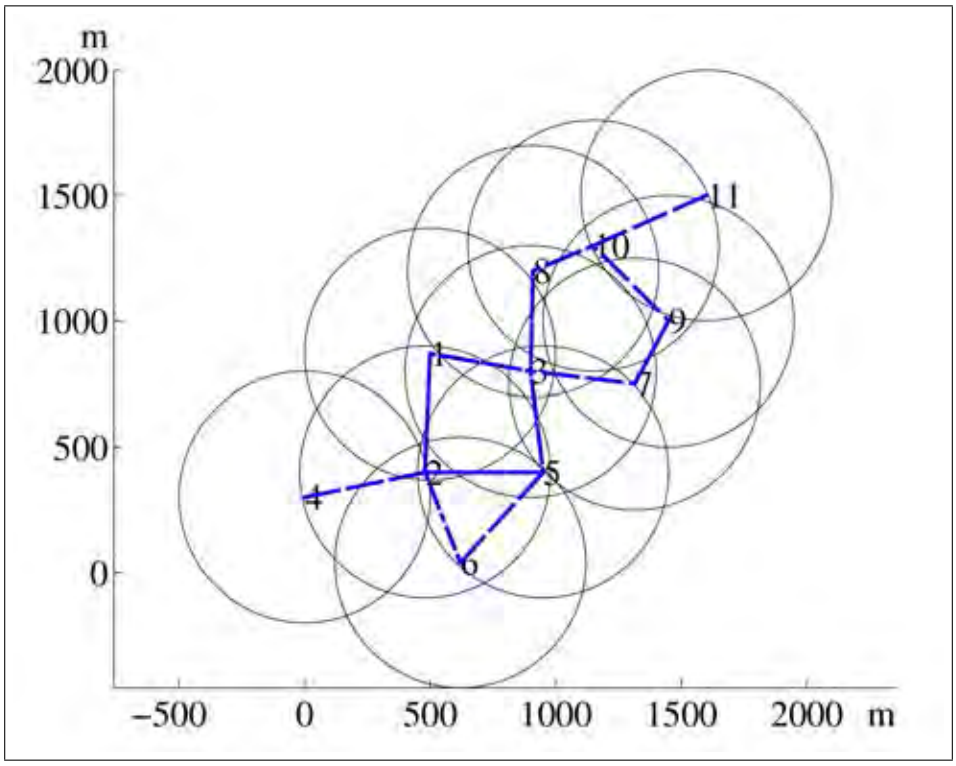
จากงานวิจัยบทที่ 3 พบว่าการประยุกต์ใช้งานวิธีมอนติคาร์โลและเร็พพิวเทชันของเส้นทางนั้นสามารถใช้งานได้ในระบบขนาดเล็ก โดยพบว่าการสร้างเส้นทางไม่ว่าจะเกิดจากกระบวนการแบบโพรแอคทีฟหรือรีแอคทีฟนั้นไม่ส่งผลกระทบต่อการทำงานของวิธีการมอนติคาร์โล ทั้งนี้เนื่องจากการกำหนดให้ขนาดของปริภูมิสถานะถูกพิจารณาจากเส้นทางที่เป็นไปได้ จึงทำให้วิธีการมอนติคาร์โลนั้นพิจารณาผลที่เกิดหลังจากที่เส้นทางถูกสร้างขึ้น โดยความแตกต่างนี้จะแตกต่างจากกระบวนการวิธีมอนติคาร์โลกับเร็พพิวเทชันของโนด คือกระบวนการตัดสินใจด้วยการใช้เร็พพิวเทชันนั้นตัวโนดตรงกลาง (immediate node) จะเป็นตัวตัดสินใจเพื่อเลือกที่จะสร้างเส้นทางต่อไป ดังนั้นด้วยวิธีการนี้จะทำให้ไม่สามารถรับประกันปัญหาของโนดที่ไม่พร้อมที่จะส่งข้อมูลได้เนื่องจากมีโอกาสที่จะเกิดความผิดพลาดจากการตัดสินใจจากโนดตรงกลางตลอดเวลา สำหรับการวางโนดในระบบนั้นเป็นการวางโนดอย่างกระจายตัวและไม่นิยมใช้วิธีวางเซนเซอร์โนดแบบหนาแน่นในพื้นที่ที่พิจารณาเพื่อให้เกิดค่าใช้จ่ายในการลงทุนที่ต่ำ จึงทำให้ขนาดของปริภูมิการกระทำนั้นมีขนาดที่จำกัดและจากข้อจำกัดทางด้านความสามารถสูงสุดของการส่งข้อมูลที่จะทำให้ระบบนั้นอยู่ในสภาวะที่ดีจะต้องไม่เกิน 4 ฮอพ และด้วยเหตุผลทั้งหมดที่กล่าวมาแล้วข้างต้นจึงทำให้การประยุกต์ใช้งานมอนติคาร์โลในระบบโครงข่ายขนาดใหญ่สามารถนำมาพิจารณาต่อไป

ในบทนี้จะเน้นการศึกษากระบวนการทำงานเพื่อนำไปประยุกต์ใช้เพิ่มเติมจากบทที่ 3 บนพื้นฐานการเรียนรู้จากประสบการณ์ที่เคยกระทำซ้ำๆเป็นฉากหรือรอบการเรียนรู้โดยระบบจะพยายามปรับปรุงนโยบายในการตัดสินใจเพื่อให้ได้ผลรางวัลในระยะยาวที่สูงสุดซึ่งเรียกวิธีการเรียนรู้แบบมอนติคาร์โล บทที่ผ่านมามีการวัดประสิทธิภาพจากการจัดสรรเส้นทางแบบโปรแอคทีฟและรีแอคทีฟซึ่งทั้งสองโพรโทคอลที่นำเสนอภายใต้การเลือกเส้นทางแบบมอนติคาร์โลที่มีค่าเร็พพิวเทชันของเส้นทางเป็นเกณฑ์หนึ่งในการตัดสินใจในสภาวะแวดล้อมคงที่ ซึ่งเป็นสถานการณ์ที่ระบบเปลี่ยนแปลงตัวแบบเชิงกำหนดหรือเปลี่ยนแปลงภายใต้กฎเกณฑ์ที่แน่นอน (deterministic) แต่ในบทนี้การนำเสนอวิธีการมอนติคาร์โลที่มีค่าเร็พพิวเทชันของเส้นทางในสภาวะแวดล้อมไม่คงที่ ซึ่งหมายถึงสถานการณ์ที่ข้อมูลที่เกี่ยวข้องที่ใช้ในการจำลองแบบมีความไม่แน่นอน โดยรูปแบบทางคณิตศาสตร์ของปริภูมิสถานะเป็นตัวแปรหนึ่งที่มีการเปลี่ยนแปลงในบทนี้ การเปลี่ยนรูปแบบดังกล่าวเป็นสิ่งหนึ่งที่พิสูจน์ว่าวิธีการมอนติคาร์โลที่นำเสนอสามารถปรับปรุงระบบให้ดีขึ้นในขณะที่สิ่งแวดล้อมของระบบเกิดการเปลี่ยนแปลง

การตัดสินใจเลือกกระทำของบทที่ผ่านมา จะเลือกจากประสบการณ์ในอดีตโดยไม่มีการพิจารณาถึงตัวแปรของสถานะ เช่น พลังงานที่เหลือและค่าเร็พพิวเทชันของโนด เพื่อทำให้เห็นผลกระทบของวิธีการมอนติคาร์โลที่มีการพิจารณาค่าเร็พพิวเทชันของเส้นทางไม่สามารถทำงานตามที่คาดไว้ บทนี้จึงมีการกำหนดปริภูมิสถานะใหม่และนำเสนอวิธีการที่จะสามารถทำให้ระบบดำเนินการได้ดีในสถานการณ์ที่แตกต่างกัน เช่น สภาวะที่ทอพอโลยีไม่มีการเปลี่ยนแปลง สภาวะที่ทอพอโลยีเกิดการเปลี่ยนแปลงและ สภาวะที่มีการเปลี่ยนแปลงของสภาพแวดล้อมที่เกิดขึ้นจริง ซึ่งจะนำวิธีการเลือกเส้นทางที่ได้นำเสนอดังกล่าวเปรียบเทียบกับวิธีการที่มีอยู่เดิมและนำเสนอในบทนี้ ส่วนในบทสุดท้ายจะเป็นข้อเสนอแนะความเป็นไปได้ของการพัฒนาผลงาน

เนื้อหาของบทนี้ประกอบด้วย หัวข้อ 4.1 สมมติฐานของแบบจำลองโดยจะกล่าวถึงโครงรูปของโครงข่ายที่ใช้ในการสมมติฐานแบบจำลอง หัวข้อ 4.2 การนิยามปัญหาโดยจะกล่าวถึงการจำลองแบบในรูปของวิธีการออนโพลิซีมอนติคาร์โล การกำหนดพารามิเตอร์ที่ใช้ประกอบในการเรียนรู้ที่มีการปรับเปลี่ยนจากบทที่แล้ว เช่น สถานะของโครงข่าย การกระทำที่เกิดขึ้นในระบบ และการกำหนดและพิจารณาผลรางวัล หัวข้อ 4.3 การวิเคราะห์ผลการทดลองที่มีการแบ่งประเภทกันพิจารณาออกเป็นช่วงการทดลอง เช่น ช่วงกำลังเรียนรู้ ช่วงที่เรียนรู้แล้ว และช่วงที่เกิดการเปลี่ยนแปลงโครงข่าย หัวข้อสุดท้าย 4.4 จะเป็นหัวข้อสรุปของการทดลองการจำลองแบบของหัวข้อนี้

4.1 โครงข่ายที่พิจารณา

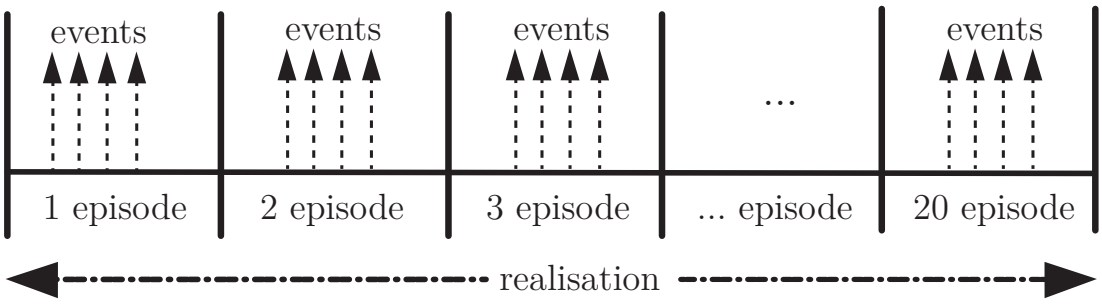


รูปที่ 4.1: ทอพอโลยีที่ใช้ในการจำลองแบบ

ในบทที่ 3 ทอพอโลยีประกอบด้วยเซนเซอร์โนดที่ไม่สามารถเคลื่อนที่ได้ มีโครงรูปที่คงที่เสมอ ในขณะที่สภาวะจริงถึงแม้เซนเซอร์โนดที่ลอยติดทุ่นบนผิวน้ำแต่โนดยังสามารถเคลื่อนที่ได้อยู่ตลอดเวลาภายในรัศมีจากตำแหน่งของโนด ฉะนั้นในบทนี้การประยุกต์ใช้เซนเซอร์ไร้สายในการแจ้งเตือนเหตุอุทกภัยจึงมีการพิจารณากรณีที่แต่ละโนดสามารถเคลื่อนที่ได้และยังสามารถสร้างข้อมูลที่ใช้ในการสื่อสารได้ด้วยตนเอง (self-generated traffic) ซึ่งคุณลักษณะเช่นนี้สามารถนำมาคำนวณค่าเร็พพิวเทชันสำหรับแต่ละเซนเซอร์โนดจากค่าจำนวนแพ็กเกตข้อมูลที่ส่งผ่านโนดนั้นๆสำเร็จในแต่ละเหตุการณ์ ดังรูปที่ 4.1 ซึ่งเป็นลักษณะทอพอโลยีที่ใช้ในการจำลองแบบในบทนี้ โดยมีการเชื่อมต่อของแต่ละโนดแสดงในลักษณะของเส้นทึบและเส้นประ

4.2 การนิยามปัญหาในรูปออนโพลิซีมอนติคาร์โล

หัวข้อนี้กล่าวถึงการนิยามปัญหาบนพื้นฐานของวิธีการออนโพลิซีมอนติคาร์โลเช่นเดียวกับหัวข้อ 3.3 แต่จะมุ่งเน้นถึงการนิยามปัญหาเพื่อนำมาใช้ในการจำลองแบบของบทนี้ กำหนดการส่งข้อมูลแต่ละครั้งจากโนดต้นทางไปยังโนดปลายทางเรียกว่าเหตุการณ์ (event) และการจำลองแบบที่มีการทำงานเป็นฉากหรือรอบเรียกว่าเอพพิโซด (episode) ซึ่งแต่ละเอพพิโซดจะมีการเกิดของเหตุการณ์ที่โนดต้นทางส่งข้อมูลไปยังโนดปลายทาง 1500 ครั้ง และจะมีการจำลองแบบรอบการทำงานซ้ำๆเรียกว่ารีอะไลเซเชิน (realisation) โดยแต่ละรีอะไลเซเชินจะมีการจำลองการทำงานในลักษณะเอพพิโซด 20 ครั้ง ดังรูปที่ 4.2



รูปที่ 4.2: นิยามรอบการทำงานของวิธีการมอนติคาร์โล

4.2.1 กำหนดนิยามสถานะของโครงข่าย (state)

สถานะของสิ่งแวดล้อมในบทนี้ถูกนิยามใหม่เพื่อให้ระบบมีข้อมูลที่เพียงพอสำหรับการตัดสินใจเลือกเส้นทาง โดยสมมุติให้โนดต้นทางและโนดปลายทางเชื่อมโยงกันด้วยชุดของเส้นทาง L ฉะนั้นจึงกำหนดสถานะของโครงข่ายในที่นี้คือสถานะพลังงานและเร็พพิวเทชันของโนด ซึ่งการค้นหาเส้นทางจะพิจารณาถึงความสมดุลระหว่างสองปัจจัยนี้ ดังนั้นปริภูมิสถานะกำหนดจากเวคเตอร์ของสถานะของสถานะเซนเซอร์โนดดังสมการ (4.1)

$$\mathcal{S} = \{\mathbf{s}(t) = [s_i(t)], \forall i\}, i = 1, 2, 3, ..., I$$
$$s_i(t) = (x_i(t); b_i(t)) \tag{4.1}$$

ซึ่ง $x_i(t)$ คือสถานะของพลังงานในเซนเซอร์โนดที่เวลา t และ $b_i(t)$ สถานะของค่าเร็พพิวเทชันของเซนเซอร์โนดที่เวลา t เนื่องจากค่าพลังงานที่สูญเสียและค่าเร็พพิวเทชันของโนดบนแต่ละเส้นทางมีค่าเป็นลักษณะแบบต่อเนื่อง ดังนั้นลักษณะเช่นนี้จึงส่งผลให้ลักษณะของปริภูมิสถานะมีค่าแบบต่อเนื่อง เพื่อลดปัญหาที่จะเกิดจากการเพิ่มขึ้นอย่างรวดเร็วของจำนวนสถานะ (state explosion) และเพื่อลดขนาดหรือเวลาที่ใช้ในการประมวลผลของการประมวลผลในแต่ละโนด งานวิจัยนี้จึงมีการแปลงค่าของพลังงานที่สูญเสียและค่าเร็พพิวเทชันให้มีรายละเอียดต่ำลง โดยทำการแบ่งนับ

การแบ่งนับของพลังงานที่สูญเสียของเซนเซอร์โนด ดังสมการ (4.2)

$$\bar{x}_i(t) = \left\lceil \frac{\alpha_i(t)}{\max \alpha_i(0)} q_\varepsilon \right\rceil \tag{4.2}$$

การแบ่งนับของค่าเร็พพิวเทชันของเซนเซอร์โนด ดังสมการ (4.3)

$$\bar{b}_i(t) = \left\lceil \frac{r_i(t)}{\max r_i(0)} q_r \right\rceil \tag{4.3}$$

ขนาดของปริภูมิสถานะ $|\mathcal{S}| = q_\varepsilon \times q_r$

การคำนวณพลังงานที่สูญเสีย $x_{i\in l}$ จะสมมุติให้แต่ละโนดรู้ตำแหน่งของโนดข้างเคียงจากระบบกำหนดตำแหน่งบนโลกหรือจีพีเอส ซึ่งแต่ละโนดจะเลือกโนดข้างเคียงจากระยะการส่งสูงสุดและความเสถียรของการเชื่อมโยง กระบวนการค้นหาเส้นทางจะเริ่มที่โนดต้นทางสร้างแพ็กเกตการร้องขอเส้นทาง (route request หรือ RREQ) และทำการแพร่กระจายแพ็กเกต RREQ ไปยังทุกๆโนดข้างเคียง ซึ่งโนดข้างเคียงนั้นๆจะทำการส่งแพ็จเกจดังกล่าวต่อไปซึ่งจะมีการลบค่าของจำนวนฮอพที่ละฮอพและเมื่อใดที่ฮอพมีค่าเท่ากับศูนย์ หมายความว่าแพ็กเกตดังกล่าวส่งถึงโนดปลายทางแล้วโนดต้นทางรอคอยการตอบกลับ (route reply หรือ RREP) เมื่อโนดปลายทางได้รับแพ็กเกต RREQ จะมีการคำนวณพลังงานที่สูญเสียและค่าเร็พพิวเทชันตลอดทั้งเส้นทาง หลังจากนั้นจะผนวกข้อมูลดังกล่าวกับแพ็กเกต RREP และส่งกลับไปยังโนดต้นทางผ่านเส้นทางเดิม ซึ่งตลอดเส้นทางจะมีการผนวกข้อมูลของระดับพลังงานที่เหลืออยู่ของแต่ละโนดเข้ากับแพ็กเกต RREP เมื่อแพ็กเกตดังกล่าวถึงโนดต้นทางทำให้โนดต้นทางทราบถึงปริมาณของพลังงานที่สูญเสียและค่าเร็พพิวเทชันในแต่ละเส้นทางที่มีการเชื่องโยงของโนดต้นทางและโนดปลายทาง หลังจากนั้นโนดต้นทางจะสามารถคำนวณการแบ่งนับของพลังงานที่สูญเสียและค่าเร็พพิวเทชันแล้วนำมาเป็นสถานะของระบบดังสมการ (4.1)

4.2.2 กำหนดนิยามการกระทำของโครงข่าย (action)

เมื่อได้ข้อมูลของสถานะที่ทำการแบ่งนับ โนดต้นทางจะทำการตัดสินใจเลือกเส้นทางจากเส้นทางทั้งหมดที่มี ซึ่งกำหนดการกระทำได้จากเส้นทางทั้งหมด

$$a(s_{i^*}(t)) = \mathbf{l}_{i^*}(t) \tag{4.4}$$

โดย $\mathbf{l}_{i^*}(t) = [l_{i^*,1}(t), l_{i^*,2}(t), l_{i^*,3}(t), ..., l_{i^*,n}(t)] = [l_{i^*,n}(t)]$

4.2.3 กำหนดผลรางวัลของโครงข่าย (reward)

จากกรอบมาตรฐานของวิธีการเรียนรู้แบบเสริมแรง (reinforcement learning) ดังแสดงในรูปที่ 2.1 จะพบว่า เริ่มจากการที่สถานะของระบบถูกการกระทำใดๆ ผ่านตัวการตัดสินใจ (decision maker) หรือตัวกระทำการตัดสินใจของระบบ (agent) ส่งผลให้เกิดการเปลี่ยนแปลงในระบบ ทำให้สถานะของระบบเกิดการเปลี่ยนแปลงไปจากเดิม โดยกรอบการเรียนรู้นี้ จะทำให้เกิดตัวแปรป้อนกลับไปยังตัวกระทำการตัดสินใจของระบบเพื่อบอกว่าผลการเลือกการกระทำในครั้งล่าสุดนั้น ดีหรือไม่ดี อย่างไร ผ่านฟังก์ชัน $f(s_i, a_i)$ โดยค่าของฟังก์ชันนี้ ขึ้นกับค่าของตัวแปรได้แก่ ผลรวมของ

พลังงานที่เหลืออยู่ของเซนเซอร์โนดในเส้นทางที่ถูกเลือก l พลังงาน($\bar{\varepsilon}$), อายุการใช้งาน $\bar{\ell}$ และค่าเร็พพิวเทชัน $\bar{R}$ เช่นเดียวกับในบทที่ 3 ค่าของฟังก์ชัน $f(s_i, a_i)$ จะถูกกำหนดโดย

$$f(s_{i^*}(t), a(s_{i^*}(t))) = w_{\mathcal{E}}\overline{\mathcal{E}}_l(t) + w_\ell \bar{\ell}_l(t) + w_R \bar{R}_l(t) \tag{4.5}$$

โดยที่ค่าของฟังก์ชันผลรางวัลนี้จะขึ้นกับค่าของน้ำหนักที่ให้ในแต่ละฟังก์ชันย่อยผ่านตัวแปร w_e w_l และ w_r โดยที่

$$\bar{\varepsilon}_l(t) \quad = \quad \frac{\varepsilon_l(t)}{\max \varepsilon_l(t)} q_\varepsilon = \frac{\sum\limits_{\forall i \in l} \alpha_{i,l}(t) \mid_{t \geq 1}}{\sum\limits_{\forall i \in l} \alpha_{i,l}(t) \mid_{t=0}} q_\varepsilon \tag{4.6}$$

$$\bar{\ell}_l(t) = \frac{\arg\max\limits_{l}(\min \ell_{i,l}(t))}{\ell_i(0)} q_\ell \tag{4.7}$$

$$\bar{R}_l(t) \quad = \quad \frac{R_l(t)}{\max R_l(0)} q_r \tag{4.8}$$

หลังจากนั้นจะมีการเฉลี่ยของค่าผลรางวัลโดยการใช้วิธีการเอพพิโซดิก- ท๊าสค์ ซึ่งเขียนแมททิกฟังก์ชันของผลรางวัลได้ว่า

$$Q(s_i(t), a(s_i(t))) = E[f(s_i(t), a(s_i(t))] \tag{4.9}$$

โดย $E[.]$ ค่าคาดหวังผลรางวัลเฉลี่ยในแต่ละสถานะ s_i และการกระทำ a_i

ความแตกต่างของค่าผลรางวัลในบทนี้กับบทที่ 3 จะแตกต่างกันอย่างสิ้นเชิง โดยในบทที่ 3 นั้น ค่าของผลรางวัลจะขึ้นกับเงื่อนไขบังคับ โดยเป็นค่าจำกัดของทั้ง 3 ฟังก์ชัน ซึ่งจะต้องผ่านเงื่อนไขที่วางเอาไว้ หรืออีกชื่อหนึ่งคือนโยบายการตัดสินใจที่ตั้งไว้ (hard policy decision) โดย จากการทดลองในบทที่ 3 นั้นได้ทำการทดสอบการวางค่าจำกัดในสามลักษณะ พบว่า การใช้นโยบายที่ตั้งไว้นั้น ทำให้การเรียนรู้ของระบบผ่าน Q ฟังก์ชันเป็นไปได้อย่างมีประสิทธิภาพน้อย ทั้งนี้ เนื่องจากตัวแปรสถานะในบทที่ 3 นั้น ไม่ได้พิจารณาในส่วนของระบบสามารถพัฒนาและปรับปรุงตนได้เข้าไปด้วย จึงทำให้ค่าของฟังก์ชัน $f(s_i, a_i) = \{0, 1\}$ เพื่อใช้ในการแจ้งป้อนกลับไปยังตัวการตัดสินใจ

สำหรับในบทที่ 4 นี้ การนำเอาทั้ง 3 ฟังก์ชันมาให้ตัวกระทำการตัดสินใจของระบบ เลือกข้อมูลที่มีรายละเอียดของทั้ง 3 เรื่องที่ต้องการให้เป็นเงื่อนไขบังคับ ทั้งในส่วนของพลังงาน อายุการใช้งาน และค่าเร็พพิวเทชันนั้น จะสามารถกระทำได้ด้วยกระบวนการแบ่งกระจายภาระงาน (load balancing) โดยแบ่งความสำคัญของทั้งสามฟังก์ชันอย่างเท่าเทียมกัน (ในหัวข้อผลการทดลองได้มีการพิจารณาในส่วนของการไม่พิจารณากระบวนการแบ่งกระจายภาระงานและการให้สัดส่วนที่เหมาะสมผ่านการให้ค่าน้ำหนักไว้แล้ว พบว่ากระบวนการแบ่งกระจายภาระงานแบบเท่าเทียมกันของทั้งสามฟังก์ชัน เป็นทางเลือกมีเหตุผลและการคำนวณไม่ซับซ้อนมากจนเกินไป ดังนั้นหมายความว่าในบทนี้ จะไม่มีการใช้ค่าจำกัดเพื่อช่วยวิธีการมอนติคาร์โลในการตัดสินใจ แต่จะให้วิธีนั้นตัดสินใจผ่าน Q ฟังก์ชันที่มีความรู้ผ่านตัวแปรสถานะเพิ่มขึ้นมากกว่าบทที่ 3

4.3 กระบวนการทำงานของวิธีมอนติคาร์โลที่มีการพิจารณาเร็พพิวเทชันของเส้นทาง

งานวิจัยนี้นำเสนอวิธีการจัดสรรเส้นทางที่คำนึงถึงพลังงานและเร็พพิวเทชันสำหรับโครงข่ายเซนเซอร์ไร้สาย ที่มีการประยุกต์วิธีการมอนติคาร์โลที่มีการจำลองแบบการทำงานในลักษณะเอพพิโซดเพื่อประมาณค่าแอคชั่น-แวลู่ฟังก์ชัน (หัวข้อ 2.3.1) ซึ่งจะเฉลี่ยค่าผลรางวัล จากนั้นนำผลรางวัลดังกล่าวมาเป็นตัวตัดสินใจเลือกกระทำสำหรับคู่สถานะและการกระทำ ณ เอพพิโซดถัดไป $(Q(s,a), \forall s \in \mathcal{S}, \forall a \in A)$ ซึ่งแอคชั่น-แวลู่ฟังก์ชันจะทำงานภายใต้นโยบาย $(\pi : \mathcal{S} \rightarrow A)$ ดังกระบวนการทำงานของวิธีมอนติคาร์โลดังตารางที่ 4.1

ตารางที่ 4.1: กระบวนการทำงานของวิธีมอนติคาร์โลที่มีการพิจารณาเร็พพิวเทชันของเส้นทาง

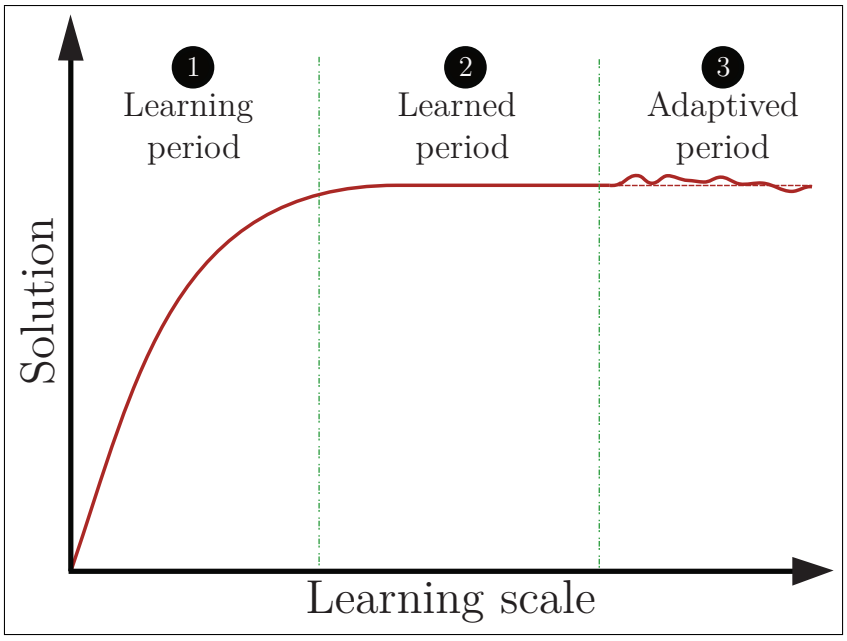
1. Initialisation, for all $s \in \mathcal{S}$, $a \in A(s)$:
2. $\quad Q(s,a) \leftarrow$ arbitrary
3. $\quad Returns(s,a) \leftarrow$ empty list
4. $\quad \pi \leftarrow$ an arbitrary $\epsilon-$soft policy
5. Randomly deployed sensor nodes in a segment of river e.g. 3 kilometres.
6. Creating the actor nodes as the existing water flooding report stations along the river segment.
7. For each alarming of sensor nodes t.
8. Using reactive protocol to construct all possible paths.
9. Calculate the quantized level of energy consumption and path reputation
10. Evaluate the remaining energy, network lifetime and path reputation.
11. Monte carlo with greedy method chooses the best possible path with high probability $1 - \epsilon$.
12. Evaluate the return value $\mathcal{R}$. The return value consists of three objective functions. Append $\mathcal{R}$ to $(Returns(s,a))$
13. Repeat step 8–11 until the end of maximum event t_{max}.
14. Update the Q value $Q(s,a)$ for next simulation.
    $Q(s,a) \leftarrow$ average$(Returns(s,a))$
15. Repeat step 7–13 until the end of simulation.

จากกระบวนการทำงานของวิธีมอนติคาร์โลที่มีการพิจารณาเร็พพิวเทชันของเส้นทาง เริ่มต้นกำหนดคู่สถานะและการกระทำของแต่เซนเซอร์โนด (หัวข้อ 4.2) กำหนดแอคชั่น- แวลู่ฟังก์ชันเริ่มต้นกำหนดค่าผลรางวัลเป็นศูนย์สำหรับแต่ละคู่สถานะและการกระทำและกำหนดนโยบายจากนโยบายละโมบ (บรรทัด 1-4) กำหนดทอพอโลยีตามสภาพแวดล้อมที่พิจารณา (บรรทัด 5-6) เมื่อโนดต้นทางต้องการส่งข้อมูลแจ้งเตือน t โนดต้นทางจะค้นหาเส้นทางไปยังโนดปลายทางในลักษณะรีแอคทีฟโพรโทคอล ซึ่งจะได้เส้นทางทั้งหมดที่เป็นไปได้ หลังจากนั้นจะมีการคำนวณการแบ่งนับของสถานะ

(บรรทัด 7-10) และมีการตัดสินใจเลือกเส้นทางจากเส้นทางที่ดีที่สุดจากนโยบายละโมบ (บรรทัด 11) ภายหลังการตัดสินใจเลือกเส้นทางระบบจะเก็บผลรางวัลที่มีการพิจารณาพลังงาน อายุการใช้งานและค่าเร็พพิวเทชันแล้ว (บรรทัด 12) ระบบจะกระทำเหตุการณ์ซ้ำเช่นนี้ 1500 เหตุการณ์ หลังจากนั้นระบบจึงนำค่าผลรางวัล ณ เวลานั้นมาคำนวณผลรางวัลสะสมที่คาดว่าจะได้รับ โดยเฉลี่ยผลตอบแทนทุกๆครั้ง ที่ระบบอยู่ในสถานะและการกระทำใดๆในแต่ละเอพพิโซด แล้วจึงนำข้อมูลดังกล่าวมาเป็นค่านโยบายในเหตุการณ์แรกของเอพพิโซดถัดไป (บรรทัด 13-15)

4.4 การวิเคราะห์ผลการจำลองแบบ

หัวข้อนี้นำเสนอการวิเคราะห์ผลการทดลองเพื่อใช้ในการทดสอบสมมติฐานการนำกรอบทฤษฎีทางคณิตศาสตร์ของวิธีการมอนติคาร์โลด้วยค่าเร็พพิวเทชันของเส้นทาง นำไปใช้ในสถานการณ์ที่แตกต่างกันโดยแบ่งลำดับของการทดลองทั้งหมดออกเป็น 3 ช่วงใหญ่คือช่วงที่ระบบกำลังเรียนรู้ ช่วงที่ระบบเรียนรู้แล้วและช่วงที่มีการปรับปรุงระบบ โดยการทดลองนี้จะถูกนำไปประยุกต์ใช้กับระบบเตือนอุทกภัยที่มีลักษณะของทอพอโลยีดังแสดงในรูป 4.1 โดยวัตถุประสงค์เริ่มแรกของงานวิจัยนี้มุ่งที่จะนำเสนอการพิจารณากรอบทฤษฎีทางคณิตศาสตร์ที่มีสามตัวแปรสถานะซึ่งก็คือพลังงานที่เหลืออยู่ อายุการใช้งานที่เหลืออยู่และค่าเร็พพิวเทชันของเส้นทาง แต่อย่างไรงานวิจัยนี้ได้เลือกพิจารณาใช้เพียงแค่พลังงานที่เหลืออยู่และค่าเร็พพิวเทชันมาเป็นตัวแปรปริภูมิสถานะเพื่อบรรเทาความซับซ้อนในการคำนวณที่มีการเพิ่มขึ้นในรูปแบบทวีคูณ (exponential growth) ซึ่งในท้ายที่สุดจะส่งผลถึงค่าพลังงานที่จะต้องใช้ที่เพิ่มขึ้นเพื่อใช้ในการประมวลผล โดยที่จะมีการอธิบายในหัวข้อย่อยที่ 4.4.1 เพื่อรับประกันว่าวิธีการที่ได้นำเสนอนั้นทำงานได้ดีและมีประสิทธิภาพเพียงพอที่จะนำไปประยุกต์ใช้ในสถานการณ์จริงที่เก็บค่ามาจากการวัดในแม่น้ำ คำอธิบายสำหรับลำดับของการทดลองทั้ง 3 นั้นสามารถอธิบายได้ดังรูปที่ 4.3



รูปที่ 4.3: 3 ช่วงการเรียนรู้ของการจำลองแบบ

โดยสรุปสามารถแยกอธิบายความหมายของทั้ง 3 ช่วงได้ดังนี้

1. ช่วงที่ระบบกำลังเรียนรู้ (Learning period) - ช่วงเวลาที่ใช้สำหรับการเรียนรู้และจับพฤติกรรมของระบบ
2. ช่วงที่ระบบเรียนรู้แล้ว (Learned period) - ช่วงเวลาสำหรับการทดสอบฟังก์ชันผลรางวัลและการเปลี่ยนแปลงอื่นๆ
3. ช่วงที่มีการปรับปรุงระบบ (Adaptive period) - ช่วงเวลาสำหรับการทดสอบระบบในสภาพแวดล้อมจริงและพยายามแก้ปัญหาในสถานการณ์นั้นๆด้วยวิธีการมอนติคาร์โลด้วยเร็พพิวเทชันของเส้นทาง

ผลการทดลองการจำลองแบบได้ดำเนินการบนพื้นฐานการทำงานเป็นเอพพิโซด ซึ่งแต่ละเอพพิโซดจะประกอบไปด้วยเหตุการณ์หลายๆเหตุการณ์โดยแต่ละเหตุการณ์จะเริ่มตั้งแต่เซนเซอร์โนดต้นทางต้องการส่งข้อมูลแจ้งเตือนไปยังแอคเตอร์โนดปลายทาง ช่วงที่ระบบกำลังเรียนรู้แอคชั่น-แวลู่ฟังก์ชันหรือคิวฟังก์ชันจะถูกสร้างขึ้นเพื่อรอเก็บข้อมูลตั้งแต่เอพพิโซดแรกและระบบจะดำเนินต่อไปซึ่งจะมีการปรับปรุงคิวฟังก์ชันเพียงครั้งเดียวเมื่อสิ้นสุดเอพพิโซดนั้นๆ ซึ่งฟังก์ชันของพลังงานจะลดทอนตามอัตราส่วนของระยะทางและขนาดของข้อมูล ในงานวิจัยนี้กำหนดขนาดของข้อมูล โดยคำนวณจากส่วนหัวของแพ็จเกจ 20 ไบต์ (การจัดสรรเส้นทางและชั้น MAC) ส่วนเนื้อหาข้อมูล 8 ไบต์ (อุณหภูมิ ความเร็ว PH และระดับความสูงของน้ำ) ดังนั้นจำนวนไบต์ทั้งหมดสำหรับการส่งข้อมูลต่อหนึ่งแพ็กเกตมีขนาดประมาณ 30 ไบต์ ซึ่งพลังงานที่สูญเสียในการส่งข้อมูลสามารถคำนวณได้ดังสมการ (3.3)

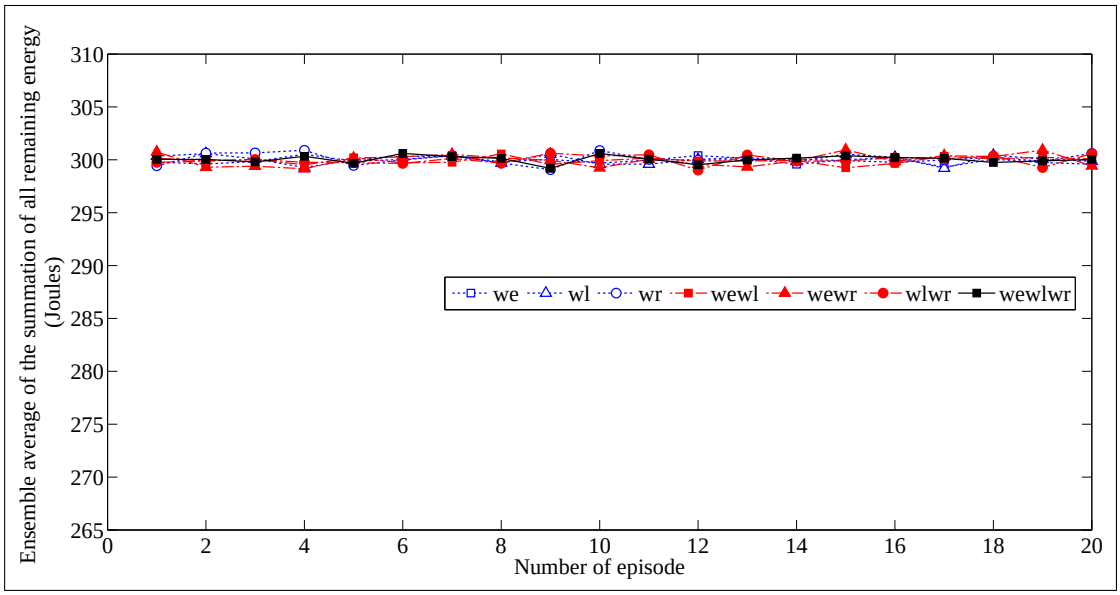
4.4.1 ช่วงที่ระบบกำลังเรียนรู้ (Learning period)

หัวข้อนี้มีวัตถุประสงค์เพื่อนำเสนอผลของการนำมอนติคาร์โลมาใช้กับเร็พพิวเทชันของเส้นทางสำหรับการค้นหาการแก้ปัญหาที่ดีที่สุดที่เป็นไปได้ในสถานการณ์ที่กำหนด ทอพอโลยีจะถูกสร้างจากหัวข้อ 4.1 ซึ่งประกอบด้วยเซนเซอร์ไร้สายสองชนิดคือเซนเซอร์โนดและแอคเตอร์โนด จากทอพอโลยีดังรูปที่ 4.1 กำหนดแอคเตอร์โนดคือโนดที่ 6 และโนดที่ 11 เซนเซอร์โนดคือโนดที่ 1 2 3 4 5 7 8 9 และ 10 ระบบที่พิจารณาจะเป็นการทดลองการส่งข้อมูลจากเซนเซอร์โนดต้นทางไปยังแอคเตอร์โนดปลายทาง โดยเหตุผลพื้นฐานสำหรับการทดสอบในช่วงเวลาการเรียนรู้เพื่อให้แน่ใจว่าวิธีการที่นำเสนอมีประสิทธิภาพดีสำหรับการประยุกต์ใช้เตือนภัยน้ำท่วม นอกจากนี้ยังมีการพิจารณาเพื่อเปรียบเทียบผลการดำเนินงานของขนาดปริภูมิสถานะที่เปลี่ยนแปลงไป

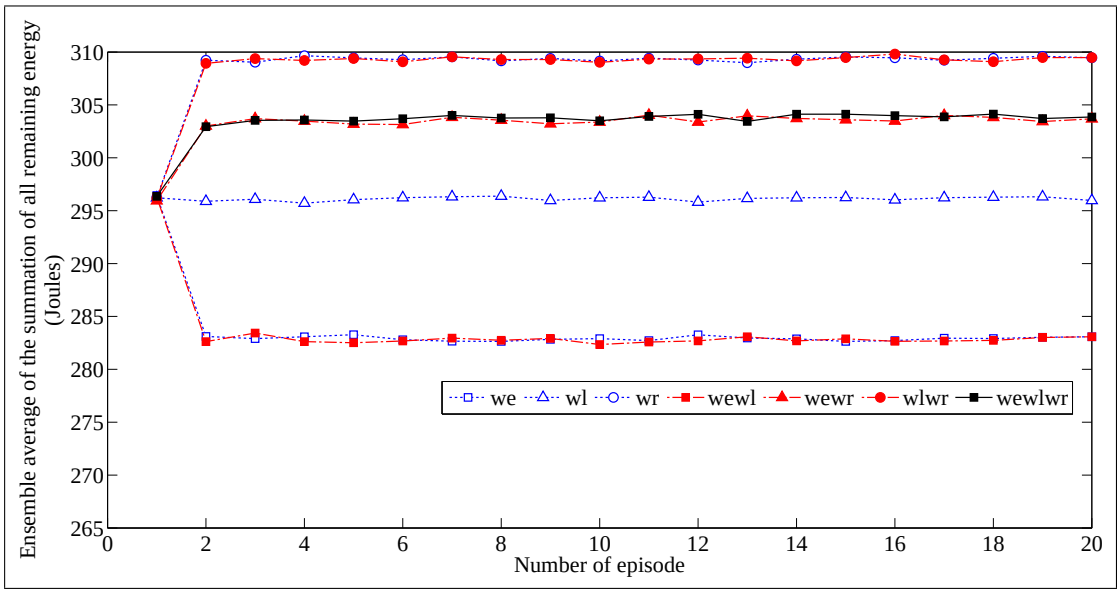
การตั้งค่าการจำลองแบบในสถานการณ์นี้คือมีเหตุการณ์เกิดขึ้น 1500 เหตุการณ์และมีรอบการทำงาน 20 เอพพิโซด ตัวแปรสำหรับการกำหนดสถานะมีการกำหนดไว้ข้างต้น (หัวข้อ 4.2.1) โดยมีปริภูมิการกระทำนิยามจากเส้นทางทั้งหมดที่เป็นไปได้ที่ถูกสร้างขึ้นมาด้วยโพรโทคอลรีแอคทีฟที่มีจำนวนฮอพที่มากที่สุด 4 ฮอพ (หัวข้อ 4.2.2) ฟังก์ชันผลรางวัลนิยามจากการคำนวณและเปรียบเทียบประสิทธิภาพของพลังงานที่เหลืออยู่ อายุการใช้งานและค่าเร็พพิวเทชันของเส้นทาง (หัวข้อ 4.2.3) สำหรับแต่ละสถานการณ์ในการทดสอบช่วงความเชื่อมั่น (confidence interval) ให้อยู่ในช่วง 95% ซึ่งในการจำลองแบบนี้จะทำการทดลอง 40 รีอะไลเซเชินเพื่อใช้ลดค่าความคลาดเคลื่อนที่เกิดขึ้นจากการทดลองหลายๆค่าและจากหลายๆการทดลองได้

ในชุดการทดสอบของช่วงที่ระบบกำลังเรียนรู้นี้จะเริ่มจากการพิจารณาหนึ่งตัวแปรสถานะ สองตัวแปรสถานะ และสามตัวแปรสถานะ ตามลำดับ การพิจารณาดังกล่าวเพื่อให้เห็นถึงผลกระทบ

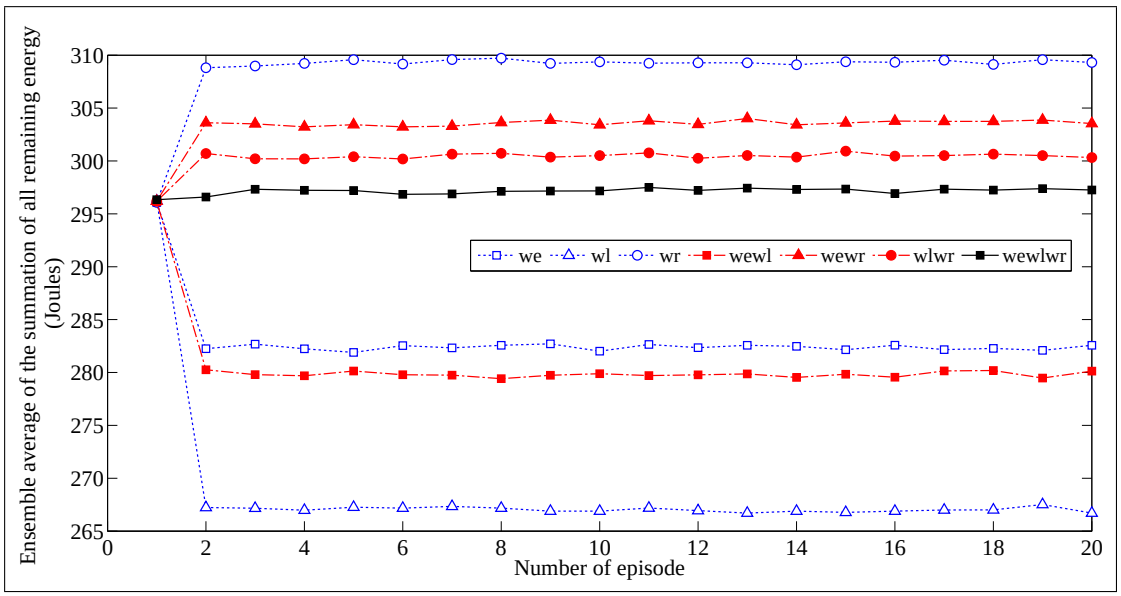
ของการเปลี่ยนแปลงปริภูมิสถานะในสถานการณ์ที่แตกต่างกัน นอกจากนี้ยังพิจารณาถึงผลกระทบของการให้ค่าน้ำหนักกับฟังก์ชันผลรางวัลอีกด้วย สำหรับการทดสอบเบื้องต้นและเพื่อให้ง่ายต่อการพิจารณางานวิจัยนี้จึงกำหนดผลรวมของการให้ค่าน้ำหนักต้องมีค่าเป็น 1 แต่อย่างไรก็ตามค่าของผลรวมดังกล่าวไม่จำเป็นต้องมีค่าเท่ากับ 1



รูปที่ 4.4: 1 ตัวแปรสถานะ (พลังงานที่เหลืออยู่)



รูปที่ 4.5: 2 ตัวแปรสถานะ (พลังงานที่เหลืออยู่และค่าเร็พพิวเทชัน)



รูปที่ 4.6: 3 ตัวแปรสถานะ (พลังงานที่เหลืออยู่ อายุการใช้งานและค่าเร็พพิวเทชัน

จากผลการทดลองดังรูปที่ 4.4 4.5 และ 4.6 สามารถวิเคราะห์ผลการเปรียบเทียบได้ 2 ประเด็นหลัก ดังนี้

1. ผลการวิเคราะห์ค่าน้ำหนักฟังก์ชันผลรางวัล

จากผลการทดลองเห็นได้ว่าฟังก์ชันผลรางวัลจะส่งผลให้ระบบประสิทธิภาพและความสามารถยิ่งขึ้นเมื่อฟังก์ชันผลรางวัลมีการพิจารณาค่าน้ำหนักของฟังก์ชันเร็พพิวเทชันร่วมด้วย ในทางกลับกันถ้าไม่มีการพิจารณาค่าน้ำหนักของฟังก์ชันเร็พพิวเทชันแล้ว ผลลัพธ์ที่ได้ในด้านพลังงานที่เหลืออยู่จะมีค่าน้อยเสมอเหตุผลมาจากช่วงของพลังงาน อายุการใช้งานและค่าเร็พพิวเทชันที่นำมาพิจารณาเป็นผลรางวัลนั้นมีความละเอียดของแต่ละฟังก์ชันที่ต่างกันจึงส่งผลถึงการพิจารณาการให้ความสำคัญของฟังก์ชันนั่นๆในผลรางวัล ยกตัวอย่างการพิจารณาความแตกต่างของช่วงฟังก์ชันผลรางวัลแยกตามฟังก์ชันย่อยดังนี้

การพิจารณาช่วงฟังก์ชันของพลังงานที่เหลืออยู่ที่เป็นไปได้

$$\bar{\varepsilon}_l(t) \quad = \quad \frac{\varepsilon_l(t)}{\max \varepsilon_l(t)} q_\varepsilon = \frac{\sum\limits_{\forall i \in l} \alpha_{i,l}(t) \mid_{t \geq 1}}{\sum\limits_{\forall i \in l} \alpha_{i,l}(t) \mid_{t=0}} q_\varepsilon \tag{4.10}$$

ยกตัวอย่าง กำหนดให้พลังงานเริ่มต้นมีค่า 100 จูล เซนเซอร์โนดตัวที่ 1 2 3 มีค่า 50 68 และ 58 ตามลำดับ ฉะนั้นค่าการแบ่งนับของฟังก์ชันของผลรวมพลังงานในเส้นทางที่ถูกเลือก l ใดๆ มีค่า $\frac{176}{300} \times 10 = 5.87$ ดังนั้นถ้าพิจารณาช่วงพลังงานที่ระดับ 5 ถึงระดับ 6 จะมีระดับพลังงานที่เป็นไปได้ 10 ค่า ด้วยเหตุนี้ถ้าพิจารณาการแบ่งนับของฟังก์ชันของผลรวมพลังงานตั้งแต่ 0 - 10 จะมีช่วงคำตอบที่เป็นไปได้ $|\varepsilon| = 10 \times 10 = 100$

การพิจารณาช่วงฟังก์ชันของอายุการใช้งานที่เหลืออยู่ที่เป็นไปได้

$$\bar{\ell}_l(t) = \frac{arg\max_i \ell_{i,l}(t)}{\ell_i(0)} q_\ell \tag{4.11}$$

ยกตัวอย่าง กำหนดให้อายุการใช้งานเริ่มต้นมีค่า 100 ครั้ง เซนเซอร์โนดตัวที่ 1 2 3 มีค่า 50 68 และ 58 ตามลำดับ ฉะนั้นค่าการแบ่งนับของฟังก์ชันของผลรวมอายุการใช้งานในเส้นทางที่ถูกเลือก l ใดๆ มีค่า $\frac{68}{100} \times 10 = 6.8$ ดังนั้นถ้าพิจารณาช่วงอายุการใช้งานที่ระดับ 5 ถึงระดับ 6 จะมีระดับพลังงานที่เป็นไปได้ 10 ค่า ด้วยเหตุนี้ถ้าพิจารณาการแบ่งนับของฟังก์ชันของผลรวมพลังงานตั้งแต่ 0 - 10 จะมีช่วงคำตอบที่เป็นไปได้ $|\ell| = 10 \times 10 = 100$ เช่นเดียวกับช่วงฟังก์ชันของพลังงาน แต่ในขณะที่ช่วงของค่าเร็พพิวเทชันของเส้นทาง

การพิจารณาช่วงฟังก์ชันของค่าเร็พพิวเทชันของเส้นทางที่เป็นไปได้

$$\bar{R}_l(t) = \frac{R_l(t)}{\max R_l(0)} q_r \tag{4.12}$$

ยกตัวอย่าง กำหนดให้ค่าเร็พพิวเทชันเริ่มต้นมีค่า 1 เซนเซอร์โนดตัวที่ 1 2 3 มีค่า 0.7 0.6 และ 0.5 ตามลำดับ ฉะนั้นค่าการแบ่งนับของฟังก์ชันของค่าเร็พพิวเทชันในเส้นทางที่ถูกเลือก l ใดๆ มีค่า $0.21 \times 10 = 2.1$ แต่เมื่อพิจารณาในกรณีที่ค่าเร็พพิวเทชันในเส้นทางมีค่าต่ำสุดคือทุกเซนเซอร์โนดมีค่า 0.1 ค่าการแบ่งนับของฟังก์ชันของฟังก์ชันที่เกิดขึ้น $0.001 \times 10 = 0.01$ ฉะนั้นช่วงคำตอบที่เป็นไปได้ของค่าเร็พพิวเทชัน $|R| = 100 \times 100 = 10000$

ดังนั้นเหตุผลของระบบเกี่ยวกับช่วงคำตอบที่เป็นไปได้ของแต่ละค่าตัวแปรของฟังก์ชันผลรางวัล ถ้าจะขยายความให้ชัดเจนกว่านั้นหมายความว่าค่าของน้ำหนักแต่ละตัวนี้จะถูกคูณด้วยค่าคงที่กับฟังก์ชันของตนเอง เช่นผลรวมของพลังงานทั้งหมด, ค่าต่ำสุด-สูงสุดของอายุการใช้งานโนด (การหาคอขวดของอายุการใช้งาน) และค่าเร็พพิวเทชันของเส้นทาง โดยแต่ละฟังก์ชันนั้นจะขึ้นกับเส้นทางที่ถูกเลือก l อีกด้วย นอกจากนั้นการพิจารณาค่าในแต่ละฟังก์ชันจะพบว่าค่าของเร็พพิวเทชันของเส้นทางนั้นมีช่วงการเปลี่ยนแปลงสูงที่สุดหลังจากการทำการแบ่งนับออกมาแล้ว หากพิจารณาเทียบกับอีกสองฟังก์ชันที่เหลือ ไม่ว่าจะเป็นฟังก์ชันผลรวมพลังงานหรือจะเป็นค่าต่ำสุด-สูงสุดของอายุการใช้งานโนด ดังนั้นการที่มีช่วงการเปลี่ยนแปลงสูงจะส่งผลให้เกิดความแตกต่างทางการคำนวณและส่งผลต่อการเลือกของการกระทำในเหตุการณ์ถัดไป โดยในกรณีนี้ค่าความแตกต่างในระดับทศนิยมสามารถแยกความแตกต่างของระบบออกมาได้ (ในทางตรงกันข้ามความแตกต่างที่เกิดขึ้นหากผลรางวัลมีค่าสูงมากจนทำให้ค่านัยสำคัญของตัวเลขมีน้อยมาก ซึ่งตัวตัดสินใจเลือกกระทำอาจไม่สามารถแยกความแตกต่างของผลรางวัลได้เช่นกัน เช่น 100000 กับ 100001 เป็นต้น)

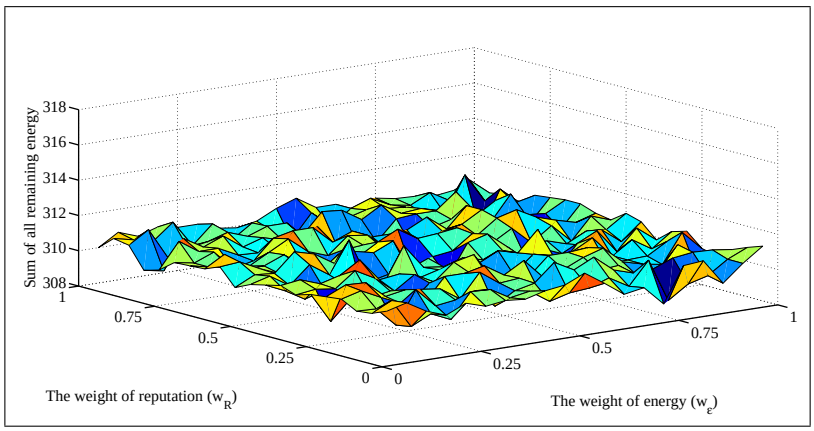
สำหรับในส่วนของฟังก์ชันผลรางวัลจะเห็นว่ามีความแตกต่างในส่วนที่พิจารณาเครื่องหมายปัดขึ้น (ceiling) ออกไป ซึ่งจะแตกต่างจากกรณีที่นำไปใช้ในตัวแปรสถานะ ด้วยการทำแบบนี้ ตัวของฟังก์ชันผลรางวัลจะมีประสิทธิภาพดีกว่าเพราะสามารถแยกความแตกต่างได้ดีมากกว่ากรณีที่เป็นจำนวนเต็มจากผลการปัดขึ้นของเลขทศนิยม ดังนั้นจะเห็นได้ว่าหากไม่นำค่าเร็พพิวเทชันของเส้นทางเข้ามาคำนวณหรือทำให้ค่าน้ำหนักของเร็พพิวเทชัน w_r มีค่าเป็นศูนย์ จะพบว่าผลที่ได้ออกมาผิดไปจากคำตอบที่ควรจะเป็นเพราะตัวกระทำการตัดสินใจของระบบไม่สามารถแยกความแตกต่างของค่าฟังก์ชันผลรางวัลในระดับทศนิยม 1 ตำแหน่งกับค่าของจำนวนเต็มได้ ดังนั้นจะเห็นได้อย่างชัดเจนว่าการมีทั้งสามฟังก์ชันเข้ามาเป็นส่วนประกอบจะทำให้วิธีการที่นำเสนอมีความสามารถในการควบคุมค่าของเงื่อนไขที่ต้องการได้และสำหรับในกรณีอื่นๆนั้น จะพบว่าการใช้จำนวนของฟังก์ชันผลรางวัลยิ่งเพิ่มขึ้นก็จะส่งผลให้มีค่าของพลังงานเหลือรวมในระบบสูงขึ้นตามลำดับ

2. ผลการวิเคราะห์ขนาดของตัวแปรสถานะ

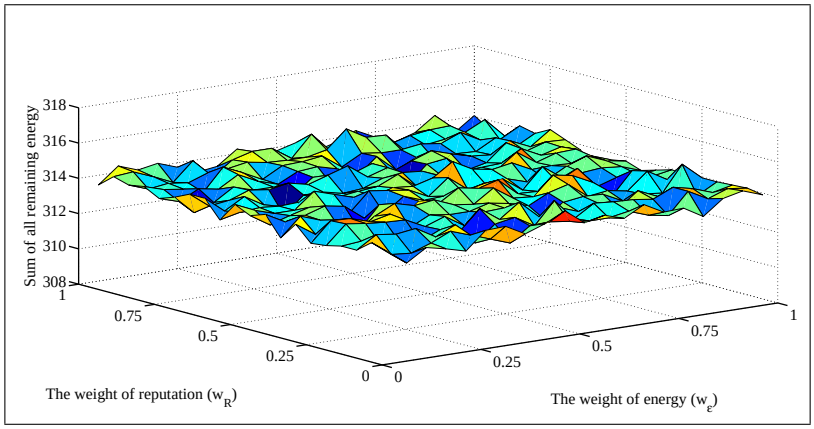
จากผลการทดลอง 4.5 พบว่าหากใช้เพียงแค่พลังงานที่เหลืออยู่และค่าเร็พพิวเทชัน เป็นตัวแปรสถานะก็สามารถทำให้ระบบมีสมรรถนะที่ดีอยู่ได้ ซึ่งบางครั้งค่าฟังก์ชันผลรางวัลอาจจะดูไม่แตกต่างกัน แต่หากพิจารณาในเทอมของค่าความจุตัวแปร (computational space) สถานะในการใช้สองตัวแปรจะใช้ความจุน้อยกว่าจากที่จะต้องใช้ $|S| \times |S| \times |S|$ จะเหลือเพียง $|S| \times |S|$ ไม่เฉพาะในส่วนของการคำนวณค่าความจุตัวแปร แต่รวมถึงระยะเวลาที่ใช้ในการหาคำตอบอีกด้วย เพราะฉะนั้นข้อสรุปของในช่วงระบบกำลังเรียนรู้นี้จะนำพารามิเตอร์ที่ถูกหาไปใช้ในช่วงถัดไป โดยมีประเด็นของการใช้สองตัวแปรสถานะกับค่าของฟังก์ชันผลรางวัลแบบ 3 ฟังก์ชันจะถูกนำมาใช้ในช่วงที่ระบบกำลังเรียนรู้และในช่วงถัดไปจะมีการทดสอบผลกระทบของฟังก์ชันผลรางวัลที่เปลี่ยนไปต่อระบบ รวมถึงการทดสอบค่าที่ดีที่สุดของการให้ค่าน้ำหนักของทั้ง 3 ตัวแปรฟังก์ชันผลรางวัล

4.4.2 ช่วงที่ระบบเรียนรู้แล้ว (Learned period)

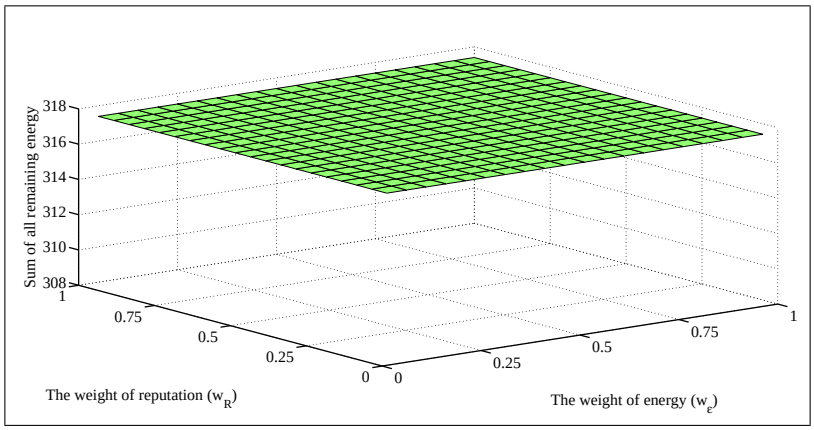
หัวข้อก่อนหน้านี้ ได้ทำการเริ่มต้นพิจารณาวิธีการมอนติคาร์โล ในส่วนที่เป็นช่วงของการเรียนรู้และได้พบว่าสถานะที่เลือกใช้เพียงแค่สองตัวแปรสถานะที่เป็นพลังงานและค่าเร็พพิวเทชันโดยใช้คู่กับสามตัวแปรฟังก์ชันผลรางวัลนั้น เพียงพอต่อการนำไปใช้งานในช่วงถัดไป โดยในทางปฏิบัติแล้วเมื่อระบบมีการเรียนรู้เสร็จสิ้นในสภาวะจำลอง ก็สามารถที่จะนำระบบไปใช้งานได้เลย แต่ในงานวิจัยนี้ เราได้ทำการศึกษาเพิ่มเติมถึงผลกระทบของการเปลี่ยนแปลงของค่าน้ำหนักของผลรางวัลต่อระบบ และเพื่อให้เห็นความแตกต่างอย่างชัดเจน โดยจะทำการเลือกให้มีการใช้ทอพอโลยีเดิม ในระบบแบบคงที่ โดยให้มีค่า $\epsilon = 0.1$ ตั้งต้น แต่เนื่องจากการใช้ระบบที่มีสามตัวแปรฟังก์ชันผลรางวัลนั้นจะทำให้เกิดการสร้างมิติที่ 4 ซึ่งไม่สามารถมองภาพได้ด้วยการแสดงผลผ่านกราฟ ดังนั้น ในเบื้องต้น จึงขอนำเสนอผลการทดลองอย่างง่ายเบื้องต้นด้วยการพิจารณาสองตัวแปรฟังก์ชันผลรางวัลที่เป็นค่าของพลังงานรวมที่เหลือทั้งหมดในระบบและค่าเร็พพิวเทชันของเส้นทาง โดยนำไปเทียบกับค่าพลังงานจริงที่เหลือออกมาจากระบบนั่นเอง โดยในการตั้งค่าการทดลองกำหนดให้มีการพิจารณาเหตุการณ์ที่ 1500 ครั้ง และมีระดับการลดพลังงานลงเป็นไปได้ดังสมการที่ (3.3) ซึ่งขึ้นกับระยะทางทีเปลี่ยนไป หลังจากเมื่อผลการทดลองที่ได้ออกมา จากสองตัวแปรฟังก์ชันผลรางวัลแล้ว เราจะพิจารณาค่าของคำตอบสุดท้ายที่เป็นจุดสมดุลด้วยการทดลองหาค่าที่ดีที่สุดที่เกิดขึ้นในสภาวะดังกล่าว



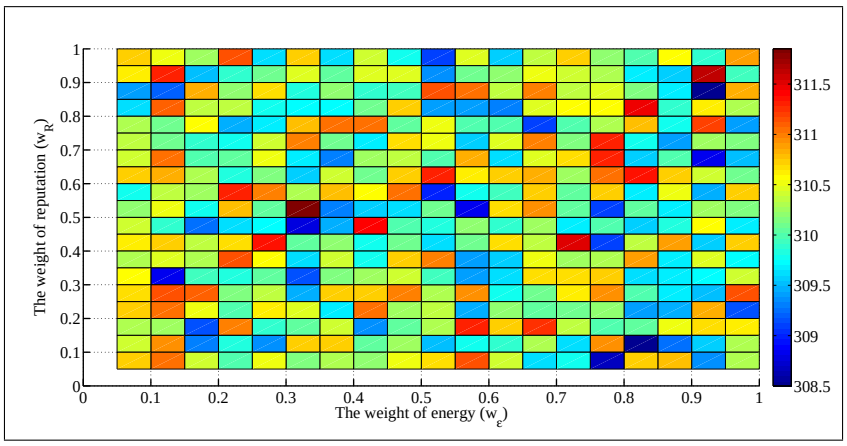
รูปที่ 4.7: พิจารณาฟังก์ชันผลรางวัลโดย ϵ มีค่า 0.2



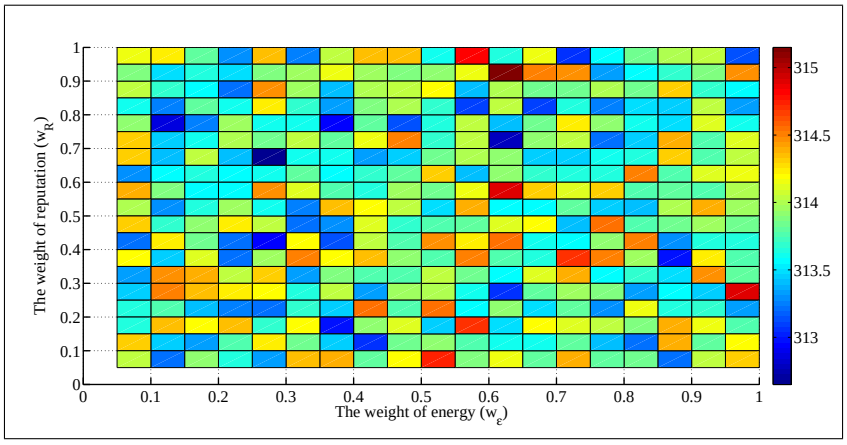
รูปที่ 4.8: พิจารณาฟังก์ชันผลรางวัลโดย ϵ มีค่า 0.1



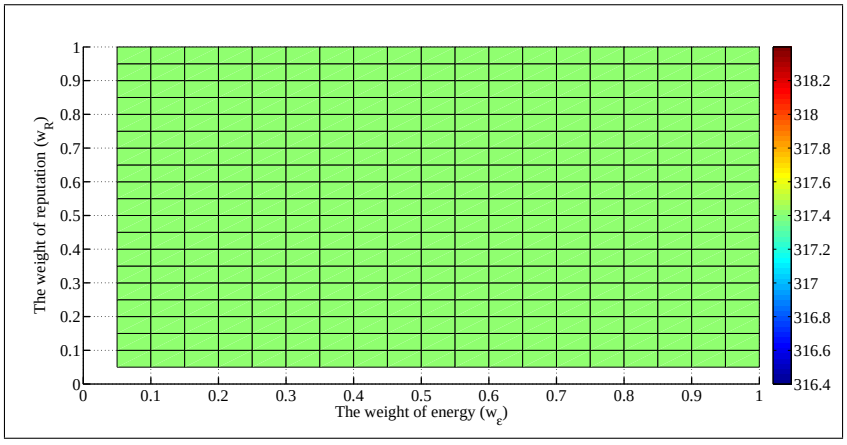
รูปที่ 4.9: พิจารณาฟังก์ชันผลรางวัลโดย ϵ มีค่า 0



รูปที่ 4.10: พิจารณาระนาบ x-y ฟังก์ชันผลรางวัล ϵ มีค่า 0.2



รูปที่ 4.11: พิจารณาระนาบ x-y ฟังก์ชันผลรางวัล ϵ มีค่า 0.1



รูปที่ 4.12: พิจารณาระนาบ x-y ฟังก์ชันผลรางวัล ϵ มีค่า 0

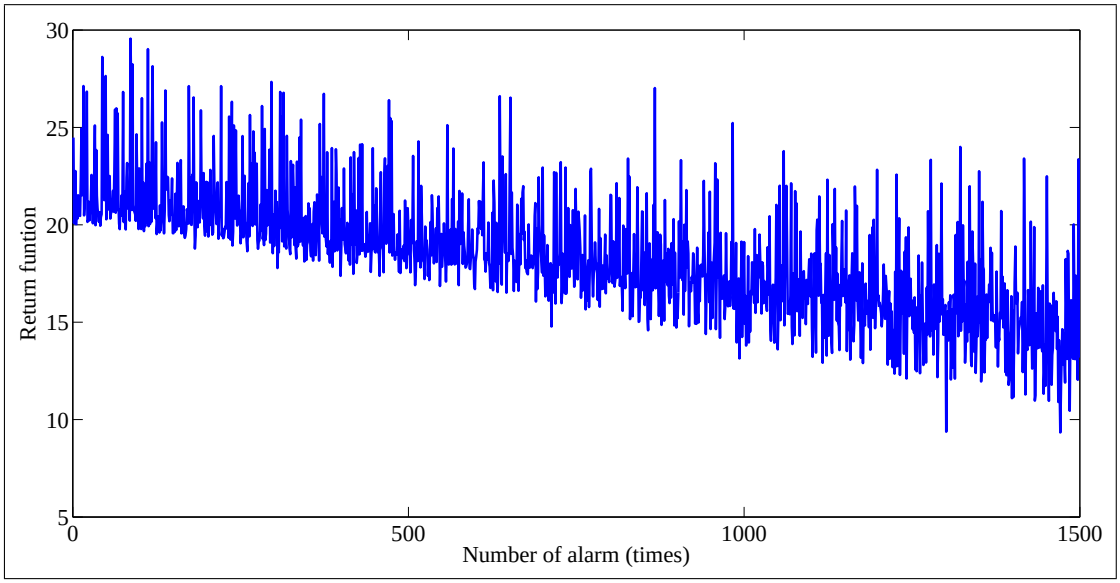
ดังรูปที่ 4.7 - 4.9 แสดงให้เห็นถึงภาพ 3 มิติของค่าพลังงานที่เกิดขึ้นจริงในระบบโดยตั้งค่า

ϵ ที่แตกต่างกัน โดยมีค่า ϵ ที่ลดลง เท่ากับ 0.2 0.1 และ 0 จะเห็นว่าค่าของพลังงานรวมของระบบจะค่อยๆ ดีขึ้น จนเป็นระนาบคงที่ดังแสดงในรูปที่ 4.9 แต่เนื่องจากการทดสอบเริ่มต้นในช่วงของระบบกำลังเรียนรู้นั้น ค่าของ Q ฟังก์ชันนั้นเหมาะสมกับลักษณะของระบบแล้ว แต่ในการทดลองนี้ได้มีการพยายามเปลี่ยนแปลงค่าของการเลือกการกระทำผ่าน Q ฟังก์ชันที่ได้จาก ϵ-กรีดี้ ดังนั้นจะพบว่าค่าของกรณีที่ไม่มีการเปลี่ยนแปลง ϵ จะเป็นคำตอบที่ดีที่สุด คำตอบเดียวกับที่ได้มาจากการทดลองในช่วงที่ระบบกำลังเรียนรู้ ซึ่งสมมติฐานเบื้องต้นนี้ ไม่ได้ผิดไปจากการคาดเดาตั้งต้น โดยที่สมมติฐานของการเปลี่ยนการกระทำแท้ที่จริงแล้วควรจะส่งผลกระทบถึงการเปลี่ยนแปลงของ Q ฟังก์ชัน ไปในทิศทางที่ดีขึ้น ซึ่งสมมติฐานนี้ จะได้รับการอธิบายในหัวข้อถัดไป โดยสาเหตุที่การเปลี่ยนแปลงการกระทำผ่าน Q ฟังก์ชันของการทดลองนี้ ให้ผลที่ไม่ดีทั้งนี้สาเหตุหลักเกิดจากการที่ระบบไม่มีการปรับปรุงค่าของ Q ฟังก์ชันที่เปลี่ยนไปตามการเปลี่ยนแปลงของระบบหรือตามสถานการณ์นั่นเอง สำหรับในรูปที่ 4.10 - 4.12 จะพบว่ามีคำตอบที่ใกล้เคียงกันอยู่หลายตำแหน่ง โดย ณ จุดต่างๆ เหล่านั้นมีค่านัยสำคัญน้อยมากโดยคำตอบนั้นแสดงถึงระดับพลังงานรวมที่เหลืออยู่ในระบบเช่นเดียวกับในการทดลองที่ผ่านมา และในช่วงสุดท้ายของช่วงที่ระบบเรียนรู้แล้วนี้ ในงานวิจัยนี้จะพยายามหาค่าของค่าถ่วงน้ำหนักที่เหมาะสมที่สุด เพื่อนำไปใช้ในการทดลองถัดไปรวมถึงการทดลองอื่นในสภาวการณ์อื่นๆ แต่สำหรับงานวิจัยที่กำลังพิจารณาสถานการณ์การเตือนอุทกภัยนั้นในบางครั้ง หากเกิดการสุ่ม (randomness) ที่เปลี่ยนไปจะทำให้ค่าของการถ่วงน้ำหนักที่เหมาะสมที่สุดนั้นเปลี่ยนแปลงได้ ขึ้นกับการให้ค่า (seed) ของการสุ่มที่แตกต่างกันในระหว่างการจำลองเหตุการณ์ อาทิเช่น ค่า $w_{\mathcal{E}}$ w_{ℓ} และ w_R ที่เป็นไปได้ ควรจะมีค่า 0.3 0.3 และ 0.3 ตามลำดับ โดยทั้งนี้ ค่าที่ออกมาจริงระหว่างความพยายามหาผลลัพธ์ที่เหมาะสมที่สุด จะเป็น 0.1 0.4 และ 0.5 ตามลำดับดังแสดงในภาคผนวก แต่อย่างไรก็ตามผลการหาผลลัพธ์ที่เหมาะสมที่สุดที่กล่าวมานั้นจะคงที่ตราบใดก็ตามที่ระบบมีการสุ่มที่ลดลง หรือ มีค่าน้อยมาก ซึ่งจะส่งผลให้ Q ฟังก์ชันมีค่าคงตัวในที่สุด แต่ในงานวิจัยนี้การสุ่มสุดท้ายที่เหลืออยู่คือการให้ค่าหรือพิจารณาของการเกิดเหตุการณ์จริง ซึ่งอาจจะเกิดการเปลี่ยนแปลงจากระบบเดิมได้ตลอดเวลา ดังนั้นการทำงานในส่วนนี้จะถูกนำไปพิจารณาในหัวข้อถัดไปเพื่อใช้ในการพิจารณาการเปลี่ยนแปลงของ ระบบเทียบกับความคงทนของวิธีการมอนติคาร์โล ดังนั้นจากข้อสรุปในเบื้องต้น พบว่าการใช้ค่าคงตัวของฟังก์ชันผลรางวัลมีค่า 0.3 กำหนดไว้ให้ผลที่ดี และยังสามารถรักษาค่าของของความเป็นธรรม (fairness) ของระบบไว้ได้อีกด้วย ดังนั้นค่า $w_{\mathcal{E}}$ w_{ℓ} และ w_R ที่มีค่า 0.3 จะถูกนำมาเสนอต่อไป และดังแสดงในส่วนของการทดลองช่วงที่ระบบกำลังเรียนรู้ที่มีหัวข้อการวิเคราะห์ค่าฟังก์ชันผลรางวัล จะเห็นได้ชัดเจนว่าการพิจารณาร่วมกันของ 3 ตัวแปรฟังก์ชันผลรางวัลแบบพร้อมกัน จะส่งผลให้ระบบเกิดความสมดุลได้ดีที่สุดเพื่อนำไปใช้ในสถานการณ์ต่างๆได้ดังแสดงในหัวข้อถัดไป

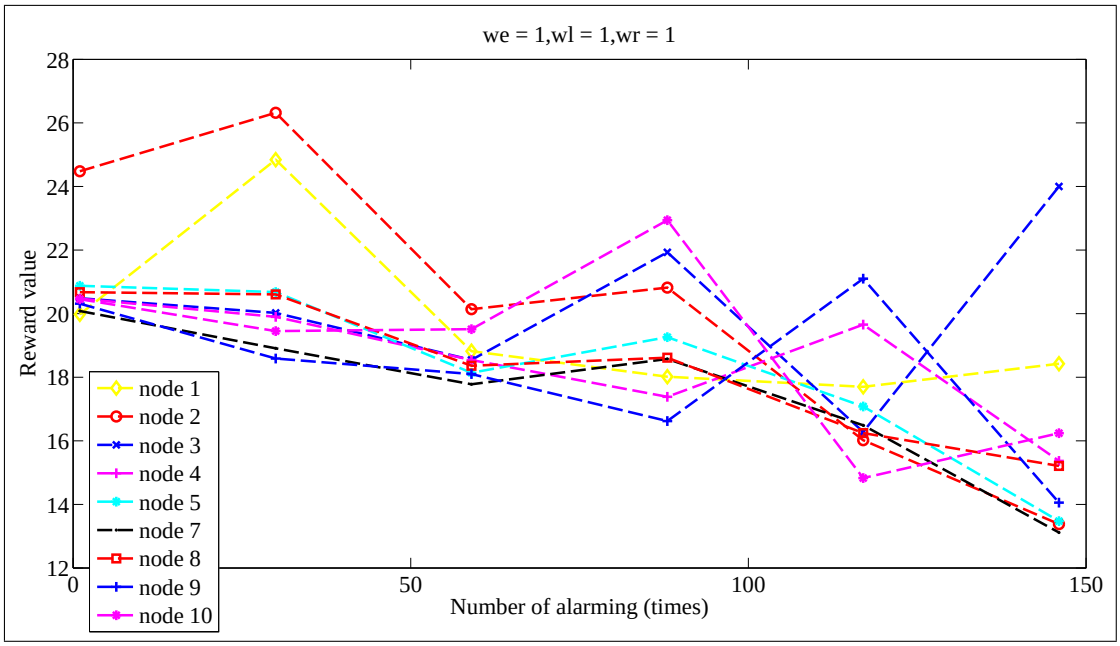
1. การวิเคราะห์ระบบ ณ จุดคอขวด (Bottleneck Analysis)

สำหรับการพิจารณาระบบโครงข่ายเซนเซอร์ไร้สายนั้น นอกจากที่ได้ทำการพิจารณาในภาพรวมแล้ว ปัจจัยที่มีผลกระทบต่อระบบมากที่สุดอีกตำแหน่งหนึ่งคือตำแหน่งที่เรียกว่าคอขวด (bottleneck) ซึ่งในงานวิจัยนี้ มีตำแหน่งหรือโนดคอขวดทั้งหมดสองโนด ได้แก่ โนดที่ 2 และโนดที่ 10 ซึ่งจะเป็นตำแหน่งที่ข้อมูลมีการส่งผ่านสูงที่สุด เนื่องจากเป็นโนดที่มีการเชื่อมต่อได้หลายทางจากการเกิดการเตือนภัยในโนดข้างเคียงใดๆ โดยในหัวข้อย่อยนี้จะเลือกพิจารณาโนดย่อยที่ 2 ในการทดลองนี้จะแบ่งชุดการทดลองออกเป็นสองช่วง โดยการทดลองช่วงแรกจะพิจารณาส่วนของฟังก์ชันรีเทิร์น (return function) และการเปรียบเทียบการลู่เข้า (convergence) ในสองมุมมองของทั้งพลังงานและอายุการใช้งาน สำหรับการทดลองในช่วงที่สองจะทำการพิจารณาที่ละเอียดขึ้น โดยเจาะลึกถึงผลกระทบที่มีการเปลี่ยนแปลงของค่าน้ำหนักของฟังก์ชันผลรางวัลต่อการปรับปรุงตัวของอายุการใช้งานในโนดที่

2 และพลังงานเฉลี่ยทั้งหมดในภาพรวม



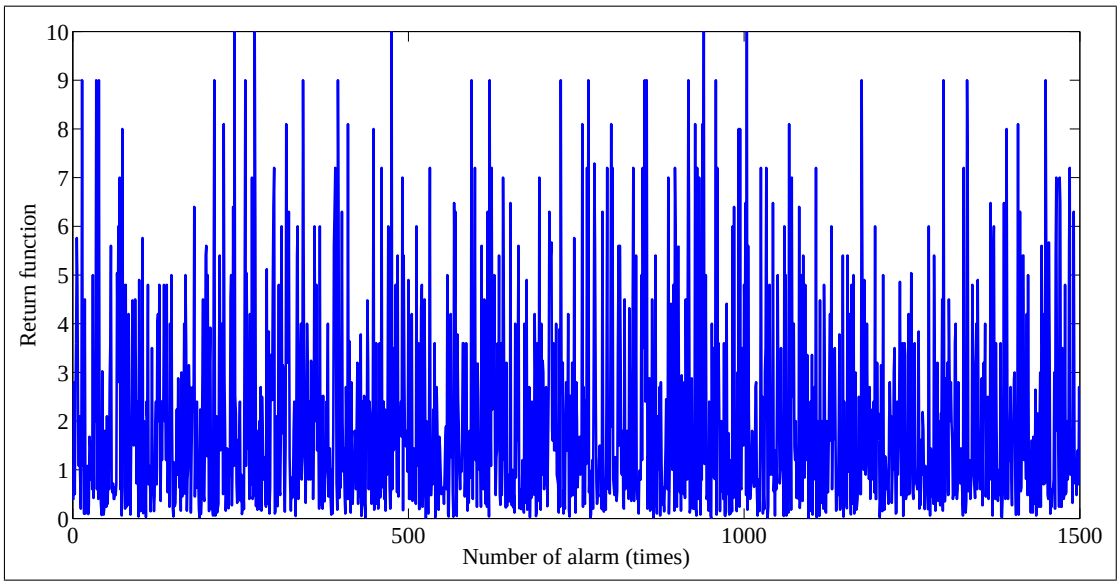
รูปที่ 4.13: ภาพรวมของฟังก์ชันรีเทิร์นต่อการแจ้งเตือน กรณี $w_{\mathcal{E}}, w_{\ell}, w_R$



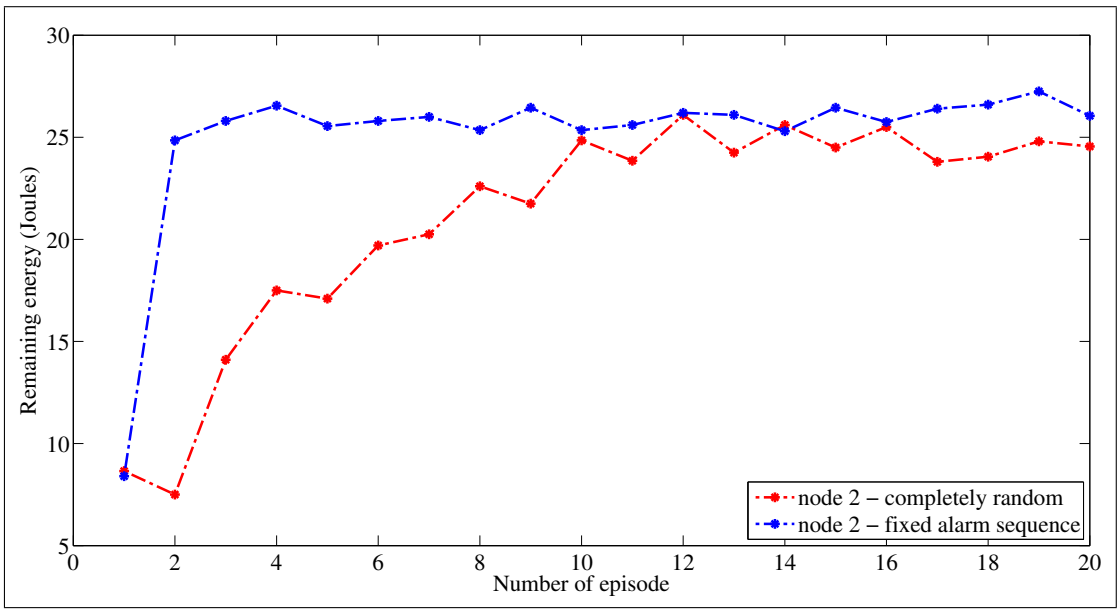
รูปที่ 4.14: ภาพรวมของฟังก์ชันรีเทิร์นที่แสดงค่าผลรางวัลแยกตามโนดการแจ้งเตือน

ดังรูปที่ 4.13, 4.14 จะพบว่าในกรณีนี้ที่ทำการพิจารณาให้ค่าของน้ำหนักเท่ากันทั้งสามฟังก์ชัน ($w_{\mathcal{E}}, w_{\ell}, w_R$ มีค่าเท่ากับ 1) จะพบว่ากราฟมีลักษณะเป็นฟังก์ชันลดลง ตามค่าของ $w_{\mathcal{E}}$ และ w_{ℓ} ที่มีอิทธิพล (dominate) ต่อค่าของฟังก์ชันรีเทิร์นอยู่ แต่ในทางตรงกันข้ามหากพิจารณาให้ค่าของน้ำหนักฟังก์ชัน w_R นั้นพียงอย่างเดียวรูปแบบของฟังก์ชันดังแสดงในรูปที่ 4.15 ก็จะแสดงให้เห็นถึงการกระจายตัวอย่างสม่ำเสมอ ทั้งนี้ขึ้นกับค่าของค่าเร็พพิวเทชัน ณ เวลานั้นๆ นั่นเอง

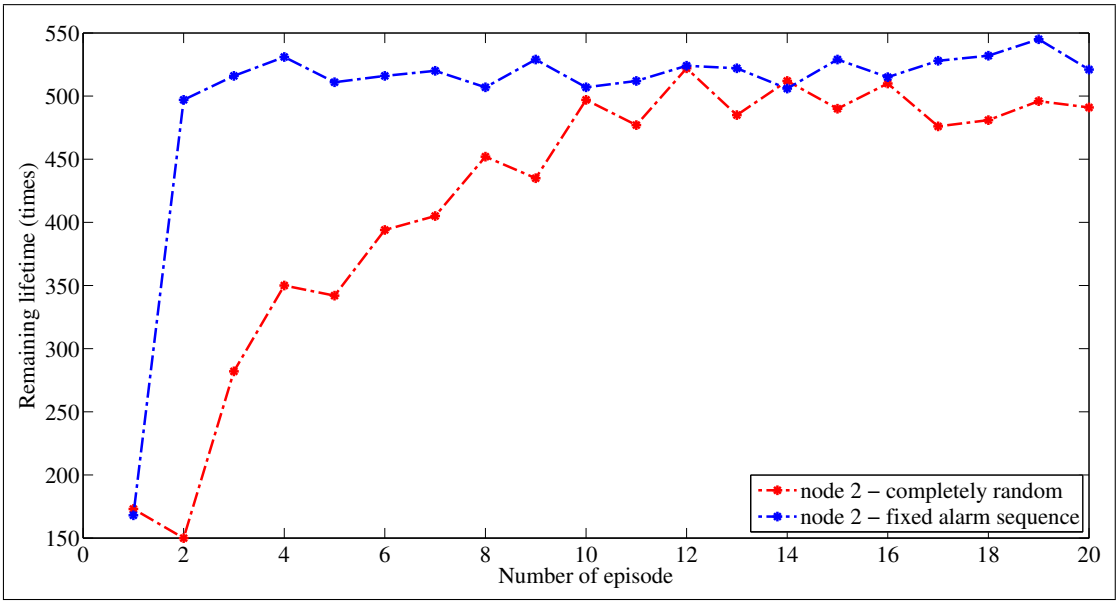
ในรูปที่ 4.14 นั้นแสดงการแยกพิจารณาฟังก์ชันรีเทิร์นแบบแยกพิจารณาตามการแจ้งเตือนของแต่ละโนด ซึ่งก็จะเห็นว่าเป็นรูปแบบของฟังก์ชัดลดที่ชัดเจนตามเวลา โดยในแกน x คือ ครั้งที่เกิดการแจ้งเตือนและแกน y คือค่าของฟังก์ชันรีเทิร์น (ไม่มีหน่วย)



รูปที่ 4.15: ภาพรวมของฟังก์ชันรีเทิร์นต่อการแจ้งเตือน กรณี w_R



รูปที่ 4.16: ค่าของพลังงานที่เหลืออยู่ ณ โนดคอขวด



รูปที่ 4.17: ค่าของอายุการใช้งานที่เหลืออยู่ ณ โนดคอขวด

สำหรับในรูปที่ 4.16 และ 4.17 ที่แสดงถึงค่าของระดับพลังงานที่เหลืออยู่และค่าของอายุการงานของโนดสองนั้น เป็นการนำระบบเดียวกันมาพิจารณาในสองลักษณะด้วยกันคือ การแจ้งเตือนของทุกโนดในระบบ ที่เกิดขึ้นในแต่ละโนด เป็นกระบวนการเฟ้นสุ่มอย่างสมบูรณ์ (complete random) และ ในอีกกรณีหนึ่งคือเป็นการเกิดลำดับเหตุการณ์ที่คงตัวเหมือนกัน (fix alarm sequence) ในทุกเอพิโซด โดยจากผลการทดลองแรกดังแสดงในรูปที่ 4.16 ถึงค่าของระดับพลังงานที่สามารถเพิ่มขึ้นได้ถึง 174.55% ไม่ว่าจะเป็นกรณีของการเฟ้นสุ่มหรือจะเป็นการเกิดลำดับเหตุการณ์แบบคงตัว ซึ่งแสดงถึงขีดความสามารถของวิธีการมอนติคาร์โลที่ทำงานได้อย่างมีประสิทธิภาพสูงมาก ซึ่งจะใกล้เคียงกับการเพิ่มขึ้นของอายุการใช้งานของเซนเซอร์โนดที่เพิ่มขึ้นถึง 181.78% โดยทั้งสองกราฟนี้จะมีรูปร่างลักษณะใกล้เคียงกันทั้งนี้เนื่องจากระดับการลดพลังงานลงนั้น มีลักษณะเป็นเชิงเส้นจึงเกิดความคล้ายเคียงกันของสองลักษณะกราฟดังแสดงในรูปที่ 4.16 และ 4.17

สำหรับการทดลองในช่วงที่ 2 นั้น จะเป็นการพิจารณาถึงผลกระทบของค่าน้ำหนักต่อระดับของพลังงานและอายุการใช้งานในโนดคอขวดโดยมี การเปลี่ยนแปลงดังแสดงในตารางที่ 4.2

ตารางที่ 4.2: ผลกระทบของค่าน้ำหนักต่อระดับของพลังงานและอายุการใช้งานในโนดคอขวด

Weight			Completely random (bottle neck)		Fixed alarm sequence (bottle neck)		Avg.Remain.Energy
$w_\mathcal{E}$	w_ℓ	w_R	mean energy	mean lifetime	mean energy	mean lifetime	
1/2	1/2	0	20.66	413.20	23.09	461.70	284.98
0	1/2	1/2	17.99	359.75	22.35	446.90	299.67
1/2	0	1/2	21.59	431.75	24.57	491.30	303.42
1/3	1/3	1/3	20.98	419.50	25.65	512.90	296.90

4.4.3 ช่วงที่มีการปรับปรุงระบบ (Adaptive period)

ในหัวข้อนี้จะแสดงถึงความสามารถของมอนติคาร์โลในการเอาไปใช้กับค่าเร็พพิวเทชันของเส้นทางในสามสถานการณ์ ซึ่งก็คือทอพอโลยีแบบคงที่ ทอพอโลยีแบบไม่คงที่ตลอดเวลาและทอพอโลยีแบบไม่คงที่ตลอดเวลาแล้วในโครงข่ายมีโนดปัญหาอยู่ ยิ่งไปกว่านั้น ผลการทดลองที่ออกมาจะทำการเปรียบเทียบสองประเภทวิธีการคือวิธีการแบบฉลาดหรือวิธีการที่ระบบสามารถพัฒนาและปรับปรุงตน (intelligent) กับวิธีการแบบไม่ฉลาดหรือวิธีการที่ระบบไม่สามารถพัฒนาและปรับปรุงตน (non-intelligent) โดยที่วิธีการแบบไม่ฉลาดนั้นจะเลือกนำมาเปรียบเทียบ 4 กรณีดังต่อไปนี้

1. วิธีเลือกเส้นทางที่สั้นที่สุด (shortest-path) คือการพิจารณาเส้นทางที่สั้นที่สุดหรือจำนวนฮอพน้อยที่สุด โดยไม่พิจารณาพลังงาน อายุการใช้งาน ฯลฯ

2. วิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน (max-min) คือการพิจารณาพลังงานที่เหลืออยู่ (สมการ (3.3)) โดยพิจารณาที่โนดที่มีพลังงานที่น้อยที่สุดของแต่ละเส้นเชื่อมโยง จากนั้นเลือกเส้นทางที่ประกอบไปด้วยเส้นเชื่อมโยงของโนดที่มีพลังงานเหลือมากที่สุดจากชุดของเส้นทางที่น้อยที่สุด

3. วิธีเลือกเส้นทางแบบสุ่มเอกรูป (uniform random) คือการเลือกเส้นทางแบบสุ่มคงตัว

4. วิธีเลือกเส้นทางแบบมอนติคาร์โลแบบที่ไม่มีการปรับปรุงตนเอง (no-update monte carlo) คือการเลือกเส้นทางด้วยวิธีมอนติคาร์โลแบบไม่มีการปรับปรุงค่า Q ฟังก์ชัน ที่ $\epsilon = 0.1$

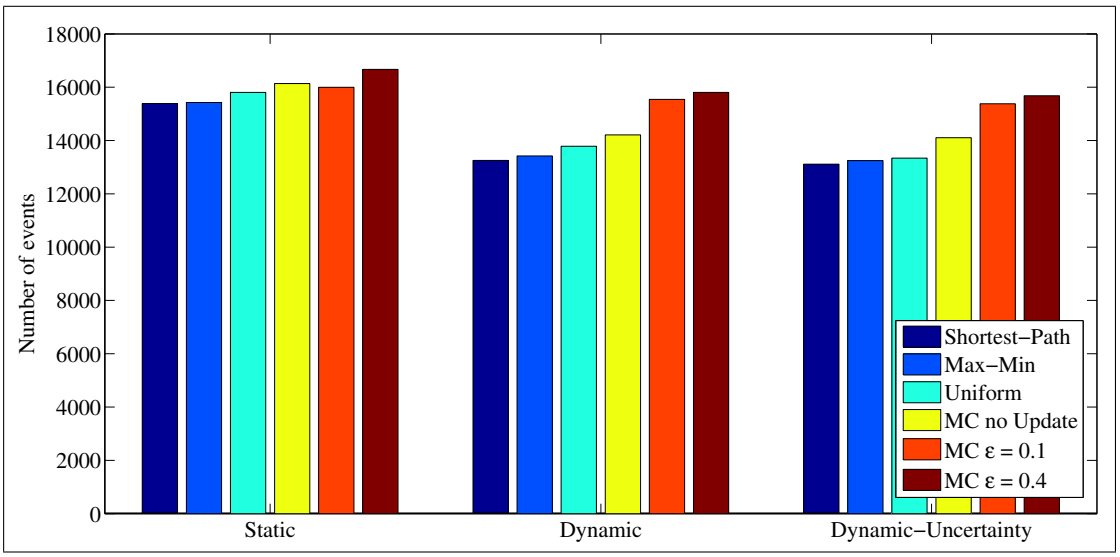
สำหรับในประเภทของวิธีการแบบฉลาด งานวิจัยนี้ได้เลือกการใช้ความสามารถของทักษะการเลือกเส้นทางแบบเดิมซ้ำๆ (exploit ability) กับ การเปลี่ยนทักษะการเพิ่มโอกาสในการเลือกตัดสินใจของมอนติคาร์โลด้วยการขยายโอกาสความน่าจะเป็นของการเลือกการกระทำใหม่ (explore ability) ให้มากขึ้น ซึ่งสำหรับความหมายของมอนติคาร์โลในมุมมองของวิธีการของระบบที่มีความฉลาดกับระบบที่ไม่ฉลาด หมายถึงลักษณะของระบบมีการปรับปรุงค่าแอคชั่น- แวลู่ฟังก์ชันและลักษณะของระบบไม่มีการปรับปรุงค่าแอคชั่น-แวลู่ฟังก์ชันตามลำดับ โดยการปรับปรุงนั้น จะมีค่าที่แตกต่างกันตามค่าของ ϵ-กรีดี้ ทำการทดสอบ 2 กรณีดังต่อไปนี้

1. วิธีเลือกเส้นทางแบบมอนติคาร์โลแบบที่มีการปรับปรุงตนเอง ($\epsilon = 0.1$) คือการเลือกเส้นทางด้วยวิธีมอนติคาร์โลแบบมีการปรับปรุงค่า Q ฟังก์ชันที่ทุกเหตุการณ์ที่เปลี่ยนไป ที่ $\epsilon = 0.1$ ซึ่งจะทำให้ระบบมีความพยายามในการเลือกการกระทำใหม่อยู่ที่ 10% (ไม่มีแบบจำลองคาดการณ์)

2. วิธีเลือกเส้นทางแบบมอนติคาร์โลแบบที่มีการปรับปรุงตนเอง ($\epsilon = 0.4$) คือการเลือกเส้นทางด้วยวิธีมอนติคาร์โลแบบมีการปรับปรุงค่า Q ฟังก์ชันที่ทุกเหตุการณ์ที่เปลี่ยนไป ที่ $\epsilon = 0.4$ ซึ่งจะทำให้ระบบมีความพยายามในการเลือกการกระทำใหม่อยู่ที่ 40% (ไม่มีแบบจำลองคาดการณ์)

โดยสำหรับกรณีของทอพอโลยีแบบคงที่ เซนเซอร์โนดจะไม่มีการขยับเลยโดยจะแตกต่างกับกรณีของทอพอโลยีแบบไม่คงที่ ตรงที่มีการปล่อยอิสระให้เซนเซอร์โนดขยับได้มากขึ้นและสำหรับในกรณีสุดท้ายจะพิจารณากรณีที่ความสามารถในการส่งของโนดเพิ่มเติม ซึ่งแต่ละโนดมีโอกาสที่ไม่สามารถส่งผ่านข้อมูลต่อไปยังโนดรอบข้างได้ในช่วงเวลาที่พิจารณาอยู่ (iid process)

การตั้งค่าการทดลองในครั้งนี้ เกิดจากผลการทดลองในช่วงที่ระบบเรียนรู้แล้วโดยที่การให้ค่าความสำคัญของน้ำหนักแต่ละฟังก์ชันที่พิจารณาเป็นผลรางวัลออกมาเป็น 1/3 ทั้งหมดทุกฟังก์ชันตัวแปร ทั้งนี้ เพื่อให้ระบบสามารถพิจารณาทุกด้านได้อย่างมีประสิทธิภาพ ตามที่ได้พิจารณาความแตกต่างที่เกิดขึ้นจากในผลการทดลองข้างต้นที่ผ่านมา สำหรับผลการทดลองในหัวข้อนี้จึงมีการตั้งค่าของการทดลองให้มีค่าของเหตุการณ์ที่จะเกิดขึ้นในระบบทั้งหมดเท่ากับ 30000 ครั้ง โดยมีค่าของการจบสิ้นการทดสอบก็ต่อเมื่อ ค่าของพลังงานโนดใดโนดหนึ่งในระบบหมดลง โดยระดับการลดลงของพลังงานนั้นจะลดลงตามค่าการคำนวณที่เกิดขึ้นตามสมการที่ (3.3) สำหรับในช่วงของโอกาสที่โนดจะมีความผิดพลาดนั้นคือไม่สามารถส่งผ่านข้อมูลต่อไปยังโนดรอบข้างได้ในช่วงเวลาที่พิจารณาอยู่นั้น โนดแต่ละตัวกำหนดให้มีความน่าจะเป็นของโอกาสที่จะส่งสำเร็จอยู่ในช่วง 0.5 - 1



รูปที่ 4.18: การเปรียบเทียบสมรรถนะเชิงระบบ

1. การวิเคราะห์ผลการจำลองแบบของทอพอโลยีแบบคงที่

จากผลการทดลองดังรูปที่ 4.18 พบว่าในช่วงทอพอโลยีแบบคงที่นั้น เมื่อพิจารณาวิธีเลือกเส้นทางที่สั้นที่สุดจะเลือกจากเส้นทางที่สั้นที่สุดและมีค่าต่ำสุดนั้นหมายถึงความสามารถในการรักษาระบบเพื่อคงขีดความสามารถและความคงทนในการทำให้ระบบนั้นไม่หลุดออกจากกัน จะมีอยู่ในช่วงจำกัด ดังนั้น การเลือกเส้นทางที่สั้นที่สุด ไม่ใช่คำตอบที่ง่ายและดีเสมอไป เนื่องจากคุณภาพต้องแลกมาด้วยทักษะบางอย่างที่สูญเสียออกไป เช่น ข้อจำกัดขีดความสามารถทางการเชื่อมต่อสื่อสาร และการคงสภาพโครงข่ายในระบบสื่อสารสาธารณะ ที่จะลดลง เช่นเดียวกับในกรณีของทั้ง 3 กรณี ไม่รวมกรณีเลือกเส้นทางแบบมอนติคาร์โลแบบที่ไม่มีการปรับปรุงตนเอง ดังนั้นเมื่อพิจารณาอีกครั้งในช่วงของทอพอโลยีแบบคงที่ จะเห็นได้ว่าวิธีเลือกเส้นทางที่สั้นที่สุดจะมีค่าของอายุการใช้งานและรักษาระดับพลังงานในระบบที่มีค่าต่ำที่สุดและเมื่อพิจารณาที่กราฟของวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงานจะมีค่าสูงกว่าหรือมีความสามารถที่มากกว่าวิธีเลือกเส้นทางที่สั้นที่สุดอยู่ประมาณ 0.25% แม้กระนั้นค่าที่เพิ่มขึ้นดังกล่าวของวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดก็ยังมีนัยสำคัญที่น้อยมากเมื่อเทียบกับวิธีเลือกเส้นทางแบบสุ่มเอกรูป ที่มีเปอร์เซ็นต์ความสามารถของการยืดอายุการใช้งานได้เพิ่มขึ้น ถึง 2.72% เนื่องจากวิธีเลือกเส้นทางแบบสุ่มเอกรูปมีโอกาสในการเลือกเส้นทางที่สูงกว่าวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุด โอกาสที่สูงกว่าในที่นี้หมายถึงการพิจารณาเลือก

เส้นทางจากค่าสูงสุดของค่าต่ำสุดจะเป็นการพิจารณาเส้นทางที่แย่ที่สุดแล้วเลือกที่ดีที่สุดของกลุ่มนั้น และหากพิจารณาวิธีเลือกเส้นทางแบบมอนติคาร์โลแบบที่ไม่มีการปรับปรุงตนเองค่าที่ได้จะเพิ่มถึง 4.86% ทั้งหมดที่กล่าวมาคือมุมมองของระบบที่ไม่ฉลาด แต่ในมุมมองของระบบที่มีความฉลาดจะทำการทดสอบค่าที่นำมาใช้ปรับปรุงระบบโดยมีค่า ϵ-กรีดี้ = 0.1 และ 0.4 ในกรณีของกราฟที่ค่า ϵ-กรีดี้ ที่ 0.1 แม้ระบบจะสามารถปรับปรุงและพัฒนาตนแต่ในกรณีนี้ผลที่ได้รับคล่อนข้างต่ำกว่ามอนติคาร์โลแบบที่ไม่มีการปรับปรุงตนเอง ทั้งนี้เพราะว่าการปรับปรุงในระบบที่เกิดขึ้น ไม่มีการปรับปรุงและเปลี่ยนแปลงแบบทันทีทันใด (online) หรือจะกล่าวได้ว่า ทอพอโลยีของระบบเป็นแบบคงที่ การปรับค่า Q ใหม่ ในสภาวะที่ระบบดีอยู่แล้วทำให้ผลลัพธ์ที่ได้แย่ลง ส่วนผลสุดท้ายของทอพอโลยีแบบคงที่ ϵ ที่ 0.4 มีการพิจารณาแบบปรับปรุงทันทีทันใดจะมีประสิทธิภาพดีที่สุดโดยเพิ่มจากกรณีการเลือกเส้นทางที่สั้นที่สุด ถึง 8.32% ดังตารางที่ 4.3

ตารางที่ 4.3: จำนวนรอบสูงสุดของระบบที่ทำงานได้ในสภาวะทอพอโลยีแบบคงที่

avg.number of event	Shortest Path	Max-Min	Randomize	MC no update	MC $\epsilon = 0.1$	MC $\epsilon = 0.4$
	15388	15427	15807	16136	15997	16669
% improvement		0.25	2.72	4.86	3.96	8.32

2. การวิเคราะห์ผลการจำลองแบบของทอพอโลยีแบบไม่คงที่

สำหรับในส่วนของการทดสอบในส่วนที่สองที่เกิดขึ้นเมื่อเป็นกรณีของทอพอโลยีแบบไม่คงที่ตลอดเวลา โดยมีการจำลองการเหวี่ยงของโนดที่ลอยอยู่ในกระแสน้ำผ่านทุ่นลอยน้ำตลอดเวลาในช่วงที่พิจารณาตั้งแต่ t ถึง T ซึ่งมีการเหวี่ยงของโนดจากเดิมที่มีการพิจารณาแบบคงที่ไป ± 50 เมตร และจะทำให้เกิดการเปลี่ยนแปลงของการสร้างเส้นทางบ่อยครั้ง แต่ท้ายที่สุดวิธีการเลือกเส้นทางแบบมอนติคาร์โลจะเป็นตัวตัดสินใจถึงการเลือกเส้นทางที่ดีที่สุด ผ่านเส้นทางที่ได้ทำการหามาจากการสร้างเส้นทางในชั้น MAC นั่นเอง ผลการทดลองดังรูปที่ 4.18 และ ตารางที่ 4.4 แสดงให้เห็นความแตกต่างว่าในกรณีของทอพอโลยีแบบไม่คงที่นั้นจะมีเปอร์เซ็นต์ที่เปลี่ยนแปลงเพิ่มขึ้นมากน้อยแต่อย่างใด โดยจะพบว่ากรณีนี้ จะมีค่าของจำนวนรอบอายุที่ทำงานได้เพิ่มขึ้นเมื่อเทียบออกมาเป็นเปอร์เซ็นต์ จะเห็นความสามารถของวิธีการเลือกเส้นทางแบบมอนติคาร์โลเพิ่มขึ้นอีกอย่างเด่นชัดขึ้นเมื่อทอพอโลยีมีการเปลี่ยนแปลง เหตุผลที่วิธีการเลือกเส้นทางแบบมอนติคาร์โลจะทำงานได้ดียิ่งขึ้นเมื่ออยู่ในสภาวะที่ระบบที่มีการเปลี่ยนแปลงตลอดเวลานั้น เพราะว่าการมีค่าความน่าจะเป็นของโอกาสในการเลือกการกระทำใหม่สูงๆ นั้นจะทำให้เกิดการเข้าถึงทุกๆความเป็นไปได้ของทุกคู่สถานะและการกระทำทั้งหมด ซึ่งการที่เข้าถึงคู่สถานะและการกระทำทั้งหมดหมายถึงการที่ฟังก์ชัน Q มีความรู้หรือมีข้อมูลอย่างสมบูรณ์ของในระบบ

ตารางที่ 4.4: จำนวนรอบสูงสุดของระบบที่ทำงานได้ในสภาวะทอพอโลยีแบบไม่คงที่

avg.number of event	Shortest Path	Max-Min	Randomize	MC no update	MC $\epsilon = 0.1$	MC $\epsilon = 0.4$
	13253	13421	13787	14211	15543	15805
% improvement		1.27	4.03	7.23	17.28	8.32

3. การวิเคราะห์ผลการจำลองแบบของทอพอโลยีแบบไม่คงที่และโนดปัญหาอยู่

ในกรณีสุดท้ายที่ทำการทดสอบได้ทำการพิจารณาถึงการนำมาประยุกต์ใช้ในสภาวะความเป็นจริง ในระบบเตือนอุทกภัยว่า อาจจะมีบางจังหวะที่ทำให้ค่าของการส่งข้อมูลไม่อาจเป็นไปได้ โดยกำหนดให้ความน่าจะเป็นในการส่งสำเร็จของแต่ละเซนเซอร์โนดมีค่า 0.8 ดังนั้น ในการทดลองสุดท้ายนี้จึงได้ใส่โอกาสที่โนดจะเกิดความไม่คงตัวหรือมีข้อจำกัดบางประการอย่างเช่น การสุ่มให้แต่ละเซนเซอร์โนดไม่สามารถส่งข้อมูลได้ ดังนั้นจากผลการทดลองดังรูปที่ 4.18 และ ตารางที่ 4.5 โดยจะเห็นได้ว่า ระบบที่การเปลี่ยนแปลงสูงๆ เป็นการยืนยันว่า วิธีการเลือกเส้นทางแบบมอนติคาร์โลที่นำเสนอนี้สามารถนำมาใช้งานได้ในสถานการณ์ของการเตือนอุทกภัย

สำหรับการทดลองในช่วงที่มีการปรับปรุงระบบ อาจจะพบว่าด้วยวิธีเลือกเส้นทางแบบสุ่มเอกรูปให้ผลของฟังก์ชันผลรางวัลที่ดีกว่า ในกรณีวิธีเลือกเส้นทางที่สั้นที่สุดและวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน แต่กระบวนการเลือกเส้นทางแบบสุ่มเอกรูปนี้เอง อาจมีปัญหาในส่วนที่ไม่สามารถหลีกเลี่ยงปัญหาเรื่องของการชนกันของข้อมูลได้ หรือในบางครั้งเส้นทางที่ถูกเลือก อาจจะเป็นเส้นทางที่เกิดปัญหาที่ระบบใช้งานอยู่ ซึ่งหากดูในแง่มุมของค่าพลังงานที่เหลืออยู่ในระบบนี้เพียงแง่มุมเดียว วิธีเลือกเส้นทางแบบสุ่มเอกรูปอาจจะเป็นทางเลือกที่ดี แต่หากเทียบประสิทธิภาพในด้านของการส่งแพ็กเกต (packet deliverable) แล้ว จะพบว่ากรณีของการเลือกเส้นทางแบบสุ่มเอกรูปเป็นกระบวนการหาเส้นทางที่แย่ที่สุด เพราะไม่สามารถหลีกเลี่ยงการชนกันของแพ็กเกตข้อมูลได้

ตารางที่ 4.5: จำนวนรอบสูงสุดของระบบที่ทำงานได้ในสภาวะทอพอโลยีแบบไม่คงที่และมีโนดปัญหา

avg.number of event	Shortest Path	Max-Min	Randomize	MC no update	MC $\epsilon = 0.1$	MC $\epsilon = 0.4$
	13112	13247	13341	14107	15378	15681
% improvement		1.03	1.75	7.59	17.28	19.59

4.4.4 การคำนวณความซับซ้อนของระบบ

ในหัวข้อย่อยนี้จะนำเสนอการวิเคราะห์การคำนวณความซับซ้อนของระบบโดยมีการเปรียบเทียบกับเทคนิคการหาคำตอบ 5 วิธีด้วยกัน ประกอบด้วย วิธีเลือกเส้นทางที่สั้นที่สุด วิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน วิธีเลือกเส้นทางแบบสุ่มเอกรูป วิธีเลือกเส้นทางแบบมอนติคาร์โลทั่วไปและวิธีการมอนติคาร์โลแบบปรับตัวที่นำเสนอหรือวิธีฮิวริสทิคมอนติคาร์โล (heuristic monte

carlo) ในเทอมของจำนวนของปริมาณการคำนวณที่ต้องการใน 1 เหตุการณ์ในเหตุการณ์ของแจ้งเตือนอุทกภัยหรืออีกนัยหนึ่งคือกระบวนการการส่งข้อมูลจากต้นทางไปปลายทางเสร็จสิ้นอย่างสมบูรณ์ โดยกำหนดให้ ขนาดของปริภูมิสถานะถูกกำหนดโดย $|\mathcal{X}|$ และขนาดของปริภูมิการกระทำถูกกำหนดโดย $|\mathcal{A}|$ ดังนั้นสำหรับวิธีการต่างๆ เราสามารถ แยกปริมาณความต้องการในการเก็บค่าของตัวแปรได้ดังต่อไปนี้

ตารางที่ 4.6: การคำนวณความซับซ้อนของระบบ

	Storage (Bytes)	Computational Complexity (Space)
Shortest path	3640	$O(\|A\|)$
Max-min	7280	$O(\|A\|)$
Uniform random	8	$O(a)$
Classic monte carlo	16320	$O(\|\mathcal{X}\|)$
Adaptive monte carlo	16320	$O(\|\mathcal{X}\|)$

ดังแสดงในตารางที่ 4.6 จะเห็นว่าในปริมาณขนาดของระบบที่พิจารณา จะมีความแตกต่างของปริมาณที่ใช้ในการเก็บค่าไม่เท่ากัน เช่นสำหรับค่าของความจุของวิธีเลือกเส้นทางที่สั้นที่สุดจะมีค่าน้อยกว่าวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน ทั้งนี้เพราะวิธีการของวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงานนั้นมีการคำนวณสองรอบเท่ากับจำนวนของการกระทำหรือเส้นทาง แต่สำหรับในการคำนวณโดยการใช้ฟังก์ชันบิ๊กโอ (big O function) สำหรับวิธีเลือกเส้นทางที่สั้นที่สุดนั้น มีค่าเท่ากับ $|A|$ ส่วนสำหรับ กรณีของวิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงานนั้นจะมีค่าเท่ากับ $|A| + |A| = 2|A| = |A|$ ดังนั้นจะพบว่าทั้งสองฟังก์ชันนี้จะขนาดเพิ่มขึ้นพร้อมกันขึ้นกับขนาดของการกระทำเป็นหลัก และสำหรับปริมาณความจุที่ใช้น้อยที่สุดนั้น คือ การใช้วิธีเลือกเส้นทางแบบสุ่มเอกรูป โดยจะพบว่าใช้เพียงปริมาณสเกลาร์ a ก็เพียงพอต่อการเก็บข้อมูล แต่อย่างไรก็ตามการใช้วิธีเลือกเส้นทางแบบสุ่มเอกรูปนั้น ไม่สามารถควบคุมเงื่อนไขใดๆ ได้ ทั้งนี้เพราะการเลือกนั้นเป็นอิสระอย่างสมบูรณ์ต่อตัวระบบดังแสดงในหัวข้อย่อยช่วงที่มีการปรับปรุงระบบ (adaptive period) สำหรับกรณีที่ใช้วิธีการมอนติคาร์โลในการควบคุมนั้น ไม่ว่าจะเป็นแบบทั่วไปหรือแบบปรับตัวนั้นจะมีปริมาณของการเก็บข้อมูล เท่ากับ $|\{X\}| = |\mathcal{S}| \times |\mathcal{S}|$ โดยที่ $\mathcal{S}$ คือขนาดการทำการแบ่งนับของปริภูมิสถานะ (quantised state) เพราะทั้งนี้เนื่องจาก การปรับปรุงของค่าฟังก์ชัน Q นั้นเป็นเพียงการปรับปรุงในการคำนวณรอบสุดท้าย จึงไม่ส่งผลกระทบต่อปริมาณของการเก็บค่าฟังก์ชัน Q

4.5 สรุป

ในบทนี้นำเสนอการทดสอบการนำวิธีการมอนติคาร์โลไปใช้ในสถานการณ์จริง โดยวิธีการมอนติคาร์โลที่ได้นำเสนอในบทนี้ได้ทำการสร้างกรอบความคิดเชิงคณิตศาสตร์ขึ้นมาใหม่ที่มีตัวแปรสถานะสองตัว และมีการทำกระบวนการแบ่งกระจายภาระงานผ่านฟังก์ชันผลรางวัล โดยได้ทำการสร้างแบบจำลองโครงข่ายแม่น้ำขึ้นเพื่อใช้ในการทดสอบ โดยฟังก์ชันผลรางวัลในบทนี้จะมีความแตกต่างจากบทที่ 3 ที่นำเอาเงื่อนไขบังคับ (constraints) เดิมมาเปลี่ยนเป็นการทำกระบวนการแบ่งกระจายภาระงานแทน โดยมีการนำวิธีการมอนติคาร์โลไปฝึกในช่วงที่ระบบกำลังเรียนรู้ (learning period)

ก่อนที่จะนำไปใช้จริงในช่วงที่มีการปรับปรุงระบบ (adaptive period) นอกจากนั้นงานวิจัยนี้ ยังได้พิจารณาเพิ่มเติมในส่วนของผลกระทบของฟังก์ชันผลรางวัลของช่วงที่ระบบเรียนรู้แล้ว (learned period)

การเปรียบเทียบสมรรถนะเชิงระบบนั้นได้เลือกวิธีที่ใช้ในการเลือกเส้นทางเหล่านี้มาเปรียบเทียบ ได้แก่ วิธีเลือกเส้นทางที่สั้นที่สุด วิธีเลือกเส้นทางจากค่าสูงสุดของค่าต่ำสุดพลังงาน วิธีเลือกเส้นทางแบบสุ่มเอกรูป และวิธีการมอนติคาร์โลทั้งแบบที่ไม่มีการปรับปรุง (classic) และมีการปรับปรุง (adaptive) ค่าของ Q ฟังก์ชันและพบว่าในกรณีที่มีการปรับปรุงนั้น วิธีการมอนติคาร์โลสามารถเพิ่มระยะเวลาการทำงานของทั้งระบบเมื่อเทียบกับในกรณีอย่างง่ายแบบวิธีเลือกเส้นทางที่สั้นที่สุดนั้นสูงถึง 19.59% และการพิจารณาสุดท้ายคือความซับซ้อนของระบบซึ่งจะพบว่าวิธีการมอนติคาร์โลเองมีความซับซ้อนสูงกว่าในกรณีอื่นๆ แต่ด้วยเหตุผลของการสูญเสียพลังงานของเซนเซอร์โนด การส่งแพ็กเกตข้อมูลหนึ่งครั้งจะมีการสูญเสียพลังงานมากกว่าพลังงานที่สูญเสียอันเนื่องมาจากการใช้ประมวลผลของเซนเซอร์โนดมาก [25] ดังนั้นการใช้วิธีการมอนติคาร์โลยังคงเป็นทางเลือกในการตัดสินใจเลือกเส้นทางที่ดีกว่าแบบอื่นๆ นอกจากนี้วิธีการมอนติคาร์โลยังสามารถควบคุมการทำกระบวนการแบ่งกระจายภาระงานที่ทำให้การใช้งานของโนดทั้งหมดในระบบ โดยภาพรวมอยู่ในสภาพที่ดีกว่าวิธีอื่นๆ ที่ไม่สามารถควบคุมทั้งในส่วนของพลังงาน ค่าอายุการใช้งานและค่าเร็พพิวเทชันของระบบได้อย่างมีประสิทธิภาพ

บทที่ 5

บทสรุปและข้อเสนอแนะ

5.1 บทสรุปผลการวิจัย

งานวิจัยนี้นำเสนอกรอบความคิดเชิงคณิตศาสตร์สำหรับวิธีการมอนติคาร์โล เพื่อไปประยุกต์ใช้กับระบบการเตือนอุทกภัยที่สามารถทำกระบวนการแบ่งกระจายภาระงานเพื่อยืดอายุการใช้งานของเซนเซอร์โนดในภาพรวมของทั้งระบบได้ โดยงานวิจัยนี้สามารถแบ่งออกได้เป็นสองส่วนหลักคืองานในส่วนของการเริ่มพิจารณาการนำวิธีการมอนติคาร์โลและค่าเร็พพิวเทชันของเส้นทางในโครงข่ายอย่างง่ายและแบบจำลองคณิตศาสตร์อย่างง่ายซึ่งได้อธิบายในบทที่ 3 สำหรับงานวิจัยในบทที่ 4 นั้นเป็นการต่อยอดขึ้นมาสู่โครงข่ายที่มีคุณลักษณะใกล้เคียงกับของจริงโดยผลการทดลองนี้ได้อ้างอิงการวัดค่าจริงจากเซนเซอร์โนดระหว่างริมฝั่งแม่น้ำ และ เพื่อให้การทำงานของกระบวนการแบ่งกระจายภาระงานเป็นไปได้อย่างมีประสิทธิภาพ ในบทที่ 4 จึงได้มีการปรับปรุงกรอบความคิดเชิงคณิตศาสตร์ให้อยู่ในรูปแบบที่ดีขึ้นและเหมาะสมด้วยการเลือกใช้ค่าพลังงานประจำโนดและค่าเร็พพิวเทชันมาเป็นตัวแปรสถานะ โดยสำหรับค่าของกระบวนการแบ่งกระจายภาระงาน พิจารณาด้วยค่าของ 3 ฟังก์ชันผลรางวัล โดยสิ่งที่น่าสนใจของแต่ละบทในงานวิจัยเล่มนี้สามารถสรุปได้ดังนี้

5.1.1 สรุปผลการวิจัยในการทดลองกับโครงข่ายขนาดเล็ก

จากที่ได้กล่าวมาแล้วข้างต้น ในบทที่ 3 ได้นำเสนอการนำวิธีการมอนติคาร์โลมาใช้ในสถานการณ์การเตือนอุทกภัยโดยนำเร็พพิวเทชันของเส้นทางมาพิจารณาร่วมในเงื่อนไข เพื่อทำให้ระบบหลีกเลี่ยงโอกาสที่จะเจอโนดที่ผิดพลาดได้ โดยงานวิจัยในบทนี้ได้พิจารณาความเป็นไปได้เบื้องต้นในการประยุกต์ใช้งานด้วยการใช้ค่าของตัวแปรสถานะแบบง่าย รวมถึงการใช้ฟังก์ชันผลรางวัลในเบื้องต้นเพื่อลดความซับซ้อนในการใช้งาน แต่อย่างไรก็ตาม ถึงแม้จะเป็นเพียงผลรางวัลเบื้องต้น แต่ก็ได้มีการพิจารณารวมถึงสามตัวแปรฟังก์ชันผลรางวัลที่เป็นค่าของพลังงานที่เหลือทั้งหมดในระบบ ค่าอายุการใช้งานและค่าเร็พพิวเทชัน เข้ามาพิจารณาในลักษณะของเงื่อนไขบังคับแทน ซึ่งในบทที่ 4 ค่าทั้งหมดนั้นจะมีการเปลี่ยนจากเงื่อนไขบังคับเป็นฟังก์ชันกระบวนการแบ่งกระจายภาระงานที่รวมทั้ง 3 องค์ประกอบเข้ามาพิจารณาร่วมกัน โดยผลการทดลองเบื้องต้นที่ได้มานี้พบว่า ด้วยวิธีการของมอนติคาร์โลที่นำเสนอ สามารถทำงานได้ดีกว่าวิธีอื่นๆ ด้วยขีดความสามารถที่สามารถควบคุมตามเงื่อนไขจำกัดที่ต้องการได้และความซับซ้อนของการใช้งานวิธีการมอนติคาร์โลในเร็พพิวเทชันของเส้นทางนั้นก็ไม่ได้แตกต่างไปจากกรณีอื่นที่เลือกใช้ ดังนั้นการเลือกใช้งานวิธีการมอนติคาร์โลในสถานการณ์การเตือนอุทกภัยก็น่าจะสามารถทำงานได้จากผลการทดลองในบทที่ 3

5.1.2 สรุปผลการวิจัยในการทดลองกับโครงข่ายขนาดใหญ่

จากการทดลองใช้งานเบื้องต้นของวิธีการมอนติคาร์โลกับค่าเร็พพิวเทชันของเส้นทาง ในบทที่ 3 กับโครงข่ายอย่างง่าย ดังนั้นการทดลองที่จะเกิดขึ้นในบทนี้ รวมถึงกรอบความคิดเชิงคณิตศาสตร์

เดิมที่เป็นรูปแบบอย่างง่าย จะได้นำมาปรับเปลี่ยนเพื่อทำให้ระบบออกมาดีกว่าเดิม ด้วยการเริ่มต้นพิจารณาจากการปรับขยายตัวแปรสถานะจากอย่างง่ายเป็นตัวแปรสถานะของระบบที่มีการพิจารณาถึงค่าของคุณลักษณะของระบบ อันได้แก่ ค่าพลังงานประจำโนด และค่าความน่าเชื่อถือของโนด และสำหรับฟังก์ชันผลรางวัลในบทนี้ ได้รวมสามฟังก์ชันผลรางวัลเข้ามาเพื่อใช้ในการทำกระบวนการแบ่งกระจายภาระงานและยิ่งไปกว่านั้น การปรับเปลี่ยนตัวแปรสถานะในครั้งนี้ได้ใช้วิธีการแบ่งระดับเพื่อใช้ในการสร้างความแตกต่างกันของสถานะในระดับต่างๆที่เกิดขึ้นของระบบที่แต่ละหน่วยเวลาที่ต่างกันและใช้การตัดสินใจผ่านแอคชั่น-แวลู่ฟังก์ชันหรือค่า Q ฟังก์ชันที่บรรจุค่าของตัวแปรสถานะที่เวลานั้นๆ

กระบวนการทดสอบสำหรับความเป็นไปได้ในการนำวิธีการมอนติคาร์โลมาประยุกต์ใช้ ถูกแบ่งออกเป็นสามช่วงเพื่อใช้ในการพิจารณา ได้แก่ช่วงที่ระบบกำลังเรียนรู้ ช่วงที่ระบบเรียนรู้แล้วและช่วงที่มีการปรับปรุงระบบตามลำดับ เพื่อใช้สำหรับการเรียนรู้พฤติกรรมจริงของระบบ ปรับเปลี่ยนค่าการทดสอบเพื่อดูพฤติกรรมการเปลี่ยนแปลงหลังจากระบบเข้าสู่สมดุลและใช้กับระบบจริงที่มีความผันผวนสูง ดังนั้นมุมมองของงานวิจัยนี้ได้ศึกษาความเป็นไปได้ก่อนที่จะนำไปใช้ในสถานการณ์จริง โดยมีการเปรียบเทียบสมรรถนะเชิงระบบกับกระบวนการค้นหาเส้นทางอื่นๆ โดยที่กระบวนการมอนติคาร์โลกับเร็พพิวเทชันของเส้นทางที่นำเสนอนั้นสามารถรักษาความสามารถโดยรวมของระบบในเชิงพลังงานรวมก่อนที่ระบบจะไม่สามารถส่งข้อมูลได้อีก การใช้กระบวนการเรียนรู้จำเป็นที่ต้องใช้หน่วยความจำในการเก็บข้อมูล ทั้งนี้รวมถึงการพิจารณาความซับซ้อนของระบบ โดยจะเห็นได้ว่าในกระบวนการอื่นๆที่นำมาเปรียบเทียบด้วยนั้น อาจจะมีความซับซ้อนต่ำกว่ากระบวนการที่ได้นำเสนอในวิทยานิพนธ์นี้ แต่อย่างไรก็ตามความซับซ้อนที่ต่ำ ไม่ได้การันตีความสามารถในการควบคุมฟังก์ชันผลรางวัลเพื่อให้บรรลุวัตถุประสงค์ที่ต้องการได้ ดังนั้นงานวิจัยนี้จึงมีขีดความสามารถที่จะนำไปใช้ในทางปฏิบัติได้

5.2 ข้อเสนอแนะและงานวิจัยในอนาคต

5.2.1 การติดตั้งวิธีการมอนติคาร์โลเพื่อใช้ในทางปฏิบัติ

เพื่อทดสอบวิธีการมอนติคาร์โลนั้น ในทางปฏิบัติแล้วมีกระบวนการมาตรฐานด้วยกันทั้งสองวิธี ได้แก่ การจำลองระบบแบบเครื่องจักรเสมือนหรืออีมูเลเตอร์ (emulator) และ การใช้ระบบทดสอบ (testbed) เป็นวิธีทดสอบหลัก ปัจจุบันกระบวนการจำลองแบบอีมูเลเตอร์นั้น จะใช้งานโปรแกรมมาตรฐานการจำลองโครงข่าย 3 (network simulator 3 : NS-3) เพื่อใช้จำลองชั้นการสื่อสารต่างๆในโครงข่ายเซนเซอร์ไร้สายได้ และยิ่งไปกว่านั้นโพรโทคอล ที่ใช้ในการควบคุมชั้นของการสื่อสาร เป็นโพรโทคอลมาตรฐานที่ยอมรับกันอย่างแพร่หลายในวงการวิจัยและสำหรับในส่วนของการทดสอบแบบระบบทดสอบนั้นเป็นกระบวนการสุดท้ายเพื่อใช้ในการปฏิบัติจริงและมักจะเป็นงานวิจัยที่มีนัยสำคัญต่อสังคม โดยทั้งนี้จะเป็นผลมาจากการมีรากฐานทฤษฎีที่ดีและการวางแผนการทดสอบที่ดีตามลำดับ

5.2.2 โครงข่ายที่มีขนาดใหญ่และมีจำนวนโนดเป็นปริมาณมาก

สำหรับการใช้งานวิจัยนี้ในสถานการณ์จริงอาจมีมูลเหตุปัจจัยความแตกต่างทางกายภาพตามแต่ลักษณะภูมิประเทศที่เปลี่ยนไป หากโครงข่ายใหญ่มากจะส่งผลกระทบต่อการคำนวณในการเลือกเส้นทางในแต่ละรอบอย่างชัดเจนเพราะขนาดของระบบที่ใหญ่จำเป็นที่จะต้องใช้เวลาในการสร้าง

เส้นทางรวมถึงการหาเส้นทางที่ดีที่สุดอีกด้วย สำหรับในงานวิจัยนี้ หากมีระบบที่มีจำนวนโนดเพิ่มขึ้น ปริมาณความซับซ้อนของระบบก็จะเพิ่มขึ้นเป็นแบบไม่เชิงเส้นซึ่งจะส่งผลให้เกิดปัญหาใหญ่ที่สุดตามมาเนื่องจาก ในโครงข่ายเซนเซอร์ไร้สายนั้นการรักษาพลังงานเป็นหัวใจสำคัญที่สุด เพื่อที่จะรักษาระบบให้สามารถทำงานได้เป็นระยะเวลานานที่สุด แต่หากการสร้างเส้นทางนั้นเกิดขึ้นตลอดเวลาและ มีขนาดใหญ่เพื่อที่จะใช้ในการเก็บข้อมูลไว้ในหน่วยข้อมูลสำรอง อาจจะเป็นทางเลือกที่จะต้องมาพิจารณาอีกครั้ง ดังนั้นการพิจารณาสร้างปริภูมิสถานะใหม่ เพื่อให้รองรับกับขนาดที่เปลี่ยนไปของระบบ นับเป็นอีกเรื่องหนึ่งที่อาจจะนำมาพัฒนาให้เกิดขึ้นในลำดับถัดไป

5.2.3 การสร้างเส้นทางสื่อสารด้วยโพรโทคอลการจัดสรรเส้นทางแบบผสม (Hybrid Protocol)

งานวิจัยสำหรับการสร้างโครงข่ายไร้สายปัจจุบันนั้น นอกจากจะต้องพิจารณาในส่วนของการควบคุมการใช้พลังงานในการสื่อสารในแต่ละครั้งแล้ว ในทางปฏิบัติจำเป็นที่จะต้องควบคุมในส่วนของการสร้างความน่าเชื่อถือให้เกิดขึ้นในระบบ ดังนั้นการสร้างเส้นทางสื่อสารด้วยวิธีแบบผสมหรือไฮบริดจ์ จึงเป็นหนึ่งในทางเลือกซึ่งเกิดจากการรวมเอาข้อดีของการจัดเส้นทางสองประเภท ทั้งวิธีการจัดสรรเส้นทางโปรแอคทีฟและวิธีการจัดสรรเส้นทางรีแอคทีฟเข้ามาผสมผสานใช้งานรวมกัน โดยการใช้วิธีการจัดสรรเส้นทางโปรแอคทีฟที่ถูกนิยามลักษณะการจัดสรรเส้นทางภายในโซนผ่านวิธีการไฮบริดจ์นั้น สามารถที่จะการันตีความน่าเชื่อถือของเส้นทางในการส่งข้อมูลไปยังโนดปลายทางหรือแอคเตอร์โนด ซึ่งเป็นส่วนสำคัญที่สุดในการรวบรวมข้อมูลรอบข้างและในขณะเดียวกันเซนเซอร์โนดที่อยู่กลางแม่น้ำเอง จะถูกพิจารณาเป็นการจัดสรรเส้นทางแบบรีแอคทีฟ ซึ่งสามารถแก้ไขปัญหาเมื่อในบางสถานะเกิดการเปลี่ยนแปลงบ่อยและยังสามารถยืดอายุการใช้งานของเซนเซอร์โนดที่ทำงานอยู่ในแม่น้ำได้ เนื่องจากการทำงานด้วยการจัดสรรเส้นทางแบบรีแอคทีฟนั้นมีลักษณะเฉพาะตัวในด้านการสร้างเส้นทางเมื่อมีความต้องการที่จะส่งข้อมูลเกิดขึ้นเท่านั้น ทำให้พลังงานของโนดแต่ละตัวในระบบจะสูญเสียน้อยกว่าเมื่อเทียบกับในบริเวณที่มีการจัดสรรเส้นทางแบบโปรแอคทีฟ ดังนั้นด้วยเทคนิคการรวมกันของวิธีการจัดสรรเส้นทางทั้งสองตัว จึงเป็นหัวข้อที่น่าสนใจอีกหัวข้อหนึ่งที่สามารถทำการวิจัยต่อไปได้

5.2.4 ผลกระทบของแอคเตอร์โนดต่อระบบ

นอกจากในส่วนของการจัดสรรเส้นทางแบบผสมหรือไฮบริดจ์จะเป็นงานที่น่าสนใจแล้ว ในทางปฏิบัติสำหรับการติดตั้งระบบด้วยวงเงินจำกัด เพื่อให้มีประสิทธิภาพเป็นส่วนสำคัญที่แปรผันได้ แต่ในส่วนของการจัดสรรเส้นทางแบบผสมหรือไฮบริดจ์นั้น ความน่าสนใจโดยเฉพาะที่ตัวของแอคเตอร์โนดนั้น พบว่าแอคเตอร์โนดเอง เป็นโนดที่มีความเปราะบาง (vulnerable) เพราะนอกจากจะต้องรับข้อมูลจากเซนเซอร์โนดรอบข้างแล้ว ยังจะต้องรักษาความสามารถในการส่งต่อข้อมูลอีกด้วย ดังนั้นหากพิจารณาผลกระทบของแอคเตอร์โนดต่อการตัดสินใจต่อระบบ นับว่าเป็นหัวข้อที่มีความสำคัญอีกประการหนึ่ง ทั้งนี้ในการพิจารณาผลกระทบอาจรวมถึงความสามารถในการส่งข้อมูล การเลือกโนดบางตัวในระบบแม่น้ำที่ไม่สามารถชาร์ทพลังงานหรือยากต่อการเข้าถึงเพื่อปรับปรุงและซ่อมแซมโนดนั้นๆ ปัญหาทางกายภาพของเซนเซอร์ อาทิเช่นความร้อนหรือความผิดพลาดที่ไม่อาจคาดเดา ซึ่งทั้งหมดนี้เป็นหัวข้อในการพิจารณาปัญหาที่ไม่เพียงเฉพาะแอคเตอร์โนด แต่ในบางครั้งอาจรวมถึงเซนเซอร์โนดด้วย

5.2.5 กระบวนการเรียนรู้แบบอื่นในสภาวะปรับปรุงระบบ (adaptive)

การเรียนรู้แบบเสริมแรงหรืออินฟอร์สเมนท์เลิร์นนิ่ง (reinforcement learning) นั้นเป็นงานวิจัยที่อาศัยการเรียนรู้จากประสบการณ์ตรงที่มีกับระบบ แต่ในงานวิจัยนี้เลือกกระบวนการเรียนรู้ที่มีพื้นฐานในลักษณะของการหาคำตอบแบบการหาจุดที่ดีที่สุด (global optimal) มาใช้ในระบบการเตือนอุทกภัยนั้นด้วยสภาวะพื้นฐานของระบบ เป็นระบบที่มีความแปรผันเชิงเวลาสูง เนื่องจากสภาพของแม่น้ำที่อาจมีการเปลี่ยนแปลงได้ตลอดเวลา ดังนั้นวิธีการอื่นๆ ของการเรียนรู้แบบเสริมแรงที่นิยมนำมาใช้กับสภาพระบบที่มีความแปรผันเชิงเวลาสูง ยกตัวอย่างเช่นการเรียนรู้แบบ Q ที่ใช้ลักษณะการทำนายล่วงหน้า เพื่อหาจุดสมดุลที่ดีที่สุดต่อการเปลี่ยนแปลงเชิงเวลาจึงน่าจะเป็นทางเลือกที่เหมาะสม หรือการนำวิธีการเรียนรู้ของเครื่อง (machine learning) แบบอื่น มาใช้งานก็สามารถทำได้ แต่จะต้องพิจารณาถึงความซับซ้อนที่เกิดขึ้นในระบบด้วยเนื่องจากจะส่งผลกระทบโดยตรงต่อการใช้พลังงานในระบบนั่นเอง

5.2.6 การเลือกใช้การเรียนรู้แบบเสริมแรงชนิดอื่นใน สำหรับระบบที่มีการเปลี่ยนแปลงสูง

ในวิทยานิพนธ์ฉบับนี้ได้นำเสนอการประยุกต์ใช้การเรียนรู้แบบเสริมแรงด้วยวิธีการมอนติคาร์โล ซึ่งอาศัยหลักการกระทำซ้ำและปรับปรุงนโยบาย เพื่อใช้ในการตัดสินใจในกระบวนการตัดสินใจเลือกเส้นทางที่ดีที่สุดในระบบเตือนภัยอุทกภัยจะพบว่าการนำวิธีการมอนติคาร์โลมาใช้ในระบบที่มีการเปลี่ยนแปลงสูงนั้นในบางครั้งระบบไม่สามารถทำงานได้อย่างมีประสิทธิภาพ ทั้งนี้เนื่องจากการปรับปรุงค่าของนโยบายนั้นจะเกิดจากผลการกระทำที่เกิดขึ้นก่อนหน้า จึงทำให้การเลือกการกระทำผ่านตัวกระทำการตัดสินใจของระบบ (agent) นั้นอยู่ในรูปแบบตามหลังเหตุการณ์ที่เกิดขึ้นนั่นเอง ดังนั้นการพิจารณานำการเรียนรู้แบบเสริมแรงที่มีความสามารถในการคาดเดาเหตุการณ์ล่วงหน้าจากประสบการณ์ที่เกิดขึ้นและสามารถปรับตัวเข้ากับระบบที่มีการเปลี่ยนแปลงสูงนั้นควรทีจะนำมาพิจารณาเพื่อใช้ในการปรับปรุงนโยบายของการเลือกให้เข้ากับลักษณะของระบบเตือนภัยอุทกภัยมากขึ้น ตัวอย่างอัลกอริธึมที่นิยมนำมาใช้ในระบบที่มีการเปลี่ยนแปลงสูงคือการเรียนรู้แบบคิว (Q-learning) ที่ใช้หลักการของสมการเบลล์แมน (Bellman's equation) และการคาดเดาสถานการณ์ที่จะเกิดล่วงหน้าก่อนที่จะกระทำการตัดสินใจลงไปดังนั้นด้วยวิธีการเรียนรู้แบบคิวจะมีนัยสำคัญแตกต่างจากวิธีการมอนติคาร์โล เพราะทำการคาดเดาและตัดสินใจล่วงหน้า แต่ในขณะที่มอนติคาร์โลนั้นจะเก็บข้อมูลความผิดพลาดมาปรับปรุงใหม่ ดังนั้นจะเห็นได้ว่าการเรียนรู้แบบคิวจะมีการคาดเดาล่วงหน้าซึ่งในบางครั้งการคาดเดาล่วงหน้าอาจจะผิดพลาดแต่ในท้ายที่สุดหากพิจารณาที่ระยะยาวในระบบที่มีการเปลี่ยนแปลงสูงแล้วการเรียนรู้แบบคิวจะมีความสามารถในการทำงานที่ดีกว่าการใช้วิธีการมอนติคาร์โลนั่นเอง

รายการอ้างอิง

[1] Paul, N.R., Tripathy, L., and Mishra, P.K. Analysis and improvement of DSDV protocol. International Journal of Computer Science Issues 8, 5. (September 2011) : 408-410.

[2] De-Rango, F., and Fotino, M. Energy efficient OLSR performance evaluation under energy aware metrics. Proceedings of Symposium (SPECTS), pp. 193–198. July 2009. Istanbul : 2009.

[3] Yu-hong, L., Jian-xin, W., and Song-qiao, C. An energy-efficient AODV routing protocol based on link stability. Journal of Circuits and Systems 6. (2008) : 141–147.

[4] Ihbeel, A.A., and Sigiuk, H.I. Performance evaluation of dynamic source routing protocol on WSN. International Journal of Computing and Digital Systems 1. (2012) : 19–24.

[5] Servetto, S.D., and Barrenechea, G. Constrained random walks on random graphs: routing algorithms for large scale wireless sensor networks. Proceedings of ACM International Workshop on Wireless Sensor Networks and Applications (WSNA), pp. 12-21. New York : 2002.

[6] Krishnaveni, P., and Nasrin, S. Energy balanced shortest path routing for wireless sensor networks. International Journal of Engineering Research Technology 2, 6. (June 2013) : 397-403.

[7] Haider, R., Javed, M.Y., and Khattak, N.S. Design and implementation of energy aware algorithm using greedy routing for sensor networks. International Journal of Security and its Applications 2, 2. (April 2008) : 71-86.

[8] Luo, D., Zhu, X., Wu, X., and Chen, G. Maximizing lifetime for the shortest path aggregation tree in wireless sensor networks. Proceedings of INFOCOM, pp. 1566-1574. April 2011. Shanghai : 2011.

[9] Marin-Perianu, M., and Havinga, P. D-FLER–a distributed fuzzy logic engine for rule-based wireless sensor networks. Proceedings of Symposium (SPECTS), pp. 86-101. November 2007. Tokyo : 2007.

[10] Cheng, D., Xun, Y., Zhou, T., and Li, W. An energy aware ant colony algorithm for the routing of wireless sensor networks. Proceedings of Intelligent Computing and Information Science Communications in Computer and Information Science, pp. 395-401. January 2011. Chongqing : 2011.

[11] Nehra, N.K., Kumar, M., and Patel, R.B. Neural network based energy efficient clustering and routing in wireless sensor networks. Proceedings of Networks and Communications (NETCOM), pp. 27-29. December 2009. Chennai : 2009.

[12] Elshqeirat, B., Sieteng, S., Lazarescu, M., and Rai, S. Dynamic programming for minimal cost topology with two terminal reliability constrain. Proceedings of Asia-Pacific Conference on Communications (APCC), pp. 740-745. August 2013. Denpasar : 2013.

[13] Sutton, R.S., and Barto, A.G. Reinforcement learning: an introduction. MA : The MIT Press, 1998.

[14] Usaha, W., and Maneenil, K. Identifying malicious nodes in mobile ad hoc networks using a reputation scheme based on reinforcement learning. Proceedings of TENCON, pp. 14-17. November 2006. Hong Kong : 2006.

[15] Naruephiphat, W., and Usaha, W. Balancing tradeoffs for energyefficient routing in MANETs based on reinforcement learning. Proceedings of Vehicular Technology Conference, pp. 2361-2365. May 2008. Singapore : 2008.

[16] Nurmi, P. Reinforcement learning for routing in ad hoc networks. Proceedings of Modeling and Optimization in Mobile, Ad-Hoc,and Wireless Networks (WiOpt), pp. 1-8. April 2007. Limassol : 2007.

[17] Chong, K.P., Kreucher, M., and Alfred, O. Monte-carlo-based partially observable markov decision process approximations for adaptive sensing. Proceedings of International Workshop on Discrete Event Systems (WODES), pp. 173–180. May 2008. Goteborg : 2008.

[18] Naveed, M., Kitchin, D., Crampton, A., Chrpa, L., and Gregory, P. A monte-carlo path planner for dynamic and partially observable environments. Proceedings of Computational Intelligence and Games (CIG), pp. 211–218. September 2012. Granada : 2012.

[19] Chettibi, S., and Chikhi, S. Adaptive maximum-lifetime routing in mobile ad-hoc networks using temporal difference reinforcement learning. Journal of Evolving Systems 5, 2. (June 2014) : 89–108.

[20] Usaha, W., and Barria, J. A reinforcement learning ticket-based probing path discovery scheme for MANETs. Journal of Ad Hoc Networks 2, 3. (July 2004) : 319-334.

[21] Duarte-Melo, E.J., and Mingyan, L. Analysis of energy consumption and lifetime of heterogeneous wireless sensor networks. Proceedings of GLOBECOM, pp. 21–25. November 2002. Taipei : 2002.

[22] Dewan, P., Dasgupta, P., and Bhattacharya, A. On using reputations in ad hoc networks to counter malicious nodes. Proceedings of Parallel and Distributed Systems (ICPADS), pp. 665 – 672. July 2004. California : 2004.

[23] Khawsaard, N., and Saivichit, C. Prolonging network lifetime and energy reduction using path reputation based on monte carlo algorithm for wsns. Proceedings of Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI), pp. 1764-1768. May 2013. Krabi : 2013.

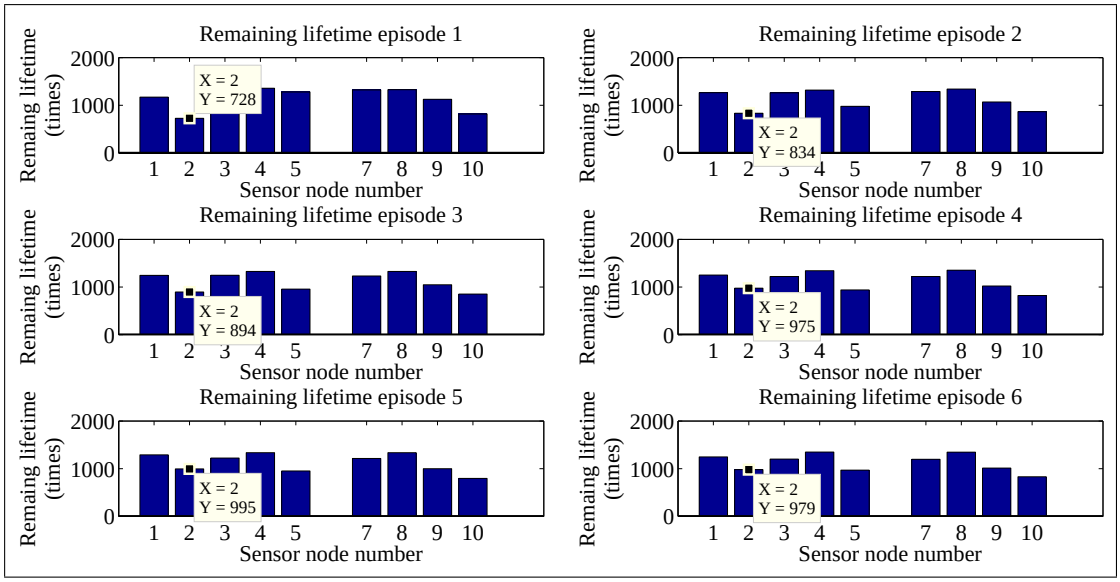
[24] Khawsaard, N., and Saivichit, C. Path-reputation based technique in reactive AODV ad hoc sensor networks routing for flood warning application. Proceedings of Intern-International Computer Science and Engineering Conference (ICSEC), pp. 284-287. Se-
ptember 2013. Bangkok : 2013.

[25] Torres, M.G. Energy consumption in wireless sensor networks usig GSP. Master's thesis, Universidad Pontificia Bolivariana, Medellin, Colombia, 2006.

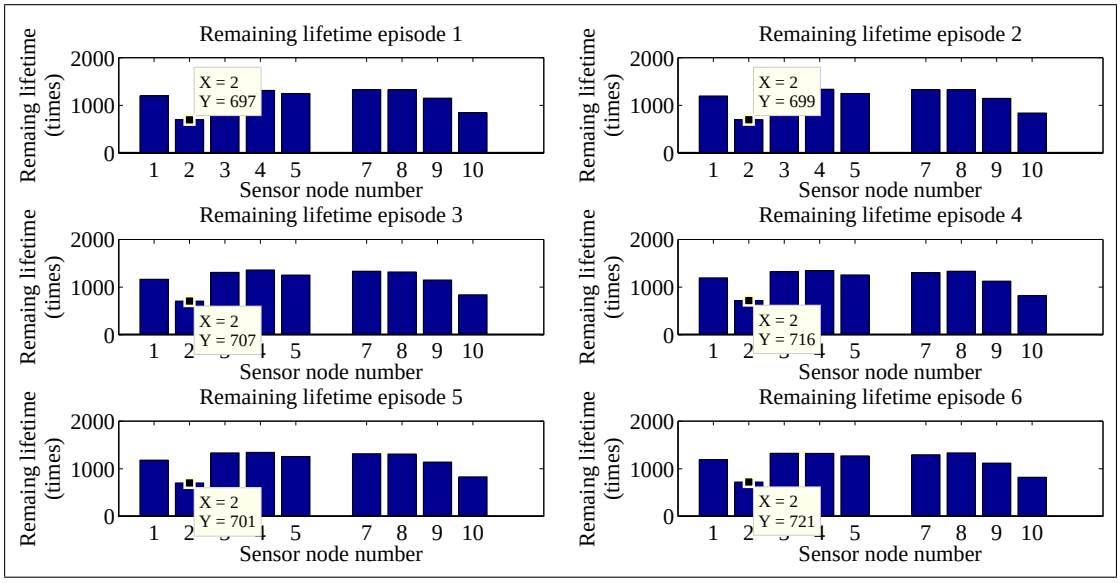
ภาคผนวก

ก อายุการใช้งานที่เหลืออยู่ของแต่ละเซนเซอร์โนด

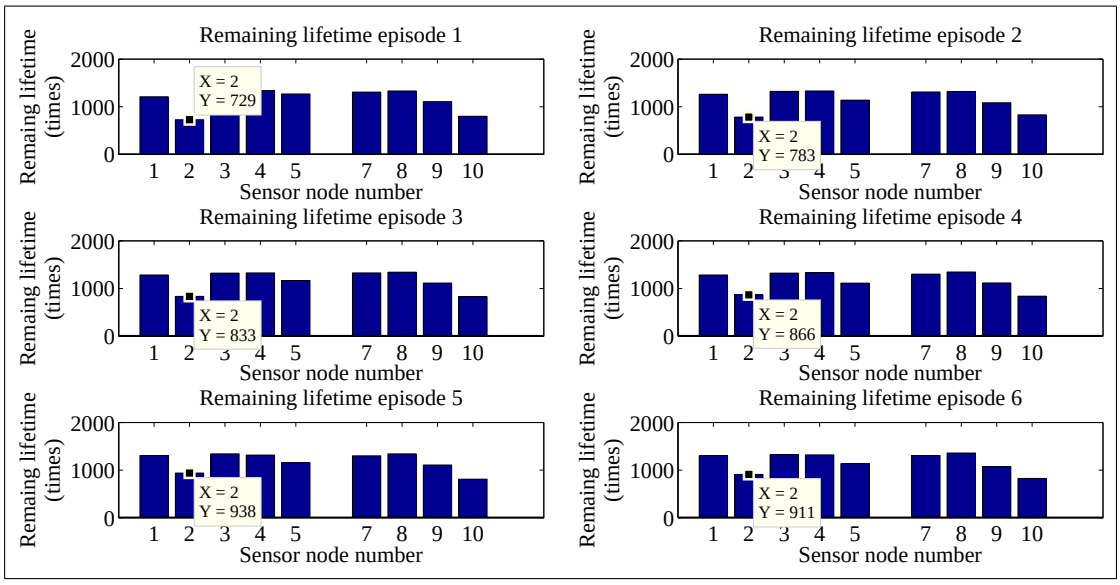
ก.1 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน)



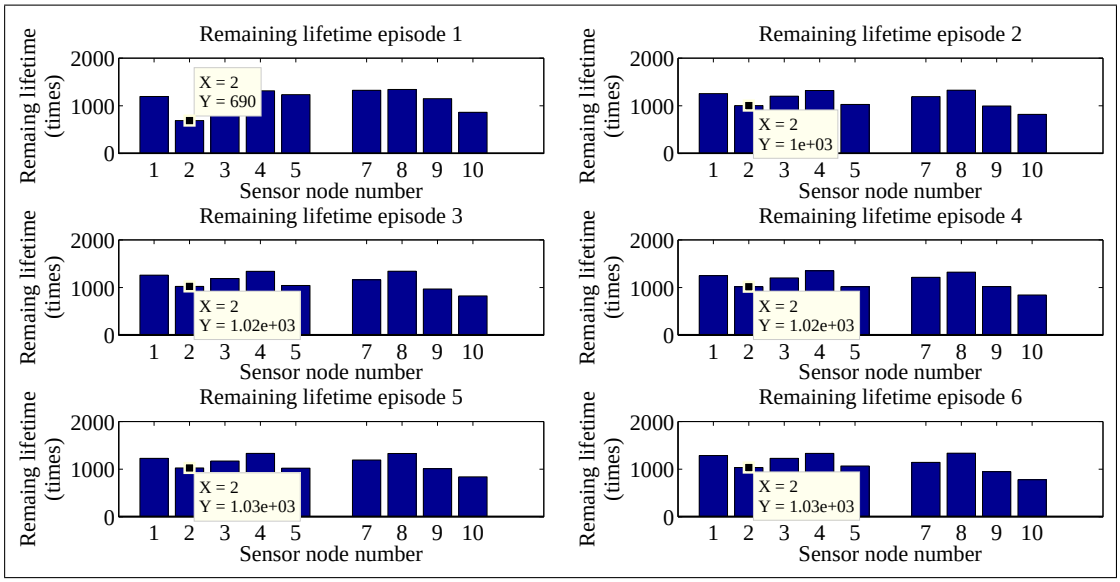
รูปที่ 1: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีพลังงานเป็นตัวแปรฟังก์ชันผลรางวัล



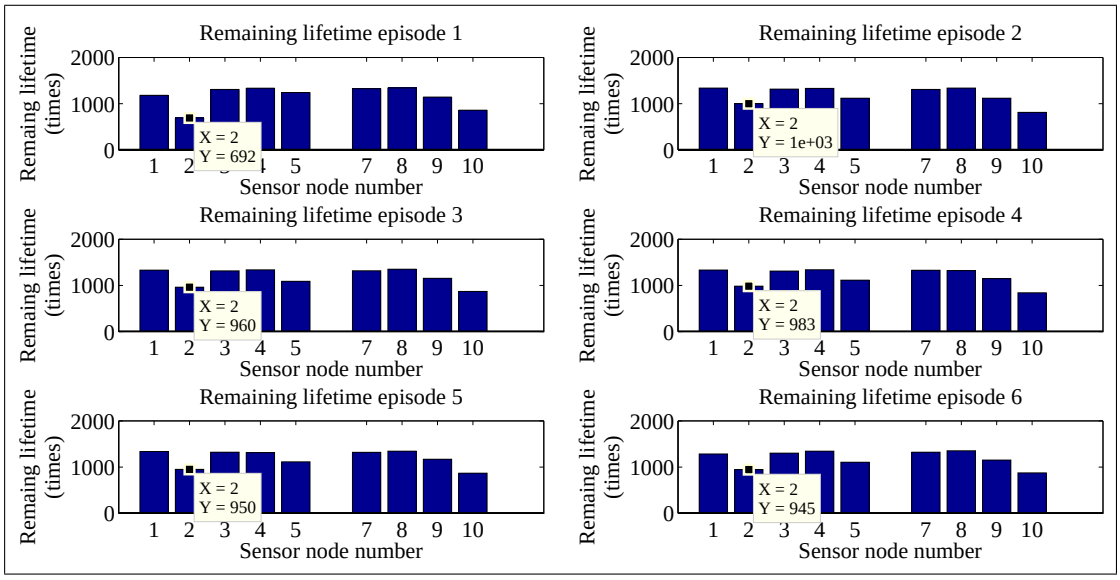
รูปที่ 2: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีอายุการใช้งานเป็นตัวแปรฟังก์ชันผลรางวัล



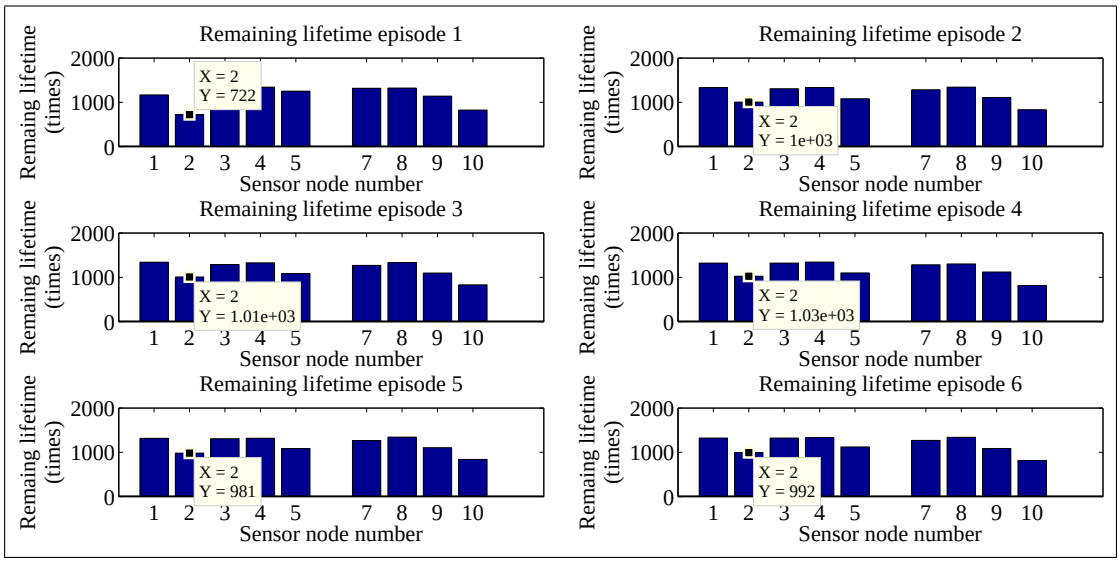
รูปที่ 3: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีเร็พพิวเทชันเป็นตัวแปรฟังก์ชันผลรางวัล



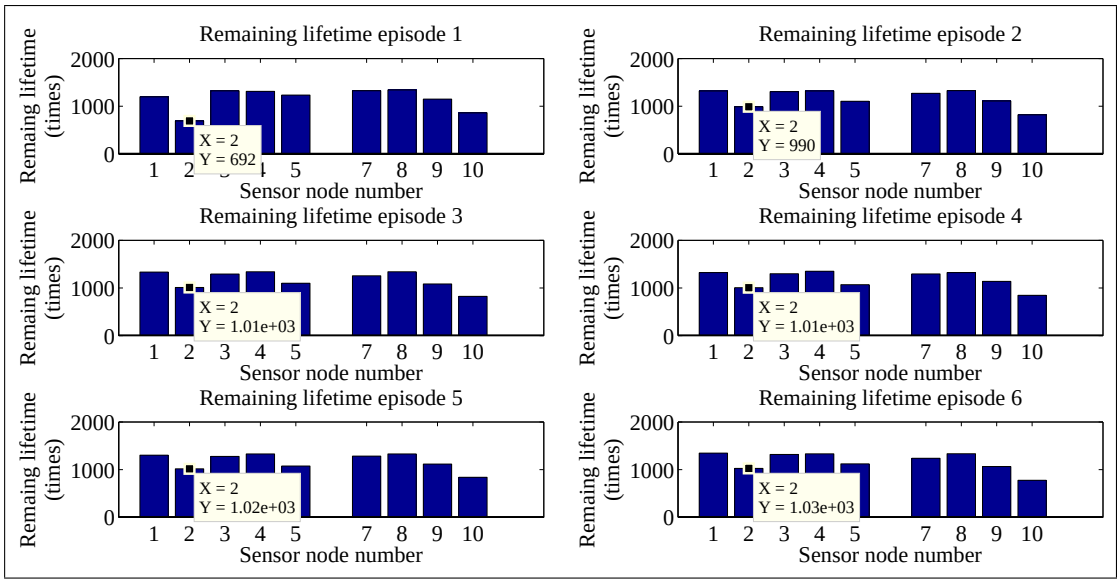
รูปที่ 4: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีพลังงานและอายุการใช้งานเป็นตัวแปรฟังก์ชันผลรางวัล



รูปที่ 5: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีอายุการใช้งานและเร็พพิวเทชันเป็นตัวแปรฟังก์ชันผลรางวัล



รูปที่ 6: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีพลังงานและเร็พพิวเทชันเป็นตัวแปรฟังก์ชันผลรางวัล



รูปที่ 7: 2 ตัวแปรสถานะ (พลังงานและค่าเร็พพิวเทชัน) โดยมีสามตัวแปรฟังก์ชันผลรางวัล

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวณัฐธิดา ขาวสะอาด เกิดเมื่อวันที่ 7 สิงหาคม พ.ศ. 2531 กรุงเทพมหานคร เป็นบุตรของนายสาลี ขาวสะอาด และ นางอ้อยทิพย์ ขาวสะอาด สำเร็จการศึกษาระดับปริญญาหลักสูตรวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมอิเล็กทรอนิกส์ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เมื่อปีการศึกษา 2553 และเข้าศึกษาต่อในหลักสูตรวิศวกรรมศาสตรมหาบัณฑิตในปีการศึกษาถัดมา ณ ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย สังกัดห้องปฏิบัติการวิจัยโทรคมนาคม ได้รับทุนศิษย์ก้นกุฎิขณะศึกษาระดับมหาบัณฑิตและมีบทความวิชาการจากวิทยานิพนธ์ในระหว่างการศึกษาในระดับมหาบัณฑิตดังนี้

[1] Khawsaard, N., and Saivichit, C. Prolonging network lifetime and energy reduction using path reputation based on monte carlo algorithm for wsns. Proceedings of Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI), pp. 1764-1768. May 2013. Krabi : 2013.

[2] Khawsaard, N., and Saivichit, C. Path-reputation based technique in reactive AODV ad hoc sensor networks routing for flood warning application. Proceedings of International Computer Science and Engineering Conference (ICSEC), pp. 284-287. September 2013. Bangkok : 2013.