KKU Res.j. 2014; 19(2) : 261-221
http://resjournal.kku.ac.th

การระบุตำแหน่งของอุปกรณ์ไร้สายภายในอาคารด้วยอัลกอริธึมทางเหมืองข้อมูลโดยใช้ 2 ตัวส่งสัญญาณ

Indoor-positioning of Wireless Devices by Using Data Mining Algorithms with 2 Access Points

*ชัชชล เปรมชัยสวัสดิ์, นรารัตน์ เรืองชัยจตุพร**
*Shutchon Premchaisawatt, Nararat Ruangchaijatupon**

ภาควิชาวิศวกรรมไฟฟ้า คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น
**Correspondent author: nararat@kku.ac.th*

บทคัดย่อ

งานวิจัยครั้งนี้มีจุดประสงค์เพื่อศึกษาวิธีการระบุตำแหน่งของอุปกรณ์ไร้สายภายในบริเวณอาคาร โดยใช้อัลกอริธึมการทำเหมืองข้อมูล ได้แก่ ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบส์ (Naive Bayes) โครงข่ายประสาทเทียม (Artificial Neural Network) และเค-มีน (K-Means) โดยอาศัยความแรงของสัญญาณคลื่นวิทยุจาก 2 ตัวส่งสัญญาณในการระบุตำแหน่งของอุปกรณ์รับสัญญาณที่อยู่ภายในอาคาร เพื่อเปรียบเทียบหาอัลกอริธึมที่มีประสิทธิภาพเหมาะแก่การนำไปพัฒนาต่อยอดในการพัฒนาซอฟต์แวร์ระบุตำแหน่งภายในอาคาร โดยในการเปรียบเทียบประสิทธิภาพเราคำนึงถึงความถูกต้องของการจำแนกตำแหน่งจากข้อมูลความแรงสัญญาณของอุปกรณ์ไร้สายที่วัดได้ ความเที่ยงตรงของระยะทางที่จำแนกได้ ความซับซ้อนของการสร้างแบบจำลองเพื่อการจำแนกตำแหน่ง และผลกระทบของจำนวนข้อมูลที่ใช้ในการเรียนรู้ต่อผลลัพธ์ของการจำแนกตำแหน่ง ผลที่ได้ คืออัลกอริธึมต้นไม้การตัดสินใจ และนาอีฟเบส์ที่มีความถูกต้อง ความเที่ยงตรง และมีความเร็วเป็นที่ยอมรับได้ เหมาะที่จะนำไปพัฒนาเพื่อสร้างซอฟต์แวร์การระบุตำแหน่งต่อไป

Abstract

This research aims to study wireless device indoor positioning methods by using machine learning algorithms, i.e; Decision Tree, Naive Bayes, Artificial Neural Networks, and K-Means by exploiting signal strength from 2 access points. The performance comparison is done in terms of accuracy of classification of positions, precision of distance classified, complexity of distinguish modeling, and effects of classification of positions on results from quantity of learning data in order to find the suitable algorithm. The result of this study can suggest that the Decision Tree algorithm and the Naive Bayes algorithm are suitable for future indoor-positioning software development.

คำสำคัญ: การระบุตำแหน่งภายในอาคาร การทำเหมืองข้อมูล การเรียนรู้ของเครื่อง
Keywords: indoor positioning, data mining, machine learning

1. บทนำ

การระบุตำแหน่งมีส่วนสำคัญกับการอำนวยความสะดวกในการใช้ชีวิตในยุคใหม่ ปัจจุบันการระบุตำแหน่งที่บริเวณกลางแจ้งสามารถทำได้และมีเครื่องมือเป็นที่ยอมรับ เช่น ระบบจีพีเอส (GPS: Global Positioning System) ซึ่งนำไปสู่ประโยชน์หลากหลายด้วยกัน เช่น การนำทางในการเดินทางไปในที่ใหม่ การระบุตำแหน่งเพื่อง่ายต่อการอ้างอิง รวมไปถึงความสามารถในการบริการเชิงพื้นที่ (Local Based Services) ที่จะเป็นประโยชน์และอำนวยความสะดวกสบายมาสู่การใช้ชีวิตมากขึ้น Liu และคณะ(1) ได้กล่าวว่า ปัจจุบันการระบุตำแหน่งในบริเวณภายในอาคารยังทำได้ไม่ดีถึงจนเป็นที่ยอมรับ ในเรื่องความเที่ยงตรงและความง่ายเป็นเพราะภายในอาคารเกิดการสะท้อน สอดแทรก และเบี่ยงเบนของคลื่นสัญญาณ อันเนื่องมาจากสภาพแวดล้อมภายในอาคาร ซึ่งเครื่องมือระบุตำแหน่งอย่างจีพีเอส ไม่สามารถใช้บริเวณภายในอาคารได้เพราะคลื่นสัญญาณระหว่างเครื่องรับจีพีเอสและดาวเทียมถูกขัดขวางโดยกำแพงของอาคาร จึงมีความคิดที่จะหาวิธีระบุตำแหน่งภายในอาคารขึ้นมาหลากหลายวิธี โดยอ้างอิงจากงานวิจัย (1-3) สามารถสรุปเป็นกลุ่มได้ ดังนี้

ไตรแองกูลาชั่น(Triangulation) เป็นวิธีที่มักจะต้องใช้ภาคส่งอย่างน้อยสามแหล่งในการระบุตำแหน่ง ซึ่งทั้งสามเสาอากาศนี้จะมีการปล่อยสัญญาณคลื่นวิทยุออกไป แล้วสามารถคำนวณหาตำแหน่งของอุปกรณ์ไร้สายจากความแรงสัญญาณที่รับได้ที่จุดที่อุปกรณ์นั้นตั้งอยู่ และสามารถติดตั้งเครื่องมือพิเศษเพื่อหามุมของวัตถุเพื่อใช้ช่วยในการระบุตำแหน่งโดยอ้างอิงมุมและระยะทางกับเสาส่งเพื่อการระบุตำแหน่งที่มีความแม่นยำยิ่งขึ้น แต่ข้อเสียของวิธีนี้คือมีค่าใช้จ่ายสูง และขอบเขตของระบุตำแหน่งขึ้นอยู่กับกำลังในการส่งของเสาส่งสัญญาณ

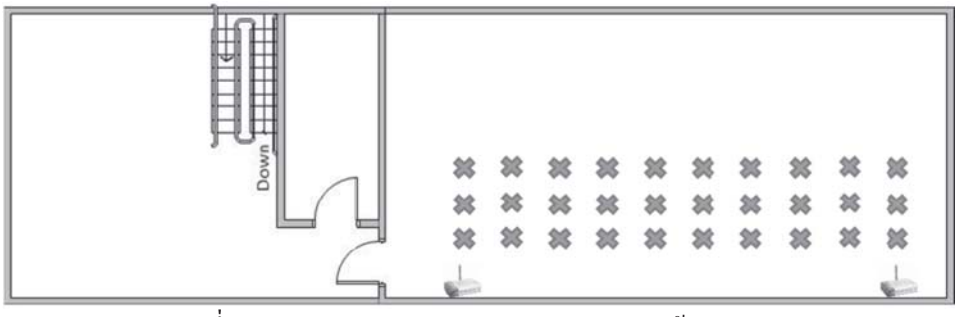
การใช้ฮาร์ดแวร์เฉพาะ เป็นวิธีที่สร้างเครื่องมือเฉพาะมาเพื่อระบุตำแหน่งในบริเวณพื้นที่เฉพาะ ซึ่งฮาร์ดแวร์จะถูกปรับแต่งเพื่อให้เหมาะสมกับการระบุตำแหน่งบริเวณนั้น โดยที่ Mautz (4) ได้ยกตัวอย่าง เช่น จีพีเอสเทียม (Pseudo GPS) เลเซอร์แทรคเกอร์ (Laser Tracker) เลเซอร์อินฟราเรดความแม่นยำสูง (Resection Infrared Laser) ซึ่งมีความแม่นยำสูงแต่ก็มีค่าใช้จ่ายสูงเช่นกัน

การใช้ความแรงสัญญาณ (Signal Strength) หลักการคล้ายไตรแองกูลาชั่นที่เป็นการหาระยะทางโดยใช้การลดทอนของความแรงสัญญาณ หรือจะใช้เปรียบเทียบความแรงสัญญาณกับตำแหน่งอ้างอิงเพื่อบอกตำแหน่งซึ่งวิธีนี้มักจะเน้นที่อัลกอริธึมในการระบุตำแหน่ง การใช้ความแรงสัญญาณจึงเป็นวิธีที่มีค่าใช้จ่ายต่ำกว่าสองวิธีที่กล่าวถึงก่อนหน้านี้

บทความนี้จะอ้างถึงการระบุตำแหน่งโดยใช้ความแรงสัญญาณ โดยจะใช้วิธีฟิงเกอร์พริ้นติ้ง (Finger Printing) ที่ดัดแปลงจากการทดลองของ Mok และ Retscher (3) โดยสร้างแผนที่ความแรงสัญญาณ (Radio Map) แล้วใช้อัลกอริธึมการทำเหมืองข้อมูลระบุตำแหน่งจากสัญญาณที่รับได้ โดยคำนวณจากข้อมูลสัญญาณที่อยู่ในแผนที่ความแรงสัญญาณนั้น โดยจะเปรียบเทียบกันในแง่ต่างๆ ระหว่าง 4 อัลกอริธึม คือ อัลกอริธึมที่ใช้ในการจำแนก (Classification) ได้แก่ ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบส์ (Naive Bayes) โครงข่ายประสาทเทียม (Artificial Neural Network) และอัลกอริธึมที่ใช้ในการแบ่งกลุ่ม (Clustering) คือ เค-มีน (K-Means) ซึ่งทั้งสองแตกต่างกันตรงที่การจำแนกเป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) แต่การแบ่งกลุ่มเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) โดยจะแตกต่างกันด้วยขั้นตอนของการเรียนรู้ (5) แต่ในงานนี้จะใช้เค-มีนในการจำแนกซึ่งเป็นการเรียกว่าการจำแนกโดยอาศัยการแบ่งกลุ่ม (Classification Based Clustering) เมื่อเปรียบเทียบจึงขอใช้คำว่าจำแนกแทนทั้งหมดเพื่อใช้ในการระบุพื้นที่ ซึ่งทั้ง 4 อัลกอริธึมนี้เป็นที่นิยมในการใช้ระบุตำแหน่ง (2) โดยได้เปรียบเทียบอัลกอริธึมเพื่อระบุตำแหน่งในหัวข้อของความถูกต้อง ความเที่ยงตรง ความซับซ้อน และความยืดหยุ่น งานวิจัยนี้จึงต้องการจะเปรียบเทียบถึงข้อดีหรือข้อเสียในหัวข้อต่างๆ เพื่อใช้ในการอ้างอิงและพัฒนาในงานอื่นต่อไป

2. วิธีการวิจัย

ในการทดลองได้ทำการทดลองภายในห้องซึ่งมีความกว้าง 6 เมตรและยาว 12 เมตร เพดานสูง 2.8 เมตร โดยในการทดสอบนี้ได้วางเสาส่งสัญญาณไว้ชิดผนังเพื่อจะได้คำนวณด้านเดียวของสัญญาณที่แผ่ออกจะทำให้ลด



รูปที่ 1 แสดงแผงผังสถานที่ทดลองและตำแหน่งของจุดส่งสัญญาณและจุดอ้างอิงสัญญาณ

ตารางที่ 1 แสดงคุณลักษณะของกลุ่มข้อมูลที่ใช้ในการทดสอบ

ประเภท	คุณลักษณะของข้อมูล
จำนวนของคุณลักษณะที่เป็นอินพุต	2 คุณลักษณะ (2 เสาสัญญาณ)
ค่าสูงสุดของสัญญาณ	-46 dBm
ค่าต่ำสุดของสัญญาณ	-59 dBm
จำนวนของคลาสตำแหน่งอ้างอิง	30 คลาส (ตำแหน่งอ้างอิง)
จำนวนข้อมูล	300 (10 ตัวอย่างต่อหนึ่งตำแหน่งอ้างอิง)

ปัญหาตำแหน่งอ้างอิงมีความแรงสัญญาณเหมือนกันได้ โดยวางเสาส่งสัญญาณชิดกำแพงที่ระยะทางเมตรที่ 1 และเมตรที่ 11 ตามลำดับ จากนั้นทำการเก็บข้อมูลความแรงสัญญาณที่จุดต่างๆ ภายในบริเวณ 3 x 10 ตารางเมตร ระหว่างระยะทางจากเสาทั้งสอง 10 เมตร ดังรูปที่ 1 ได้ตัวอย่างข้อมูลทั้งหมด 3,000 ตัวอย่าง

โดยทำซ้ำ 10 ครั้ง ทำให้ได้ข้อมูล 10 ชุด ชุดละ 300 ตัวอย่างข้อมูล โดยจะใช้วิธีการทำ10-โฟลด์ครอสวาลิเดชั่น (10-fold cross-validation) ในการทดสอบประสิทธิภาพการจำแนกแต่ละชุดข้อมูล ซึ่งจะเป็นข้อมูลความแรงสัญญาณที่รับได้ในแต่ละจุดอ้างอิงจากทั้งสองเสาส่ง โดยเมื่อตรวจดูคุณลักษณะของข้อมูลแต่ละชุดจะมีลักษณะคล้ายกันดังตารางที่ 1 แต่จะต่างกันที่ค่าเฉลี่ย (Mean) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ส่วนตารางที่ 2 เป็นตัวอย่างบางส่วนของข้อมูลที่ใช้ โดย RSS1, RSS2 คือ ความแรงสัญญาณจากเสาส่งที่ 1 และเสาส่งที่ 2 (หน่วย dBm) ID คือตำแหน่งของจุดอ้างอิงที่ใช้วัดสัญญาณ

เมื่อได้ข้อมูลที่จะใช้ในการเรียนรู้ของอัลกอริธึมแล้ว Witten และคณะ (5) ได้แนะนำว่า ก่อนจะนำข้อมูล

ตารางที่ 2 แสดงตัวอย่างลักษณะข้อมูลที่ใช้ในการทดสอบ

RSS1	RSS2	ID
-48	-59	1
-49	-58	2
-51	-58	3
-52	-57	4
-54	-56	5
-56	-54	6
-57	-52	7
-58	-51	8
-58	-49	9
-59	-48	10

ไปสอนอัลกอริธึมควรจะต้องผ่านการเตรียมข้อมูลก่อน ในการทดลองนี้เตรียมข้อมูลโดย ตรวจรูปแบบของข้อมูลว่าเป็นตัวเลขทั้งหมด หรือไม่ และตัดข้อมูลที่มีค่าผิดปกติ (Outliers) ที่มากกว่าหรือน้อยกว่าข้อมูลปกติ 2 เท่าขึ้นไป และตัดข้อมูลที่ไม่สมบูรณ์ออก จากนั้นนำไปประมวลผลด้วยอัลกอริธึมแต่ละตัวที่มีการกำหนดคุณลักษณะดังนี้

ต้นไม้การตัดสินใจ (6) อาศัยหลักการของเอนโทรปี (Entropy) ในการจำแนกข้อมูลซึ่งในการทดลองใช้ต้นไม้การตัดสินใจแบบไอดีทรี (ID3) ไม่ได้มีกำหนดค่าในการปรับแต่ง (Configuration) แต่ต้องผ่านการเตรียมข้อมูลมาก่อน โดยนำข้อมูลที่ไม่สมบูรณ์ออกจึงจะนำมาสร้างต้นไม้การตัดสินใจได้

นาอีฟเบส์ (7) ใช้หลักการทางด้านความน่าจะเป็นจำแนกข้อมูลโดยไม่ได้มีกำหนดค่าในการปรับแต่ง แต่ต้องผ่านการเตรียมข้อมูลมาก่อนเช่นเดียวกัน

โครงข่ายประสาทเทียม (8) เป็นแบบมัลติเลเยอร์เพอร์เซปตรอน (Multi-layer Perceptron) ที่มีชั้นซ่อน (Hidden Layer) ตามที่ (5) แนะนำซึ่งเท่ากับจำนวนของคุณสมบัติ (Attribute) ที่ใช้ในการคำนวณในที่นี้คือจำนวนของเสาส่งคือสอง คูณด้วยจำนวนของคลาส (Class) คือสามสิบจุดอ้างทั้งหมดอิงส่วนด้วยสองตัวส่งสัญญาณที่มีอัตราการเรียนรู้ (Learning Rate) เท่ากับ 0.3 และมีโมเมนตั้ม (Momentum) เท่ากับ 0.2 ที่มีรอบการเรียนรู้ 500 รอบ

เค-มีน (9) มีจำนวนของกลุ่มที่จะแบ่งกลุ่ม (Number of Cluster) เท่ากับ 30 และมีจำนวนซีด (Seed) เท่ากับ 30 มีรอบการทำงานเท่ากับ 500 รอบ

อัลกอริทึมทั้งหมดถูกดำเนินการโดยใช้เครื่องมือวิเคราะห์ข้อมูล ซึ่งซอฟต์แวร์ที่ใช้ในการทดลองคือ เวก้า (WEKA : Waikato Environment for Knowledge Analysis) ซึ่งเป็นซอฟต์แวร์ทางด้านการทำเหมืองข้อมูลที่เป็นที่นิยมและมีมาตรฐานสามารถอ้างอิงได้ในงานวิจัย (5)

ในการทดสอบความสามารถในการจำแนกวิธีหนึ่งที่เป็นที่นิยมใช้คือเทคนิค เค-โฟลด์ครอสวาลิเดชั่น (k-fold cross-validation) หรือเรียกว่าครอสวาลิเดชั่น เป็นวิธีการวัดประสิทธิภาพในการจำแนกข้อมูลของแบบจำลอง (10) โดยพื้นฐานของเทคนิคนี้คือการสุ่มตัวอย่าง (re-sampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆหรือเรียกว่าโฟลด์ (fold) และนำบางส่วนจากชุดข้อมูลนั้นมาทดสอบผลลัพธ์จากการทำนายข้อมูลทดสอบของแบบจำลอง กรณีการเลือกสุ่มข้อมูลแบบความเที่ยงตรง k กลุ่ม เราจะแบ่งข้อมูลออกเป็น k ชุดเท่าๆกัน และทำการคำนวณค่าความถูกต้องจากการจำแนก k รอบ ในการทดลองนี้ใช้ k เท่ากับ 10 ซึ่งเป็นจำนวนที่เป็นที่นิยมใช้หรือเรียกว่า 10-โฟลด์ครอสวาลิเดชั่น (10-fold cross-validation) โดยแต่ละรอบจะมีการสร้างแบบจำลองการจำแนกประเภทหนึ่งตัว จากข้อมูลเรียนรู้ k-1 ชุด และใช้ข้อมูลทดสอบ 1 ชุด (ชุดที่ไม่ได้นำมาเรียนรู้) จากรูปที่ 2 ในการทำงานรอบแรก ข้อมูลในชุดที่ 1 จะใช้เป็นข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 2 ถึง 10 จะนำมาใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ ซึ่งจะได้แบบจำลองจำแนกประเภท 1 ตัว ต่อมารอบที่สอง ก็จะใช้ข้อมูลในชุดที่ 2 เป็นข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 1 และ 3 ถึง 10 จะนำมาใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ ซึ่งจะได้แบบจำลองการ

รูปที่ 2 แสดงขั้นตอนการแบ่งข้อมูลของ10-โฟลด์ครอสวาลิเดชั่น

จำแนกประเภทอีก 1 ตัว จะมีการทำงานลักษณะนี้ไปเรื่อยๆ จนถึงรอบที่สิบ จะใช้ข้อมูลในชุดที่ 10 เป็นชุดข้อมูลทดสอบ ส่วนข้อมูลในชุดที่ 1 ถึง 9 จะนำมาใช้เป็นชุดข้อมูลสำหรับการเรียนรู้ และจะได้แบบจำลองจำแนกประเภทอีก 1 ตัว หลังจากนั้นทำการเฉลี่ยความถูกต้องก็จะได้ค่าความถูกต้องหลังทดสอบด้วยวิธีครอสวาลิเดชั่น

จากนั้นจะนำผลการจำแนกหลังทดสอบด้วยครอสวาลิเดชั่นไปทดสอบคุณสมบัติต่างๆ เพื่อใช้ในการวัดประสิทธิภาพของอัลกอริธึมทางด้านการทำเหมืองข้อมูล ได้แก่ ความถูกต้อง ความเที่ยงตรง ความซับซ้อน และความยืดหยุ่น ซึ่งมีรายละเอียดดังนี้

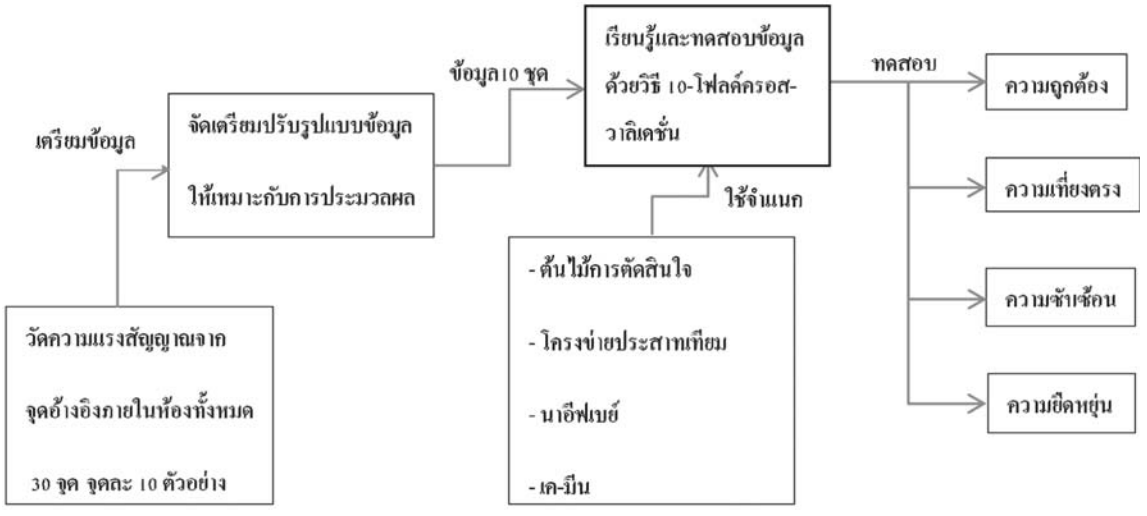
2.1) ความถูกต้อง (Accuracy) จะวัดจากผลจากการจำแนกกับค่าจริงว่ามีค่าเหมือนกันหรือไม่ จำนวนที่จำแนกถูกจะถือเป็นความถูกต้อง โดยจะทำการสอนอัลกอริธึมจากข้อมูลที่ใช้สอน หลังจากนั้นจะนำไปทดสอบการจำแนกกับข้อมูลความแรงสัญญาณที่เก็บมาอีกสิบชุด

2.2) ความเที่ยงตรง (Precision) คำนวณได้จากค่าที่จำแนกเปรียบเทียบกับค่าจริง โดยวัดระยะจากค่าที่จำแนกไปถึงค่าจริงโดยใช้ระยะทางยูคลิด (Euclidean Distance) แล้วคำนวณค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ถ้าค่าเบี่ยงเบนมาตรฐานมีค่าน้อย จะแสดงถึงความเที่ยงตรงสูง

2.3) ความซับซ้อน (Complexity) จะอาศัยค่าเวลาการคำนวณ (Computing Time) โดยอ้างอิงกับคอมพิวเตอร์ที่ทดสอบที่มีหน่วยประมวลกลาง (CPU: Intel® Core™ i7-2635QM 2.GHz) และมีหน่วยความจำ (RAM: 8 GB) สำหรับอ้างอิงถึงเวลาในที่แต่ละอัลกอริธึมใช้ในการเรียนรู้

2.4) ความยืดหยุ่น (Scalability) คือ ความสามารถระบุตำแหน่งได้ถูกต้องเมื่อพารามิเตอร์ของข้อมูลเปลี่ยนไป ซึ่งในการศึกษานี้จะทำการลดจำนวนตัวอย่างข้อมูลที่ใช้ในการเรียนรู้ อัลกอริธึมที่ยังมีความถูกต้องมากที่สุดจะถือว่ามีความยืดหยุ่นมากที่สุด โดยทำการสุ่มลดตัวอย่างค่าสัญญาณจาก 10 ตัวอย่างต่อจุดอ้างอิง เป็น 9, 7, 5, 3 และ 1 ตัวอย่างต่อจุดตามลำดับ

จากขั้นตอนดังกล่าวทั้งหมดสามารถสรุปเป็นแผนภาพดังรูปที่ 3

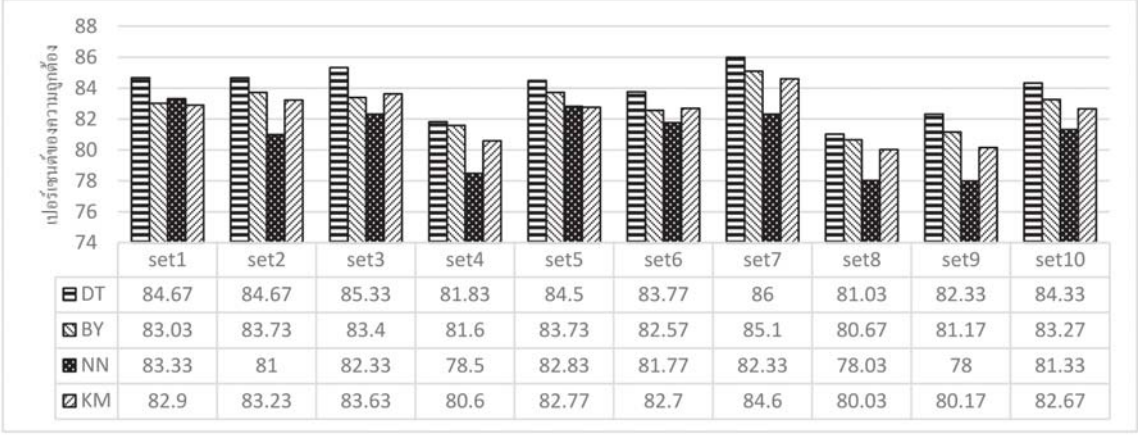


รูปที่ 3 แสดงขั้นตอนการทดสอบเพื่อเปรียบเทียบประสิทธิภาพอัลกอริธึม

3. ผลการวิจัยและอภิปราย

จากการทดสอบทั้งหมด 4 การทดสอบ โดยในการนำเสนอผลจะขอใช้อักษรย่อดังต่อไปนี้เพื่อความสะดวก ผลของต้นไม้การตัดสินใจ เขียนแทนด้วย “DT” นาอีฟเบส์ จะเขียนแทนด้วย “BY” โครงข่ายประสาทเทียมจะเขียนแทนด้วย “NN” และ เค-มีนเขียนแทนด้วย “KM” และได้ผลการทดสอบดังนี้

จากรูปที่ 4 ผลของการเปรียบเทียบความถูกต้องในการจำแนกตำแหน่ง ในการทดลองนี้จะบอกถึงความสามารถที่จะจำแนกว่าอุปกรณ์ไร้สายนี้มาจากตำแหน่งอ้างอิงใด ซึ่งจากรูปที่ 4 จะเห็นได้ว่าประสิทธิภาพในการจำแนกของอัลกอริธึมทั้งหมดมีค่าใกล้เคียงกัน แต่มีข้อสังเกตคือ แทบทุกข้อมูลในการทดลองนี้ NN จะมีค่าความถูกต้องของการจำแนกน้อยที่สุดต่างจากแบบจำลองอื่นอย่างสามารถสังเกตได้



	set1	set2	set3	set4	set5	set6	set7	set8	set9	set10
DT	84.67	84.67	85.33	81.83	84.5	83.77	86	81.03	82.33	84.33
BY	83.03	83.73	83.4	81.6	83.73	82.57	85.1	80.67	81.17	83.27
NN	83.33	81	82.33	78.5	82.83	81.77	82.33	78.03	78	81.33
KM	82.9	83.23	83.63	80.6	82.77	82.7	84.6	80.03	80.17	82.67

รูปที่ 4 แสดงเปอร์เซ็นต์ความถูกต้องในการการจำแนกพื้นที่ของอัลกอริธึมที่ใช้ในการทดสอบ

การทดสอบความเที่ยงตรง ประเมินจากค่าเบี่ยงเบนมาตรฐานที่มีค่าน้อยที่สุด ซึ่งคำนวณจากความเบี่ยงเบนของระยะทางยูคลิดจากตำแหน่งที่จำแนกเทียบตำแหน่งอ้างอิง จากตารางที่ 3 จะสังเกตได้ว่า KM แม้ว่าจะสามารถบอกจุดที่จำแนกได้อย่างถูกต้องจำนวนมาก แต่เมื่อถึงกรณีที่จำแนกผิด ระยะทางที่จำแนกผิดของ KM กลับมีค่าสูงกว่าของ DT และ BY ดังนั้นจากนิยามที่ให้ไว้ในของความเที่ยงตรง KM จึงมีความเที่ยงตรงน้อยที่สุด

การทดสอบความความซับซ้อน ในการทดลองนี้จะวัดความเร็วในการสร้างแบบจำลองของแต่ละอัลกอริธึมจากข้อมูลเดียวกันและเครื่องคอมพิวเตอร์เดียวกัน

จากตารางที่ 4 แสดงให้เห็นถึงความซับซ้อนทางการคำนวณของ NN ที่ใช้ระยะเวลามากกว่าอัลกอริธึมอื่นอย่างเห็นได้ชัด แม้กระทั้ง KM ที่กำหนดรอบของการเรียนรู้ไว้ 500 รอบ เหมือนกับ NN แต่กลับใช้ระยะเวลาน้อยกว่าหลายเท่าตัว อันเนื่องมาจากการหาคำตอบที่ดีที่สุดได้แล้วจึงหยุดการทำงานเร็วกว่า NN แต่อย่างไรก็ตาม DT กลับเป็นอัลกอริธึมที่ใช้เวลาน้อยที่สุดในการทดสอบนี้ เป็นเหตุมาจากกลไกในการคำนวณของต้นไม้การตัดสินใจจะพิจารณาถึงจำนวนคุณลักษณะของข้อมูล ถ้ามีจำนวนน้อยจะทำให้คำนวณได้เร็วกว่าจำนวนมาก ซึ่งในการทดลองนี้มีแค่สองคุณลักษณะคือข้อมูลจากสองเสาส่งทำให้มีความเร็วสูงกว่าการคำนวณแบบอื่นๆ

การทดสอบความยืดหยุ่นของอัลกอริธึมต่อจำนวนข้อมูล การทดสอบนี้จะเปลี่ยนจำนวนของข้อมูลตัวอย่างที่ใช้สอน ซึ่งจากการทดสอบที่ผ่านมาจะใช้ 10 ตัวอย่างต่อหนึ่งตำแหน่งอ้างอิงจะลดเป็น 9, 7, 5, 3 และน้อยที่สุดคือ 1 ตามลำดับ ซึ่งจะวัดผลกระทบที่เกิดขึ้นจากค่าความถูกต้อง ซึ่งผลที่ได้ดังรูปที่ 5 จะเห็นได้ว่า ความอ่อนไหวต่อการเปลี่ยนแปลงของข้อมูลมีค่าต่างกัน แบบจำลองในการจำแนกทั้งหมด เมื่อมีจำนวนตัวอย่างต่อพื้นที่อ้างอิงเท่ากับ 1 ไม่สามารถจำแนกเลย แต่เมื่อมีจำนวนตัวอย่างต่อพื้นที่อ้างอิงเท่ากับ 3 อัลกอริธึม DT BY KM มีความถูกต้องในการจำแนกเป็น 60 เปอร์เซ็นต์ ในขณะที่ NN มีความ

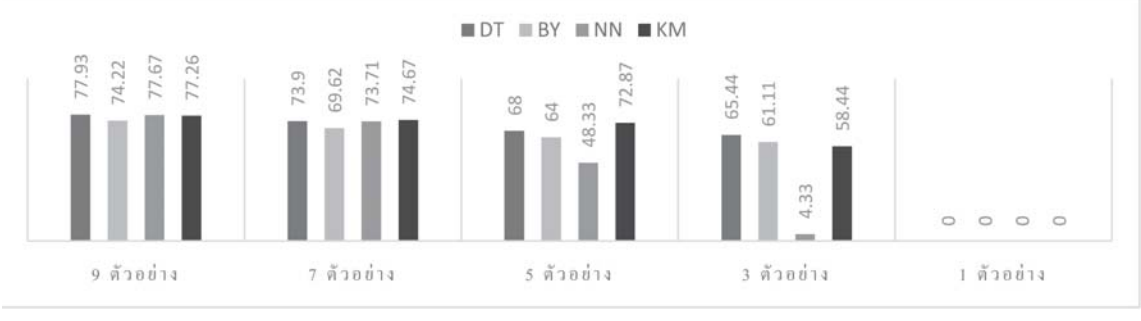
ถูกต้องเพียง 4 เปอร์เซ็นต์เท่านั้น เมื่อมี 5 ตัวอย่าง NN ถึงจะมีความถูกต้องสูงขึ้นมาถึงประมาณ 50 เปอร์เซ็นต์ แต่ก็ยังน้อยกว่าแบบจำลองอื่นเมื่อมีข้อมูล 3 ตัวอย่างอยู่ดี แต่พอมีตั้งแต่ 7 ตัวอย่างเป็นต้นไป ค่าความถูกต้องของการจำแนกของแต่ละแบบจำลองจะแตกต่างกันน้อยมาก ทำให้บอกได้แค่ว่าเมื่อจำนวนข้อมูลมีค่ามากขึ้นแบบจำลองจำแนกจะค่อยๆเพิ่มความถูกต้อง แต่ NN มีความยืดหยุ่นน้อยกว่าแบบจำลองอื่น เพราะต้องการจำนวนตัวอย่างข้อมูลที่มากกว่าในการจำแนกให้ได้ผลดีเท่าแบบจำลองอื่น

ตารางที่ 3 แสดงค่าส่วนเบี่ยงเบนมาตรฐานของระยะทางที่จำแนกได้กับระยะทางอ้างอิง

อัลกอริธึม	ค่าส่วนเบี่ยงเบนมาตรฐานของระยะ
ต้นไม้การตัดสินใจ (DT)	1.72
นาอีฟเบส์ (BY)	2.24
โครงข่ายประสาทเทียม (NN)	3.23
เค-มีน (KM)	4.32

ตารางที่ 4 แสดงเวลาที่ใช้ในการคำนวณเพื่อสร้างแบบจำลองของการจำแนกตำแหน่ง

อัลกอริธึม	เวลาที่ใช้คำนวณ (ไมโครวินาที)
ต้นไม้การตัดสินใจ (DT)	35
นาอีฟเบส์ (BY)	75
โครงข่ายประสาทเทียม (NN)	22, 1522
เค-มีน (KM)	2,670



รูปที่ 5 แสดงเปอร์เซ็นต์ความถูกต้องของแต่ละอัลกอริธึม เมื่อมีจำนวนตัวอย่างที่ใช้ในการเรียนรู้เปลี่ยนไป

4. สรุป

จากการศึกษาอัลกอริธึมทางการทำเหมืองข้อมูล เช่น ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบส์ (Naive Bayes) โครงข่ายประสาทเทียม (Artificial Neural Network) และ เค-มีน (K-Means) เมื่อนำมาใช้ระบุตำแหน่งโดยอาศัยความแรงสัญญาณจาก 2 ตัวส่งสัญญาณ โดยเมื่อทดสอบในกรณีต่างๆ แล้วจะเห็นได้ว่าอัลกอริธึมต่างๆ ทุกอัลกอริธึมต่างมีความสามารถในการจำแนกเพื่อระบุตำแหน่งได้ถูกต้องใกล้เคียงกันและต่างมีจุดเด่นและจุดด้อยแตกต่างกัน เช่น โครงข่ายประสาทเทียม มีความสามารถในการจำแนกข้อมูลเพื่อระบุตำแหน่งได้ดี แต่มีความซับซ้อนในการสร้างแบบจำลองมาก หรือเค-มีนที่มีถูกต้องสูง แต่ก็มีความเที่ยงตรงน้อยที่สุดในกลุ่ม แต่ที่โดดเด่นคือ ต้นไม้การตัดสินใจที่ให้ผลลัพธ์ที่เป็นที่ยอมรับได้ทุกการทดสอบ ซึ่งอาจจะเป็นเพราะลักษณะของข้อมูลชุดนี้เหมาะกับกลไกลการคำนวณของต้นไม้การตัดสินใจ และรองลงมาคือ นาอีฟเบส์ที่มีความเร็วและความถูกต้องสูงเช่นกัน ดังนั้นการพัฒนาต่อไปในพื้นที่ใหญ่ขึ้นควรจะสนใจสองอัลกอริธึมนี่ก่อน อีกสิ่งที่สามารถสังเกตได้จากการทดลองนี้จะเห็นได้ว่าความถูกต้องในการจำแนกโดยเฉลี่ยอยู่ที่ประมาณ 80 เปอร์เซ็นต์ ซึ่งเกิดความท้าทายที่จะพัฒนาให้สูงกว่านี้ได้

5. เอกสารอ้างอิง

(1) Liu H, Darabi H, Banerjee P, Liu J. Survey of wireless indoor positioning techniques and systems. Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2007; 37(6):1067-1080.

(2) Lin T N, Lin P C. Performance comparison of indoor positioning techniques based on location fingerprinting in wireless networks, Wireless Networks Communications and Mobile Computing. 2005; 1569–1574.

(3) Mok E, Retscher G. Location determination using WiFi fingerprinting versus Wi-Fi trilateration. Journal of Location Based Services. 2007; 1(2):145–159.

(4) Mautz R. Overview of current indoor positioning systems. Geodezija ir kartografija. 2009; 35(1): 18-22.

(5) Witten I H, Frank E, Hall M A. Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. 3rd ed. San Francisco: Morgan Kaufman; 2011.

(6) Quinlan J R. Induction of decision trees. Machine Learning. 1986; 1(1): 81–106.

(7) Sandberg R, Winberg G, Bränden C I, Kaske A, Ernberg I, Cöster J. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Research. 2001; 11(8): 1404-1409.

(8) Bigus J P. Data mining with Artificial Neural Networks. 1st ed. New York: McGraw-Hill; 1996. p. 23-47.

(9) Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery. 1998; 2(3): 283-304.

(10) Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI. 1995; 14(2):1137-1145.