ตัวกรองคัดเลือกลักษณะเฉพาะสำหรับการจำแนกข้อมูลเพื่อการประยุกต์ใช้บนระบบอินเทอร์เน็ตของทุกสิ่ง*
Filter-Based Feature Selection for Data Classification in IoT

พาสน์ ปราโมกข์ชน[1**] , **พันธุ์ปิติ เปี่ยมสง่า**[2]
Part Pramokchon[1**] , Punpiti Piamsa-nga[2]

[1]*สาขาวิชานวัตกรรมเทคโนโลยีดิจิทัล คณะวิทยาศาสตร์ มหาวิทยาลัยแม่โจ้*
เลขที่ 63 หมู่ 4 ตำบลหนองหาร อำเภอสันทราย จังหวัดเชียงใหม่ 50290
[1]*Digital Technology Innovation Program, Faculty of Science, Maejo University*
63 Moo 4, Nonghan Subdistrict, Sansai District, Chiang Mai 50290

[2]*ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์*
เลขที่ 50 ถนนงามวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพฯ 10900
[2]*Department of Computer Engineering, Faculty of Engineering, Kasetsart University*
50 Ngam Wong Wan Rd, Ladyaow Chatuchak Bangkok 10900

บทคัดย่อ

ระบบอินเตอร์เน็ตที่เชื่อมต่อในทุกสิ่งเป็นเทคโนโลยีที่ถูกนำมาใช้ในชีวิตประจำวันด้านต่างๆ อย่างกว้างขวาง ซึ่งระบบได้ก่อให้เกิดข้อมูลดิจิทัลปริมาณมหาศาลซึ่งมีสารสนเทศที่มีประโยชน์ซ่อนอยู่ เทคนิคด้านเหมืองข้อมูลจึงถูกนำมาประยุกต์ใช้เพื่อวิเคราะห์และค้นหาสารสนเทศเหล่านี้ การจำแนกข้อมูลเป็นเทคนิคด้านเหมืองข้อมูลที่ถูกประยุกต์ใช้กับการพัฒนระบบอัจฉริยะ การคัดเลือกลักษณะเฉพาะเป็นการเตรียมชุดข้อมูลตัวอย่างที่เหมาะสมสำหรับการพัฒนาตัวจำแนกข้อมูลอย่างมีประสิทธิภาพในด้านการเรียนรู้ข้อมูลเพื่อสร้างโมเดลและการจำแนกข้อมูลใหม่ บทความนี้นำเสนอตัวกรองคัดเลือกลักษณะเฉพาะเพื่อการจำแนกข้อมูลกรณีที่มีจำนวนลักษณะเฉพาะจำนวนมาก ซึ่งเป็นธรรมชาติของข้อมูลที่ได้จากระบบอินเตอร์เน็ตที่เชื่อมต่อในทุกสิ่ง ด้วยการประยุกต์ใช้เทคนิคทางสถิติเพื่อประเมินค่าขีดแบ่งเพื่อเลือกกลุ่มของลักษณะเฉพาะที่ดีที่สุดต่อการจำแนกข้อมูล วิธีการที่นำเสนอสามารถหลีกเลี่ยงการวนทำซ้ำหลายครั้ง

---

* ปรับปรุงเพิ่มเติมเนื้อหาจากบทความ เรื่อง Effective threshold estimation for filter-based feature selection ที่นำเสนอในการประชุมวิชาการ 2016 International Computer Science and Engineering Conference (ICSEC 2016) ณ จังหวัดเชียงใหม่ ประเทศไทย วันที่ 14-17 ธันวาคม พศ. 2559

** ผู้เขียนหลัก
อีเมล: part@mju.ac.th

และช่วยทำให้ขั้นตอนการวิจัยและพัฒนาระบบอัจฉริยะสามารถทำได้อย่างรวดเร็วและแม่นยำมากยิ่งขึ้น บทความได้ทดลองเปรียบเทียบการคัดเลือกลักษณะเฉพาะที่ได้นำเสนอกับวิธีการที่มีอยู่เดิมและชุดข้อมูลมาตรฐานที่มีจำนวนลักษณะเฉพาะจำนวนมากและมีหลายกลุ่มคำตอบ ผลการทดลองแสดงให้เห็นว่าขั้นตอนวิธีการที่ได้นำเสนอสามารถคัดเลือกลักษณะเฉพาะจำนวนน้อยและมีนัยสำคัญต่อประสิทธิภาพการจำแนกข้อมูลและสามารถใช้ทดแทนวิธีการคัดเลือกแบบเดิมได้อย่างมีประสิทธิภาพ

คำสำคัญ

ระบบอัจฉริยะ ระบบเหมืองข้อมูล การจำแนกข้อมูล การคัดเลือกลักษณะเฉพาะ ตัวกรอง

Abstract

The Internet of Things (IoT) is a new important technology that is widely used in various fields today. This technology has generated and captured an enormous amount of data. Several techniques of data mining have been applied to analyze and search for valuable information hidden from these data in order to improve IoT smarter. Data Classification is one of mining techniques which has played a role in the development of intelligent systems by using valuable information from IoT. Feature selection is an important process for improving the efficiency of classification both in terms of data learning to construct the model and of classifying a new instance. This paper presents a filter-based feature selection method for analysis highly dimensional data which is the particular characteristic of IoT data. The proposed feature selection effectively estimates the statistical cut-off to select the optimal feature subset for classification. This proposed method can avoid iterative empirical process, thus, this will help the tasks of research and development of intelligent systems in terms of speed-up and correctness. The classification performance of the proposed feature selection method on the highly dimensional dataset is compared with the existing method. The results show that the proposed method can select a small feature subset which has effective performance. It means that the intelligent system in IoT can use the proposed feature selection method instead of the traditional feature selection.

Keywords

Intelligent System, Data Mining, Data Classification, Feature Selection, Filtering

บทนำ

ในปัจจุบันนวัตกรรมเทคโนโลยีดิจิทัลได้เข้ามามีบทบาทสำคัญในชีวิตประจำวันเป็นอย่างมาก ซึ่งหนึ่งในนั้นคือ เทคโนโลยีอินเตอร์เน็ตที่เชื่อมต่อในทุกสิ่ง (Internet of Things, IoT) เรียกโดยย่อว่า เทคโนโลยีไอโอที เป็นเทคโนโลยีสำหรับการเชื่อมต่อระหว่างเครือข่ายคอมพิวเตอร์พื้นฐานทั่วไปและอุปกรณ์ดิจิทัลใหม่ๆ ให้สามารถอำนวยความสะดวกแก่ผู้ใช้ในด้านต่างๆ แนวความคิดพื้นฐานของไอโอทีคือ ความพยายามให้ทุกๆ สิ่ง ได้แก่ อุปกรณ์คอมพิวเตอร์ ผู้ใช้ วัตถุต่างๆ และ สิ่งอื่นใดๆ ในโลกสามารถเชื่อมต่อระหว่างกันได้ เพื่อส่งระบุตัวตน ส่งข้อมูลและสร้างการตัดสินใจได้ด้วยตัวเอง เช่น แนวคิดบ้านอัจฉริยะ (Smart Home) มีการใช้เครือข่ายอุปกรณ์ตรวจจับ (Sensor Network) เพื่อเก็บข้อมูลสภาพแวดล้อมรอบตัวผู้ใช้เพื่ออำนวยความสะดวกในการดำรงชีพภายในบ้าน (Tsai, Lai, Chiang, Yang, 2014, 77-97)

องค์ประกอบหนึ่งที่สำคัญของเทคโนโลยีไอโอทีคือ อุปกรณ์ที่มีลักษณะความเป็นอัจฉริยะ หรือ สมาร์ทออบเจ็ค (Smart Object) ที่จะต้องมีความสามารถในการระบุตัวตน (Identification) ตรวจจับเหตุการณ์ (Sensing Event) และปฏิสัมพันธ์ระหว่างอุปกรณ์ด้วยกัน (Object Interacting) สร้างการตัดสินใจ (Making Decision) ได้ด้วยตัวเอง (Gubbi, Buyya, Marusic, Palaniswami, 2013, 1645-1660) งานวิจัยเพื่อพัฒนาอุปกรณ์อัจฉริยะเหล่านี้จึงมีความสำคัญต่อเทคโนโลยีไอโอทีเป็นอย่างมาก ซึ่งนักวิจัยมักจะใช้แนวคิดของระบบเหมืองข้อมูล (Data Mining) เพื่อหาอัลกอริธึมที่เพิ่มความสามารถของอุปกรณ์ให้สามารถตัดสินใจได้อย่างแม่นยำและรวดเร็วมากยิ่งขึ้น การพัฒนาระบบจำแนกข้อมูล (Data Classification) เป็นหนึ่งในเทคนิคสำคัญของระบบเหมืองข้อมูลที่มักจะถูกนำมาใช้ในงานที่มีลักษณะเป็นการตัดสินใจอัตโนมัติ โดยระบบจำแนกข้อมูลจะมีการนำข้อมูลที่ระบบไอโอทีรวบรวมไว้มาเป็นชุดข้อมูลตัวอย่าง (Training Dataset) สร้างตัวจำแนกข้อมูล (Classifier) ที่สามารถระบุกลุ่มของข้อมูลที่เข้ามาใหม่ (Unknown Data) ได้อย่างมีประสิทธิภาพ (Chen et al., 2015, 1-14) ตัวอย่างงานด้านไอโอทีที่ต้องมีการพัฒนาตัวจำแนกข้อมูล ได้แก่ ระบบอาคารอัจฉริยะที่จะต้องสามารถจำแนกเหตุการณ์ต่างๆ ของผู้ใช้ภายในอาคาร และสามารถตอบสนองการใช้ชีวิตของผู้ใช้ภายในอาคารได้โดยการควบคุมอุปกรณ์อำนวยความสะดวกรวมถึงอุปกรณ์สาธารณูปโภคต่างๆ ภายในอาคาร เช่น ระบบไฟส่องสว่าง หรือ เครื่องปรับอากาศได้อย่างสอดคล้องเหมาะสม

โดยปกติอุปกรณ์ดิจิทัลที่ใช้ภายในเทคโนโลยีไอโอทีผลิตข้อมูลดิจิทัลจำนวนมาก และมีประเภทของข้อมูลที่หลากหลายอยู่ตลอดเวลา ข้อมูลเหล่านี้อาจเป็นข้อมูลภายในของตัวอุปกรณ์เอง เช่น สถานะ ตำแหน่ง หมายเลขของอุปกรณ์ ฯลฯ หรือ ข้อมูลภายนอกอุปกรณ์ เช่น ข้อมูลที่อุปกรณ์ตรวจจับวัดได้ เช่น อุณหภูมิ ความชื้น การเคลื่อนไหว ฯลฯ ข้อมูลเหล่านี้จะถูกเรียกว่าลักษณะเฉพาะ (Feature) ของอุปกรณ์แต่ละตัว อุปกรณ์ที่เกี่ยวข้องในระบบไอโอทีจะสร้างข้อมูลอยู่ตลอดเวลาซึ่งอาจมีขนาดหลายเทระไบต์ขึ้นไปต่อวัน จึงมีงานวิจัยที่พยายามลดความซับซ้อนของข้อมูลสำหรับการเข้าเพื่อสร้างตัวจำแนก ซึ่งการคัดเลือกลักษณะเฉพาะ (Feature Selection) ได้รับการยอมรับว่าเป็นวิธีการหนึ่งในการจัดการลดความซับซ้อนของข้อมูลปริมาณมหาศาลนั้น สามารถเพิ่มประสิทธิภาพของการวิเคราะห์และการสร้างตัวจำแนกข้อมูลเพื่องานพัฒนาเทคนิคไอโอทีได้เป็นอย่างดี (Tsai, Lai, Chiang, Yang, 2014, 77-97) ทั้งในด้านการลดความซับซ้อนการคำนวณ

และเพิ่มความแม่นยำของตัวจำแนกข้อมูล การคัดเลือกลักษณะเฉพาะใช้แนวคิดว่าในฐานข้อมูลทั้งหมดอาจมีเพียงบางลักษณะเฉพาะที่สัมพันธ์กับกลุ่มคำตอบการจำแนก (Relevant Feature) ซึ่งมีประโยชน์ต่อการสร้างตัวจำแนก ดังนั้นสิ่งจำเป็นคือ การเลือกเพียงลักษณะเฉพาะบางอันที่จำเป็นมาใช้ในกระบวนการสร้างตัวจำแนกสำหรับขั้นต่อไป (Yu & Liu, 2004, 125-1224) ข้อมูลที่มีมิติขนาดมหาศาลทั้งในด้านจำนวนลักษณะเฉพาะนี้เป็นปัญหาสำคัญในขั้นตอนการพัฒนาระบบจำแนกข้อมูล เพราะการมีลักษณะเฉพาะที่ไม่มีนัยยะสำคัญ (Irrelevant Feature) ต่อประสิทธิภาพของตัวจำแนกในฐานข้อมูลอยู่เป็นจำนวนมากจะก่อให้เกิดความล่าช้าทั้งในขั้นตอนการพัฒนาตัวจำแนกและการจำแนกข้อมูลใหม่ นอกจากนี้ยังลดความแม่นยำของตัวจำแนกอีกด้วย ดังนั้นหากสามารถคัดเลือกลักษณะเฉพาะมากลุ่มเล็กๆ จะเป็นการลดขนาดของข้อมูลซึ่งเป็นผลดีต่อการวิจัยและพัฒนาตัวจำแนกข้อมูลต่อไป

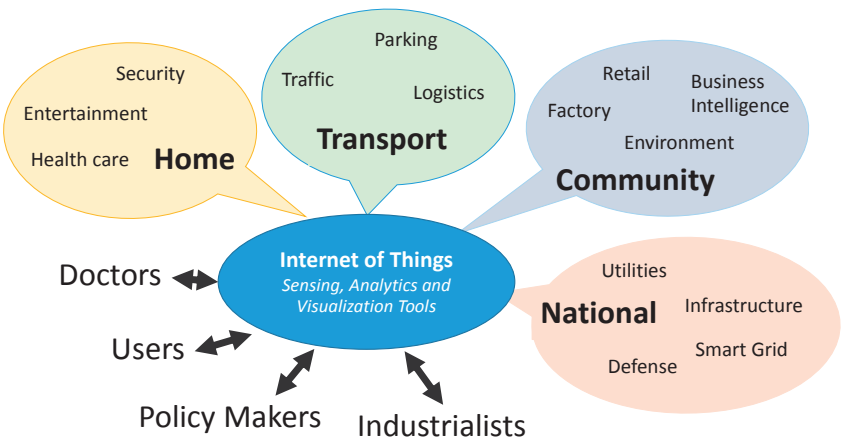
การคัดเลือกลักษณะเฉพาะแบ่งเป็น 2 ประเภทคือ แบบตัวกรอง (Filter-Based) และแบบห่อหุ้ม (Wrapper-Based) (Liu & Yu, 2005, 491-502) วิธีการคัดเลือกแบบตัวกรองใช้การประเมินความมีประโยชน์ของแต่ละลักษณะเฉพาะต่อการจำแนกข้อมูล โดยเรียกว่า ค่าคะแนนลักษณะเฉพาะ (Feature Score) ขั้นตอน วิธีตัวกรองจะเลือกลักษณะเฉพาะที่มีค่าคะแนนสูง (Top Score Feature) มากกว่าลักษณะเฉพาะที่มีคะแนนต่ำ วิธีการคัดเลือกแบบห่อหุ้มจะใช้อัลกริธึมการค้นหา (Searching Algorithm) และอัลกอริธึมการเรียนรู้ (Learning Algorithm) ที่เตรียมไว้แล้วสำหรับกลุ่มของลักษณะเฉพาะที่เหมาะสมกับอัลกอริธึมการเรียนรู้นั้น ทำให้เป็นวิธีการแบบห่อหุ้มสามารถเลือกลักษณะเฉพาะที่ส่งผลต่อความแม่นยำการจำแนกข้อมูลได้ดีกว่าแบบตัวกรอง แต่อย่างไรก็ตามวิธีการแบบห่อหุ้มก็จะมีความยุ่งยากและซับซ้อนต้องใช้ทรัพยากรในการคำนวณมากกว่าวิธีแบบตัวกรอง ดังนั้นในด้านความรวดเร็วและความง่ายในการนำไปใช้วิธีการแบบตัวกรองจึงได้รับความนิยมมากกว่าซึ่งรวมไปถึงงานวิจัยและพัฒนาตัวจำแนกในเทคโนโลยีไอโอทีด้วยเช่นกัน มีงานวิจัยที่เสนอวิธีการแบบผสม (Hybrid-Based) ที่รวมข้อดีของทั้งวิธีการแบบตัวกรอง และวิธีการแบบห่อหุ้ม โดยจะมีขั้นตอนการทำงานที่เร็วขึ้นกว่าวิธีห่อหุ้มและคัดเลือกได้ลักษณะเฉพาะที่ช่วยให้ตัวจำแนก ทำงานได้อย่างแม่นยำขึ้น อย่างไรก็ตามวิธีการแบบผสมนี้ยังคงต้องใช้เวลาการประมวลผลเพื่อคัดเลือกกลุ่มของลักษณะเฉพาะที่สูงมาก (Combarro, Montanes, Diaz, Ranilla & Mones, 2005, 1223-1232) เพราะต้องเสียเวลาวนรอบทำงานเพื่อหาจำนวนของลักษณะเฉพาะที่เหมาะสม และยังไม่มีเทคนิคหรือทฤษฎีสำหรับการคาดคะเนจำนวนของลักษณะเฉพาะได้อย่างมีประสิทธิภาพ

บทความนี้จึงนำ เสนอวิธีการประมาณค่าขีดแบ่ง (Cut-Off or Threshold) สำหรับวิธีการคัดเลือกลักษณะเฉพาะแบบตัวกรองบนพื้นฐานเทคนิคทางสถิติ วิธีการที่นำเสนอไม่ได้ยึดติดกับอัลกอริธึมการเรียนรู้ใดแต่ใช้การประมาณค่าจากคะแนนของลักษณะเฉพาะทั้งหมดในฐานข้อมูล วิธีการที่นำเสนอนี้อาศัยแนวคิดด้านการระบุข้อมูลที่มีลักษณะผิดปกติ (Outlier Identification) ที่พิจารณาว่าลักษณะเฉพาะที่ไม่มีประโยชน์มักมีค่าคะแนนต่ำๆ และมีค่าใกล้เคียงกัน ดังนั้นลักษณะเฉพาะใดมีค่าคะแนนที่สูงกว่าลักษณะเฉพาะอื่นอย่างผิดปกติแสดงว่าเป็นลักษณะเฉพาะที่มีความสัมพันธ์กับกลุ่มคำตอบมากกว่าลักษณะเฉพาะอื่นๆ ทั้งหมด บทความจะนำเสนอวิธีการประมาณค่าขีดแบ่งเพื่อระบุลักษณะเฉพาะที่มีค่าคะแนนที่ผิดปกตินั้น แล้ววิธีการ

แบบตัวกรองจะคัดเลือกลักษณะเฉพาะที่มีค่าคะแนนผิดปกตินั้นไปใช้สำหรับขั้นตอนการพัฒนาตัวจำแนกต่อไป การประมาณค่าขีดแบ่งที่นำเสนอนี้จะใช้พารามิเตอร์ทางสถิติเพียงตัวเดียวและไม่จำเป็นต้องมีการวนทำซ้ำใดๆ นอกจากนี้วิธีการที่เสนอนี้ยังสามารถปรับตัวตามชุดข้อมูลตัวอย่างต่างๆ ได้ ผู้วิจัยได้ทดสอบวิธีการที่นำเสนอกับชุดข้อมูลตัวอย่างมาตรฐานที่มีมิติขนาดใหญ่ ทั้งด้านจำนวนลักษณะเฉพาะและจำนวนตัวอย่าง โดยใช้ประสิทธิภาพของตัวจำแนกเป็นตัวชี้วัดประสิทธิภาพของการคัดเลือกลักษณะเฉพาะและเปรียบเทียบกับวิธีการคัดเลือกลักษณะเฉพาะแบบผสมที่นิยมใช้งานกันในงานวิจัยด้านการจำแนกข้อมูล ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถคัดเลือกลักษณะเฉพาะมากลุ่มหนึ่งที่มีจำนวนน้อยมากเมื่อเทียบกับจำนวนลักษณะเฉพาะทั้งหมด แต่ตัวจำแนกข้อมูลยังสามารถให้ผลลัพธ์ความถูกต้องของการจำแนกข้อมูลที่มีมิติขนาดใหญ่ได้ในระดับที่น่าพอใจ

ทบทวนวรรณกรรม

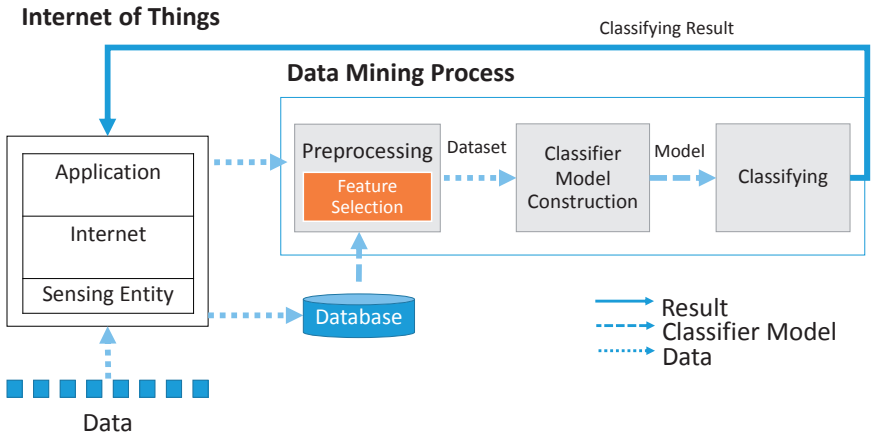
จากที่ได้กล่าวไว้เบื้องต้น ดังภาพที่ 1 เป็นการอธิบายเพิ่มเติม ระบบไอโอที ซึ่งประกอบด้วยระบบอุปกรณ์เซนเซอร์ (Sensing Device) เพื่อการตรวจจับข้อมูลสภาพแวดล้อม ระบบวิเคราะห์ข้อมูล (Data Analytics System) เพื่อการตัดสินใจอย่างชาญฉลาด และการนำเสนอข้อมูล (Data Visualization) เพื่อสนับสนุนการตัดสินใจ เหล่านี้เพื่อประโยชน์ด้านต่างๆ เช่น การอำนวยความสะดวกภายในบ้านและอาคาร การใช้งานเพื่อการขนส่ง การเพิ่มศักยภาพภายในชุมชน ภาคธุรกิจและสังคม การบริหารและพัฒนาประเทศ โดยอาจมีบุคคลที่เกี่ยวข้องได้แก่ ผู้ใช้ในบ้าน แพทย์ ผู้บริหารองค์กร ผู้บริหารประเทศ เป็นต้น



ภาพที่ 1: ระบบอินเตอร์เน็ตที่เชื่อมต่อในทุกสิ่งหรือระบบไอโอที
ที่มา: Gubbi, Buyya, Marusic & Palaniswami, 2013, 1645-1660

การประยุกต์ใช้ระบบไอโอทีที่ได้รับความสนใจอย่างกว้างขวาง คือ การพัฒนาระบบอัจฉริยะบนพื้นฐานการจำแนกข้อมูลเพื่ออำนวยความสะดวกต่อผู้ใช้งาน ภาพที่ 2 เป็นการอธิบายเพิ่มเติม ขั้นตอนพื้นฐานในการพัฒนาระบบอัฉริยะด้วยข้อมูลจากระบบไอโอที โดยเริ่มต้นจากการเก็บข้อมูลจากหลายๆ แหล่งที่แตกต่างกันภายในระบบไอโอทีลงในฐานข้อมูล (Database) กระบวนการพัฒนาตัวจำแนกข้อมูลจะเริ่มจากการนำข้อมูลในฐานข้อมูลมาผ่านการเตรียมข้อมูลเบื้องต้น (Preprocessing) เพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมกับการสร้างโมเดลการจำแนกข้อมูล เรียกว่า ชุดข้อมูลตัวอย่างสำหรับการฝึกหัด (Training Dataset) ซึ่งข้อมูลตัวอย่างทั้งหมดจะถูกแทนในรูปแบบตารางเมทริกซ์ (Marix) ขนาด m x n โดย m คือ จำนวนข้อมูลตัวอย่าง และ n คือ จำนวนลักษณะเฉพาะ ทุกๆ ข้อมูลตัวอย่างจะต้องมีการระบุผลเฉลยคำตอบว่า ตัวอย่างนั้นๆ อยู่ในกลุ่มคำตอบ (Class) ใด ยกตัวอย่างเช่น ในระบบควบคุมอุณหภูมิภายในบ้าน ข้อมูลตัวอย่างประกอบด้วยลักษณะเฉพาะ คือค่าอุณหภูมิ ความชื้น จำนวนคน การเคลื่อนไหวของบุคคลภายในบ้าน เป็นต้น และกลุ่มคำตอบคือช่วงอุณหภูมิที่เหมาะสม ซึ่งมี 3 กลุ่มคำตอบได้แก่ อบอุ่น หรือ ปกติเพื่อการประหยัดพลังงาน หรือ เย็นสบาย เป็นต้น ชุดข้อมูลตัวอย่างที่เตรียมไว้จะถูกนำไปยังขั้นตอนการสร้างตัวจำแนกโดยใช้กระบวนการเรียนรู้เครื่องจักร (Machine Learning) เรียนรู้รูปแบบของสารสนเทศภายในข้อมูลเพื่อสร้างตัวจำแนก (Classifier) ที่สามารถวินิจฉัยข้อมูลที่เข้ามาในระบบและจำแนกกลุ่มคำตอบของข้อมูลนั้น

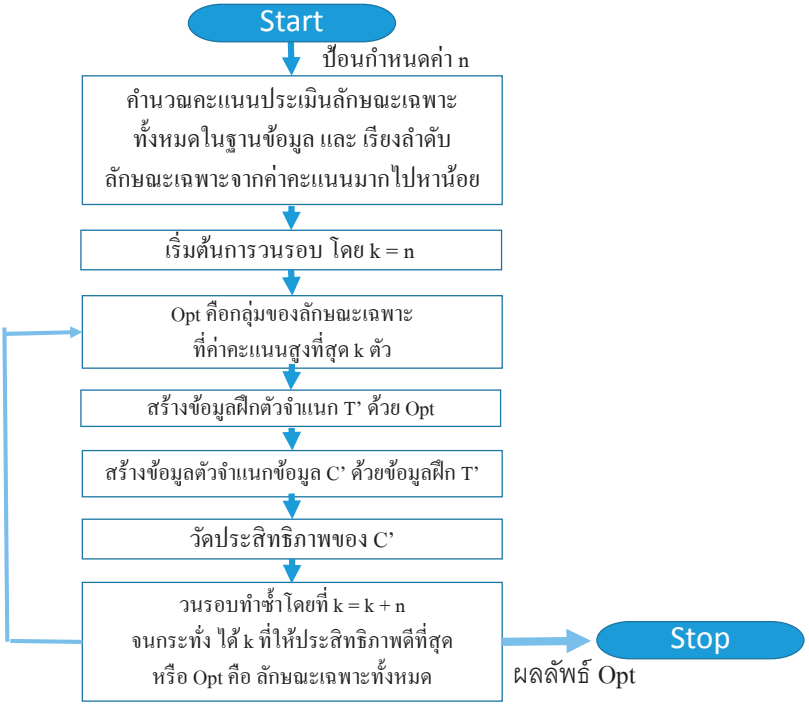
การคัดเลือกลักษณะเฉพาะเป็นส่วนหนึ่งของการเตรียมข้อมูลเบื้องต้นเพื่อลดจำนวนลักษณะเฉพาะให้มีปริมาณที่เหมาะสมต่อการนำไปสร้างโมเดลการจำแนกข้อมูล ตัวอย่างเช่น อุปกรณ์ต่างๆ ในระบบไอโอทีอาจมีการตรวจจับ วัดค่า หรือสร้างข้อมูลใดๆได้มากมาย ข้อมูลเหล่านี้จะเป็นข้อมูลลักษณะเฉพาะที่มีจำนวนมหาศาล กลุ่มของลักษณะเฉพาะที่ถูกเลือกจะเรียกว่า กลุ่มลักษณะเฉพาะที่เหมาะสม (Optimal Feature Set) ที่สามารถเป็นตัวแทนของสารสนเทศที่มีอยู่ในข้อมูลตัวอย่างทั้งหมดได้



ภาพที่ 2: ขั้นตอนวิธีการพัฒนาระบบจำแนกข้อมูลในระบบไอโอที
ที่มา: Tsai, Lai, Chiang & Yang, 2014, 77-97

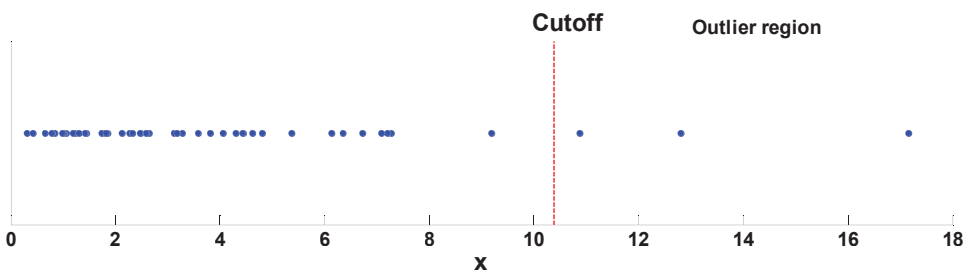
การคัดเลือกลักษณะเฉพาะด้วยการเรียงลำดับและการห่อหุ้ม (Ranking Wrapper with Sequential Forward, RW-SF) (Ruiz, Riquelme, Aguilar-Ruiz & Garcia-Torres, 2012, 11094-11102) เป็นหนึ่งในวิธีการคัดเลือกลักษณะเฉพาะที่ใช้เป็นมาตรฐานในขั้นตอนการวิจัยและพัฒนาระบบจำแนกข้อมูล ขั้นตอนวิธีมีขั้นตอนดังแสดงในภาพที่ 3 ขั้นแรก คือ การคำนวณคะแนนความมีประโยชน์ของลักษณะเฉพาะทั้งหมดที่มีในฐานข้อมูล และเรียงลำดับลักษณะเฉพาะตามค่าคะแนนจากมากไปน้อย ต่อมาคือ การเลือกลักษณะเฉพาะที่มีค่าคะแนนสูงมา k จำนวน (k Top-Ranked) เรียกว่า กลุ่มของลักษณะเฉพาะที่เลือกไว้ (Opt) ขั้นที่สามจะใช้อัลกอริธึมการเรียนรู้ที่งานวิจัยเลือกไว้ร่วมกับชุดข้อมูลตัวอย่างที่พิจารณาเพียงลักษณะเฉพาะที่เลือกไว้ในสร้างตัวจำแนกขึ้นมา และวัดประสิทธิภาพของตัวจำแนกนั้น เก็บค่าประสิทธิภาพของกลุ่มลักษณะเฉพาะที่เลือกไว้ ต่อจากนั้นเพิ่มจำนวนลักษณะเฉพาะอีก n จำนวน กระบวนการจะวนทำซ้ำในขั้นตอนที่สามใหม่อีกครั้งหนึ่ง และเก็บค่าประสิทธิภาพทุกๆ ค่า k ที่เปลี่ยนไป กระบวนการจะวนทำซ้ำเพิ่มค่า k ด้วยค่า n ไปเรื่อยๆ จนกระทั่งครบทุกๆ ลักษณะเฉพาะ สุดท้ายจะมีการพิจารณาว่า จำนวนลักษณะเฉพาะเท่าใดที่ให้ค่าการประเมินประสิทธิภาพตัวจำแนกสูงที่สุด แล้วใช้กลุ่มของลักษณะเฉพาะนั้นในการนำไปพัฒนาระบบจำแนกข้อมูลจริง ต่อไป

การคำนวณคะแนนประเมินลักษณะเฉพาะ (Feature Scoring) เป็นการประเมินประโยชน์ของลักษณะเฉพาะที่มีต่อประสิทธิภาพของตัวจำแนกโดยส่วนใหญ่แล้วจะเป็นวัดระดับสหสัมพันธ์ (Correlation) ระหว่างลักษณะเฉพาะกับกลุ่มคำตอบของการจำแนก ลักษณะเฉพาะที่มีค่าคะแนนสูงแสดงว่ามีความสัมพันธ์กับกลุ่มคำตอบเป็นอย่างมาก ซึ่งจะทำให้เมื่อนำไปใช้เป็นข้อมูลการเรียนรู้จะให้ตัวจำแนกที่มีความแม่นยำ คะแนนประเมินลักษณะเฉพาะที่นิยมใช้กัน ได้แก่ อินฟอร์เมชันเกน (Information Gain, IG) ไคสแคว์ (CHI-square) และ จีนีอินเด็กซ์ (GINI) เป็นต้น (Yang, Liu, Zhu, Liu & X.Zhang, 2012, 741-754) IG วัดสารสนเทศของลักษณะเฉพาะที่มีในแต่ละกลุ่มคำตอบ CHI วัดความเป็นอิสระ (Independence) ระหว่างลักษณะเฉพาะกับแต่ละกลุ่มคำตอบ และ GINI วัดความบริสุทธิ์ (Purity) ของลักษณะเฉพาะในแต่ละกลุ่มคำตอบ ผลวิจัยที่ผ่านมาชี้ว่า GINI ส่งผลต่อประสิทธิภาพของการจำแนกข้อมูลโดยรวมได้ดีที่สุด ส่วน IG นั้นเหมาะสำหรับการจำแนกข้อมูลที่มีหลายกลุ่มคำตอบ (Multi-Class) และ CHI เหมาะสำหรับการจำแนกข้อมูลที่มี 2 กลุ่ม (Binary-Class)



ภาพที่ 3: การคัดเลือกลักษณะเฉพาะด้วยการเรียงลำดับและการห่อหุ้ม
ที่มา: Ruiz, Riquelme, Aguilar-Ruiz & Garcia-Torres, 2012, 11094-11102

การระบุข้อมูลที่ผิดปกติ (Outlier Identification) เป็นเทคนิคทางสถิติสำหรับการระบุข้อมูลตัวอย่างที่มีค่าข้อมูลที่แตกต่างไปจากค่าข้อมูลที่เหลือ (Leys, Ley, C. , Klein, Bernard & Licata, 2013, 764-766) แนวคิดการระบุข้อมูลที่ผิดปกติจากข้อมูลทั้งหมดสามารถอธิบายด้วยภาพที่ 4 กำหนดให้ข้อมูลตัวอย่าง คือ แต่ละจุดบนแผนภูมิ จำนวน 50 ตัวอย่าง คือ 50 จุดข้อมูล การระบุข้อมูลจะกำหนดขอบเขตบริเวณเพื่อระบุการเป็นข้อมูลผิดปกติ (Outlier Region) หากจุดข้อมูลมีค่าเกินขีดแบ่งที่กำหนดไว้ (Predetermined Cut-off) นี้จะถูกระบุว่าเป็นข้อมูลผิดปกติ ดังนั้นจากภาพที่ 4 ข้อมูล 3 จุด ที่มีค่าของข้อมูลสูงว่าค่าขีดแบ่งจะถูกระบุว่าเป็นข้อมูลผิดปกติไปจากข้อมูลที่เหลือ



ภาพที่ 4: การระบุข้อมูลผิดปกติด้วยการกำหนดขอบเขตข้อมูล
ที่มา: Seo, 2006

วิธีการที่ได้นำเสนอ

ในงานวิจัยนี้มุ่งเน้นการปรับปรุงวิธีการคัดเลือกลักษณะเฉพาะแบบตัวกรองเพื่อให้เหมาะสมกับการนำไปใช้ในการพัฒนาตัวจำแนกข้อมูลในเทคโนโลยีไอโอทีในส่วนของการพยายามเพิ่มประสิทธิภาพของการเลือกลักษณะเฉพาะ โดยงานวิจัยได้ใช้แนวคิดการแทนค่าคะแนนของลักษณะเฉพาะ (IG หรือ CHI หรือ GINI) คือจุดข้อมูลแต่ละจุด ลักษณะเฉพาะที่มีค่าคะแนนสูงจนผิดปกติ ก็คือลักษณะเฉพาะที่มีสหสัมพันธ์กับกลุ่มคำตอบของชุดข้อมูลตัวอย่างมากๆ และจะมีประโยชน์มากต่อการสร้างตัวจำแนก ดังนั้น หากสามารถคำนวณค่าหาเส้นขีดแบ่งสำหรับขอบเขตการระบุข้อมูลผิดปกติได้อย่างมีประสิทธิภาพ ก็จะสามารถระบุว่าลักษณะเฉพาะใดมีข้อมูลที่ผิดปกติไปจากลักษะเฉพาะที่เหลือนั่นเอง และกลุ่มของลักษณะ เฉพาะที่มีค่าคะแนนสูงผิดปกตินั้นก็จะถูกเลือกไปใช้พัฒนาตัวจำแนกข้อมูลต่อไป โดยในงานวิจัยได้นำเสนอเทคนิคการคำนวณค่าเส้นขีดแบ่งด้วยวิธีทางสถิติ 3 วิธี (Seo, 2006) คือ

1. **วิธีคำนวณด้วยค่าส่วนเบี่ยงเบนมาตรฐาน** (Standard Deviation Method: SD)

เป็นวิธีพื้นฐานคำนวณค่าขีดแบ่งเพื่อระบุข้อมูลผิดปกติซึ่งมักจะถูกใช้บ่อยครั้งที่สุดในงานวิจัยเชิงสถิติ ตั้งอยู่บนสมมติฐานว่าข้อมูลมีลักษณะการกระจายแบบปกติ ค่าขีดแบ่ง (Cut-off, θ) สามารถคำนวณด้วยสมการ $\theta = \mu + \alpha\ \sigma$ โดยที่ μ คือค่าเฉลี่ย (Mean) ของคะแนนของลักษณะเฉพาะทั้งหมดในชุดข้อมูลตัวอย่าง σ คือ ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard deviation) ของคะแนนของลักษณะเฉพาะทั้งหมดในชุดข้อมูลตัวอย่าง และ α คือค่าสัมประสิทธิ์ความเชื่อมั่น (Confidence Coefficient) ซึ่งเป็นเพียงพารามิเตอร์เดียวที่ต้องกำหนดในวิธีการที่นำเสนอนี้ ซึ่งส่วนใหญ่มักใช้ตัวเลข 2 หรือ 2.5 หรือ 3 จำนวนของลักษณะเฉพาะที่ถูกเลือกจะแปรตามค่าของ α ซึ่งจะขึ้นกับลักษณะของข้อมูลในแต่ละแอพพลิเคชั่น ซึ่งบทความจะอธิบายเพิ่มเติมในส่วนของผลการทดลองต่อไป

2. **วิธีคำนวณด้วยขอบเขตอินเตอร์ควอร์ไทล์** (Interquartile Range: IQR)

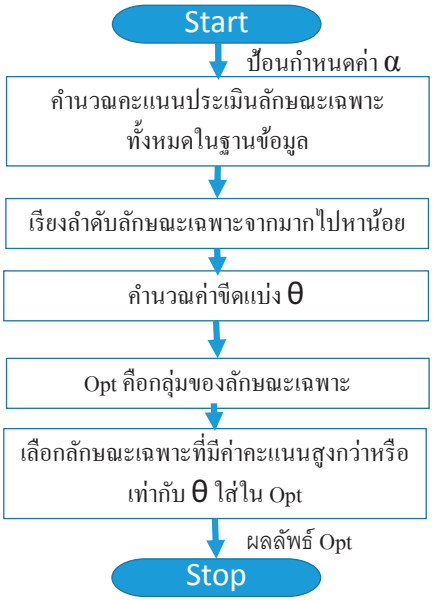
ในทางสถิติเชิงพรรณนาค่าขอบเขตอินเตอร์ควอร์ไทล์ คือการวัดการกระจายโดยหาค่าความแตกต่างระหว่างควอร์ไทล์อันดับที่ 3 (Q3) กับควอร์ไทล์อันดับที่ 1 (Q1) คำนวณจาก IQR = Q3-Q1 ซึ่ง IQR นี้สามารถใช้วัดว่าค่าของข้อมูลมีการกระจายตัวอย่างไรโดยไม่อยู่บนสมมติฐานการกระจายแบบปกติ นักวิจัยสามารถใช้ IQR เพื่อบอกว่า ข้อมูลใดที่มีลักษณะห่างจากค่ากลางของข้อมูลมากๆ คือ ข้อมูลผิดปกติ ดังนั้น ค่าขีดแบ่งจึงคำนวณจากสมการดังนี้ θ = Q3 + α IQR ซึ่งโดยปกติเลขสัมประสิทธิ์ความเชื่อมั่น α มักใช้ค่าตัวเลขในช่วง 1.5 ถึง 3

3. **วิธีคำนวณด้วยค่ามัธยฐานของส่วนเบี่ยงเบนสัมบูรณ์** (Median Absolute Deviation: MAD)

ค่าเฉลี่ยของกลุ่มตัวอย่าง (Sample Mean) และค่าส่วนเบี่ยงเบนมาตรฐานของกลุ่มตัวอย่าง (Sample Standard Deviation) มักจะได้รับผลกระทบจากข้อมูลบางกลุ่มที่มีค่าผันผวนมากๆ ดังนั้นในทางสถิติจึงหลีกเลี่ยงการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐาน โดยเสนอการใช้ค่ามัธยฐาน (Median) และวัดการกระจายข้อมูลด้วย MAD แทน โดยกำหนดให้ MAD = 1.483 median ($|x_i - M|$) ซึ่ง M คือ ค่ามัธยฐานของของข้อมูล x_i ทั้งหมด ดังนั้นค่าขีดแบ่งจึงคำนวณจากสมการดังนี้ θ = M + α MAD ซึ่งโดยปกติเลขสัมประสิทธิ์ความเชื่อมั่น α มักใช้ค่าตัวเลขในช่วง 1.5 ถึง 5

ค่าขีดแบ่งเพื่อระบุการเป็นข้อมูลผิดปกติที่ถูกคำนวณได้ทั้ง 3 วิธีนี้จะถูกใช้ในขั้นตอนวิธีคัดเลือกลักษณะเฉพาะด้วยแนวคิดการระบุค่าข้อมูลผิดปกติ ดังขั้นตอนในภาพที่ 5 เริ่มต้นจากการคำนวณค่าคะแนน

ประเมินของลักษณะเฉพาะแต่ละตัว (โดยใช้ IG หรือ CHI หรือ GINI แล้วแต่ข้อมูลที่ผู้วิจัยสนใจ) ขั้นตอนต่อมาคือ การคำนวณค่าขีดแบ่งด้วยค่าคะแนนของลักษณะเฉพาะทั้งหมดในฐานข้อมูล โดยในขั้นตอนนี้จะมีการกำหนดค่าสัมประสิทธิ์ตามข้อมูลที่ใช้ในงานวิจัย สุดท้ายเมื่อได้ค่าขีดแบ่งแล้วก็นำไปใช้คัดเลือกลักษณะเฉพาะที่มีค่าสูงหรือเท่ากับค่าขีดแบ่ง รวบรวมลักษณะเฉพาะเหล่านั้นเพื่อนำไปในขั้นตอนการพัฒนาตัวจำแนกต่อไป



ภาพที่ 5: ขั้นตอนวิธีการคัดเลือกลักษณะเฉพาะด้วยค่าขีดแบ่งข้อมูลผิดปกติ

วิธีการที่เสนอนี้สามารถเลือกลักษณะเฉพาะที่มีค่าการมีประโยชน์ต่อการสร้างตัวจำแนกในระดับที่สูงที่สุดมากลุ่มหนึ่งโดยจะมีจำนวนลักษณะเฉพาะเพียงเล็กน้อยเมื่อเทียบกับจำนวนลักษณะเฉพาะทั้งหมดที่มีในฐานข้อมูล ซึ่งขั้นตอนวิธีการเลือกลักษณะเฉพาะเป็นแบบตัวกรองมีค่าใช้จ่ายในการคำนวณ (Computation Cost) ที่น้อยกว่าวิธีแบบห่อหุ้มและแบบผสม ในภาพที่ 3 อย่างชัดเจน เพราะไม่มีการวนรอบทำซ้ำ และไม่ใช้อัลกอริธึมการเรียนรู้ใดๆ ในการเลือกกลุ่มลักษณะเฉพาะที่ดีที่สุด

ผลการวิจัย

งานวิจัยได้ประเมินประสิทธิภาพของวิธีการคัดเลือกลักษณะเฉพาะแบบตัวกรองที่ได้นำเสนอโดยใช้ประสิทธิภาพของตัวจำแนกที่สร้างจากลักษณะเฉพาะที่ถูกเลือกเป็นตัวชี้วัดคุณภาพของวิธีการที่นำเสนอโดยใช้การประเมินประสิทธิภาพแบบข้ามกลุ่มจำนวน 10 กลุ่ม (10 Fold Cross-Validation) ร่วมกับชุดข้อมูลมาตรฐานที่มีลักษณะเป็นข้อมูลที่มีหลายกลุ่มคำตอบและมีมิติขนาดใหญ่ ทั้งด้านลักษณะเฉพาะและข้อมูล

ตัวอย่างจำนวนมาก ซึ่งได้รับความนิยมใช้ในงานวิจัยที่เกี่ยวข้อง (Yang et al., 2012, 741-754) ตารางที่ 1 แสดงรายละเอียดของชุดข้อมูลมาตรฐานที่ใช้ในการทดลอง ซึ่งข้อมูลมีลักษณะเป็น 8 กลุ่มคำตอบ (8 Classes Dataset) และมีมากกว่า 2,000 ลักษณะเฉพาะ โดยแต่ละกลุ่มจะมีจำนวนข้อมูลตัวอย่างในช่วง 79 ถึง 351 ตัวอย่างและมีค่าสัมประสิทธิ์การกระจาย (Coefficient of Variation: CV) 0.45

ตารางที่ 1

จำนวนตัวอย่างในชุดข้อมูลการทดลอง

กลุ่มคำตอบ (Class)	จำนวนตัวอย่าง	เปอร์เซ็นต์	กลุ่มคำตอบ (Class)	จำนวนตัวอย่าง	เปอร์เซ็นต์
A	351	17.87%	E	179	9.11%
B	345	17.57%	F	176	8.96%
C	343	17.46%	G	148	7.54%
D	343	17.46%	H	79	4.02%
			รวมทั้งหมด	1964	100.00%

ในส่วนของตัวจำแนกที่ใช้สำหรับการวัดประสิทธิภาพ ในงานวิจัยนี้ได้เลือกใช้ตัวจำแนกแบบซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) หรือ เอสวีเอ็ม (SVM) (Cortes & Vapnik, 1995, 273-297) ซึ่งเป็นตัวจำแนกข้อมูลที่ได้รับความนิยมในการประยุกต์ใช้เพื่อพัฒนาระบบอัจฉริยะต่างๆ เป็นจำนวนมาก เนื่องจากให้ผลลัพธ์ความแม่นยำการจำแนกสูงและทนทานต่อข้อมูลที่มิติขนาดใหญ่ได้เป็นอย่างดี นอกจากนี้เพื่อเปรียบเทียบความสามารถในการคัดเลือกลักษณะเฉพาะที่เหมาะสม งานวิจัยจึงต้องมีการวัดประสิทธิภาพของตัวจำแนกที่สร้างจากข้อมูลที่ประกอบด้วยลักษณะเฉพาะที่คัดเลือกไว้ โดยมาตรวัดประสิทธิภาพของตัวจำแนกที่ใช้ในการทดลองแบบเอฟวัน (F1-measure) ซึ่งค่าเอฟวันประกอบด้วยการคำนวณค่าความแม่นยำ (Precision, P) และค่ารีคอล (Recall, R) ซึ่งตัวชี้วัดประสิทธิภาพมีการคำนวณ ดังนี้

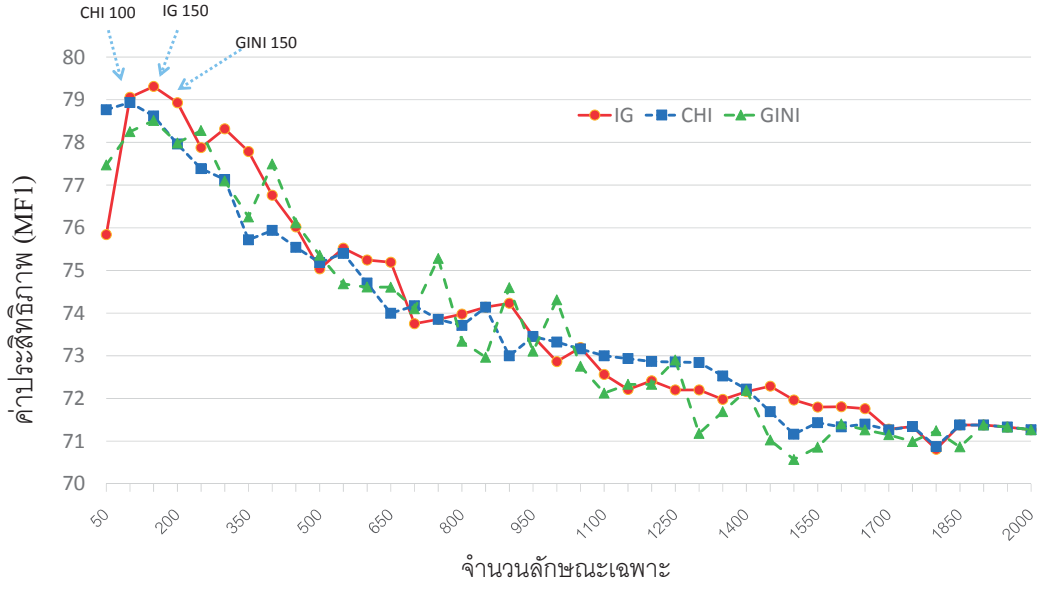
สำหรับข้อมูลกลุ่มที่มีเฉลยคำตอบ Class A

$$F1_A = \frac{2\times P_A \times R_A}{(P_A + R_A)},\ P_A = \frac{TP_A}{(TP_A + FP_A)},\ R_A = \frac{TP_A}{(TP_A + FN_A)}$$

โดยที่ TP_A (True Positive) หมายถึง จำนวนของข้อมูลตัวอย่างที่ถูกจำแนกให้อยู่กลุ่ม A อย่างถูกต้อง FN_A (False Negative) หมายถึง จำนวนของข้อมูลตัวอย่างที่อยู่ในกลุ่ม A แต่ถูกจำแนกผิดไปเป็นกลุ่มอื่นๆ (B จนถึง H) และ FP_A หมายถึง จำนวนข้อมูลตัวอย่างกลุ่มที่ไม่ใช่กลุ่ม A แต่ถูกจำแนกผิดไปอยู่กลุ่ม A เนื่องจากชุดข้อมูลทดสอบมาตรฐานในงานวิจัยนี้มีลักษณะเป็นข้อมูลที่มีหลายกลุ่มคำตอบ ดังนั้นต้อง

มีการประเมินประสิทธิภาพรวมของทุกๆ กลุ่มคำตอบ งานวิจัยจึงใช้การวัดประสิทธิภาพรวมทุกกลุ่มแบบเฉลี่ยค่าเอฟวันทุกกลุ่ม (Macro-averaged F1, MF1) โดย MF1 สามารถคำนวณได้จากสมการ $MF1 = \frac{\sum F1_{C_a}}{|C|}$ โดยที่ C คือเซตของกลุ่มคำตอบทั้งหมดในฐานข้อมูล {A,..,H} และ |C| คือจำนวนของกลุ่มคำตอบที่เป็นไปได้ทั้งหมด ซึ่งในการทดลองนี้คือ 8 กลุ่ม

นอกจากนี้ในการทดลองจำเป็นต้องมีการหาจำนวนลักษณะเฉพาะที่ดีที่สุดไว้สำหรับเปรียบเทียบกับจำนวนลักษณะเฉพาะที่คัดเลือกโดยวิธีการที่นำเสนอ ซึ่งในบทความนี้ใช้วิธีการคัดเลือกลักษณะเฉพาะแบบผสมการเรียงลำดับและการห่อหุ้ม (RW-SF) เป็นวิธีการบรรทัดฐาน (Baseline) และจะเปรียบเทียบประสิทธิภาพของการจำแนกด้วยกลุ่มลักษณะเฉพาะที่เลือกโดยวิธีการนำเสนอ กับกลุ่มลักษณะเฉพาะที่เลือกโดยวิธีการบรรทัดฐาน ดังนั้นผู้วิจัยจึงได้เริ่มต้นดำเนินการวิจัยโดยใช้วิธีการคัดเลือกแบบผสมกับชุดข้อมูลตัวอย่างทำการทดลองกับคะแนนลักษณะเฉพาะ 3 แบบคือ IG CHI และ GINI โดยเริ่มตั้งแต่จำนวนลักษณะสำคัญที่สูงที่สุด 50 ตัว เพิ่มจำนวนทีละ 50 ตัว ไปจนครบ 2000 ตัว ดำเนินการประเมินประสิทธิภาพแบบข้ามกลุ่มจำนวน 10 กลุ่ม และทำซ้ำ 5 ครั้งและหาค่าเฉลี่ยของ MF1ของทั้ง 5 ครั้ง ผลการดำเนินการในขั้นตอนนี้เป็นดังภาพที่ 6 ซึ่งจะเห็นได้ว่า ด้วยวิธีการคัดเลือกลักษณะเฉพาะแบบผสมการเรียงลำดับคะแนนและการห่อหุ้มกับตัวจำแนก (RW-SF) ด้วยคะแนนลักษณะเฉพาะแบบ IG จะได้จำนวนลักษณะเฉพาะเท่ากับ 150 และ CHI เท่ากับ 100 และ GINI เท่ากับ 150 ซึ่งจากภาพจะเห็นได้ว่า การใช้ลักษณะเฉพาะทั้งหมดในฐานข้อมูลไม่ได้ช่วยให้ประสิทธิภาพตัวจำแนกดีที่สุด และการเลือกเพียงบางกลุ่มของลักษณะเฉพาะที่มีประโยชน์จะช่วยให้ได้ตัวจำแนกที่มีประสิทธิภาพ อย่างไรก็ตามผลลัพธ์จะมีการเปลี่ยนแปลงไปตามคะแนนประเมินประสิทธิภาพเพื่อการคัดเลือกลักษณะเฉพาะ



ภาพที่ 6: ค่าประสิทธิภาพ MF1ของแต่ละคะแนนลักษณะเฉพาะ

ต่อจากนั้นผู้วิจัยได้ใช้วิธีคัดเลือกลักษณะเฉพาะแบบตัวกรองด้วยการประเมินค่าขีดแบ่งข้อมูลผิดปกติทั้ง 3 แบบ ที่ได้นำเสนอเรียกว่า วิธีการแบบ Outlier-Based กับข้อมูลทดสอบมาตรฐานเช่นเดียวกัน โดยใช้ค่าพารามิเตอร์ α ที่แตกต่างกันหลายๆ ค่าเพื่อหาผลลัพธ์ที่ดีที่สุด จำนวนลักษณะเฉพาะที่สูงผิดปกติจะแสดงในตารางที่ 2 และเปรียบเทียบกับจำนวนลักษณะเฉพาะที่เลือกโดยวิธี RW-SF ก่อนหน้านี้ ซึ่งจากผลการทดลองจะเห็นได้ว่า วิธีการที่เสนอสามารถเลือกกลุ่มลักษณะเฉพาะที่มีจำนวนน้อยกว่ากลุ่มที่เลือกโดยวิธีบรรทัดฐาน

ตารางที่ 2
เปรียบเทียบจำนวนลักษณะเฉพาะที่คัดเลือกโดยวิธีต่างๆ

คะแนนลักษณะเฉพาะ (Feature Score)	**วิธีการเลือกลักษณะเฉพาะ** (Featuer Selection Method)			
	วิธี RW-SF	**วิธี** Outlier-Based **โดยใช้การประมาณค่าขีดแบ่ง 3 แบบ**		
		SD (α = 2)	IQR (α = 3)	MAD (α = 5)
IG	150	55	95	111
CHI	100	50	98	135
GINI	150	67	107	172

แม้ว่าวิธีการที่นำเสนอจะสามารถคัดเลือกกลุ่มลักษณะเฉพาะที่มีขนาดเล็กกว่า แต่ยังคงต้องทดสอบว่าลักษณะเฉพาะนี้สามารถนำไปใช้พัฒนาตัวจำแนกข้อมูลที่มีหลายกลุ่มคำตอบได้อย่างมีประสิทธิภาพหรือไม่ โดยจะต้องเปรียบเทียบกับกลุ่มลักษณะเฉพาะที่คัดเลือกด้วยวิธี RW-SF การทดลองใช้การประเมินประสิทธิภาพแบบข้ามกลุ่ม จำนวน 10 กลุ่ม และหาค่าเฉลี่ยจากการทำซ้ำ 5 ครั้งเช่นเดียวกัน ตารางที่ 3 แสดงค่า MF1 ของแต่ละกลุ่มลักษณะเฉพาะที่เลือกด้วยวิธีต่างๆ และใช้สถิติทดสอบเปรียบเทียบค่าเฉลี่ยของ 2 กลุ่มตัวอย่างแบบที (Two Sample Mean T-test) เพื่อพิจารณาความแตกต่างอย่างมีนัยสำคัญทางสถิติระหว่างประสิทธิภาพของวิธีแบบ Outlier-Based กับ วิธี RW-SF ในตารางที่ 3 นี้ ค่า p-Val หมายถึงค่าความน่าจะเป็นที่แสดงว่าค่าเฉลี่ยของประสิทธิภาพของทั้งสองวิธีมีความแตกต่างกันมากหรือน้อย ถ้าค่าใกล้เคียงกับ 1 แสดงว่า มีความแตกต่างกันน้อย แต่ถ้าค่า p-Val เข้าใกล้ 0 แสดงว่าความแตกต่างกันยิ่งมากขึ้น ในส่วนของสัญลักษณ์ + และ – หมายถึงประสิทธิภาพของวิธี Outlier-Based สูงกว่า RW-SF หรือไม่ ถ้าใช้สัญลักษณ์ + แสดงว่า วิธีการ Outlier-Based มีประสิทธิภาพดีกว่าวิธีการ RW-SF โดยจากการทดลองแสดงให้เห็นว่าประสิทธิภาพของวิธี Outlier-Based โดยใช้ IQR และ MAD ไม่มีความแตกต่างกับ RW-SF อย่างมีนัยสำคัญโดยส่วนใหญ่ วิธีการคัดเลือกแบบ Outlier Based ที่ใช้วิธีประเมินเส้นขีดแบ่งแบบ IQR ให้ผลลัพธ์กับคะแนนลักษณะเฉพาะแบบ GINI ได้ดีที่สุด ขณะเดียวกันก็มีจำนวนลักษณเฉพาะที่น้อยกว่าวิธี RW-SF ดังนั้นผลการทดลองจึงสรุปได้ว่า วิธีการคัดเลือกลักษณะเฉพาะแบบ Outlier Based

ร่วมกับการประมาณค่าขีดแบ่งที่ได้นำเสนอสามารถลดจำนวนลักษณะเฉพาะได้น้อยกว่าวิธีการแบบ RW-SF แต่ก็ยังให้ผลการจำแนกข้อมูลได้น่าพอใจและไม่แตกต่างอย่างมีนัยยะสำคัญทางสถิติ

ตารางที่ 3

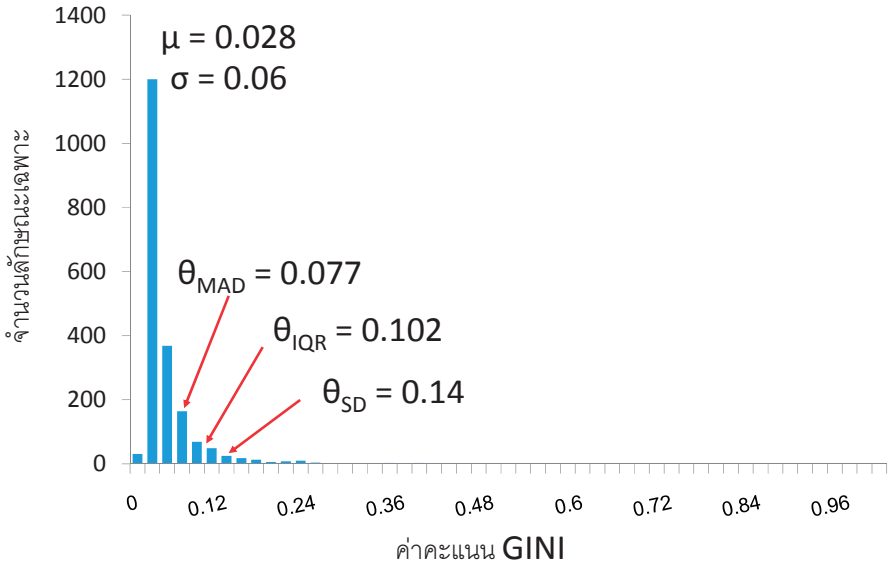
เปรียบเทียบค่าประเมินประสิทธิภาพ MF1 ของตัวจำแนก SVM กับกลุ่มลักษะเฉพาะที่เลือกโดยวิธีต่างๆ

คะแนนลักษณะเฉพาะ (Feature Score)	**วิธีการเลือกลักษณะเฉพาะ** (Featuer Selection Method)						
	วิธี RW-SF	**วิธี** Outlier-based **โดยใช้การประมาณค่าขีดแบ่ง 3 แบบ**					
		SD (α = 2)		IQR (α = 3)		MAD (α = 5)	
	MF1	MF1	p-Val	MF1	p-Val	MF1	p-Val
IG	79.31	78.09	0.01-	79.03	0.65	**79.34**	0.96
CHI	78.94	78.77	0.86	**79.55**	0.52	79.09	0.87
GINI	78.52	78.17	0.79	**79.70**	0.26	78.87	0.71

อย่างไรก็ตามเพื่ออธิบายเพิ่มเติมในส่วนของการคำนวณค่าขีดแบ่งเพื่อระบุลักษณะเฉพาะที่ผิดปกติ ในกรณีที่ใช้วิธีคำนวณแบบ SD ให้ผล MF1 ที่ต่ำกว่าวิธีการคัดเลือกแบบ RW-SF เป็นผลมาจากกลุ่มของลักษณะเฉพาะที่ถูกคัดเลือกมีจำนวนน้อยเกินไปสำหรับข้อมูลที่มีหลายกลุ่มคำตอบ ทำให้ขั้นตอนการพัฒนาตัวจำแนกข้อมูลไม่สามารถสกัดสารสนเทศในฐานข้อมูลทั้งหมดได้ ในขณะที่วิธีการคำนวณแบบ IQR และ MAD สามารถเลือกลักษณะเฉพาะในจำนวนที่เหมาะสมมากกว่า และเมื่อพิจารณาจากภาพที่ 7 ซึ่งแสดงฮิสโตแกรมการกระจายของค่าคะแนน GINI ของลักษณะเฉพาะที่มีอยู่ในฐานข้อมูลทั้งหมด จะเห็นว่าการกระจายไม่ได้มีลักษณะเป็นการกระจายปกติ แต่สมมติฐานของวิธีการคำนวณแบบ SD คือข้อมูลมีการกระจายแบบปกติ ดังนั้นวิธีการคำนวณแบบ IQR และ MAD ซึ่งเหมาะสมกับการกระจายข้อมูลลักษณะนี้สามารถกำหนดค่าขีดแบ่งเพื่อระบุข้อมูลผิดปกติได้ดีกว่าจึงเลือกลักษณะเฉพาะได้ดีกว่านั่นเอง และค่าคะแนนแบบ CHI และ IG ของลักษณะเฉพาะทั้งหมดก็จะมีลักษณะเช่นเดียวกันกับค่าคะแนน GINI นั่นเอง

จากผลการทดลองที่นำเสนอไปทั้งหมด จึงสามารถสรุปได้ว่า วิธีการแบบตัวกรองเพื่อคัดเลือกลักษณะเฉพาะโดยใช้แนวคิดการระบุข้อมูลผิดปกติที่ได้นำเสนอนี้สามารถเลือกลักษณะเฉพาะที่มีความสัมพันธ์กับกลุ่มคำตอบของข้อมูลตัวอย่างได้เป็นอย่างดี และเมื่อเทียบกับวิธีการที่เป็นบรรทัดฐานและใช้ข้อมูลทดสอบมาตรฐาน วิธีการที่นำเสนอก็สามารถลดขนาดของข้อมูลตัวอย่างเพื่อนำไปสู่การพัฒนาตัวจำแนกข้อมูลได้อย่างมีประสิทธิภาพไม่แตกต่างกันแต่ขนาดของข้อมูลตัวอย่างมีขนาดเล็กกว่า ดังนั้นงานวิจัยด้านการพัฒนาตัวจำแนกข้อมูลเพื่อระบบอัจฉริยะสามารถนำวิธีการนี้ไปใช้ในขั้นตอนการเตรียมข้อมูลตัวอย่าง

แทนวิธีการมาตรฐานแบบเดิมซึ่งจะทำให้การดำเนินการวิจัยทำได้โดยสะดวกและมีประสิทธิภาพมากยิ่งขึ้น เพราะวิธีการที่นำเสนอเป็นแบบตัวกรองที่ใช้เพียง 1 ค่าพารามิเตอร์จึงไม่ต้องเสียค่าใช้จ่ายในการคำนวณ ทั้งจากขั้นตอนวิธีการเรียนรู้ การวนทำซ้ำเพื่อทดลองปรับค่าพารามิเตอร์ และใช้ได้กับทุกๆ ค่าคะแนนลักษณะเฉพาะและตัวจำแนกข้อมูล



ภาพที่ 7: ฮิสโตแกรมของค่าคะแนน GINI ของลักษณะเฉพาะ 2000 ตัว

สรุป

ในงานวิจัยนี้ได้นำเสนอตัวกรองเพื่อคัดเลือกลักษณะเฉพาะที่ใช้การประมาณค่าขีดแบ่งบนพื้นฐานของเทคนิคทางสถิติและแนวคิดการระบุข้อมูลที่ผิดปกติ ตัวกรองที่นำเสนอนี้สามารถนำไปใช้กับการเตรียมข้อมูลเพื่อการพัฒนาระบบจำแนกข้อมูลที่มีมิติของข้อมูลขนาดมหาศาล ทั้งทางด้านลักษณะเฉพาะและปริมาณข้อมูลตัวอย่าง โดยตัวกรองที่นำเสนอนี้สามารถลดขนาดมิติข้อมูลเหลือขนาดที่เล็กลงได้อย่างมีนัยสำคัญ วิธีการนี้จึงเหมาะสมกับการนำไปประยุกต์ใช้ในการพัฒนาระบบอัจฉริยะต่างๆ ในเทคโนโลยีไอโอทีที่กำลังได้รับความสนใจเป็นอย่างมาก บทความได้มีการทดลองเพื่อพิสูจน์ประสิทธิภาพของวิธีการที่ได้นำเสนอกับข้อมูลที่มีมิติข้อมูลขนาดใหญ่ ผลการทดลองแสดงให้เห็นว่า วิธีการที่นำเสนอสามารถเลือกเพียงลักษณะเฉพาะกลุ่มหนึ่งที่มีสารสนเทศที่สนับสนุนประสิทธิภาพของตัวจำแนกข้อมูล โดยมีจำนวนลักษณะเฉพาะเพียงเล็กน้อยเมื่อเทียบกับปริมาณลักษณะเฉพาะทั้งหมดในฐานข้อมูล แต่ประสิทธิภาพของตัวจำแนกที่พัฒนาจากฐานข้อมูลที่ลดขนาดแล้วนี้กลับสูงเป็นที่น่าพอใจ นอกจากนี้วิธีการที่นำเสนอยังมีขั้นตอนการทำงานที่น้อยกว่าวิธีการคัดเลือกลักษณะเฉพาะอื่นๆ ดังนั้นจึงสามารถสรุปได้ว่า วิธีการคัดเลือกลักษณะเฉพาะที่ได้นำเสนอนี้มีความเหมาะสมในการนำไปประยุกต์ใช้กับการพัฒนาตัวจำแนกข้อมูลสำหรับระบบอัจฉริยะในเทคโนโลยีไอโอที ในส่วนของงานวิจัยต่อเนื่องในอนาคตจะเป็นการทดสอบประสิทธิภาพ

ของวิธีตัวกรองคัดเลือกลักษณะเฉพาะกับตัวจำแนกอื่นๆ และข้อมูลจากระบบเทคโนโลยีไอโอทีด้านอื่นๆ ให้หลากหลายมากยิ่งขึ้น หรือกับข้อมูลที่มีลักษณะขนาดของกลุ่มคำตอบบางกลุ่มไม่สมดุลซึ่งเป็นปัญหาวิจัยด้านเหมืองข้อมูลที่กำลังได้รับความสนใจ เพื่อก่อให้เกิดแนวทางในการพัฒนาความสามารถและเพิ่มประสิทธิภาพการทำงานของระบบไอโอทีต่อไป

เอกสารอ้างอิง

Chen, F ; Deng, P. ; Wan, J. ; Zhang, D ; Vasilakos, A. V. & Rong, X. (2015). Data mining for the Internet of Things: Literature review and challenges. **International Journal of Distributed Sensor Networks**. 15, 1-14.

Combarro, E. F. ; Montanes, E. ; Diaz, I. ; Ranilla, J. & Mones, R. (2005). Introducing a family of linear measures for feature selection in text categorization. **IEEE Transactions on Knowledge and Data Engineering**. 17, 1223-1232.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. **Machine Learning**. 20, 273-297.

Gubbi, J. ; Buyya, R. ; Marusic, S. & Palaniswami, M. . (2013). Internet of Things (IoT): A vision, architectural elements, and future directions. **Future Generation Computer Systems**. 13(29), 1645-1660.

Leys, C. ; Ley, C. ; Klein, O ; Bernard, P. & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. **Journal of Experimental Social Psychology**. 49, 764-766.

Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. **IEEE Transactions on Knowledge and Data Engineering**. 17, 491-502.

Ruiz, R. ; Riquelme, R. C. ; Aguilar-Ruiz, J. S. & Garcia-Torres, M. (2012). Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches. **Expert Systems with Applications**. 39, 11094-11102.

Seo, S. (2006). **A Review and comparision of methods for detecting outliers in univariate data sets**. Master of Science, Graduate School of Public Health, University of Pittsburgh.

Tsai, Chun-Wei ; Lai, Chin-Feng ; Chiang, Ming-Chao ; Yang, L. T. (2014). Data mining for Internet of Things: A survey. **IEEE Communications Surveys & Tutorials**. 16(1), 77-97.

Yang, J ; Liu, Y. ; Zhu, X. ; Liu, Z. & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. **Information Processing & Management**. 48, 741-754.

Yu, L. & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. **J. Mach. Learn. Res**. 5, 1205-1224.