วารสารวิจัยสหวิทยาการไทย
Journal of Thai Interdisciplinary Research
Volume 11, Number 4, Pages 41 - 48

การเลือกคุณลักษณะแบบอัตราส่วนสัญญาณต่อสัญญาณรบกวน สำหรับการจำแนกประเภทข้อมูลหลายกลุ่ม

Signal-to-Noise Ratio feature selection for multi-class classification

สุพจน์ เฮงพระพรหม[1,*] และสถาพร ประณิธานวิทยา[1,2]
Supoj Hengpraprohm[1,*] and Sataporn Pranithanwitthaya[1,2]

[1]หน่วยวิจัยอัจฉริยภาพแห่งเครื่องจักร สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม
[2]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏนครปฐม
[1]Machine Intelligence Research Unit, Research and Development Institute,
Nakhon Pathom Rajabhat University
[2]Program in Computer Science, Faculty of Science and Technology,
Nakhon Pathom Rajabhat University

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการเลือกคุณลักษณะแบบอัตราส่วนสัญญาณต่อสัญญาณรบกวน (signal-to-noise ratio: SNR) สำหรับการจำแนกประเภทข้อมูลหลายกลุ่ม โดยได้ทดสอบกับชุดข้อมูลเกณฑ์มาตรฐาน (benchmark) จำนวน 6 ชุดข้อมูล แบ่งเป็นข้อมูลสำหรับปัญหาการจำแนกประเภทสองกลุ่มจำนวน 3 ชุด และข้อมูลสำหรับปัญหาการจำแนกประเภทหลายกลุ่มจำนวน 3 ชุด การทดลองเริ่มจากการทดสอบกับปัญหาการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม โดยเปรียบเทียบวิธีการเลือกคุณลักษณะ 3 แบบ คือ สหสัมพันธ์โคไซน์ (cosine correlation: CC) ระยะห่างยูคลิเดียน (Euclidean distance: ED) และ SNR ตัวจำแนกประเภทที่ใช้ ประกอบด้วย วิธีเบย์อย่างง่าย (Naïve Bayes) และวิธีเพื่อนบ้านใกล้ที่สุดเค (k-nearest neighbor: KNN) โดยใช้โปรแกรม WEKA จากผลการทดลองพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดสำหรับการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม แต่เนื่องจากวิธี SNR จะใช้สำหรับปัญหาการจำแนกประเภท 2 กลุ่มเท่านั้น ดังนั้นงานวิจัยนี้จึงได้พัฒนาวิธีการเลือกคุณลักษณะแบบ SNR สำหรับการจำแนกประเภทข้อมูลหลายกลุ่ม โดยใช้วิธีการคำนวณค่า SNR จาก กลุ่ม 1 กับกลุ่มที่เหลือ และวนสลับไปจนครบทั้ง n กลุ่ม จากนั้นนำค่า SNR ที่ได้ทั้งหมดรวมกัน เป็นค่าคะแนนของคุณลักษณะนั้น ๆ ซึ่งทดลองพบว่า วิธีการที่นำเสนอสามารถให้ค่าประสิทธิภาพในการจำแนกประเภทข้อมูลหลายกลุ่มที่ดีที่สุด

คำสำคัญ: การเลือกคุณลักษณะ, อัตราส่วนสัญญาณต่อสัญญาณรบกวน, การจำแนกประเภทข้อมูลหลายกลุ่ม

Abstract

This research aims to develop the Signal-to-Noise Ratio (SNR) feature selection for multi-class classification. Six benchmark datasets are used to test the performance of the proposed method. The datasets are divided into 2 groups: 1) 3 datasets for 2-class classification problem, and 2) 3 datasets for multi-class classification problem. First of all, the experiment starts with 2-class classification problem. Three feature selection techniques are used to compare the performance: Cosine Correlation (CC), Euclidean Distance (ED) and SNR. Two algorithms: Naïve Bayes and K-Nearest Neighbor (KNN) are used as the classifier with WEKA software. The experimental results show that the SNR offers the best result for 2-class classification. Since the SNR can be only used for the 2-class classification problem, this research

*Corresponding author; e-mail: supojn@yahoo.com
DOI 10.14456/jtir.2016.15

tries to develop an SNR based feature selection for multi-class classification. The SNR value is the summation of the SNR which is calculated from comparison results of the group 1 and the rest through to n groups. The experimental results show that the proposed method yields the best performance.

Keywords: feature selection, Signal-to-Noise Ratio (SNR), multi-class classification

Article history: Received 19 April 2016, Accepted 31 August 2016

1. บทนำ

การทำเหมืองข้อมูล (data mining) เป็นกระบวนการค้นหารูปแบบ แนวทาง หรือความสัมพันธ์ที่ซ่อนอยู่ภายในชุดข้อมูล อาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ กระบวนการทำเหมืองข้อมูลประกอบไปด้วย 6 ขั้นตอน คือ (1) การทำความเข้าใจปัญหา (2) การรวบรวมข้อมูลที่เกี่ยวข้อง (3) การเตรียมข้อมูล (4) การสร้างแบบจำลอง การวิเคราะห์ข้อมูลด้วยเทคนิคเหมืองข้อมูล (5) การประเมินประสิทธิภาพของแบบจำลอง และ (6) การนำไปใช้งาน [1]

การเลือกคุณลักษณะ (feature selection) [2] เป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล ในกระบวนการต้องปรับข้อมูลจากข้อมูลเดิมให้มีขนาดเล็กลง โดยให้สูญเสียคุณลักษณะสำคัญของข้อมูลน้อยที่สุดและสามารถเป็นตัวแทนของข้อมูลส่วนใหญ่ได้ การลดขนาดของข้อมูลจะช่วยให้การจำแนกประเภทข้อมูล สามารถทำงานได้ถูกต้องและรวดเร็วมากยิ่งขึ้น จากการศึกษาพบว่า เทคนิคที่นิยมใช้การเลือกคุณลักษณะข้อมูลมีหลายวิธี ส่วนใหญ่จะใช้กับปัญหาการจำแนกประเภทข้อมูล 2 กลุ่ม ซึ่งมีหลายวิธี เช่น สหสัมพันธ์โคไซน์ (cosine correlation: CC) ระยะห่างยูคลิเดียน (Euclidean distance: ED) และอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio: SNR) เป็นต้น

จากการศึกษาพบว่า การเลือกคุณลักษณะแบบอัตราส่วนสัญญาณต่อสัญญาณรบกวนให้ประสิทธิภาพที่ดีสำหรับการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม [3] แต่ยังมีข้อจำกัด คือ ไม่สามารถเลือกคุณลักษณะการจำแนกประเภทข้อมูลหลายกลุ่มได้ ดังนั้นในการวิจัยนี้จึงได้ศึกษาเพื่อเพิ่มประสิทธิภาพและพัฒนาวิธีการเลือกคุณลักษณะแบบอัตราส่วนสัญญาณต่อสัญญาณรบกวนให้สามารถรองรับการจำแนกประเภทแบบหลายกลุ่มได้

2. ทฤษฎีที่เกี่ยวข้อง

2.1 วิธีการเลือกคุณลักษณะ (feature selection)

2.1.1 อัตราส่วนสัญญาณต่อสัญญาณรบกวน (signal-to-noise ratio: SNR)

SNR เป็นวิธีการทางสถิติเพื่อวัดประสิทธิภาพของคุณลักษณะในการจำแนกประเภทข้อมูลจากข้อมูลกลุ่มหนึ่งออกจากข้อมูลกลุ่มอื่น ๆ [3] การคำนวณหาค่า SNR แสดงดังสมการ (1) ค่ายิ่งมากหมายถึง คุณลักษณะนั้นมีความสำคัญต่อการจำแนกประเภทข้อมูลมาก

$$SNR_F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \quad (1)$$

โดยที่ μ_1 และ μ_2 คือ ค่าเฉลี่ยของข้อมูลกลุ่มที่ 1 และกลุ่มที่ 2 σ_1 และ σ_2 คือ ค่าส่วนเบี่ยงเบนมาตรฐานของข้อมูลในแต่ละกลุ่ม

2.1.2 สหสัมพันธ์โคไซน์ (cosine correlation: CC)

วิธีนี้เป็นการวัดความคล้ายคลึงระหว่าง 2 เวกเตอร์ โดยการวัดมุมโคไซน์ของเวกเตอร์ทั้งสอง ซึ่งคำนวณได้จากสมการ (2)

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (2)$$

โดยที่ $A = \{a_1, a_2, a_3, \ldots, a_n\}$ และ $B = \{b_1, b_2, b_3, \ldots, b_n\}$ คือ 2 เวกเตอร์ที่ต้องการนำมาเปรียบเทียบ ซึ่งในปัญหาการเลือกคุณลักษณะนั้น เวกเตอร์ A คือ คุณลักษณะที่ i ใด ๆ ที่ต้องการคำนวณหาค่าความสำคัญ ส่วนเวกเตอร์ B คือ คุณลักษณะของคำตอบเป้าหมาย (class) ของชุดข้อมูล โดยค่าที่คำนวณได้จะอยู่ระหว่าง 0 – 1 ค่าของคุณลักษณะใดที่คำนวณได้มีค่าเข้าใกล้ 1 จะหมายถึงคุณลักษณะนั้นมีความสอดคล้องกับคำตอบเป้าหมายมากและถือว่าเป็นคุณลักษณะที่มีความสำคัญมาก

2.1.3 ระยะห่างยูคลิเดียน (Euclidean distance: ED)

วิธีนี้เป็นการวัดระยะห่างปกติระหว่างจุด 2 จุดในแนวเส้นตรง ซึ่งอาจวัดได้ด้วยอุปกรณ์วัดระยะที่ได้มาจากทฤษฎีพีทาโกรัส ปัญหาการเลือกคุณลักษณะ จะต้องมีการกำหนดตัวแปรอุดมคติเช่นเดียวกับสหสัมพันธ์โคไซน์ ระยะห่างยูคลิเดียน ระหว่างจุด x_i และจุด x_j แสดงด้วย $dist(x_i, x_j)$ คำนวณได้ดังสมการ (3)

$$dist(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{i,k} - x_{j,k})^2} \quad (3)$$

โดยที่ $x_{i,k}$ คือ คุณสมบัติตัวที่ *i* ของข้อมูลตัวที่ k ซึ่งในปัญหาการเลือกคุณลักษณะนั้น x_i คือ คุณลักษณะที่ *i* ใด ๆ ที่ต้องการคำนวณหาค่าความสำคัญ ส่วน x_j คือ คุณลักษณะของคำตอบเป้าหมายของชุดข้อมูล ถ้าค่าที่คำนวณได้มีค่าเข้าใกล้ 0 จะหมายถึงคุณลักษณะนั้นมีความใกล้เคียงกับคำตอบเป้าหมายมากและถือว่าคุณลักษณะนั้นมีความสำคัญมาก

2.2 วิธีการจำแนกประเภทข้อมูล (data classification)

2.2.1 วิธีเพื่อนบ้านใกล้ที่สุด k (k-nearest neighbor algorithm: KNN)

วิธีเพื่อนบ้านใกล้ที่สุดเค [1] เป็นวิธีการจำแนกประเภทข้อมูลที่ใช้วิธีการหาระยะห่างระหว่างคุณลักษณะของแต่ละข้อมูล ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข โดยวิธีการเพื่อนบ้านใกล้ที่สุดเค มีขั้นตอนโดยสรุปดังนี้

1) กำหนดจำนวนเพื่อนบ้าน *k* (นิยมกำหนดให้เป็นเลขคี่)

2) คำนวณระยะห่าง (distance) ของข้อมูลที่ต้องการพิจารณากับชุดข้อมูลสอน โดยสามารถคำนวณได้จากสมการระยะทางยูคลิเดียน (Euclidean distance) ดังสมการ (3)

3) จัดลำดับของระยะห่างจากน้อยไปมากและเลือกชุดข้อมูลที่น้อยที่สุด ตามจำนวน *k* ที่กำหนดไว้

4) กำหนดให้คำตอบของข้อมูลที่ต้องการทำนายคือ กลุ่มที่มีจำนวนมากที่สุดในกลุ่มของชุดข้อมูล *k* ตัวแรก

2.2.2 เบย์อย่างง่าย (Naïve Bayes)

ตัวจำแนกประเภทเบย์อย่างง่าย [4] เป็นตัวจำแนกประเภทในกลุ่มของการเรียนรู้แบบเกียจคร้าน (lazy learning) วิธีนี้เหมาะกับกรณีของข้อมูลที่มีจำนวนมากและคุณลักษณะ (attribute) ไม่ขึ้นต่อกัน สมมติให้ $a_1, a_2, ..., a_n$ เป็นคุณลักษณะของชุดข้อมูล จะได้ว่า ค่า (ประเภท) ที่น่าจะเป็นที่สุดของตัวอย่าง *x* คำนวณได้จากสมการ (4)

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, ..., a_n \mid v_j) P(v_j) \quad (4)$$

เนื่องจากการคำนวณค่าของ $P(a_1, a_2, ..., a_n|v_j)$ จำเป็นต้องมีชุดข้อมูลสอนในปริมาณมากและครอบคลุมทุกกรณี ซึ่งโดยทั่วไปมักเป็นไปได้ยาก ดังนั้น สมมติฐานของตัวจำแนกประเภทเบย์อย่างง่าย คือ กำหนดให้คุณลักษณะของข้อมูลแต่และตัวไม่ขึ้นต่อกัน (เป็นอิสระต่อกัน) กับคุณลักษณะอื่น ๆ ซึ่งทำให้สามารถเขียนแทน $P(a_1, a_2, ..., a_n|v_j)$ ด้วยผลคูณของความน่าจะเป็น ดังสมการ (5)

$$P(a_1, a_2, ..., a_n \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j) \quad (5)$$

ดังนั้น ตัวจำแนกประเภทเบย์อย่างง่าย จึงคำนวณตามสมการ (6)

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i=1}^{n} P(a_i \mid v_j) \quad (6)$$

3. การทดลองจำแนกประเภทข้อมูล 2 กลุ่ม

เพื่อยืนยันประสิทธิภาพของวิธีการเลือกคุณลักษณะแบบ SNR สำหรับการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม เราได้ทดลองเปรียบเทียบประสิทธิภาพกับวิธีการเลือกคุณลักษณะแบบสหสัมพันธ์โคไซน์ และแบบระยะห่างยูคลิเดียน กับชุดข้อมูลเกณฑ์มาตรฐานสำหรับปัญหาการจำแนกประเภทข้อมูลแบบ 2 กลุ่ม โดยใช้ตัวจำแนกประเภทแบบเบย์อย่างง่าย และ เพื่อนบ้านใกล้ที่สุดเค ด้วยโปรแกรม WEKA ซึ่งมีรายละเอียดดังนี้

3.1 ชุดข้อมูลที่ใช้ในการทดลองสำหรับปัญหาการจำแนกประเภทข้อมูล 2 กลุ่ม

ข้อมูลที่มีคุณลักษณะจำนวนมากส่วนมากมักจะเป็นข้อมูลทางด้านการแพทย์ ในการทดลองนี้ได้ใช้ข้อมูลจำนวน 3 ชุดข้อมูล ดังนี้

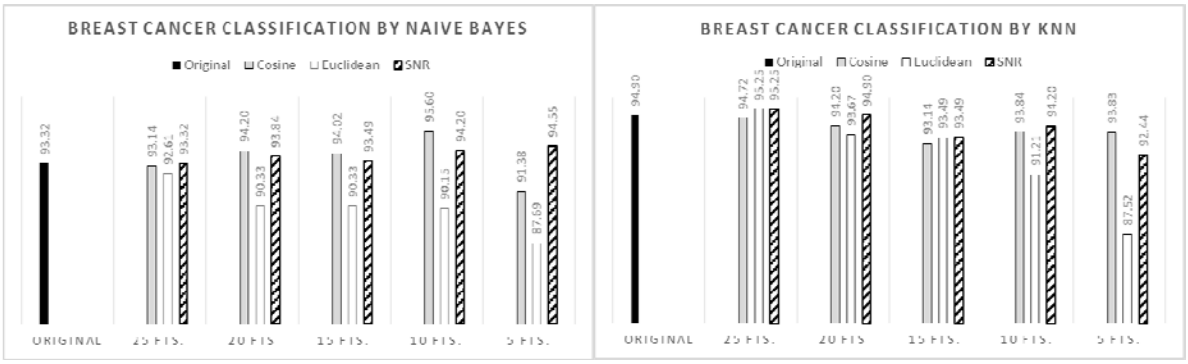
3.1.1 ข้อมูลมะเร็งเต้านม (WDBC) ประกอบด้วยข้อมูลจำนวน 569 ตัวอย่าง มี 30 คุณลักษณะ มี 2 ประเภท เป็นข้อมูลตัวเลขจำนวนจริง จาก UC Irvine Machine Learning Repository database [5]

3.1.2 ข้อมูลมะเร็งลำไส้ (COLON) ประกอบด้วยข้อมูลจำนวน 62 ตัวอย่าง มี 2000 คุณลักษณะ มี 2 ประเภท เป็นข้อมูลตัวเลขจำนวนจริง [6]

3.1.3 ข้อมูลมะเร็งต่อมน้ำเหลือง (DLBCL) ประกอบด้วยข้อมูลจำนวน 47 ตัวอย่าง มี 4026 คุณลักษณะ มี 2 ประเภท เป็นข้อมูลตัวเลขจำนวนจริง [7]

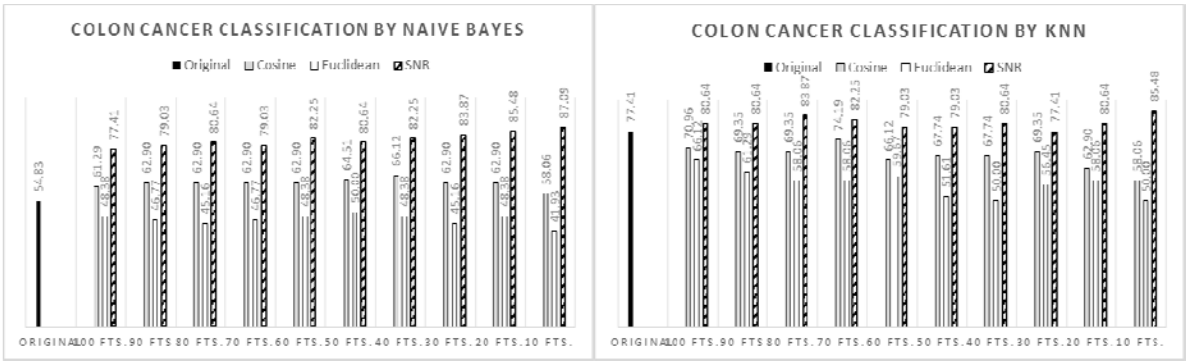
3.2 ผลการทดลองการจำแนกประเภทข้อมูล 2 กลุ่ม

3.2.1 ผลการทดลองกับชุดข้อมูลมะเร็งเต้านม (WDBC) จากการทดลองจำแนกประเภทข้อมูลมะเร็งเต้านมด้วยวิธีเบย์อย่างง่ายพบว่า การเลือกคุณลักษณะแบบสหสัมพันธ์โคไซน์ ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 10 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 95.60 (แสดงดังรูปภาพที่ 1 ก) ส่วนวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า ทั้งวิธีการเลือกคุณลักษณะแบบสหสัมพันธ์โคไซน์ และวิธี SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 25 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 95.25 (แสดงดังรูปภาพที่ 1 ข)



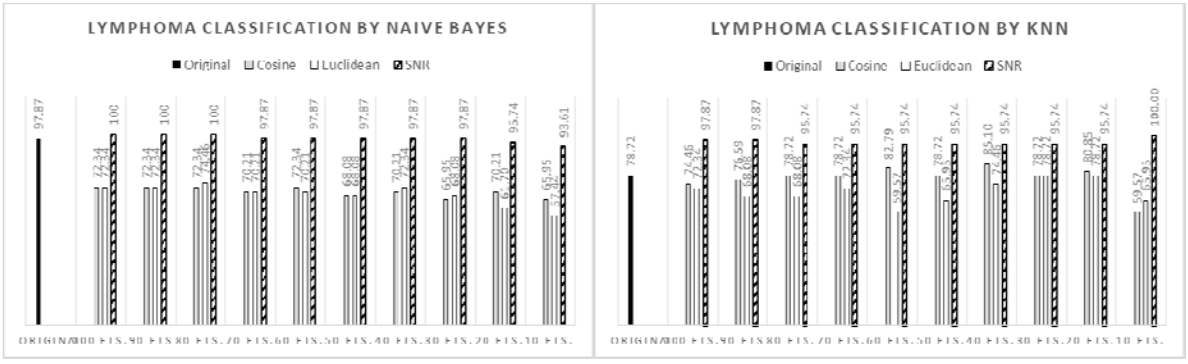
(ก) (ข)

รูปภาพที่ 1 (ก) ผลการจำแนกประเภทข้อมูลมะเร็งเต้านมด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลมะเร็งเต้านมด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค



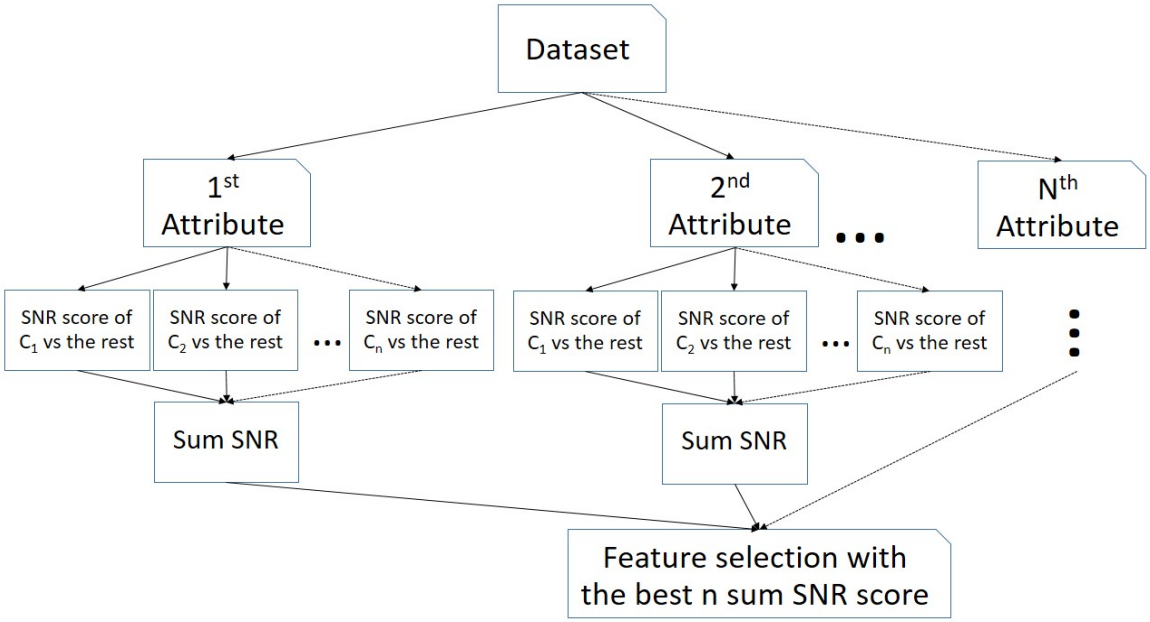
(ก) (ข)

รูปภาพที่ 2 (ก) ผลการจำแนกประเภทข้อมูลมะเร็งลำไส้ด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลมะเร็งลำไส้ด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค



(ก) (ข)

รูปภาพที่ 3 (ก) ผลการจำแนกประเภทข้อมูลมะเร็งต่อมน้ำเหลืองด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลมะเร็งต่อมน้ำเหลืองด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค



รูปภาพที่ 4 วิธีการคำนวณหาค่า SNR สำหรับการจำแนกประเภทข้อมูลหลายกลุ่มที่นำเสนอ

3.2.2 ผลการทดลองกับชุดข้อมูลมะเร็งลำไส้ (COLON) จากการทดลองจำแนกประเภทข้อมูลมะเร็งลำไส้ด้วยวิธีเบย์อย่างง่ายพบว่า การเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 10 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 87.09 (แสดงดังรูปภาพที่ 2 ก) ส่วนวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 10 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 85.48 (แสดงดังรูปภาพที่ 2 ข)

3.2.3 ผลการทดลองกับชุดข้อมูลมะเร็งต่อมน้ำเหลือง (DLBCL) จากการทดลองจำแนกประเภทข้อมูลมะเร็งต่อมน้ำเหลืองด้วยวิธีเบย์อย่างง่ายพบว่า การเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 80 – 100 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 100.00 (แสดงดังรูปภาพที่ 3 ก) ส่วนวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 10 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 100.00 (แสดงดังรูปภาพที่ 3 ข)

จากผลการทดลองจำแนกประเภทข้อมูล 2 กลุ่ม แสดงให้เห็นว่าประสิทธิภาพของวิธีการเลือกคุณลักษณะแบบ SNR จะให้ประสิทธิภาพในการจำแนกประเภทที่ดี พบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ดีที่สุดจำนวน 5 จาก 6 การทดลอง

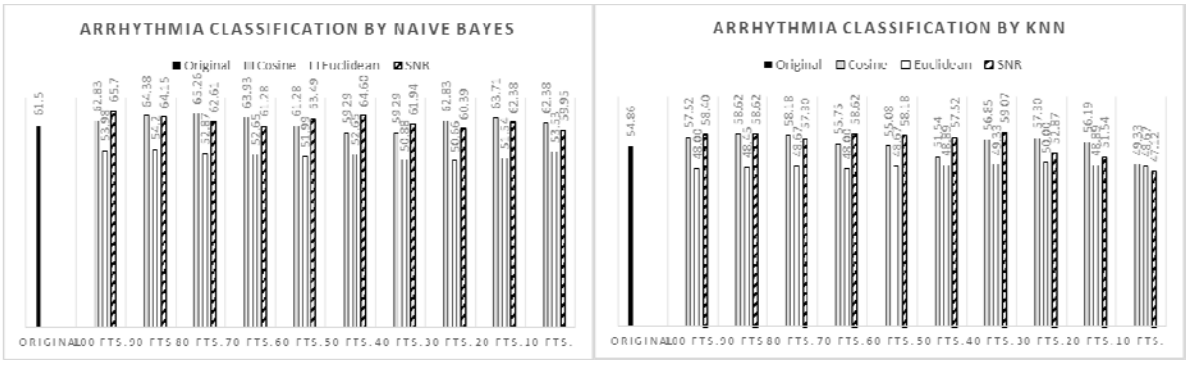
4. การจำแนกประเภทข้อมูลหลายกลุ่ม

การคำนวณค่า SNR เป็นวิธีการคำนวณสำหรับการจำแนกประเภทข้อมูล 2 กลุ่มเท่านั้น ดังนั้นผู้วิจัยจึงได้ออกแบบการคำนวณค่า SNR สำหรับปัญหาการจำแนกประเภทข้อมูลหลายกลุ่ม โดยใช้วิธีการคำนวณค่า SNR แบ่งเป็นการคำนวณค่า SNR ของกลุ่ม 1 เทียบกับกลุ่มที่เหลือ และวนสลับไปจนถึงกลุ่ม n กับกลุ่มที่เหลือ จากนั้นนำค่า SNR ที่ได้ทั้งหมดรวมกัน เป็นค่าคะแนนของคุณลักษณะนั้น ๆ ซึ่งวิธีการโดยรวมแสดงดังรูปภาพที่ 4 การทดสอบประสิทธิภาพจะใช้วิธีการเช่นเดียวกันกับการทดสอบประสิทธิภาพสำหรับปัญหาการจำแนกประเภท 2 กลุ่มที่กล่าวมาข้างต้น มีรายละเอียดดังนี้

4.1 ชุดข้อมูลที่ใช้ในการทดลองสำหรับปัญหาการจำแนกประเภทข้อมูลหลายกลุ่ม

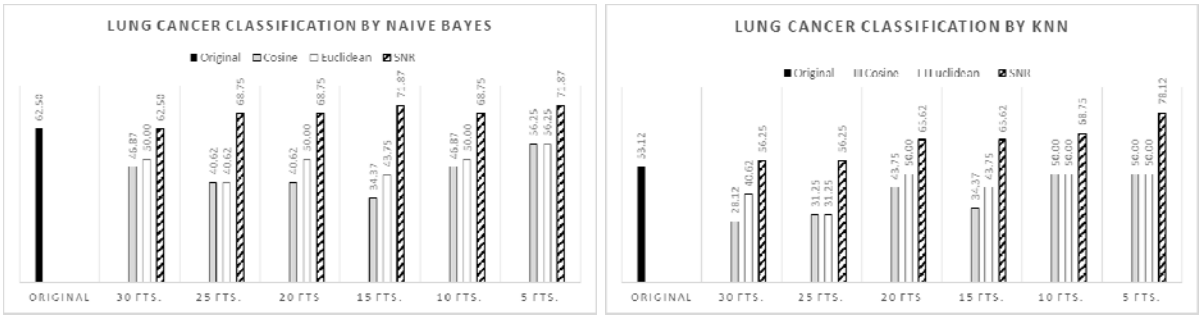
4.1.1 ข้อมูลจังหวะการเต้นของหัวใจ (arrhythmia) ประกอบด้วย ข้อมูลจำนวน 32 ตัวอย่าง มี 56 คุณลักษณะ มี 3 ประเภท เป็นข้อมูลตัวเลขจำนวนจริง

4.1.2 ข้อมูลมะเร็งปอด (lung cancer) ประกอบด้วยข้อมูลจำนวน 452 ตัวอย่าง มี 279 คุณลักษณะ มี 13 ประเภท เป็นข้อมูลตัวเลขจำนวนเต็ม



(ก) (ข)

รูปภาพที่ 5 (ก) ผลการจำแนกประเภทข้อมูลจังหวะการเต้นของหัวใจด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลจังหวะการเต้นของหัวใจด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค



(ก) (ข)

รูปภาพที่ 6 (ก) ผลการจำแนกประเภทข้อมูลมะเร็งปอดด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลมะเร็งปอดด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค

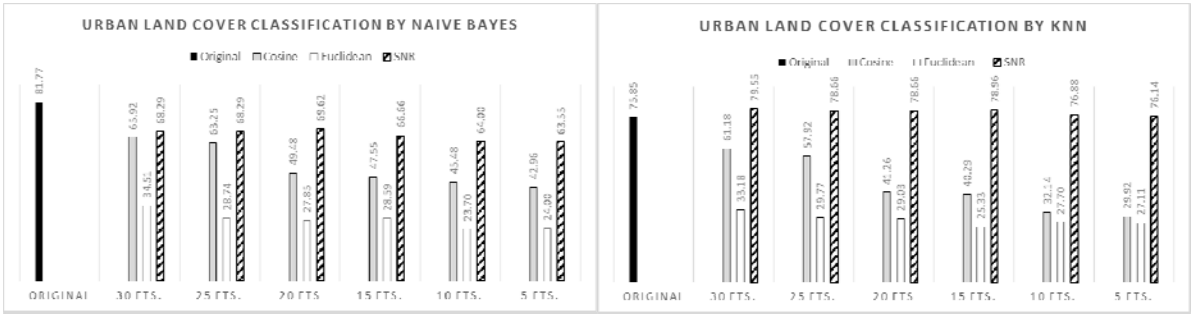
4.1.3 ข้อมูลการจำแนกประเภทจากภาพถ่ายดาวเทียม (urban land cover) ประกอบด้วย ข้อมูลจำนวน 675 ตัวอย่าง มี 279 คุณลักษณะ มี 9 ประเภท เป็นข้อมูลตัวเลขจำนวนจริง

4.2 ผลการทดลองการจำแนกประเภทข้อมูลหลายกลุ่ม

4.2.1 ผลการทดลองกับชุดข้อมูลจังหวะการเต้นของหัวใจ (arrhythmia) จากการทดลองจำแนกประเภทข้อมูลจังหวะการเต้นของหัวใจด้วยวิธีเบย์อย่างง่ายพบว่า การเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 100 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 65.70 (แสดงดังรูปภาพที่ 5 ก) ส่วนวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 70 และ 90 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 58.62 (แสดงดังรูปภาพที่ 5 ข)

4.2.2 ผลการทดลองกับชุดข้อมูลมะเร็งปอดจากการทดลองจำแนกประเภทข้อมูลมะเร็งปอดด้วยวิธีเบย์อย่างง่ายพบว่า การเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 5 และ 15 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 71.87 (แสดงดังรูปภาพที่ 6 ก) ส่วนวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 5 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 78.12 (แสดงดังรูปภาพที่ 6 ข)

4.2.3 ผลการทดลองกับชุดข้อมูลการจำแนกประเภทจากภาพถ่ายดาวเทียมจากการทดลองจำแนกประเภทข้อมูลการจำแนกประเภทจากภาพถ่ายดาวเทียมด้วยวิธีเบย์อย่างง่าย พบว่า การใช้ข้อมูลดั้งเดิม คือใช้ทั้ง 279 คุณลักษณะ ให้ประสิทธิภาพดีที่สุด ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 81.77 ส่วนการใช้วิธีการเลือกคุณลักษณะให้ประสิทธิภาพที่ต่ำลง แต่ก็พบว่าวิธีการ SNR ให้ประสิทธิภาพที่ดีกว่าวิธีการเลือกคุณลักษณะแบบอื่น ๆ ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 20 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล



(ก) (ข)

รูปภาพที่ 7 (ก) ผลการจำแนกประเภทข้อมูลภาพถ่ายดาวเทียมด้วยวิธีเบย์อย่างง่าย
(ข) ผลการจำแนกประเภทข้อมูลภาพถ่ายดาวเทียมด้วยวิธีเพื่อนบ้านใกล้ที่สุดเค

คิดเป็นร้อยละ 69.62 (แสดงดังรูปภาพที่ 7 ก) สำหรับวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 30 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 79.55 (แสดงดังรูปภาพที่ 7 ข)

4.2.3 ผลการทดลองกับชุดข้อมูลการจำแนกประเภทจากภาพถ่ายดาวเทียมจากการทดลองจำแนกประเภทข้อมูลการจำแนกประเภทจากภาพถ่ายดาวเทียมด้วยวิธีเบย์อย่างง่าย พบว่า การใช้ข้อมูลดั้งเดิม คือใช้ทั้ง 279 คุณลักษณะ ให้ประสิทธิภาพดีที่สุด ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 81.77 ส่วนการใช้วิธีการเลือกคุณลักษณะให้ประสิทธิภาพที่ต่ำลง แต่ก็พบว่าวิธีการ SNR ให้ประสิทธิภาพที่ดีกว่าวิธีการเลือกคุณลักษณะแบบอื่น ๆ ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 20 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 69.62 (แสดงดังรูปภาพที่ 7 ก) สำหรับวิธีเพื่อนบ้านใกล้ที่สุดเคพบว่า วิธีการเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพดีที่สุดเมื่อเลือกใช้ 30 คุณลักษณะ ให้ความถูกต้องในการจำแนกประเภทข้อมูล คิดเป็นร้อยละ 79.55 (แสดงดังรูปภาพที่ 7 ข)

5. สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ศึกษาและพัฒนาวิธีการเลือกคุณลักษณะแบบ SNR ให้สามารถใช้ได้กับปัญหาการจำแนกประเภทข้อมูลหลายกลุ่มได้ โดยเริ่มจากการศึกษาวิธีการเลือกคุณลักษณะสำหรับการจำแนกประเภทข้อมูลสองกลุ่มซึ่งใช้วิธีการเลือกคุณลักษณะแบบ สหสัมพันธ์โคไซน์ ระยะห่างยูคลิเดียน และ SNR โดยได้ทดลองกับ 3 ชุดข้อมูล และใช้ตัวจำแนกประเภทด้วยโปรแกรม WEKA จำนวน 2 วิธี คือ เบย์อย่างง่าย และ เพื่อนบ้านใกล้ที่สุดเค รวมเป็น 6 การทดลอง ผลการทดลองแสดงให้เห็นว่า การเลือกคุณลักษณะแบบ SNR ให้ประสิทธิภาพสูงที่สุดถึง 5 ใน 6 การทดลอง แสดงให้เห็นถึง ประสิทธิภาพของวิธีการเลือกคุณลักษณะแบบ SNR สำหรับปัญหาการจำแนกประเภทข้อมูล 2 กลุ่ม

จากนั้นได้ศึกษาหาวิธีการพัฒนาประสิทธิภาพการเลือกคุณลักษณะแบบ SNR เพื่อให้สามารถใช้ได้กับปัญหาการจำแนกประเภทข้อมูลหลายกลุ่ม โดยใช้วิธีการคำนวณค่า SNR ด้วยการแบ่งคำนวณค่า SNR ของกลุ่ม 1 เทียบกับกลุ่มที่เหลือ และวนสลับไปจนถึงกลุ่ม n กับกลุ่มที่เหลือ และนำค่า SNR ที่ได้ทั้งหมดรวมกัน เป็นค่าคะแนนของคุณลักษณะนั้น ๆ โดยได้ทดลองกับข้อมูลจำนวน 3 ชุด ผลการทดลองแสดงให้เห็นว่า การเลือกคุณลักษณะแบบ SNR ที่นำเสนอให้ประสิทธิภาพสูงที่สุดถึง 5 จาก 6 การทดลอง ซึ่งแสดงให้เห็นว่า วิธีการที่นำเสนอนี้ยังคงให้ค่าประสิทธิภาพที่ดีที่สุดสำหรับปัญหาการจำแนกประเภทข้อมูลแบบหลายกลุ่มด้วยเช่นกัน

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับการสนับสนุนทุนวิจัยจากโครงการวิจัยบูรณาการนักศึกษาและอาจารย์เพื่อการพัฒนาท้องถิ่นและความเป็นเลิศทางวิชาการ ปีงบประมาณ 2559 จากสถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏนครปฐม

เอกสารอ้างอิง

[1] Chapman P, Clinton J, Kerber R, Khabaza T, Reinart T, Shearer C, Wirth R. **CRISP-DM Step-by-step data mining guide [internet].** 2000. Available from: http://www .crisp-dm.org

[2] Guyon I, Elisseeff A. An introduction to variable and feature selection. **The Journal of Machine Learning Research**. 2003; **3**: 1157-82.

[3] Slonim DK, Tamayo P, Mesirov JP, Golub TR, Lander ES. Class prediction and discovery using gene expression data. **Proceeding of the 4th Annual International Conference on Computational Molecular Biology**. 2000: 263-72.

[4] Rish I, An empirical study of the naive Bayes classifier. **Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence**. 2001: 41-6.

[5] UC Irvine Machine Learning Repository database. Available from http://archive.ics.uci.edu/ml/

[6] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. **Proceedings of National Academy of Sciences of the United States of American**. 1999; **96**: 6745-50.

[7] Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J-JR, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. **Nature**. 2000; **403**: 503-11.