การเปรียบเทียบประสิทธิภาพการเลือกคุณลักษณะที่เหมาะสม สำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์

An Efficiency Comparison of Feature Selection Methods for Microarray Data Classification

ไพศาล จันทร์เจริญ[1*] , สุพจน์ เฮงพระพรหม[2] และ ไก้รุ่ง เฮงพระพรหม[3]

[1]สาขาวิชาเทคโนโลยีสารสนเทศ [2]สาขาวิชาวิศวกรรมซอฟต์แวร์ [3]สาขาวิชาเทคโนโลยีคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏนครปฐม
Jub_5@windowslive.com

บทคัดย่อ

งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพ วิธีการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ โดยในการทดลองจะทำการทดสอบ 2 ขั้นตอน คือ 1) การคัดเลือกคุณลักษณะที่เหมาะสม ด้วยวิธี Cosine และ SNR โดยทำการคัดเลือกคุณลักษณะที่เหมาะสมแบ่งเป็นกลุ่มย่อย ๆ ตั้งแต่ 100 ถึง 1000 คุณลักษณะ และ 2) การทดสอบประสิทธิภาพการจำแนกข้อมูล ประกอบด้วยเทคนิคการจำแนกข้อมูล 3 เทคนิค ได้แก่ ต้นไม้ตัดสินใจ วิธีเพื่อนบ้านใกล้ที่สุดเค และเบย์อย่างง่าย โดยทำการทดสอบกับข้อมูลโรคมะเร็งต่อมน้ำเหลือง จำนวน 47 ตัวอย่าง 4,026 คุณลักษณะ

ผลของการทดลองพบว่า วิธีการคัดเลือกคุณลักษณะแบบ SNR สามารถให้ประสิทธิภาพการจำแนกประเภทข้อมูลเฉลี่ยดีที่สุดที่ 94.33% โดยใช้จำนวนคุณลักษณะเพียง 100 คุณลักษณะ ขณะที่การจำแนกข้อมูลด้วยคุณลักษณะทั้งหมดให้ประสิทธิภาพที่ 86.52% จึงแสดงให้เห็นว่าการคัดเลือกคุณลักษณะที่เหมาะสมก่อนการจำแนกประเภทข้อมูลนั้นเป็นแนวทางที่เหมาะสมอย่างยิ่งที่จะทำให้เพิ่มประสิทธิภาพของการจำแนกข้อมูลได้เป็นอย่างดี เนื่องจากข้อมูลไมโครอาร์เรย์นั้น อาจมีการกระจายตัวของข้อมูลอย่างมาก ซึ่งข้อมูลบางส่วนอาจไม่สอดคล้องกัน จึงจำเป็นอย่างยิ่งที่จะต้องทำการกรองข้อมูลก่อนเบื้องต้น

คำสำคัญ : ข้อมูลไมโครอาร์เรย์, การเลือกคุณลักษณะ, SNR, ต้นไม้ตัดสินใจ, วิธีเพื่อนบ้านใกล้ที่สุดเค, เบย์อย่างง่าย

Abstract

This research aimed to compare an efficiency of feature selection methods for microarray data classification. In this experiment comprise of two phases: 1) the appropriate feature selection with Cosine and SNR method by selected them into small sets of features, ranking 100 to 1000 features. 2) The efficiency of classification, the three classification techniques including Decision Tree, K-Nearest Neighbor and Naïve Bayesian are used in this step. And, the Lymphoma with 47 instants and 4,026 features was used to test in this experiment.

The result shown that the SNR feature selection method can provide the best average classification at 94.33 % using 100 features compare with classification with all features that provide the performance at 86.52%. The data in microarray data often comprise of irrelevant feature or widely range data, thus, the appropriate feature selection before classification approach is suited to optimize the data classification.

Keywords: microarray data, feature selection, SNR, decision tree, K-Nearest Neighbor, Naïve Bayesian

1. บทนำ

ในยุคปัจจุบันเทคโนโลยีสารสนเทศมีการพัฒนาความก้าวหน้าอย่างรวดเร็วส่งผลดี เช่น สามารถติดต่อสื่อสารกันได้ทันต่อเวลาในสถานการณ์ต่างๆ แต่ในทางตรงกันข้าม ก็ก่อให้เกิดปัญหาที่อาจจะตามมา เช่น พื้นที่ในการจัดเก็บข้อมูลทางเทคโนโลยีสารสนเทศมีจำนวนจำกัด เนื่องจากข้อมูลต่างๆ มีขนาดที่ใหญ่ขึ้น โดยข้อมูลอาจมีจำนวนที่มากขึ้น ทับซ้อนกัน หรือซ้ำกัน และอาจมีข้อมูลที่ขาดหายไป ซึ่งบางครั้งข้อมูลอาจมีคุณลักษณะที่ไม่ตรงประเด็นและมีคุณลักษณะของข้อมูลจำนวนมาก ซึ่งในการนำข้อมูลมาวิเคราะห์เพื่อการตัดสินใจในด้านต่างๆ จะต้องมีการจัดหมวดหมู่หรือการจำแนกประเภทของข้อมูลเพื่อจัดเก็บ สืบค้น และง่ายต่อการเข้าถึงข้อมูล ดังนั้นข้อมูลที่มีคุณลักษณะสูงจึงใช้ทรัพยากรในการประมวลผลและใช้เวลานานในการประมวลผล อาจส่งผลให้เกิดความ คลาดเคลื่อนในการวิเคราะห์ข้อมูลที่ต้องการได้

ชุดข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่ได้รับผลกระทบจากปัญหาความไม่สอดคล้องของคุณลักษณะของข้อมูลเช่นเดียวกัน โดยเป็นชุดข้อมูลที่นิยมใช้ในการศึกษารูปแบบของสิ่งมีชีวิตระดับโมเลกุลที่ถูกแสดงออกถึงระดับการแสดงออกของยีน หลายพันยีนในเวลาเดียวกัน ในการวิเคราะห์ข้อมูลไมโครอาร์เรย์นั้นสามารถศึกษาได้หลากหลายวิธี เช่น เทคนิคทางสถิติหรือเหมืองข้อมูล (Data Mining) และการเรียนรู้ของเครื่องจักร (Machine Learning) ได้แก่ การจำแนกประเภท (Classification) และการจัดกลุ่ม (Clustering) ซึ่งในการวิเคราะห์ทางสถิติดังกล่าวมีหลายเทคนิคที่หากลดจำนวนคุณลักษณะลงแล้วสามารถทำให้การจำแนกประเภทข้อมูลมีประสิทธิภาพที่สูงขึ้นได้

จากการศึกษางานวิจัยที่เกี่ยวข้องหลายเรื่องพบว่าการใช้เทคนิคการเลือกคุณลักษณะ มาช่วยในการเตรียมข้อมูลเพื่อลดคุณลักษณะของข้อมูลก่อนการจำแนกประเภท โดยแบบแต่ละแบบมีข้อดีข้อเสียที่แตกต่างกัน เช่น การเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสม ด้วยวิธีการจัดอันดับแบบ Information Gain, Gain Ratio (GR) และ Chi-Square โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเคอร์เนลฟังก์ชันแบบเรเดียลเบสิสฟังก์ชันในการจัดหมวดหมู่เว็บเพจ จำนวน 5,399 URL ผลการทดสอบพบว่าวิธีการจัดอันดับแบบ Chi-Square ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุด (น้ำทิพย์ มากนคร และมาลีรัตน์ โสดานิล, 2557) การใช้เทคนิคแบบ Information Gain ลดคุณลักษณะของข้อมูลแบบจำลอง 3 มิติ โดยเพิ่มความแม่นยำในการจำแนก จาก 96.41 % เป็น 96.62% (นิธินันท์ มาตา, 2554) การถ่วงน้ำหนักของเทคนิค SNRw โดยไม่จำเป็นที่จะต้องกำหนดจำนวนของคุณลักษณะ ส่งผลให้การจำแนกประเภทของข้อมูลไมโครอาร์เรย์ เพิ่มขึ้น (สุพจน์ เฮงพระพรหม, 2552)ซึ่งจากการศึกษาพบว่า การใช้เทคนิคการลดคุณลักษณะของข้อมูลสามารถเพิ่มประสิทธิภาพของการจำแนกได้ โดยในงานวิจัยนี้ได้นำเสนอการเปรียบเทียบเทคนิคการเลือกคุณลักษณะ ด้วยวิธีการคัดเลือกคุณลักษณะ Cosine และ SNR โดยวิธีการเลือกคุณลักษณะที่ดีที่สุดแบ่งออกเป็นกลุ่มย่อยๆ และทดสอบประสิทธิภาพการจำแนกประเภทด้วยวิธีการ ต้นไม้ตัดสินใจ วิธีเพื่อนบ้านใกล้ที่สุดเค และ เบย์อย่างง่าย โดยทดสอบกับชุดข้อมูลไมโครอาร์เรย์ โรคมะเร็งต่อมน้ำเหลือง เพื่อแก้ไขปัญหาการจำแนกประเภทข้อมูลไมโครอาร์เรย์ให้ดีขึ้น

2. วัตถุประสงค์การวิจัย

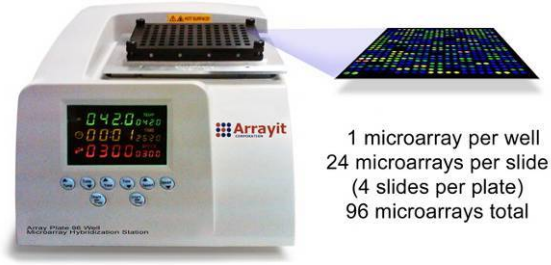
เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์

3. ทฤษฎีที่ และงานวิจัยเกี่ยวข้อง

3.1 ข้อมูลไมโครอาร์เรย์

เป็นเทคนิคในการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายๆ ยีนในช่วงเวลาเดียวกันเพื่อให้เข้าใจถึงกลไกในการทำงานหรือการพัฒนาของสิ่งมีชีวิตระดับโมเลกุล โดยปกติแล้ว ในทุกๆเซลล์ในร่างกายของมนุษย์เราจะมีกลุ่มที่เรียกว่าโครโมโซมและกลุ่มของยีน (Gene) โดยในการแสดงออกของยีนนั้น หมายถึง ความสามารถในการถอดรหัส (Transcription) ในส่วนของข้อมูลทางพันธุกรรมที่อยู่ใน DNA ซึ่งได้มาเป็นเอ็มอาร์เอ็นเอ (mRNA) แล้วนำ (mRNA) ที่ได้มาแปลรหัส (Transcription) เป็นสายโพลีไตเปปไตด์ (Polypeptide) สายหนึ่งอีกทีหนึ่ง เพื่อสร้างเป็นโปรตีนและทำหน้าที่ต่างๆ ในเซลล์ (Cell) โดยปกติแล้วข้อมูลไมโครอาร์เรย์สร้างมาจากการจับคู่กันของชิ้นส่วน DNA ซึ่งตัวอย่างที่ต้องการจะศึกษา

จะต้องทำการติดฉลากด้วยสารเรืองแสง โดยแทนด้วยสีในขนาดหรือปริมาณที่เท่ากัน และจัดเรียงเป็นแนวยาว (Array) บนแผ่นสไลด์ด้วยเครื่องจักรกล และทำการวัดปริมาณของสารเรืองแสงแต่ละสีด้วยเครื่องสแกน



ภาพที่ 1 การเตรียมข้อมูลไมโครอาร์เรย์ (เครื่องสแกน Arrayit 96 Well)
(ที่มา: www.arrayit.com ค้นหาวันที่ 19 พ.ค. 2558)

3.2 การเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณลักษณะ เป็นการลดขนาดของคุณลักษณะข้อมูลโดยการทำให้คุณลักษณะที่มีอยู่เดิมนั้นมีขนาดลดลง โดยที่สูญเสียคุณลักษณะที่สำคัญของข้อมูลน้อยที่สุด โดยเทคนิคการเลือกคุณลักษณะที่แตกต่างกันทำให้ได้คุณลักษณะที่แตกต่างกันด้วย ในการเลือกคุณลักษณะนั้นแบ่งออกเป็น 3 ประเภท ได้แก่

1) การเลือกคุณลักษณะแบบฝังตัว (Embedded)

การเลือกคุณลักษณะแบบฝังตัวนั้น จะเกิดขึ้นโดยขั้นตอนวิธีสำหรับการเรียนรู้เอง ซึ่งในขั้นตอนวิธีสำหรับการเรียนรู้เอง จะมีการเลือกคุณลักษณะที่เหมาะสมสำหรับการสร้างแบบจำลองในการแก้ไขปัญหาต่างๆ โดยไม่ต้องเพิ่มกระบวนการวิธีในการคัดเลือกคุณลักษณะที่เหมาะสมอื่นๆ เข้ามาช่วย

2) การเลือกคุณลักษณะแบบควบรวม (Wrapper)

การเลือกคุณลักษณะแบบควบรวม จะเกิดขึ้นในขั้นตอนสำหรับคัดเลือกเซตย่อยจากคุณลักษณะทั้งหมดของข้อมูลทั้งหมด โดยจะเน้นที่การค้นหาเซตย่อยของคุณลักษณะที่เหมาะสมกับขั้นตอนวิธีการเรียนรู้วิธีใดวิธีหนึ่งโดยเฉพาะ ดังนั้นวิธีการเลือกคุณลักษณะแบบควบรวม จะเป็นการเพิ่มประสิทธิภาพของขั้นตอนวิธีการเรียนรู้ได้ดีที่สุด แต่อาจข้อเสียที่จะต้องใช้เวลาในการเรียนรู้มาก และเซตย่อยของคุณลักษณะที่เลือกมา จะเหมาะกับวิธีการเรียนรู้แบบนั้นอย่างเดียว ซึ่งอาจไม่เหมาะกับวิธีการเรียนรู้แบบอื่นๆ ก็ได้

3) การเลือกคุณลักษณะแบบกรอง (Filter)

การเลือกคุณลักษณะแบบกรอง จะเป็นขั้นตอนการประเมินประสิทธิภาพของคุณลักษณะของข้อมูลในแต่ละตัวว่ามีความเหมาะสมกับข้อมูลมากน้อยเพียงใด โดยไม่ต้องคำนึงถึงขั้นตอนวิธีของการเรียนรู้แบบใดแบบหนึ่ง การเลือกคุณลักษณะแบบกรอง จะทำการจัดลำดับตามความสำคัญของคุณลักษณะแต่ละตัว และเลือกคุณลักษณะที่มีระดับความสำคัญสูงสุดตามจำนวนที่ผู้ใช้ระบุ หรืออาจระบุเป็นค่าขีดแบ่ง (Threshold) ของคุณลักษณะที่จะเลือกก็ได้ ข้อดีของการเลือกคุณลักษณะแบบนี้ คือ การประมวลผลที่รวดเร็ว และไม่ขึ้นกับขั้นตอนวิธีการเรียนรู้

ในกระบวนการดำเนินงานวิจัยในครั้งนี้ได้เลือกเทคนิคการเลือกคุณลักษณะมาทำการเปรียบเทียบดังนี้

3.2.1 การวัดค่าความเหมือน (Cosine Similarity) (จุฬาลักษณ์, 2553) เป็นการเลือกคุณลักษณะแบบกรอง เป็นการหาค่าความเหมือนหรือความคล้ายคลึงที่ได้จากค่าความต่างของมุมของวัตถุ 2 อย่างที่เกิดขึ้นบนพื้นที่เว็กเตอร์ โดยความคล้ายคลึงกันในแบบของ Cosine นั้น จะมีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น จึงเป็นวิธีการที่นิยมและมีประสิทธิภาพสูง มีสมการดังนี้

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

โดย A คือ คุณลักษณะแต่ละตัวของข้อมูลทั้งหมด
B คือ ประเภทของกลุ่มในการจำแนกของแต่ละตัวอย่าง

ดังนั้น ค่าความเหมือนหรือค่าความคล้ายคลึงของเอกสารอยู่ระหว่าง 0 ถึง 1 ถ้าค่าของความเหมือนใกล้ 1 แสดงถึงว่าข้อมูลนั้นมีความคล้ายคลึงกันมาก แต่ถ้าค่าความเหมือนใกล้เคียง 0 หมายถึง ข้อมูลนั้นไม่มีความคล้ายคลึงกันเลย

3.2.2 การวัดค่าจากสัญญาณรบกวน (SNR) เป็นการเลือกคุณลักษณะแบบ Filter คืออัตราสัญญาณจริงต่อค่าการรบกวน เป็นวิธีการทางสถิติในการใช้สำหรับการวัดประสิทธิภาพของคุณลักษณะในการจำแนกประเภทข้อมูลออกจากกลุ่มอื่นๆ ซึ่งในการคำนวณ คิดได้จากสมการต่อไปนี้

$$SNR_i = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \quad (2)$$

โดย μ_1 และ μ_2 คือ ค่าเฉลี่ยของข้อมูลกลุ่มที่ 1 และ กลุ่มที่ 2
σ_1 และ σ_2 คือ ค่าส่วนเบี่ยงเบนมาตรฐานแต่ละกลุ่มข้อมูล

3.3 การจำแนกประเภทข้อมูล (Classification)

ในการจำแนกประเภทของข้อมูลที่ต้องการศึกษานั้น เป็นการจำแนกประเภทแบบมีผู้สอน (Supervised Learning) โดยมีวัตถุประสงค์เพื่อกำหนดประเภทของข้อมูลให้กับข้อมูลใหม่ ที่ยังไม่มีการกำหนดประเภท ซึ่งในการใช้ชุดข้อมูลที่มีอยู่เป็นชุดข้อมูลที่รู้ประเภทแล้ว ซึ่งในกระบวนการดำเนินงานวิจัยนี้ ใช้เทคนิคดังต่อไปนี้

3.3.1 ต้นไม้ตัดสินใจ (Decision Tree) แบบ J48 (Quinlan, 1986) เป็นเทคนิคทางคณิตศาสตร์ที่ใช้ในการจำแนกประเภทของข้อมูล โดยพิจารณาจากคุณลักษณะของข้อมูลที่ต้องการศึกษา โดยค่าของคุณลักษณะที่สนใจนั้นจะเป็นค่าที่ไม่ต่อเนื่องกัน (Discrete Value) ในการเรียนรู้จะแทนในรูปแบบโครงสร้างของต้นไม้ โดยมีโหนดภายในเป็นคุณลักษณะต่างๆ ที่ใช้เรียนรู้ และจะมีกิ่ง เท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะนั้นๆ และมีส่วนใบเป็นประเภทที่เป็นไปได้ของข้อมูล

3.3.2 วิธีหาสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor) จะใช้วิธีในการแบ่งประเภทคลาส (Class) ซึ่งจะตัดสินใจว่าคลาสไหนที่จะแทนเงื่อนไขหรือกรณีใหม่ๆ ได้บ้าง โดยในขั้นตอนกระบวนการตรวจสอบจำนวนบางจำนวนของกรณีหรือเงื่อนไขที่เหมือนกัน หรือที่ใกล้เคียงกันมากที่สุด ในการนำเทคนิคของ K-NN ไปใช้ในนั้นเป็นการหาวิธีการวัดระยะห่างระหว่างแต่ละ คุณลักษณะในข้อมูลให้ได้ และจากนั้นคำนวณค่าออกมา ซึ่งวิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลข แต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้นก็สามารถทำได้ เพียงแต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น

3.3.3 การพยากรณ์ความน่าจะเป็นของสมาชิก (Naïve Bayesian) หรือ เบย์อย่างง่าย เป็นตัวจำแนกประเภทในกลุ่มการเรียนรู้แบบขี้เกียจ (Lazy learning) โดยเหมาะกับข้อมูลที่มีจำนวนมากๆ และมีมิติของข้อมูลที่ไม่สัมพันธ์หรือต่อเนื่องกัน Naïve Bayes เป็นเทคนิคการเรียนรู้หรือ อัลกอริทึมที่ใช้งานอย่างแพร่หลายในการจำแนกประเภท และให้ผลดี ซึ่งการใช้งานจะได้ความน่าจะเป็นของเงื่อนไขของแต่ละคุณลักษณะ เพื่อใช้งานในการทำนายคลาสของความเห็นใหม่ๆ ที่เข้ามาใช้ในการจำแนก

3.4 วิธีการวิเคราะห์ความแม่นตรงของโมเดล K-fold Cross-Validation

เป็นการตรวจสอบความไขว้กัน (Cross-Validation) เป็นวิธีการตรวจสอบค่าความผิดพลาดในการคาดการของโมเดล โดยพื้นฐานวิธีการ การตรวจสอบไขว้กัน K-Fold Cross-validation เป็นวิธีการที่แบ่งข้อมูลออกเป็นกลุ่มจำนวน K กลุ่ม (K-Fold) ในตอนแรกเลือกข้อมูลกลุ่มที่ 1 เป็นข้อมูลชุดทดสอบ และข้อมูลชุดที่เหลือจะเป็นข้อมูลชุดสอน นำข้อมูลไปจัดหมวดหมู่ จากนั้นจะสลับข้อมูล กลุ่มที่ 2 มาเป็นชุดทดสอบและข้อมูลกลุ่มอื่นๆ ที่เหลือเป็นชุดทดสอบ สลับอย่างนี้ไปเรื่อยๆ จนครบ K กลุ่ม ในขั้นตอนสุดท้ายจะหาค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของค่าความถูกต้องในแต่ละกลุ่ม วิธีการนี้ข้อมูลทุกตัวอย่างจะได้เป็นทั้งชุดทดสอบและชุดสอน

3.5 งานวิจัยที่เกี่ยวข้อง

นิธินันท์ มาตา พนิดา หล่อวงศ์ตระกูล และพยุง มีสัจ (2554) ได้นำเสนอวิธีการเปรียบเทียบการเลือกคุณลักษณะ (Feature Selection) ระหว่างเทคนิค Info Gain Attribute Evaluation, Consistency Subset Evaluation และ Wrapper Subset Evaluation พบว่า เทคนิคที่ช่วยเพิ่มประสิทธิภาพในการจำแนกประเภทของข้อมูลแบบจำลอง 3 มิติ คือ

Info Gain Attribute Evaluation โดยเพิ่มความแม่นยำในการจำแนก จาก 96.41 % เป็น 96.62% และลดเวลาในการสร้างโมเดลในการจำแนกจาก 86.06 วินาที เป็น 70.05 วินาที

น้ำทิพย์ มากนคร และมาลีรัตน์ โสดานิล (2557) ได้นำเสนอวิธีการเปรียบเทียบการเลือกคุณลักษณะที่เหมาะสมด้วยวิธี Information Gain, Gain Ratio และ Chi-Square โดยใช้ SVM ในการจัดหมวดหมู่เว็บเพจ จำนวน 5,399 URL ผลการทดสอบพบว่า วิธีการจัดอันดับแบบ Chi-Square ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุด เมื่อลดจำนวนแอตทริบิวต์ 50% ให้ค่าความถูกต้อง ร้อยละ 95.98 และเมื่อทำการปรับ ค่าพารามิเตอร์ ส่งผลให้ค่าความถูกต้องเพิ่มขึ้นโดยเพิ่มขึ้นจาก ร้อยละ 95.94 เป็น ร้อยละ 97.93 แสดงให้เห็นว่าการคัดเลือกคุณลักษณะและการปรับค่าพารามิเตอร์ของเทคนิคซัพพอร์ตเวกเตอร์แมชชีนสามารถเพิ่มประสิทธิภาพการจัดหมวดหมู่ได้

สุพจน์ เฮงพระพรหม (2552) ได้นำเสนอวิธีการถ่วงน้ำหนักของเทคนิค SNRw โดยไม่จำเป็นที่จะต้องกำหนดจำนวนของคุณลักษณะ ซึ่งส่งผลให้การจำแนกประเภทของข้อมูลเพิ่มขึ้น โดยทดสอบกับข้อมูลไมโครอาร์เรย์ จำนวน 6 ชุด ได้แก่ ข้อมูลโรคมะเร็งในเม็ดเลือด ข้อมูลมะเร็งลำไส้ใหญ่ ข้อมูลมะเร็งต่อมน้ำเหลือง ข้อมูลมะเร็งรังไข่ ข้อมูลมะเร็งต่อมลูกหมาก และข้อมูลมะเร็งปอด

นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ (2554) ได้นำเสนอวิธีการลดคุณลักษณะด้วยวิธี Information Gain และประมวลผลด้วยเครื่องจักรการเรียนรู้ บนแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย โดยวัดประสิทธิภาพการจัดหมวดหมู่เอกสารดีที่สุด พบว่า อัลกอริทึม SVM ให้ประสิทธิภาพสูงสุด คือ 94.3% และสามารถลดขนาดคุณลักษณะจากกลุ่มตัวอย่างด้วยอัลกอริทึม SVM พบว่า สามารถลดลงได้มากถึง 90% โดยการลดลงของ คุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด

4. ขอบเขตการวิจัย

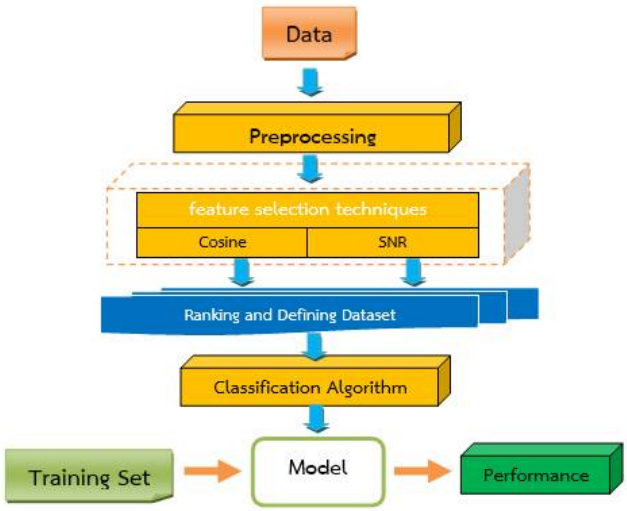
งานวิจัยนี้เป็นงานวิจัยที่เป็นเชิงทดลอง ซึ่งได้ทำการเปรียบเทียบประสิทธิภาพการคัดเลือกคุณลักษณะแบบวิธีการกรอง (Filter) เท่านั้น โดยใช้เทคนิคดังนี้ 1) การวัดค่าความเหมือน (Cosine Similarity) 2) การวัดค่าจากสัญญาณรบกวน (SNR)

ข้อมูลที่ใช้ในการทดลอง เป็นข้อมูลไมโครอาร์เรย์ โรคมะเร็งต่อมน้ำเหลือง (Lymphoma) ประกอบด้วยข้อมูล 47 ตัวอย่าง แบ่งเป็น Germinal Centre B-like จำนวน 24 ตัวอย่าง และ Activated B-like จำนวน 23 ตัวอย่าง จากจำนวนคุณลักษณะ ทั้งหมด 4,026 คุณลักษณะ

การเปรียบเทียบประสิทธิภาพของการจำแนกประเภท จะใช้เทคนิคดังนี้ 1) ต้นไม้ตัดสินใจ (Decision Tree) แบบ J48 2) วิธีเพื่อนบ้านใกล้ที่สุดเค (K-Nearest Neighbor) 3) เบย์อย่างง่าย (Naïve Bayes) โดยใช้โปรแกรม Weka ในการเปรียบเทียบประสิทธิภาพความถูกต้องของการจำแนกประเภท

5. วิธีการดำเนินการวิจัย

ขั้นตอนการดำเนินงาน



ภาพที่ 2 ขั้นตอนการดำเนินงาน

5.1.1 จัดเตรียมชุดข้อมูลข้อมูลไมโครอาร์เรย์ โรคมะเร็งต่อมน้ำเหลือง โดยตรวจความสมบูรณ์ของข้อมูล เนื่องจากข้อมูลไมโครอาร์เรย์ เป็นข้อมูลที่มีจำนวนคุณลักษณะมาก อาจทำให้ชุดข้อมูลไม่สมบูรณ์ อาจมีข้อมูลที่ขาดหายไป จึงต้องเตรียมข้อมูลเบื้องต้นก่อนคำนวนเพื่อเลือกชุดคุณลักษณะจากเทคนิคการเลือกคุณลักษณะ

5.1.2 การเลือกคุณลักษณะ เป็นขั้นตอนการลดมิติของข้อมูลโดยใช้วิธีเลือกคุณลักษณะข้อมูลด้วยเทคนิคการวัดค่าความเหมือน และ การวัดค่าจากสัญญาณรบกวน (SNR) โดยแต่ละเทคนิคจะทำการคำนวณค่าความเหมาะสมและจัดเรียงลำดับ

5.1.3 จัดลำดับข้อมูล (Rank) เป็นการเรียงลำดับของคุณลักษณะตามลำดับความเหมาะสม เพื่อกำหนดเป็นกลุ่มๆ เพื่อเปรียบเทียบประสิทธิภาพความแม่นยำในการจำแนกประเภทข้อมูลต่อไป

5.1.4 การจำแนกประเภทข้อมูล เป็นขั้นตอนการทดสอบการจำแนกประเภทข้อมูล ไมโครอาร์เรย์ หลังจากทำการเลือกคุณลักษณะของข้อมูลที่มีความเหมาะสมและสอดคล้องมากที่สุดจากเทคนิค Cosine Similarity, SNR โดยทำการเปรียบเทียบค่าเฉลี่ยความถูกต้องการจำแนกประเภทข้อมูลของแต่ละกลุ่มการทดสอบ ด้วยต้นไม้ตัดสินใจ (Decision Tree) แบบ J48, วิธีหาสมาชิกที่ใกล้ที่สุด (K-Nearest Neighbor) และเบย์อย่างง่าย

6. ผลการดำเนินงาน

6.1 ผลการเปรียบเทียบการคัดเลือกคุณลักษณะ

ในการคัดเลือกคุณลักษณะด้วยเทคนิค 2 เทคนิค คือ Cosine Similarity และ SNR ทำการเลือกคุณลักษณะที่แบ่งออกเป็นกลุ่ม โดยกลุ่มแรก เลือกคุณลักษณะที่ดีที่สุดจากการคำนวณด้วยเทคนิคที่เลือกมาจำนวน 100 คุณลักษณะ แล้วนำไปจำแนกประเภทด้วยเทคนิคการจำแนกประเภท 3 แบบ คือ ต้นไม้ตัดสินใจ แบบ J48, วิธีหาสมาชิกที่ใกล้ที่สุด และเบย์อย่างง่าย จากนั้นทำการจำแนกประเภทกับจำนวนคุณลักษณะที่ดีที่สุด 200 คุณลักษณะ โดยทำการทดลองโดยเพิ่มจำนวนข้องคุณลักษณะเพิ่มขึ้นทีละ 100 ไปถึง 1,000 คุณลักษณะที่ดีที่สุด โดยผลของการทดลองแบ่งตามจำนวนขนาดของกลุ่มคุณลักษณะมีดังนี้

คุณลักษณะที่เหมาะสมที่สุดจำนวน 100 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	68.09	74.47	76.60	73.05
SNR	85.11	97.87	100	94.33
เฉลี่ย	76.60	86.17	88.30	83.69

คุณลักษณะที่เหมาะสมที่สุดจำนวน 200 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	72.34	72.34	82.98	75.89
SNR	85.11	95.74	100	93.62
เฉลี่ย	78.72	84.04	91.49	84.75

คุณลักษณะที่เหมาะสมที่สุดจำนวน 300 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	72.34	72.34	76.60	73.76
SNR	85.11	95.74	100	93.62
เฉลี่ย	78.72	84.04	88.30	83.69

คุณลักษณะที่เหมาะสมที่สุดจำนวน 400 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	72.34	70.21	74.47	72.34
SNR	85.11	95.74	100	93.62
เฉลี่ย	78.72	82.98	87.23	82.98

คุณลักษณะที่เหมาะสมที่สุดจำนวน 500 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	70.21	68.09	76.00	71.63
SNR	85.11	95.74	100	93.62
เฉลี่ย	77.66	81.91	88.30	82.62

คุณลักษณะที่เหมาะสมที่สุดจำนวน 600 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	68.09	68.09	78.72	71.63
SNR	82.98	95.74	100	92.91
เฉลี่ย	75.53	81.91	89.36	82.27

คุณลักษณะที่เหมาะสมที่สุดจำนวน 700 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	68.09	65.96	80.85	71.63
SNR	82.98	95.74	100	92.91
เฉลี่ย	75.53	80.85	90.43	82.27

คุณลักษณะที่เหมาะสมที่สุดจำนวน 800 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	68.09	68.09	80.85	72.34
SNR	85.11	97.87	100	94.33
เฉลี่ย	76.60	82.98	90.43	83.33

คุณลักษณะที่เหมาะสมที่สุดจำนวน 900 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	63.83	68.09	82.98	71.63
SNR	82.98	89.36	100	90.78
เฉลี่ย	73.40	78.72	91.49	81.21

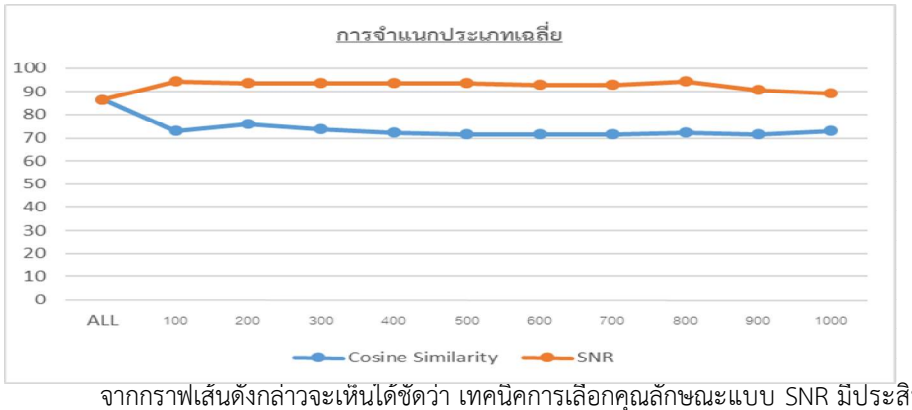
คุณลักษณะที่เหมาะสมที่สุดจำนวน 1,000 คุณลักษณะ

การเลือกคุณลักษณะ	การจำแนกประเภท			
	J48	KNN	Baye	เฉลี่ย
Cosine Similarity	63.83	68.09	87.23	73.05
SNR	82.98	87.23	97.87	89.36
เฉลี่ย	73.40	77.66	92.55	81.21

เมื่อพิจารณาในภาพรวมของการจำแนกประเภทโดยพิจารณาจากค่าเฉลี่ยของแต่ละกลุ่มที่เลือกและการจำแนกประเภทที่แตกต่างกัน โดยสามารถสรุปผลได้ทดลองได้ดังตารางดังต่อไปนี้

การเลือกคุณลักษณะ	การจำแนกประเภทเฉลี่ย											
	ทั้งหมด	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1,000**	**เฉลี่ย**
Cosine Similarity	86.52	73.05	75.89	73.76	72.34	71.63	71.63	71.63	72.34	71.63	73.05	**72.70**
SNR	86.52	**94.33**	93.62	93.62	93.62	93.62	92.91	92.91	94.33	90.78	89.36	**92.91**
เฉลี่ย		83.69	84.75	83.69	82.98	82.62	82.27	82.27	83.33	81.21	81.21	82.80

เมื่อพิจารณาจากตารางดังกล่าวจะเห็นได้ว่า เทคนิคการเลือกคุณลักษณะ แบบ SNR โดยเฉลี่ยจากการจำแนกประเภทโดย 3 การเรียนรู้แล้ว ผลการจำแนกประเภทความถูกต้องของข้อมูล ที่เลือกคุณลักษณะด้วยเทคนิค SNR ให้ผลที่สูงกว่า เทคนิคแบบ Cosine Similarity อย่างชัดเจน ซึ่งในภาพรวมโดยเฉลี่ยแล้ว การเลือกคุณลักษณะด้วยเทคนิค SNR ให้ค่าเฉลี่ยการจำแนกประเภทที่ 92.91 % ในขณะที่เทคนิคการเลือกคุณลักษณะแบบ Cosine Similarity ให้ค่าเฉลี่ยอยู่ที่ 72.70 % และเมื่อสังเกตจากจำนวนคุณลักษณะที่คัดเลือกได้พบว่า เทคนิคการเลือกคุณลักษณะแบบ SNR สามารถใช้คุณลักษณะที่มีคุณภาพที่ดีที่สุดเพียงจำนวน 100 คุณลักษณะ หรือคิดเป็นร้อยละ 0.4026 ของคุณลักษณะทั้งหมด ก็ให้ประสิทธิภาพการจำแนกประเภทได้สูงขึ้นจากเดิมที่ไม่มีการลดจำนวนคุณลักษณะ จาก 86.52 % สูงขึ้นเป็น 94.33 %



ภาพที่ 3 กราฟเส้นแสดงการจำแนกประเภทเฉลี่ย

จากกราฟเส้นดังกล่าวจะเห็นได้ชัดว่า เทคนิคการเลือกคุณลักษณะแบบ SNR มีประสิทธิภาพในการจำแนกประเภทข้อมูลเฉลี่ยดีกว่า Cosine Similarity อย่างเห็นได้ชัด และยังเห็นได้ว่ากลุ่มชุดข้อมูลที่มีจำนวนมากขึ้นส่งผลให้การจำแนกประเภทเฉลี่ยลดลง

7. สรุปผลการวิจัย

การเปรียบเทียบประสิทธิภาพของการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ โดยใช้เทคนิคการตัดเลือกคุณลักษณะ แบบ SNR และ Cosine สรุปได้ว่า ในการเลือกคุณลักษณะที่มีคุณภาพจากจำนวนคุณลักษณะทั้งหมด สามารถเพิ่มประสิทธิภาพในการจำแนกประเภทข้อมูลไมโครอาร์เรย์ได้ และผลการทดลองพบว่าเทคนิคแบบ SNR ให้ผลลัพธ์ที่ดีกว่า เทคนิคแบบ Cosine ซึ่ง SNR เลือกคุณลักษณะที่ตรงประเด็นได้มากกว่า และสูญเสียคุณลักษณะที่สำคัญได้น้อยกว่า Cosine เป็นเทคนิคที่เหมาะสมการข้อมูลชนิดตัวเลข สังเกตได้จากการเลือกคุณลักษณะที่มีคุณภาพสูงที่สุดเพียง 100 คุณลักษณะ ก็สามารถให้ประสิทธิภาพการจำแนกข้อมูลที่สูงที่สุดได้ โดยเพิ่มจากเดิม จาก 86.52% เป็น 94.33% และการวิจัยการคัดเลือกคุณลักษณะที่เหมาะสมในครั้งนี้ สามารถนำไปทิศทางในการเปรียบเทียบเทคนิคการเปรียบเทียบการเลือกคุณลักษณะอื่น ๆ อีกต่อไป

8. เอกสารอ้างอิง

ภาษาไทย

จุฬาลักษณ์ วัฒนานนท์. (2554). การหาความสัมพันธ์ของความรู้ โดยใช้กรอบความรู้การจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหสัมพันธ์. สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

นิธินันท์ มาตา. พนิดา หล่อวงศ์ตระกูล และพยุง มีสัจ. (2558). การเปรียบเทียบประสิทธิภาพเทคนิคการลดมิติของข้อมูลสำหรับค้นหาปัจจัยและสร้างโมเดลการจำแนกกลุ่มการระบายน้ำของประตูระบายน้ำ สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์.

น้ำทิพย์ มากนคร และมาลีรัตน์ โสดานิล. (2557). การเปรียบเทียบวิธีการเลือกคุณลักษณะที่เหมาะสมเพื่อการจัดหมวดหมู่เว็บเพจผิดกฎหมายโดยใช้เทคนิคการทำเหมืองข้อมูล. ภาควิชาเทคโนโลยีสารสนเทศคณะเทคโนโลยีสารสนเทศ. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

นิเวศ จิระวิชิตชัย. ปริญญา สงวนสัตย์ และพยุง มีสัจ. (2554). การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยรังสิต.

พฤฒิพงศ์ เพ็งศิริ. สุริยะ พินิจการ. ณัฐชุดา มงคลชาติ. นวพร วิสิฐพงศ์พันธ์ และพยุง มีสัจ. (2557). การลดมิติข้อมูลการวิเคราะห์ความสัมพันธ์และการประยุกต์สำหรับวิเคราะห์ข้อมูลพื้นฐานการใช้งานสมาร์โฟน. คณะเทคโนโลยีสารสนเทศ. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ภาษาอังกฤษ

Supoj Hengpraprohm. (2008). Feature Selection by Weighted-SNR for Cancer Microarray Data Classification. International Journal of Innovative Computing, Information and Control ICIC International 2008.

Molina L.C., Belanche L. and Nebot A. (2000) Feature Selection Algorithms: A Survey and Experimental Evaluation. Proceeding of the International conference on data mining, pp. 306 - 313.

Quinlan, J. Ross. (1986). Induction of decision trees. Machine learning 1.1, 81-