บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การจำแนกประเภท (Classification) เป็นการวิเคราะห์ข้อมูลสำหรับลักษณะของงานที่กำหนดกลุ่มให้กับหน่วยตัวอย่าง โดยอาศัยลักษณะของหน่วยตัวอย่าง ที่เรียกว่าตัวแปรทำนาย (Predictor Variable) ในการจำแนกกลุ่มจะต้องทราบล่วงหน้าว่ามีกลุ่มอะไรบ้าง และมีข้อมูลลักษณะของหน่วยตัวอย่างในแต่ละกลุ่มก่อน การระบุตัวแปรทำนายที่เกี่ยวข้องกับการจำแนกประเภทได้อย่างถูกต้อง จะทำให้การจำแนกประเภทข้อมูลมีความแม่นยำ

เนื่องด้วยเทคโนโลยีในยุคใหม่มีประสิทธิภาพสูง ทำให้สามารถวัดลักษณะของหน่วยตัวอย่างได้มากขึ้น โดยลักษณะเหล่านี้สามารถนำไปใช้เป็นตัวแปรสำหรับการวิเคราะห์ข้อมูล ข้อมูลที่มีตัวแปรเป็นจำนวนมากเรียกว่าเป็นข้อมูลที่มีจำนวนมิติมาก (High-Dimensional Data) ได้แก่ ข้อมูลไมโครอาร์เรย์ (Microarray Data) ซึ่งเป็นข้อมูลที่ได้จากการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน โดยยีนที่ศึกษามีจำนวนเป็นหลักพันหรือหลักหมื่น สามารถนำมาใช้ในการจำแนกเนื้อเยื่อมะเร็งและเนื้อเยื่อปกติได้ ข้อมูลสำหรับการจำแนกประเภทเอกสาร (Text Classification) ก็เป็นข้อมูลที่มีจำนวนมิติมาก โดยตัวแปรที่เป็นไปได้ในการใช้จำแนกเอกสารคือคำ หรือวลีในเอกสาร ซึ่งมีจำนวนมากในแต่ละเอกสาร (Forman, 2003) หากนำตัวแปรทั้งหมดที่มีจำนวนมากนี้ไปใช้เป็นตัวแปรทำนายสำหรับการจำแนกประเภทเอกสาร ก็อาจลดความแม่นยำในการจำแนกประเภทได้ ข้อมูลทางดาราศาสตร์ที่ได้จากกล้องโทรทรรศน์ที่ใช้เทคโนโลยีขั้นสูง ทำให้ได้ข้อมูลจากภาพของวัตถุที่ได้รับการวัดค่าตัวแปรในจำนวนหลักสิบหรือหลักร้อยตัวแปร (Zheng & Zhang, 2008) ซึ่งเป็นตัวแปรที่มีจำนวนมากหากนำไปใช้จำแนกประเภทของวัตถุ เป็นต้น

การที่ข้อมูลเหล่านี้มีมิติเป็นจำนวนมาก อาจก่อให้เกิดปัญหาหากนำไปใช้ค้นหารูปแบบโดยปราศจากการจัดการข้อมูลในเบื้องต้น เนื่องจากข้อมูลอาจเกิดการกระจาย จนทำให้ในบางจุดของปริภูมิตัวอย่าง (Sample Space) ไม่มีข้อมูลอยู่เลย หากนำตัวแปรทำนายทั้งหมดไปวิเคราะห์ข้อมูลเพื่อสร้างตัวแบบสำหรับการจำแนกประเภทข้อมูล อาจส่งผลเสียต่อความแม่นยำในการจำแนกประเภท และวิธีจำแนกประเภทบางวิธีก็ไม่สามารถรองรับการทำงานกับตัวแปรทำนายที่มีจำนวนมาก นอกจากนี้หากขนาดตัวอย่างสำหรับการวิเคราะห์ข้อมูลมีน้อยจนไม่เพียงพอต่อตัวแปรทำนายที่มีจำนวนมาก อาจทำให้ได้ค่าประมาณที่ไม่ดี รวมทั้งทำให้เกิดการสิ้นเปลืองทรัพยากร เช่น เวลา หรือหน่วยความจำ สิ่งต่าง ๆ เหล่านี้เรียกว่าเป็นปัญหาของมิติข้อมูล (Curse of Dimensionality) ดังนั้น

จึงมีความจำเป็นที่จะคัดเลือกเฉพาะตัวแปรที่มีความเกี่ยวข้อง (Relevant Variables) สำหรับนำไปใช้ในการวิเคราะห์การจำแนกประเภทข้อมูลต่อไป

ข้อมูลไมโครอาร์เรย์เป็นข้อมูลที่ได้จากการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน เพื่อศึกษากลไกที่ควบคุมการเจริญเติบโต และพัฒนาการของสิ่งมีชีวิต (ไก้รุ่ง เฮงพระพรหม, 2553, หน้า 9) ซึ่งสามารถนำข้อมูลนี้ไปใช้เป็นตัวแปรทำนายในการจำแนกประเภทการเป็นมะเร็ง เช่น ในการจำแนกประเภทเนื้อเยื่อมะเร็งและเนื้อเยื่อปกติ ข้อมูลไมโครอาร์เรย์ถือว่าเป็นข้อมูลที่มีจำนวนมิติมาก เนื่องจากยีนที่ศึกษานี้มีจำนวนเป็นหลักพันหรือหลักหมื่นนั่นเอง แต่จำนวนยีนที่มีประโยชน์ต่อการวิเคราะห์ข้อมูลอาจมีจำนวนไม่มาก เช่น ยีนที่มีประโยชน์สำหรับการจัดกลุ่มข้อมูล อาจมีเพียงประมาณ 5% ของยีนทั้งหมดที่ศึกษา (Krzanowski & Hand, 2009) ในขณะที่หน่วยตัวอย่างสำหรับข้อมูลลักษณะนี้มีจำนวนน้อย ดังนั้นกระบวนการที่ใช้ในการกรองตัวแปรที่มีจำนวนมากเหล่านี้ จึงเป็นกระบวนการที่จำเป็นก่อนที่จะใช้วิธีการวิเคราะห์ข้อมูลเพื่อจำแนกประเภทข้อมูลสำหรับช่วยลดจำนวนตัวแปร กำจัดข้อมูลที่ไม่เกี่ยวข้องและข้อมูลซ้ำซ้อน เพิ่มความแม่นยำในการเรียนรู้ (Learning) หรือการสร้างตัวแบบ เพิ่มความเข้าใจต่อตัวแบบที่ได้ และสามารถลดเวลาในการเรียนรู้ข้อมูล อีกทั้งลดความต้องการของหน่วยความจำ กระบวนการนี้เรียกว่าการคัดเลือกตัวแปร (Variable Selection)

การคัดเลือกตัวแปร ซึ่งตัวแปรนี้มีความหมายถึงตัวแปรทำนาย สามารถแบ่งออกได้เป็น 3 วิธี ได้แก่ วิธีฟิลเตอร์ (Filter Method) วิธีแรปเปอร์ (Wrapper Method) และวิธีฝังตัว (Embedded Method) สำหรับข้อมูลที่มีจำนวนมิติมากนั้น วิธีฟิลเตอร์เป็นวิธีที่ได้รับการเลือกให้ใช้ในการคัดเลือกตัวแปรเป็นส่วนใหญ่ เนื่องจากประมวลผลได้เร็วกว่าอีกสองวิธี (Saeys, Inza, & Larrañaga, 2007) วิธีฟิลเตอร์ประเมินความเกี่ยวข้องหรือความสำคัญของตัวแปรในการจำแนกประเภทข้อมูล โดยอาศัยดัชนีวัดความสำคัญของตัวแปร ตัวแปรใดที่มีค่าดัชนีสูง ตัวแปรนั้นจะได้รับการคัดเลือก แบ่งย่อยเป็นวิธีฟิลเตอร์แบบตัวแปรเดียว (Univariate Filter Method) และวิธีฟิลเตอร์แบบหลายตัวแปร (Multivariate Filter Method) โดยวิธีฟิลเตอร์แบบตัวแปรเดียวเป็นวิธีการคัดเลือกตัวแปรที่พิจารณาตัวแปรแต่ละตัวอย่างเป็นอิสระกัน โดยไม่ได้คำนึงถึงความสัมพันธ์ระหว่างตัวแปรด้วยกันในการคัดเลือกตัวแปร กลุ่มตัวแปรที่ได้รับการคัดเลือกอาจมีความสัมพันธ์กันเอง บางตัวแปรอาจไม่ได้ให้สารสนเทศเพิ่มเติมจากตัวแปรอื่นที่ได้รับคัดเลือก หรือในอีกลักษณะหนึ่ง ตัวแปรบางตัวอาจไม่ได้ให้สารสนเทศมากหากพิจารณาเดี่ยว ๆ แต่เมื่อร่วมกับตัวแปรอื่นสามารถให้สารสนเทศที่เป็นประโยชน์ต่อการจำแนกประเภทข้อมูลได้ การพิจารณาตัวแปรแต่ละตัวแปรอย่างเป็นอิสระกัน อาจทำให้ตัวแปรที่คัดเลือกมามีความแม่นยำน้อยในการจำแนกประเภทข้อมูล ซึ่งแตกต่างจากวิธีฟิลเตอร์แบบหลายตัวแปรที่คำนึงถึงความสัมพันธ์ระหว่างตัวแปรด้วยกันใน

การพิจารณาคัดเลือกตัวแปร วิธีฟิลเตอร์แบบหลายตัวแปรจึงมีแนวโน้มที่จะได้ตัวแปรที่มีสารสนเทศไม่ซ้ำซ้อนกัน และมีความเพียงพอในการจำแนกประเภทข้อมูล

Variable Influence on the Projection (VIP) เป็นดัชนีวัดความสำคัญของตัวแปรสำหรับการคัดเลือกตัวแปรด้วยวิธีฟิลเตอร์แบบหลายตัวแปร ดัชนี VIP เป็นฟังก์ชันของพารามิเตอร์ที่ได้จากการวิเคราะห์ข้อมูลด้วยการถดถอยของวิธีกำลังสองน้อยที่สุดบางส่วน (Partial Least Squares Regression: PLS-R) ซึ่งได้รับความสนใจเพิ่มมากขึ้น โดยเฉพาะอย่างยิ่งเมื่อตัวแปรมีความสัมพันธ์กันเอง (Multicollinearity) (Chong & Jun, 2005) แนวคิดของดัชนี VIP คือ ถ้าองค์ประกอบ $\mathbf{t}_a$ มีความสัมพันธ์กับตัวแปรตาม $\mathbf{y}$ และตัวแปรทำนาย $\mathbf{x}_j$ มีความสำคัญในการสร้างองค์ประกอบ $\mathbf{t}_a$ แล้ว ตัวแปรทำนาย $\mathbf{x}_j$ จะมีความสัมพันธ์กับตัวแปรตาม $\mathbf{y}$ ซึ่งดัชนี VIP อาศัยค่าน้ำหนัก $\mathbf{w}_a$ สำหรับวัดการมีส่วนร่วมของตัวแปรทำนายในองค์ประกอบ $\mathbf{t}_a$ และถ่วงน้ำหนักด้วยความสำคัญของแต่ละองค์ประกอบ โดยวัดด้วยความแปรผันของตัวแปรตามที่ได้รับการอธิบายจากแต่ละองค์ประกอบ ซึ่งก็คือค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination: R^2) ของการทำนายตัวแปรตามด้วยองค์ประกอบนั้น ๆ

PLS-R ใช้สำหรับการวิเคราะห์การถดถอย ดังนั้นตัวแปรตามจึงเป็นตัวแปรเชิงปริมาณ แต่ได้มีการนำ PLS-R มาประยุกต์กับข้อมูลที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพเพื่อนำไปใช้กับงานทางด้านการจำแนกประเภทข้อมูล เรียกว่าการวิเคราะห์การจำแนกประเภทของวิธีกำลังสองน้อยที่สุดบางส่วน (Partial Least Square Discriminant Analysis: PLS-DA) โดยแปลงตัวแปรเชิงคุณภาพให้เป็นตัวแปรดัมมี่ (Dummy Variables) ซึ่งตัวแปรดัมมี่แต่ละตัวจะมีเพียงสองค่าเท่านั้น เช่น ค่า 0 และ 1 (Alsberg, Kell, & Goodacre, 1998; Barker & Rayens, 2003; Cho, Lee, Park, Kim, & Lee, 2002; Huang & Pan, 2008; Man, Dyson, Johnson, & Liao 2004; Rajalahti & Kvalheim, 2011; Tan, Shi, Tong, Hwang, & Wang, 2004) หรือค่า -1 และ 1 (Gutkin, Shamir, & Dror, 2009) การประยุกต์วิธีดังกล่าวส่งผลให้การใช้ค่าสัมประสิทธิ์การตัดสินใจ ซึ่งใช้เป็นค่าวัดความสำคัญขององค์ประกอบในการทำนายตัวแปรตาม อาจไม่สะท้อนความสามารถที่แท้จริงขององค์ประกอบนั้น ๆ บางองค์ประกอบอาจให้ค่าสัมประสิทธิ์การตัดสินใจสูงในการทำนายค่าตัวแปรตาม ทั้ง ๆ ที่มีความสามารถในการทำนายการเป็นสมาชิกกลุ่มต่ำ เนื่องจากสัมประสิทธิ์การตัดสินใจใช้วัดความสอดคล้องของข้อมูล เมื่อตัวแปรตามเป็นตัวแปรเชิงปริมาณ การหาดัชนี VIP เพื่อวัดความสำคัญของตัวแปรทำนายที่ต้องอาศัยค่าสัมประสิทธิ์การตัดสินใจ จึงอาจไม่สะท้อนความสำคัญของตัวแปรทำนายแต่ละตัวได้เท่าที่ควร ซึ่งส่งผลให้ตัวแปรทำนายที่ได้รับการคัดเลือกมีความแม่นยำในการทำนายลดลง

นอกจากปัญหาที่ได้กล่าวข้างต้นแล้ว ดัชนีวัดความสำคัญของตัวแปรที่ได้มีการนำเสนอโดยทั่วไปหลายดัชนีไม่ได้กำหนดเกณฑ์จุดตัด (Cutoff Threshold Criterion) ที่ชัดเจนในการคัดเลือกตัวแปร ผู้ใช้อาจต้องกำหนดเกณฑ์จุดตัด หรือจำนวนตัวแปรด้วยตนเอง สำหรับเกณฑ์จุดตัดที่ใช้สำหรับคัดเลือกตัวแปรเมื่อใช้ดัชนี VIP นั้น ส่วนมากคือ 1.0 (Chen, 2008, pp. 14-15; Cho et al., 2002; Chong & Jun, 2005; Lazraq, Cléroux, & Gauchi, 2003; Sun, Yu, Liu, Xu, & Di, 2012) กล่าวคือตัวแปรใดที่ดัชนี VIP>1.0 จะถือว่าเป็นตัวแปรที่มีความสำคัญหรือเกี่ยวข้องกับการทำนายค่าตัวแปรตาม ซึ่งการกำหนดเกณฑ์จุดตัดที่เป็นค่าคงที่ อาจไม่เหมาะกับโครงสร้างข้อมูลทุกลักษณะ (Chong & Jun, 2005)

การจำแนกประเภทข้อมูล ในกรณีที่ข้อมูลมีจำนวนมิติมากให้มีความแม่นยำ ควรมีการคัดกรองตัวแปรที่ไม่เกี่ยวข้องกับการจำแนกประเภทข้อมูลออกจากการวิเคราะห์ข้อมูล วิธีการคัดเลือกตัวแปรควรเป็นวิธีที่เหมาะสมกับข้อมูลที่มีจำนวนมิติมาก ซึ่งตัวแปรมักมีความสัมพันธ์กันเอง การใช้ดัชนี VIP ที่สร้างบนพื้นฐานวิธีกำลังสองน้อยที่สุดบางส่วน ในการวัดความสำคัญของตัวแปรเพื่อการคัดเลือกตัวแปร เป็นวิธีที่เหมาะสมกับข้อมูลที่มีจำนวนมิติมาก แต่ไม่เหมาะกับการจำแนกประเภท เนื่องจากการจำแนกประเภทนั้น ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ ในขณะที่วิธีกำลังสองน้อยที่สุดบางส่วนใช้สำหรับตัวแปรตามเป็นตัวแปรเชิงปริมาณ ดังนั้นการใช้ดัชนี VIP จึงควรมีการปรับเมื่อนำมาใช้กับการจำแนกประเภท และนอกจากนี้การกำหนดจุดตัดสำหรับดัชนีในการคัดเลือกตัวแปรให้เหมาะกับข้อมูลแต่ละชุด แทนที่จะใช้เกณฑ์จุดตัดที่เป็นค่าคงที่ จะทำให้สามารถเลือกตัวแปรที่มีความเกี่ยวข้องกับการจำแนกประเภทได้ถูกต้องยิ่งขึ้น ซึ่งส่งผลถึงการจำแนกประเภทข้อมูลที่มีความถูกต้องมากขึ้น

จากเหตุผลและความสำคัญดังที่กล่าวมา ผู้วิจัยจึงสนใจพัฒนาดัชนีวัดความสำคัญของตัวแปร และเกณฑ์จุดตัดในการคัดเลือกตัวแปรบนพื้นฐานวิธีกำลังสองน้อยที่สุดบางส่วน สำหรับการจำแนกประเภทข้อมูล การคัดเลือกตัวแปรประกอบด้วยดัชนีที่พัฒนาขึ้นใช้ชื่อว่า Variable Importance Index for Classification (VIIC) และเกณฑ์จุดตัดที่พัฒนาขึ้นใช้ชื่อว่า เกณฑ์จุดตัดของแผนภาพกล่อง (Boxplot Cutoff Threshold หรือเกณฑ์จุดตัด BCT) ใช้สัญลักษณ์ว่า BCT_{VIIC} และ BCT_{VIP} สำหรับนำไปใช้เป็นเกณฑ์จุดตัดของดัชนี VIIC และดัชนี VIP ตามลำดับ การคัดเลือกตัวแปรนี้จะนำไปประยุกต์กับข้อมูลมะเร็งลำไส้ใหญ่ โดยคัดกรองยีนที่ไม่มีความเกี่ยวข้องกับการจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติ ออกจากการวิเคราะห์ข้อมูล ซึ่งจะทำให้ได้ตัวแบบที่สามารถจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติได้อย่างถูกต้องยิ่งขึ้น

วัตถุประสงค์ของการวิจัย

1. เพื่อพัฒนาดัชนีวัดความสำคัญของตัวแปรและเกณฑ์จุดตัดในการคัดเลือกตัวแปรบนพื้นฐานวิธีกำลังสองน้อยที่สุดบางส่วน สำหรับการจำแนกประเภทข้อมูล

2. เพื่อเปรียบเทียบการคัดเลือกตัวแปรบนพื้นฐานวิธีกำลังสองน้อยที่สุดบางส่วน สำหรับการจำแนกประเภทข้อมูล ระหว่างการคัดเลือกตัวแปร 3 วิธี ได้แก่ (1) การคัดเลือกตัวแปรที่ใช้ดัชนี VIP ที่มีเกณฑ์จุดตัด 1.0 (2) การคัดเลือกตัวแปรแบบใหม่ที่ใช้ดัชนี VIP และเกณฑ์จุดตัด BCT_{VIP} และ (3) การคัดเลือกตัวแปรแบบใหม่ที่ใช้ดัชนี VIIC และเกณฑ์จุดตัด BCT_{VIIC} ภายใต้ 180 สถานการณ์จำลอง (4x3x5x3) คือ เปอร์เซ็นต์ตัวแปรที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล 4 ระดับ (1%, 3%, 5% และ 10%) ความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มข้อมูลของตัวแปรที่เกี่ยวข้อง 3 ระดับ (1, 3 และ 5 หน่วย) แบบแผนความสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้อง 5 แบบแผน (ไม่สัมพันธ์กัน สัมพันธ์กันที่ระดับสัมประสิทธิ์สหสัมพันธ์คงที่ 0.5 และ 0.9 สัมพันธ์กันที่ระดับสัมประสิทธิ์สหสัมพันธ์ไม่คงที่ซึ่งกำหนดสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทำนายที่ i กับ j เท่ากับ $a^{|i-j|}$ เมื่อ a เท่ากับ 0.5 และ 0.9) ขนาดตัวอย่าง 3 ขนาด (40, 70 และ 100) โดยใช้ค่า Balanced Accuracy (BAC) ของการจำแนกประเภทตัวแปร เป็นเกณฑ์สำหรับการเปรียบเทียบ

3. เพื่อศึกษาความสามารถของยีนที่ได้จากการคัดเลือกตัวแปรในการจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติ

กรอบแนวคิดในการวิจัย

ดัชนี VIP เป็นดัชนีหนึ่งที่ใช้วัดความสำคัญหรือความเกี่ยวข้องของตัวแปรในการทำนายค่าตัวแปรตาม โดยอาศัยพารามิเตอร์จาก PLS-R ซึ่งได้รวมความสัมพันธ์ระหว่างตัวแปรทำนายเข้าไว้ในค่าที่ได้ ดัชนี VIP ได้รับการเสนอโดย Wold and Others ในปี ค.ศ. 1993 (Chong & Jun, 2005) ดัชนี VIP ของตัวแปรทำนายที่ j คำนวณดังนี้

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{a=1}^{h} Red(\mathbf{Y},\mathbf{t}_a)\left(w_{aj}/\left\|\mathbf{w}_a\right\|\right)^2}{(1/p)\sum_{a=1}^{h} Red(\mathbf{Y},\mathbf{t}_a)}} \qquad (1.1)$$

โดยที่ p แทน จำนวนตัวแปรทำนาย

h แทน จำนวนองค์ประกอบ

w_{aj} แทน น้ำหนักของตัวแปรทำนายที่ j ในองค์ประกอบ $\mathbf{t}_a$

$\left\|\mathbf{w}_a\right\|$ แทน ขนาดของเวกเตอร์น้ำหนักขององค์ประกอบ $\mathbf{t}_a$

$Red(\mathbf{Y},\mathbf{t}_a)$ แทน ความแปรผันของตัวแปรตาม $\mathbf{Y}$ ที่ถูกอธิบายด้วยองค์ประกอบ $\mathbf{t}_a$ สำหรับกรณีที่มีตัวแปรตามเพียงหนึ่งตัวนั้น

$Red(\mathbf{Y},\mathbf{t}_a) = R^2(\mathbf{Y},\mathbf{t}_a)$ และ

$$R^2(\mathbf{Y},\mathbf{t}_a) = 1 - \frac{(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})}{\mathbf{y}'\mathbf{y}}$$

การวัดความสำคัญของตัวแปรด้วยดัชนี VIP มีแนวคิดจากการพิจารณาการมีส่วนร่วมของตัวแปรทำนายที่ j ในองค์ประกอบ $\mathbf{t}_a$ ซึ่งวัดด้วยค่าน้ำหนัก w_{aj} และพิจารณาการมีส่วนร่วมขององค์ประกอบ $\mathbf{t}_a$ ในการทำนายตัวแปรตามด้วยความแปรผันที่ถูกอธิบายด้วยองค์ประกอบ $\mathbf{t}_a$ หรือ $\text{Red}(\mathbf{Y},\mathbf{t}_a)$ ในกรณีที่มีตัวแปรตามเพียงหนึ่งตัว ค่าดังกล่าวคือค่าสัมประสิทธิ์การตัดสินใจของการทำนายตัวแปรตามด้วยองค์ประกอบ $\mathbf{t}_a$

การนำ PLS-R และดัชนี VIP ซึ่งใช้สำหรับกรณีตัวแปรตามเป็นตัวแปรเชิงปริมาณมาประยุกต์กับงานการจำแนกประเภทนั้น ตัวแปรตาม $\mathbf{y}$ จะถูกแปลงค่าให้เป็นตัวแปรดัมมี่ ในกรณีที่เป็นงานการจำแนกประเภทที่มีสองกลุ่ม ตัวแปรดัมมี่จะมีเพียงหนึ่งตัว ซึ่งมีสองค่า เช่น -1 และ 1 การใช้ R^2 ในการวัดการมีส่วนร่วมหรือวัดความสอดคล้องขององค์ประกอบ $\mathbf{t}_a$ ในการทำนายตัวแปรตามซึ่งเป็นตัวแปรเชิงคุณภาพ จึงอาจไม่เหมาะสม เนื่องจาก R^2 ใช้สำหรับกรณีตัวแปรตามเป็นตัวแปรเชิงปริมาณ

Adjusted count R^2 เป็นค่าวัดที่มีกรอบแนวคิดที่คล้ายคลึงกับ R^2 ในแง่ของการสะท้อนค่าการลดลงเชิงสัดส่วนของความคลาดเคลื่อนในการทำนายตัวแปรตาม $\mathbf{y}$ แต่ใช้วัดความสอดคล้องของตัวแบบที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพ โดย Adjusted count R^2 ขององค์ประกอบ $\mathbf{t}_a$ คำนวณดังนี้ (Gordon, 2012, p. 578)

$$ACR_a = \frac{\sum_{j=1}^{2} n_{jj} - max(n_{1.}, n_{2.})}{n - max(n_{1.}, n_{2.})} \qquad (1.2)$$

โดยที่ n_{jj} แทน จำนวนหน่วยตัวอย่างของกลุ่ม j ที่ทำนายได้อย่างถูกต้องเมื่อใช้องค์ประกอบ $\mathbf{t}_a$ ในการทำนาย

$n_{1.}$ และ $n_{2.}$ แทน ขนาดตัวอย่างในกลุ่มที่ 1 และ 2 ตามลำดับ

n แทน ขนาดตัวอย่างทั้งหมด

ผู้วิจัยจึงได้นำ Adjusted Count R^2 มาเป็นค่าวัดความสอดคล้องขององค์ประกอบ $\mathbf{t}_a$ ในการทำนายตัวแปรตาม ซึ่งเป็นตัวแปรที่มีสองค่า และนำเสนอดัชนีวัดความสำคัญของตัวแปรแบบ

ใหม่สำหรับงานการจำแนกประเภทคือดัชนี VIIC ตัวแปรใดที่มีดัชนี VIIC สูง แสดงว่าความสามารถในการจำแนกกลุ่มสูงด้วย โดยดัชนี VIIC ของตัวแปรทำนายที่ j คำนวณดังนี้

$$\text{VIIC}_j = \sum_{a=1}^{h} ACR_a \cdot \left(w_{aj} / \left\| \mathbf{w}_a \right\| \right)^2 \quad (1.3)$$

โดยที่ ACR_a แทน Adjusted Count R^2 ขององค์ประกอบ $\mathbf{t}_a$

w_{aj} แทน น้ำหนักของตัวแปรทำนายที่ j ในองค์ประกอบ $\mathbf{t}_a$

$\left\| \mathbf{w}_a \right\|$ แทน ขนาดของเวกเตอร์น้ำหนักขององค์ประกอบ $\mathbf{t}_a$

ในส่วนของเกณฑ์จุดตัดในการคัดเลือกตัวแปรสำหรับดัชนี VIP โดยมากใช้เกณฑ์จุดตัดที่ดัชนี VIP มีค่ามากกว่า 1.0 (Chen, 2008, pp. 14-15; Chong & Jun, 2005; Lazraq et al., 2003; Sun et al., 2012) ซึ่งการกำหนดค่าที่แน่นอนเป็น 1.0 เมื่อใช้ดัชนี VIP นั้น แม้จะง่ายกับการใช้งาน แต่ไม่เหมาะกับทุกลักษณะของโครงสร้างข้อมูล ผู้วิจัยจึงได้นำแนวทางของ Johansson, Lindgren, and Berglund (2003) ที่ได้สร้างเกณฑ์จุดตัดโดยการเพิ่มตัวแปรทำนายซึ่งเป็นค่าที่สร้างขึ้นอย่างสุ่ม โดยสร้างจากการเรียงสับเปลี่ยน (Permutation) ตัวแปรทำนายเริ่มต้น เพื่อนำมาเป็นตัวแปรรบกวน (Noise Variable) จากนั้นคำนวณดัชนีของทั้งตัวแปรทำนายเริ่มต้น และตัวแปรรบกวน

ดัชนี VIIC และดัชนี VIP ของตัวแปรรบกวนนั้น ส่วนใหญ่มีค่าเข้าใกล้ 0 และมีลักษณะเบ้ขวา (Right-Skewed) เนื่องจากเป็นตัวแปรที่สร้างขึ้นอย่างสุ่มที่ไม่ได้มีความเกี่ยวข้องกับการทำนายตัวแปรตามหรือการจำแนกประเภทข้อมูล ผู้วิจัยจึงได้อาศัยแนวคิดการหาค่าผิดปกติ (Outliers) ของดัชนี VIIC หรือดัชนี VIP ของตัวแปรรบกวนเพื่อใช้ในการกำหนดเกณฑ์จุดตัดสำหรับคัดเลือกตัวแปร ค่าผิดปกติเป็นค่าสังเกตที่เกิดอย่างไม่สอดคล้องกับค่าสังเกตอื่นในชุดข้อมูลนั้น ๆ โดยมีโอกาสน้อยที่ค่าผิดปกติจะเกิดจากประชากรเดียวกับค่าสังเกตอื่นที่เป็นปกติ เนื่องจากดัชนี VIIC หรือดัชนี VIP ของตัวแปรรบกวนที่มีค่าปกตินั้น เป็นดัชนีที่ได้จากตัวแปรรบกวนที่สร้างขึ้นอย่างสุ่ม ซึ่งไม่ได้มีความเกี่ยวข้องกับการจำแนกประเภทข้อมูล ดังนั้น ดัชนี VIIC หรือดัชนี VIP ของตัวแปรรบกวนที่มีค่าผิดปกติ จึงมีแนวโน้มว่าจะได้มาจากตัวแปรที่มีความเกี่ยวข้องกับการจำแนกประเภทข้อมูล ตัวแปรใดที่มีดัชนี VIIC หรือดัชนี VIP สอดคล้องกับค่าผิดปกติของดัชนี VIIC หรือดัชนี VIP ของตัวแปรรบกวน จึงแสดงว่าตัวแปรนั้นมีความเกี่ยวข้องต่อการทำนายตัวแปรตาม หรือมีความเกี่ยวข้องในการจำแนกประเภทข้อมูล

เนื่องจากลักษณะการแจกแจงของดัชนี VIIC และดัชนี VIP ของตัวแปรรบกวนไม่ได้มีลักษณะเป็นการแจกแจงปรกติ Tukey ได้นำเสนอการตรวจสอบค่าผิดปกติสำหรับข้อมูลที่ไม่ได้มีการแจกแจงปรกติ ด้วยวิธีแผนภาพกล่อง (Boxplot) ในปี ค.ศ. 1977 (Acuna & Rodriguez, 2004)

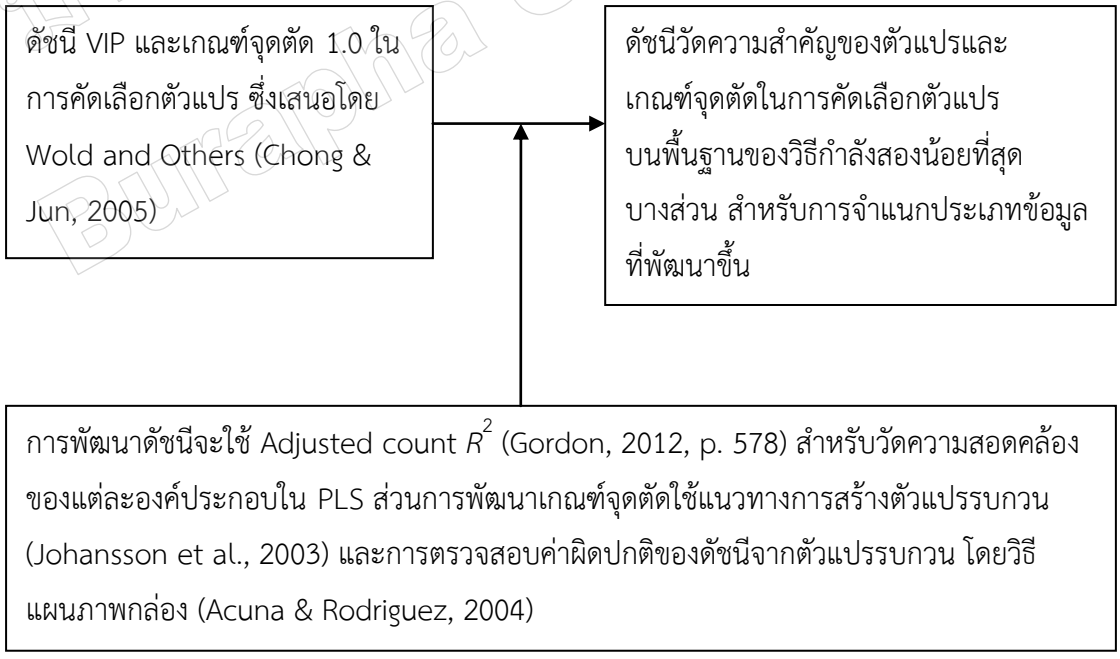
โดยค่าสังเกตที่อยู่นอกช่วง $(Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR)$ เมื่อ Q_1 และ Q_3 แทนควอร์ไทล์ที่ 1 และ 3 ตามลำดับ และ *IQR* แทนพิสัยระหว่างควอร์ไทล์ (Interquartile Range) ซึ่งคำนวณพิสัยระหว่างควอร์ไทล์ได้จาก $IQR = Q_3 - Q_1$ จะถือว่าเป็นค่าผิดปกติ แต่เนื่องจากดัชนี VIIC และดัชนี VIP ของตัวแปรใดมีค่าน้อยหมายถึง ตัวแปรนั้นมีความเกี่ยวข้องกับการทำนายตัวแปรตามน้อย การกำหนดเกณฑ์จุดตัดเพื่อคัดเลือกตัวแปรที่มีความเกี่ยวข้องกับการทำนายตัวแปรตามจึงกำหนดเฉพาะขอบเขตบน โดยที่ตัวแปรใดที่มีดัชนี (VIIC หรือดัชนี VIP) เกินกว่าเกณฑ์จุดตัด BCT ซึ่งเป็นเกณฑ์จุดตัดที่พัฒนาขึ้น ตัวแปรนั้นจะได้รับการคัดเลือก เนื่องจากเป็นตัวแปรที่มีแนวโน้มว่ามีความเกี่ยวข้องกับการจำแนกประเภทข้อมูล เกณฑ์จุดตัด BCT สำหรับดัชนี VIIC และดัชนี VIP จะแทนได้ด้วยสัญลักษณ์ BCT_{VIIC} และ BCT_{VIP} ตามลำดับ และคำนวณดังนี้

$$BCT = Q_3 + 1.5 \times IQR \qquad (1.4)$$

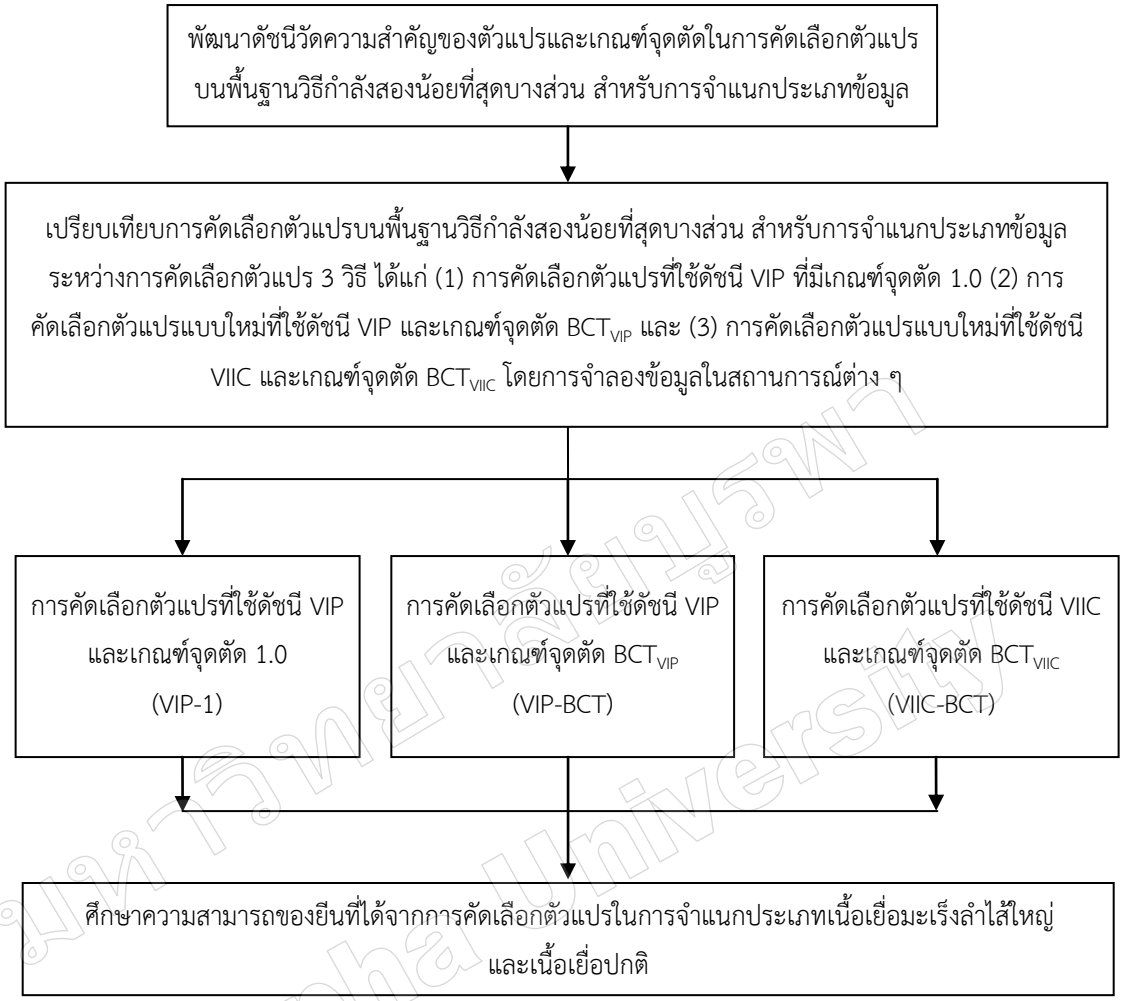
โดยที่ Q_3 แทน ควอร์ไทล์ที่ 3 ของค่าดัชนีของตัวแปรรบกวน

IQR แทน พิสัยระหว่างควอร์ไทล์ (Interquartile Range)

ดังนั้นกรอบแนวคิดการพัฒนาดัชนีวัดความสำคัญของตัวแปรและเกณฑ์จุดตัดในการคัดเลือกตัวแปรบนพื้นฐานของวิธีกำลังสองน้อยที่สุดบางส่วน สำหรับการจำแนกประเภทข้อมูล แสดงดังภาพที่ 1 และกรอบแนวคิดของการวิจัยในภาพรวมแสดงดังภาพที่ 2



ภาพที่ 1 กรอบแนวคิดการพัฒนาดัชนีวัดความสำคัญของตัวแปรและเกณฑ์จุดตัดในการคัดเลือกตัวแปร สำหรับการจำแนกประเภทข้อมูล



ภาพที่ 2 กรอบแนวคิดของการวิจัยในภาพรวม

จากกรอบแนวคิดดังกล่าว ผู้วิจัยจึงตั้งสมมุติฐานของการวิจัย ดังนี้

1. วิธีการคัดเลือกตัวแปรแบบใหม่ทั้ง 2 วิธี คือการคัดเลือกตัวแปรแบบใหม่ที่ใช้ดัชนี VIP และเกณฑ์จุดตัด BCT_{VIP} และการคัดเลือกตัวแปรแบบใหม่ที่ใช้ดัชนี VIIC และเกณฑ์จุดตัด BCT_{VIIC} มีความสามารถในการคัดเลือกตัวแปรได้แม่นยำกว่าการคัดเลือกตัวแปรแบบเดิมที่ใช้ดัชนี VIP ที่มีเกณฑ์จุดตัด 1.0 โดยวัดค่าความแม่นยำด้วยค่า Balanced Accuracy ของการจำแนกประเภทตัวแปร ภายใต้สถานการณ์ต่าง ๆ กัน ที่มีเปอร์เซ็นต์ตัวแปรที่เกี่ยวข้องกับการจำแนกประเภทข้อมูลเท่ากับ 1%, 3%, 5% และ 10% ความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มข้อมูลของตัวแปรที่เกี่ยวข้องเท่ากับ 1, 3 และ 5 หน่วย แบบแผนความสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้อง ได้แก่ ไม่สัมพันธ์กัน สัมพันธ์กันที่ระดับสัมประสิทธิ์สหสัมพันธ์คงที่ 0.5 และ 0.9 สัมพันธ์กันที่ระดับสัมประสิทธิ์สหสัมพันธ์

ไม่คงที่ โดยกำหนดสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทำนายที่ i กับ j เท่ากับ $a^{|i-j|}$ โดยที่ a เท่ากับ 0.5 และ 0.9 และขนาดตัวอย่างเท่ากับ 40, 70 และ 100 ตัวอย่าง

2. ยีนที่ได้จากการคัดเลือกตัวแปรด้วยวิธี VIP-1 วิธี VIP-BCT และวิธี VIIC-BCT สามารถจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติได้ถูกต้องมากกว่า 80%

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ได้ดัชนีวัดความสำคัญของตัวแปรและเกณฑ์จุดตัดบนพื้นฐานของวิธีกำลังสองน้อยที่สุดบางส่วนและนำไปใช้ในการคัดเลือกตัวแปร สำหรับการจำแนกประเภทข้อมูล

2. นักวิจัยสามารถนำวิธีการคัดเลือกตัวแปรที่อาศัยดัชนีวัดความสำคัญของตัวแปร และเกณฑ์จุดตัดที่พัฒนาขึ้นนี้ ไปใช้คัดกรองตัวแปรก่อนนำไปวิเคราะห์ข้อมูลในงานการจำแนกประเภทข้อมูลสำหรับข้อมูลที่มีจำนวนตัวแปรทำนายมากได้

3. ยีนที่คัดเลือกได้จากวิธีการคัดเลือกตัวแปรที่พัฒนาขึ้นสามารถนำมาใช้ในการจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติได้อย่างมีประสิทธิภาพ

ขอบเขตของการวิจัย

สำหรับข้อมูลจำลอง (วัตถุประสงค์ข้อ 2)

1. ตัวแปรทำนาย ($\mathbf{x}$) มีจำนวน 2,000 ตัว ตัวแปรตาม ($\mathbf{y}$) เป็นตัวแปรเชิงคุณภาพ ซึ่งมี 2 ค่า คือ ค่า −1 และ +1 ซึ่งแทนสมาชิกที่อยู่ในกลุ่ม −1 และ +1 ตามลำดับ และในแต่ละกลุ่มมีจำนวนที่เท่ากัน

2. ตัวแปรทำนาย แบ่งเป็นตัวแปรที่เกี่ยวข้อง และตัวแปรที่ไม่เกี่ยวข้องกับการจำแนกประเภทข้อมูล มีการแจกแจงข้อมูลดังนี้

2.1 ตัวแปรที่เกี่ยวข้องมีการแจกแจงปรกติหลายตัวแปร $X_{ijk} \sim \text{MVN}\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}\right)$ โดยที่ i=1,..., n; j=1,..., d (จำนวนตัวแปรที่เกี่ยวข้อง); k=−1, +1 และ $Var(\text{x}) = 1$ เมื่อเวกเตอร์ค่าเฉลี่ย ($\boldsymbol{\mu}_k$) และ เมทริกซ์ความแปรปรวนและความแปรปรวนร่วมเกี่ยว ($\boldsymbol{\Sigma}$) ได้จากการกำหนดความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มข้อมูลของตัวแปรที่เกี่ยวข้อง และแบบแผนความสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้อง ตามลำดับ

2.2 ตัวแปรที่ไม่เกี่ยวข้อง มีการแจกแจงปรกติ $X_{ijk} \sim \text{N}(0,1)$ โดยที่ i=1,..., n; j=1,..., 2000−d; k=−1, +1

3. ในแต่ละสถานการณ์ของการจำลองทำซ้ำ 200 ครั้ง โดยใช้โปรแกรม MATLAB

4. ศึกษา 4 พารามิเตอร์ จำนวน 180 สถานการณ์ (4×3×5×3) ดังต่อไปนี้

4.1 เปอร์เซ็นต์ตัวแปรที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล (Prel) กำหนด 4 ระดับ ได้แก่

1) 1 % (มีตัวแปรที่เกี่ยวข้องเท่ากับ 20 ตัวแปร)

2) 3 % (มีตัวแปรที่เกี่ยวข้องเท่ากับ 60 ตัวแปร)

3) 5 % (มีตัวแปรที่เกี่ยวข้องเท่ากับ 100 ตัวแปร)

4) 10 % (มีตัวแปรที่เกี่ยวข้องเท่ากับ 200 ตัวแปร)

4.2 ความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มข้อมูลของตัวแปรที่เกี่ยวข้อง (Mdif) กำหนด 3 ระดับ ได้แก่

1) 1 หน่วย ($\boldsymbol{\mu}_{-1} = (-0.5\ -0.5\ \ldots\ -0.5)'$ และ $\boldsymbol{\mu}_{+1} = (0.5\ 0.5\ \ldots\ 0.5)'$)

2) 3 หน่วย ($\boldsymbol{\mu}_{-1} = (-1.5\ -1.5\ \ldots\ -1.5)'$ และ $\boldsymbol{\mu}_{+1} = (1.5\ 1.5\ \ldots\ 1.5)'$)

3) 5 หน่วย ($\boldsymbol{\mu}_{-1} = (-2.5\ -2.5\ \ldots\ -2.5)'$ และ $\boldsymbol{\mu}_{+1} = (2.5\ 2.5\ \ldots\ 2.5)'$)

4.3 แบบแผนความสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้อง ($\boldsymbol{\Sigma}$) กำหนด 5 แบบแผน ได้แก่

1) ไม่สัมพันธ์กัน ($\boldsymbol{\Sigma}_1$)

$$\boldsymbol{\Sigma}_1 = \mathbf{I}_{d\times d}$$

2) สัมพันธ์กัน ที่ระดับสัมประสิทธิ์สหสัมพันธ์คงที่ เท่ากับ 0.5 ($\boldsymbol{\Sigma}_2$)

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0.5 & \cdots & 0.5 \\ 0.5 & 1 & \cdots & 0.5 \\ \vdots & \ddots & & \vdots \\ 0.5 & \cdots & & 1 \end{bmatrix}_{d\times d}$$

3) สัมพันธ์กัน ที่ระดับสัมประสิทธิ์สหสัมพันธ์คงที่ เท่ากับ 0.9 ($\boldsymbol{\Sigma}_3$)

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 1 & 0.9 & \cdots & 0.9 \\ 0.9 & 1 & \cdots & 0.9 \\ \vdots & \ddots & & \vdots \\ 0.9 & \cdots & & 1 \end{bmatrix}_{d\times d}$$

4) สัมพันธ์กัน ที่ระดับสัมประสิทธิ์สหสัมพันธ์ไม่คงที่ โดยกำหนดสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทำนายที่ i กับ j เท่ากับ $a^{|i-j|}$ เมื่อ a เท่ากับ 0.5 ($\boldsymbol{\Sigma}_4$)

$$\boldsymbol{\Sigma}_4 = \begin{bmatrix} 1 & (0.5) & (0.5)^2 & (0.5)^3 & \dots & (0.5)^{d-1} \\ (0.5) & 1 & (0.5) & (0.5)^2 & \dots & (0.5)^{d-2} \\ \vdots & & \ddots & & & \vdots \\ (0.5)^{d-1} & & & \dots & & 1 \end{bmatrix}_{d \times d}$$

5) สัมพันธ์กัน ที่ระดับสัมประสิทธิ์สหสัมพันธ์ไม่คงที่ โดยกำหนดสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรทำนายที่ i กับ j เท่ากับ $a^{|i-j|}$ เมื่อ a เท่ากับ 0.9 ($\boldsymbol{\Sigma}_5$)

$$\boldsymbol{\Sigma}_5 = \begin{bmatrix} 1 & (0.9) & (0.9)^2 & (0.9)^3 & \dots & (0.9)^{d-1} \\ (0.9) & 1 & (0.9) & (0.9)^2 & \dots & (0.9)^{d-2} \\ \vdots & & \ddots & & & \vdots \\ (0.9)^{d-1} & & & \dots & & 1 \end{bmatrix}_{d \times d}$$

4.4 ขนาดตัวอย่าง (n) จำนวน 3 ขนาด ได้แก่

1) 40 ตัวอย่าง

2) 70 ตัวอย่าง

3) 100 ตัวอย่าง

5. เกณฑ์ที่ใช้ในการเปรียบเทียบการคัดเลือกตัวแปร คือ Balanced Accuracy (BAC) ของการจำแนกประเภทตัวแปร (ซึ่งจำแนกเป็นตัวแปรที่เกี่ยวข้องและตัวแปรที่ไม่เกี่ยวข้อง) โดยเป็นค่าสัดส่วนความแม่นยำในการจำแนกประเภทตัวแปรที่ถ่วงน้ำหนักระหว่าง Sensitivity (Sen) กับ Specificity (Spe) ของการจำแนกประเภทตัวแปร เมื่อ Sen คือสัดส่วนตัวแปรที่เกี่ยวข้องที่ได้รับการทำนายว่าเป็นตัวแปรที่เกี่ยวข้อง (ได้รับการคัดเลือก) และ Spe คือสัดส่วนตัวแปรที่ไม่เกี่ยวข้องที่ได้รับการทำนายว่าเป็นตัวแปรที่ไม่เกี่ยวข้อง (ไม่ได้รับการคัดเลือก)

การคำนวณค่า BAC, Sen และ Spe ของการจำแนกประเภทตัวแปร อาศัยตารางดังนี้

ตารางที่ 1 ผลการจำแนกประเภทตัวแปร

ประเภทของตัวแปรที่เป็นจริง	ประเภทของตัวแปรที่ได้รับการทำนาย	
	ตัวแปรที่เกี่ยวข้อง	ตัวแปรที่ไม่เกี่ยวข้อง
ตัวแปรที่เกี่ยวข้อง	a_1	a_2
ตัวแปรที่ไม่เกี่ยวข้อง	a_3	a_4

a_1 หมายถึง จำนวนตัวแปรที่เกี่ยวข้องและได้รับการทำนายว่าเป็นตัวแปรที่เกี่ยวข้อง

a_2 หมายถึง จำนวนตัวแปรที่เกี่ยวข้องและได้รับการทำนายว่าเป็นตัวแปรที่ไม่เกี่ยวข้อง

a_3 หมายถึง จำนวนตัวแปรที่ไม่เกี่ยวข้องและได้รับการทำนายว่าเป็นตัวแปรที่เกี่ยวข้อง

a_4 หมายถึง จำนวนตัวแปรที่ไม่เกี่ยวข้องและได้รับการทำนายว่าเป็นเป็นตัวแปรที่ไม่เกี่ยวข้อง

BAC ของการจำแนกประเภทตัวแปร คำนวณดังนี้

$$\text{BAC} = \frac{Sen + Spe}{2} \tag{1.5}$$

$$\text{Sen} = \frac{a_1}{a_1 + a_2} \tag{1.6}$$

$$\text{Spe} = \frac{a_4}{a_3 + a_4} \tag{1.7}$$

BAC, Sen และ Spe มีค่าตั้งแต่ 0 ถึง 1 ถ้าการคัดเลือกตัวแปรของวิธีใดมีค่าวัดเหล่านี้สูงแสดงว่าวิธีนั้นมีความสามารถในการคัดเลือกตัวแปรได้อย่างถูกต้องมาก

สำหรับข้อมูลจริง (วัตถุประสงค์ข้อ 3)

1. ข้อมูลจริงเป็นข้อมูลไมโครอาร์เรย์เกี่ยวกับมะเร็งลำไส้ใหญ่ที่ได้จากการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน โดยเป็นข้อมูลที่เผยแพร่ในเว็บไซต์ http://genomics-pubs.princeton.edu/oncology/affydata/index.html (Alon et al., 1999)

กลุ่มตัวอย่างเป็นกลุ่มเนื้อเยื่อมะเร็งลำไส้ใหญ่จำนวน 40 ชิ้นเนื้อ และเนื้อเยื่อปกติจำนวน 22 ชิ้นเนื้อ

ตัวแปรทำนายที่ใช้ในการศึกษา คือยีนจำนวน 2,000 ยีน โดยค่าสังเกตคือค่าระดับการแสดงออกของยีน (Expression Level of Gene)

ตัวแปรตาม คือประเภทเนื้อเยื่อ (แบ่งเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติ)

2. การสร้างตัวแบบการจำแนกประเภทด้วยวิธี PLS-DA

3. เกณฑ์ที่ใช้วัดความสามารถของยีนที่ได้จากการคัดเลือกตัวแปรในการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่คือ BAC ของการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่ ซึ่งเป็นค่าสัดส่วนความแม่นยำในจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่และเนื้อเยื่อปกติ โดยถ่วงน้ำหนักระหว่าง Sen กับ Spe ของการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่ เมื่อ Sen คือสัดส่วนตัวอย่างเนื้อเยื่อมะเร็งลำไส้ใหญ่ที่ได้รับการทำนายว่าเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่ และ Spe คือสัดส่วนตัวอย่างเนื้อเยื่อปกติที่ได้รับการทำนายว่าเป็นเนื้อเยื่อปกติ

การคำนวณค่า BAC, Sen และ Spe ของการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่อาศัยตารางดังนี้

ตารางที่ 2 ผลการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่

ประเภทของเนื้อเยื่อที่เป็นจริง	ประเภทของเนื้อเยื่อที่ได้รับการทำนาย	
	เนื้อเยื่อมะเร็งลำไส้ใหญ่	เนื้อเยื่อปกติ
เนื้อเยื่อมะเร็งลำไส้ใหญ่	a_1	a_2
เนื้อเยื่อปกติ	a_3	a_4

a_1 หมายถึง จำนวนเนื้อเยื่อมะเร็งลำไส้ใหญ่และได้รับการทำนายว่าเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่

a_2 หมายถึง จำนวนเนื้อเยื่อมะเร็งลำไส้ใหญ่และได้รับการทำนายว่าเป็นเนื้อเยื่อปกติ

a_3 หมายถึง จำนวนเนื้อเยื่อปกติและได้รับการทำนายว่าเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่

a_4 หมายถึง จำนวนเนื้อเยื่อปกติและได้รับการทำนายว่าเป็นเนื้อเยื่อปกติ

BAC ของการจำแนกประเภทการเป็นมะเร็งลำไส้ใหญ่ คำนวณดังนี้

$$\text{BAC} = \frac{Sen + Spe}{2} \tag{1.8}$$

$$\text{Sen} = \frac{a_1}{a_1 + a_2} \tag{1.9}$$

$$\text{Spe} = \frac{a_4}{a_3 + a_4} \tag{1.10}$$

การคำนวณค่าในสมการที่ (1.8) ถึงสมการที่ (1.10) มีลักษณะคล้ายกับในสมการที่ (1.5) ถึงสมการที่ (1.7) ตามลำดับ แตกต่างกันที่ข้อมูลที่นำมาใช้ในการคำนวณเท่านั้น โดยในสมการที่ (1.5) ถึงสมการที่ (1.7) ใช้ผลลัพธ์จากการทำนายประเภทของตัวแปรที่ถูกจำแนกประเภทว่าเป็นตัวแปรที่เกี่ยวข้องหรือตัวแปรที่ไม่เกี่ยวข้อง ส่วนในสมการที่ (1.8) ถึงสมการที่ (1.10) ใช้ผลลัพธ์จากการทำนายประเภทของการเป็นมะเร็งลำไส้ใหญ่ว่าเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่หรือเนื้อเยื่อปกติ

นิยามศัพท์เฉพาะ

การวิจัยนี้กำหนดนิยามศัพท์เฉพาะดังต่อไปนี้

1. การคัดเลือกตัวแปร (Variable Selection) หมายถึงกระบวนการลดขนาดของตัวแปรทำนาย เพื่อให้ได้เซตย่อยของตัวแปรทำนายที่มีความเกี่ยวข้องกับการจำแนกประเภทข้อมูล เพื่อวัตถุประสงค์สำหรับการปรับปรุงประสิทธิภาพในการจำแนกประเภทข้อมูล

2. ข้อมูลที่มีจำนวนมิติมาก (High-Dimensional Data) หมายถึงข้อมูลที่ประกอบด้วยตัวแปรจำนวนมากในหลักร้อยหรือหลักพัน สำหรับการวิเคราะห์ข้อมูล

3. ข้อมูลไมโครอาร์เรย์ (Microarray Data) หมายถึงข้อมูลที่ได้จากการศึกษารูปแบบการแสดงออกของยีนของสิ่งมีชีวิตหลายยีนพร้อม ๆ กัน เพื่อศึกษากลไกที่ควบคุมการเจริญเติบโตและพัฒนาการของสิ่งมีชีวิต

4. การเรียนรู้ (Learning) หรือการเรียนรู้ของเครื่อง (Machine Learning) หมายถึงการที่คอมพิวเตอร์สังเคราะห์ความรู้ขึ้นมาจากข้อมูลจำนวนมาก โดยอาศัยอัลกอริทึมการเรียนรู้ในการระบุตัวแบบที่เหมาะที่สุดสำหรับความสัมพันธ์ระหว่างตัวแปรทำนายกับตัวแปรตาม

5. ตัวแปรที่เกี่ยวข้อง (Relevant Variable) หมายถึงตัวแปรทำนายที่มีความเกี่ยวข้องกับการทำนายตัวแปรตาม หรือตัวแปรทำนายที่มีอิทธิพลในการจำแนกประเภทข้อมูล ซึ่งควรจะถูกนำไปใช้ในการสร้างตัวแบบสำหรับการจำแนกประเภทข้อมูล

6. ตัวแปรที่ไม่เกี่ยวข้อง (Irrelevant Variable) หมายถึงตัวแปรทำนายที่ไม่ได้เกี่ยวข้องกับการทำนายตัวแปรตาม หรือตัวแปรทำนายที่ไม่มีอิทธิพลต่อการจำแนกประเภทข้อมูล ซึ่งควรกำจัดทิ้งไปก่อนที่จะสร้างตัวแบบสำหรับการจำแนกประเภทข้อมูล

7. Adjusted count R^2 หมายถึง ค่าที่ใช้วัดความสอดคล้อง (Goodness of Fit Measures) ของตัวแบบโดยวัดจากความแม่นยำของตัวแบบในการจำแนกประเภทข้อมูล ซึ่งได้รับการปรับค่าด้วยการไม่พิจารณาความแม่นยำของการจำแนกประเภทที่อาศัยการเดาว่าเป็นกลุ่มที่มีขนาดใหญ่กว่า

8. เปอร์เซ็นต์ตัวแปรที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล (Prel) หมายถึง ค่าเปอร์เซ็นต์ตัวแปรที่เกี่ยวข้องต่อตัวแปรทำนายทั้งหมด คำนวณได้ดังนี้

$$\text{Prel} = \frac{\text{จำนวนตัวแปรที่เกี่ยวข้อง}}{\text{จำนวนตัวแปรทำนายทั้งหมด}} \times 100$$

9. ความแตกต่างของค่าเฉลี่ยระหว่างกลุ่มข้อมูลของตัวแปรที่เกี่ยวข้อง (Mdif) หมายถึง ผลต่างของค่าเฉลี่ยของตัวแปรที่เกี่ยวข้องระหว่างกลุ่ม +1 กับกลุ่ม -1 คำนวณดังนี้

$$\text{Mdif} = \boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}$$

$\boldsymbol{\mu}_{+1}$ แทน ค่าเฉลี่ยของตัวแปรที่เกี่ยวข้องของกลุ่ม +1

$\boldsymbol{\mu}_{-1}$ แทน ค่าเฉลี่ยของตัวแปรที่เกี่ยวข้องของกลุ่ม -1

10. แบบแผนความสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้อง ($\boldsymbol{\Sigma}$) หมายถึงแบบแผนความสัมพันธ์ของตัวแปรที่เกี่ยวข้องที่ i กับ j โดยที่ $i, j = 1,\ldots, d$ โดยเขียนอยู่ในรูปเมทริกซ์ดังนี้

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{12} \cdots \rho_{1d} \\ \rho_{21} & 1 \quad \cdots \rho_{2d} \\ \vdots & \ddots \quad \vdots \\ \rho_{d1} \cdots & 1 \end{bmatrix}$$

ρ_{ij} แทนสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรที่เกี่ยวข้องที่ i กับ j

d แทนจำนวนตัวแปรที่เกี่ยวข้อง

11. ขนาดตัวอย่าง หมายถึงจำนวนหน่วยตัวอย่างซึ่งแต่ละหน่วยตัวอย่างประกอบด้วยข้อมูลตัวแปรทำนาย และตัวแปรตาม (กลุ่มข้อมูล)

12. Sensitivity (Sen) สำหรับการประเมินประสิทธิภาพการจำแนกประเภทตัวแปร หมายถึงสัดส่วนตัวแปรที่เกี่ยวข้องที่ได้รับการทำนายว่าเป็นตัวแปรที่เกี่ยวข้อง และ Sen สำหรับการประเมินประสิทธิภาพการจำแนกประเภทข้อมูล หมายถึงสัดส่วนเนื้อเยื่อมะเร็งลำไส้ใหญ่ที่ได้รับการทำนายว่าเป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่

$$\text{Sen} = \frac{\text{จำนวนตัวแปรที่เกี่ยวข้อง(จำนวนเนื้อเยื่อมะเร็งลำไส้ใหญ่) ที่ได้รับการทำนายว่าเป็นตัวแปรที่เกี่ยวข้อง (เป็นเนื้อเยื่อมะเร็งลำไส้ใหญ่)}}{\text{จำนวนตัวแปรทั้งหมด (จำนวนหน่วยตัวอย่างทั้งหมด)}}$$

13. Specificity (Spe) สำหรับการประเมินประสิทธิภาพการจำแนกประเภทตัวแปร หมายถึงสัดส่วนตัวแปรที่ไม่เกี่ยวข้องที่ได้รับการทำนายว่าเป็นตัวแปรที่ไม่เกี่ยวข้อง และ Spe สำหรับการประเมินประสิทธิภาพการจำแนกประเภทข้อมูล หมายถึงสัดส่วนเนื้อเยื่อปกติที่ได้รับการทำนายว่าเป็นเนื้อเยื่อปกติ

$$\text{Spe} = \frac{\text{จำนวนตัวแปรที่ไม่เกี่ยวข้อง (จำนวนเนื้อเยื่อปกติ) ที่ได้รับการทำนายว่าเป็นตัวแปรที่ไม่เกี่ยวข้อง (เป็นเนื้อเยื่อปกติ)}}{\text{จำนวนตัวแปรทั้งหมด (จำนวนหน่วยตัวอย่างทั้งหมด)}}$$

14. Balanced Accuracy (BAC) หมายถึงค่าที่ใช้วัดความแม่นยำในการจำแนกประเภทตัวแปรหรือข้อมูลที่ถ่วงน้ำหนักระหว่าง Sen และ Spe คำนวณได้ดังนี้

$$\text{BAC} = \frac{\text{Sen}+\text{Spe}}{2}$$