บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การวิจัยนี้เป็นการพัฒนาดัชนีวัดความสำคัญของตัวแปรและเกณฑ์จุดตัดในการคัดเลือกตัวแปรบนพื้นฐานวิธีกำลังสองน้อยที่สุดบางส่วน สำหรับการจำแนกประเภทข้อมูล เพื่อเปรียบเทียบกับการคัดเลือกตัวแปรแบบดั้งเดิมที่ใช้ดัชนี VIP ที่มีเกณฑ์จุดตัด 1.0 และนำไปศึกษาความสามารถของยีนที่ได้จากการคัดเลือกตัวแปรในการจำแนกประเภทเนื้อเยื่อมะเร็งลำไส้ใหญ่ และเนื้อเยื่อปกติ ซึ่งผู้วิจัยได้ศึกษาค้นคว้าเอกสารและงานวิจัยที่เกี่ยวข้อง โดยนำเสนอผลการทบทวนเอกสารและงานวิจัยที่เกี่ยวข้องเป็น 4 ตอน ได้แก่

ตอนที่ 1 การคัดเลือกตัวแปร

ตอนที่ 2 วิธีกำลังสองน้อยที่สุดบางส่วน

ตอนที่ 3 การจำแนกประเภท

ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง

ตอนที่ 1 การคัดเลือกตัวแปร

ความหมายของการคัดเลือกตัวแปร

การคัดเลือกตัวแปร (Variable Selection) เป็นที่รู้จักกันในหลายชื่อเช่น การคัดเลือกคุณลักษณะ (Feature Selection) การลดคุณลักษณะ (Feature Reduction) การคัดเลือกตามลักษณะประจำ (Attribute Selection) การคัดเลือกเซตย่อยของตัวแปร (Variable Subset Selection) หรือใช้ชื่อที่เฉพาะเจาะจงกับตัวแปร เช่น การคัดเลือกยีน (Gene Selection) การคัดเลือกความยาวคลื่น (Wavelength Selection) เป็นต้น การคัดเลือกตัวแปรได้รับการนิยามจากผู้เขียนหลายท่าน (Guyon, 2008; John, Kohavi, & Pfleger, 1994; Kira & Rendell, 1992; Koller & Sahami, n.d.; Narendra & Fukunaga, 1977) ซึ่งมีความหมายที่ครอบคลุมในประเด็นต่าง ๆ 4 ประเด็น สรุปโดย Dash and Liu (1997) เป็นดังนี้

1. คำนิยามตามความหมายอุดมคติ

การคัดเลือกตัวแปรเป็นการหาเซตย่อยของตัวแปรที่มีขนาดเล็กที่สุดที่จำเป็นและเพียงพอสำหรับการจำแนกประเภทข้อมูล

2. คำนิยามตามความหมายดั้งเดิม

การคัดเลือกตัวแปรเป็นการเลือกเซตย่อยของตัวแปรขนาด m จากกลุ่มตัวแปรที่มีขนาด p โดยที่มีค่าฟังก์ชันเกณฑ์ที่เหมาะสมที่สุดในบรรดาเซตย่อยขนาด m เมื่อ m และ p เป็น

จำนวนเต็มบวก และ $m < p$

3. คำนิยามในแง่ของการปรับปรุงความแม่นยำในการทำนาย

การคัดเลือกตัวแปรมีเป้าหมายในการเลือกเซตย่อยของตัวแปร เพื่อปรับปรุงความแม่นยำของการจำแนกประเภทข้อมูล หรือการลดขนาดของโครงสร้าง โดยไม่ลดความแม่นยำในการจำแนกประเภทอย่างมีนัยสำคัญ เมื่อใช้เฉพาะตัวแปรที่เลือกในการสร้างตัวแบบจำแนกประเภท

4. คำนิยามในแง่ของการประมาณการกระจายของกลุ่มเริ่มต้น

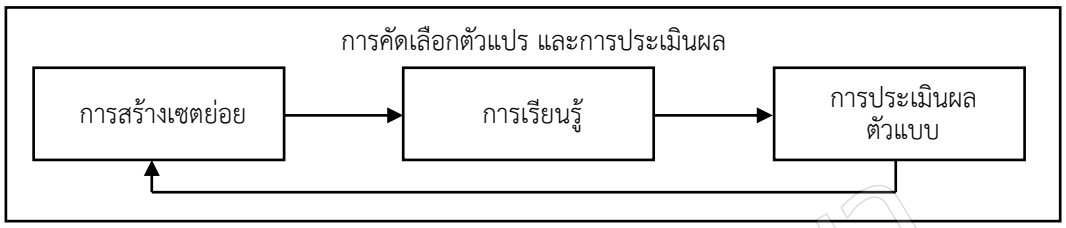
การคัดเลือกตัวแปรมีเป้าหมายเพื่อเลือกเซตย่อยที่มีขนาดเล็กที่การกระจาย ของกลุ่ม (Class Distribution) เมื่อใช้เฉพาะตัวแปรที่ถูกเลือกมีลักษณะใกล้เคียงกับการกระจายของกลุ่มเริ่มต้นเมื่อมีตัวแปรครบ

โดยสรุปแล้วการคัดเลือกตัวแปร หมายถึงกระบวนการลดขนาดของตัวแปร เพื่อให้ได้เซตย่อยของตัวแปรที่มีความเกี่ยวข้องกับการจำแนกประเภท เพื่อวัตถุประสงค์สำหรับการปรับปรุงประสิทธิภาพในการจำแนกประเภท การจัดเตรียมตัวแปรสำหรับการจำแนกประเภทที่สามารถประมวลผลได้อย่างรวดเร็วมีประสิทธิภาพ และเพิ่มความเข้าใจต่อตัวแบบที่ได้

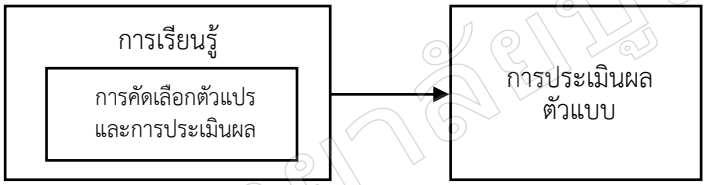
แนวทางของการคัดเลือกตัวแปร

ในบริบทของการจำแนกประเภท การคัดเลือกตัวแปรแบ่งออกได้เป็น 3 วิธี ได้แก่ (1) วิธีฟิลเตอร์ (Filter Method) (2) วิธีแรปเปอร์ (Wrapper Method) และ (3) วิธีฝังตัว (Embedded Method) ซึ่งภาพที่ 3 แสดงการเปรียบเทียบแนวคิดของทั้ง 3 วิธีดังกล่าว (Hilario & Kalousis, 2008)

(1) วิธีฟิลเตอร์



(2) วิธีแรปเปอร์



(3) วิธีฝังตัว

ภาพที่ 3 การเปรียบเทียบแนวคิดของการคัดเลือกตัวแปร (1) วิธีฟิลเตอร์ (2) วิธีแรปเปอร์ และ (3) วิธีฝังตัว

วิธีฟิลเตอร์คัดเลือกตัวแปรโดยประเมินความเกี่ยวข้องหรือวัดความสำคัญของตัวแปรต่อการจำแนกประเภทด้วยการพิจารณาคุณสมบัติในเนื้อแท้ของข้อมูลอย่างเป็นอิสระกับวิธีการจำแนกประเภท ซึ่งวิธีฟิลเตอร์จะคำนวณค่าความสำคัญของตัวแปร หรือเซตย่อยของตัวแปรจากดัชนีวัดความสำคัญ และเลือกตัวแปรหรือเซตย่อยของตัวแปรที่ให้ค่าดัชนีสูง

วิธีแรปเปอร์อาศัยวิธีการจำแนกประเภท ในการวัดความสำคัญของเซตย่อยของตัวแปร โดยเลือกเซตย่อยที่มีความแม่นยำในการจำแนกประเภทข้อมูลสูง หรือใช้ความแม่นยำในการจำแนกประเภทข้อมูลเมื่อใช้เซตย่อยนั้น ๆ ในการเป็นดัชนีวัดความสำคัญของเซตย่อย ซึ่งทำให้ได้เซตย่อยที่มีความแม่นยำในการจำแนกประเภทสูงกว่าการใช้ดัชนีวัดความสำคัญอื่น ๆ แต่เนื่องจากในการประเมินความเกี่ยวข้องของเซตย่อยแต่ละครั้งต้องเข้าสู่กระบวนการของวิธีการจำแนกประเภททำให้การคัดเลือกตัวแปรด้วยวิธีแรปเปอร์นั้นใช้เวลาในการประมวลผลค่อนข้างมากถ้าข้อมูลมีจำนวนตัวแปรมาก (Saeys et al., 2007; Yu & Liu, 2004)

วิธีฝังตัวเป็นวิธีที่การคัดเลือกตัวแปรเป็นส่วนหนึ่งในกระบวนการจำแนกประเภทด้วย โดยทำการเลือกเซตย่อยที่มีตัวแปรที่เหมาะสมในแต่ละขั้นตอนวิธี พร้อมไปกับการสร้างตัวแบบสำหรับ

การจำแนกประเภท ซึ่งวิธีฝังตัวมีลักษณะคล้ายกับวิธีแรปเปอร์ที่ว่ามีการผูกติดกับวิธีการจำแนกประเภทที่เฉพาะเจาะจง แต่ใช้เวลาในการประมวลผลน้อยกว่าวิธีแรปเปอร์ (Janecek, 2009, pp. 41-43; Saeys et al., 2007)

วิธีแรปเปอร์เป็นวิธีการคัดเลือกตัวแปรที่มีความซับซ้อนในการคำนวณสูงสุด ตามมาด้วยวิธีฝังตัว ทั้งสองวิธีดังกล่าวมีแนวทางในการคัดเลือกเซตย่อยของตัวแปรบนพื้นฐานของวิธีการจำแนกประเภทที่เฉพาะเจาะจง ซึ่งมีแนวโน้มที่จะได้ตัวแบบที่สามารถนำมาใช้ในการเรียนรู้ได้ดี แต่นำไปใช้ทำนายข้อมูลอื่นได้ไม่ดี เนื่องจากทำให้เกิดปัญหา Overfitting โดยแนวโน้มการเกิดปัญหาดังกล่าวมีมากกว่าวิธีฟิลเตอร์ซึ่งเป็นอิสระจากวิธีการจำแนกประเภท นอกจากนี้ถึงแม้ว่าวิธีแรปเปอร์ และวิธีฝังตัวจะเป็นวิธีการคัดเลือกตัวแปรที่มีความแม่นยำในการจำแนกประเภทสำหรับปัญหาที่เฉพาะเจาะจง (Janecek, 2009, pp. 41-43) แต่สำหรับข้อมูลที่มีจำนวนมิติมากนั้น วิธีฟิลเตอร์มักเป็นวิธีที่ถูกเลือกใช้ในการคัดเลือกตัวแปร เนื่องจากประมวลผลได้เร็วกว่าทั้งสองวิธี

วิธีฟิลเตอร์

วิธีฟิลเตอร์เป็นวิธีการคัดเลือกตัวแปรที่เร็วและง่ายต่อการตีความ โดยจะกำจัดตัวแปรที่ไม่เกี่ยวข้องต่อการจำแนกประเภทข้อมูล ด้วยคุณสมบัติในเนื้อแท้ของข้อมูล ซึ่งกระบวนการเป็นอิสระจากวิธีการจำแนกประเภท

วิธีการคัดเลือกตัวแปรโดยวิธีฟิลเตอร์สามารถแบ่งได้เป็น 2 ประเภท ได้แก่ วิธีฟิลเตอร์แบบตัวแปรเดียว (Univariate Filter Method) และวิธีฟิลเตอร์แบบหลายตัวแปร (Multivariate Filter Method) โดยวิธีฟิลเตอร์แบบตัวแปรเดียวจะพิจารณาความเกี่ยวข้องของตัวแปรทำนายต่อตัวแปรกลุ่ม (ตัวแปรตาม) โดยพิจารณาแต่ละตัวแปรทำนายแยกกัน ในขณะที่วิธีฟิลเตอร์แบบหลายตัวแปรจะรวมความสัมพันธ์ระหว่างตัวแปรทำนายด้วยกันให้มีผลต่อการพิจารณาคัดเลือกตัวแปร โดยส่วนใหญ่จะพิจารณาเซตย่อยของตัวแปร เพื่อที่จะนำความสัมพันธ์ระหว่างตัวแปรเข้ามาพิจารณาด้วย ซึ่งทำให้มีโอกาสได้เซตย่อยของตัวแปรที่เหมาะสมมากกว่า แต่อย่างไรก็ดีกระบวนการค้นหาเซตย่อยจะทำให้ใช้เวลาในการประมวลผลมาก และลดความสามารถในการจัดการกับข้อมูลที่มีขนาดใหญ่ Guyon and Elisseeff (2003) ได้กล่าวถึงการคัดเลือกตัวแปรที่มีลักษณะเดียวกับวิธีฟิลเตอร์แบบตัวแปรเดียวโดยเรียกว่าเป็นตัวจำแนกตัวแปรเชิงเดี่ยว (Single Variable Classifiers) ส่วนวิธีฟิลเตอร์แบบหลายตัวแปรซึ่งถูกรวมเข้ากับวิธีแรปเปอร์ และวิธีฝังตัวจะเรียกเป็นการคัดเลือกเซตย่อยของตัวแปร

วิธีฟิลเตอร์แบบตัวแปรเดียว

ในบางครั้งอาจเรียกวิธีนี้ว่าการเรียงลำดับตัวแปร (Feature Ranking) การถ่วงน้ำหนักตัวแปร (Feature Weighting) (Yu & Liu, 2003) หรือการประเมินตัวแปรเดียว (Individual Evaluation) (Yu & Liu, 2004) วิธีฟิลเตอร์แบบตัวแปรเดียวจึงเป็นวิธีการคัดเลือกตัวแปรโดย

ประเมินผลทีละตัวแปรแยกจากกัน จากนั้นเรียงลำดับความสำคัญ หรือความเกี่ยวข้องของตัวแปร โดยอาศัยดัชนีวัดความสำคัญในการเรียงลำดับตัวแปร ได้แก่ มาตรวัดระยะทาง มาตรวัดสารสนเทศ และมาตรวัดความไม่เป็นอิสระ ตัวแปรที่มีค่าดัชนีสูงสุด *m* ตัวแปร จะถูกคัดเลือก หรือเลือกตัวแปรที่มีค่าสูงกว่าเกณฑ์จุดตัด (Cutoff Threshold Criterion) *t* โดยที่ *m* และ *t* เป็นเกณฑ์ที่ผู้ใช้เป็นผู้กำหนด วิธีนี้เป็นวิธีที่ทำได้ง่าย และประมวลผลได้รวดเร็ว แต่มีข้อเสียในเรื่องการละเลยความสัมพันธ์ระหว่างตัวแปรทำนายด้วยกัน ตัวแปรที่ถูกเลือกถึงแม้เป็นตัวแปรที่มีความเกี่ยวข้องกับการจำแนกประเภท แต่ในบรรดาตัวแปรเหล่านั้นบางตัวแปรอาจไม่ได้เพิ่มสารสนเทศที่ช่วยในการจำแนกประเภท หรือเรียกว่าเป็นตัวแปรที่ซ้ำซ้อน (Redundant Variable) และนอกจากนี้การคัดเลือกตัวแปรโดยวิธีนี้อาจละเลยตัวแปรที่ไม่เกี่ยวข้องกับการจำแนกประเภทเมื่อพิจารณาเดี่ยว ๆ แต่จะมีความเกี่ยวข้องเมื่อพิจารณาร่วมกับตัวแปรอื่น

วิธีฟิลเตอร์แบบหลายตัวแปร

วิธีนี้เป็นวิธีการคัดเลือกตัวแปรที่ได้พิจารณาความสัมพันธ์ระหว่างตัวแปรทำนายด้วยกันในระดับหนึ่งด้วย ซึ่งจะทำให้ได้ตัวแปรที่ทำนายผลได้แม่นยำขึ้น ช่วยแก้ปัญหาที่เป็นข้อเสียของวิธีฟิลเตอร์แบบตัวแปรเดียว

ตอนที่ 2 วิธีกำลังสองน้อยที่สุดบางส่วน

วิธีกำลังสองน้อยที่สุดบางส่วน (Partial Least Squares Method: PLS) เป็นวิธีการวิเคราะห์ข้อมูลแบบหลายตัวแปร ในการสร้างตัวแบบความสัมพันธ์ระหว่างกลุ่มของตัวแปรสังเกตได้ (Observed Variable) โดยอาศัยตัวแปรแฝง (Latent Variable) ซึ่งเริ่มแรกอัลกอริทึมที่ใช้ใน PLS ที่มีชื่อว่า NIPALS (Non-Linear Iterative Partial Least Squares) ได้รับการพัฒนาโดย Herman Wold ใน ค.ศ.1966 (Russolillo, 2009, pp. 1-4) และ ใน ค.ศ. 1975 เขาใช้อัลกอริทึมในการจัดการกับปัญหาการสร้างตัวแบบเส้นทาง (Path Modeling) ในทางเศรษฐศาสตร์ ในช่วงปี ค.ศ. 1980-1989 Svante Wold บุตรชายของเขาและเพื่อนได้พัฒนา PLS ให้สามารถจัดการกับปัญหาการวิเคราะห์การถดถอย (Regression Analysis) ในการสร้างตัวแบบเคโมเมตริกซ์ (Chemometrics) และสเปกโทรเมตริก (Spectrometric) (Boulesteix & Strimmer, 2006) และเรียกวิธีการนี้ว่า PLS-Regression หรือ PLS-R

PLS เป็นวิธีการวิเคราะห์ข้อมูลที่ได้รับความสนใจจากนักสถิติและนักวิจัยเป็นอย่างมากใน 20 ปีที่ผ่านมา (Boulesteix & Strimmer, 2006; Mehmood, Liland, Snipen, & Sæbø, 2012) เนื่องจาก PLS สามารถจัดการกับข้อมูลที่มีจำนวนมิติมากในขณะที่มีขนาดตัวอย่างน้อยได้เป็นอย่างดี นอกจากนี้ยังไม่มีข้อสมมุติเบื้องต้นเกี่ยวกับโครงสร้างของข้อมูลซึ่งทำให้ PLS ได้รับการเรียกว่าเป็น

Soft Modelling (Manne, 1987) คือเป็นวิธีที่ยืดหยุ่นและสามารถแก้ปัญหาได้หลายด้านในการวิเคราะห์ข้อมูล PLS จึงสามารถนำไปใช้ในการวิเคราะห์ข้อมูลได้ในหลายลักษณะ

การลดมิติข้อมูลของวิธีกำลังสองน้อยที่สุดบางส่วน

PLS สามารถนำไปใช้ในการลดมิติของข้อมูล (Dimension Reduction) และเรียกว่าการลดมิติข้อมูลของวิธีกำลังสองน้อยที่สุดบางส่วน (Partial Least Squares Dimension Reduction: PLS-DR) โดยมีหลักการสร้างตัวแปรแฝงซึ่งมีจำนวนน้อยกว่าจำนวนตัวแปรเดิมจากผลรวมเชิงเส้นของตัวแปรเดิม และสามารถนำตัวแปรแฝงที่สร้างขึ้นนี้ไปเป็นตัวแปรสำหรับจำแนกประเภทข้อมูล ด้วยวิธีการจำแนกประเภทวิธีต่าง ๆ เช่น การจำแนกโลจิสติก (Logistic Discrimination: LD) การวิเคราะห์การจำแนกประเภทเชิงเส้น (Linear Discriminant Analysis: LDA) การวิเคราะห์การจำแนกประเภทกำลังสอง (Quadratic Discriminant Analysis: QDA) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เพื่อนบ้านใกล้ที่สุด K ตัว (K-Nearest Neighbor: KNN) หรือ ตัวจำแนกต้นไม้การตัดสินใจแบบ C4.5 (C4.5 Decision Tree Classifier: C4.5) เป็นต้น (Boulesteix, 2004; Nguyen & Rocke, 2002a; Nguyen & Rocke, 2002b; Zeng, Li, Wu, Yang, & Yang, 2008) PLS-DR มีลักษณะคล้ายกับวิธีการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) ในแง่ของการลดมิติของข้อมูล แต่ PLS ทำงานได้อย่างมีประสิทธิภาพกว่า PCA สำหรับการนำไปใช้ในงานการพยากรณ์ หรือการจำแนกประเภท เนื่องจากการสร้างตัวแปรแฝงของ PCA ได้จากผลรวมเชิงเส้นของตัวแปรเดิมที่ทำให้ความแปรปรวนสูงสุด แต่ PLS สร้างตัวแปรแฝงภายนอก (ตัวแปรแฝงที่ทำหน้าที่เป็นตัวแปรทำนาย) และตัวแปรแฝงภายใน (ตัวแปรแฝงที่ทำหน้าที่เป็นตัวแปรตาม) จากผลรวมเชิงเส้นของตัวแปรเดิมที่ทำให้ความแปรปรวนร่วมระหว่างตัวแปรแฝงทั้งสองนี้มีค่าสูงสุด นอกจากนี้แล้ว PCA ยังใช้เวลาในการประมวลผลมากกว่า PLS (Hilario & Kalousis, 2008) ดังนั้นในระยะหลัง PLS จึงมักถูกนำมาเปรียบเทียบกับ PCA และพบว่าเป็นเทคนิคที่มีประสิทธิภาพมากกว่า (Dai, Lieu, & Rocke, 2006; Nguyen & Rocke, 2002a)

การถดถอยของวิธีกำลังสองน้อยที่สุดบางส่วน

PLS-R (Partial Least Squares Regression) เป็นเทคนิคการวิเคราะห์การถดถอยเชิงสถิติที่มีประสิทธิภาพสามารถรับมือกับข้อมูลที่มีจำนวนมิติมาก (มากกว่าจำนวนตัวอย่าง) ตัวแปรบกวนจำนวนมาก ตัวแปรที่มีความสัมพันธ์กันเอง (Collinearity) และมีค่าสูญหายในชุดข้อมูล นอกจากนี้ยังไม่มีข้อสมมุติเบื้องต้นเกี่ยวกับการแจกแจงของความคลาดเคลื่อน ดังนั้น PLS-R จึงเหมาะสมที่จะใช้ในการวิเคราะห์ข้อมูลไมโครอาร์เรย์ (Johansson et al., 2003) Yeniay and Göktaş (2002) ได้เปรียบเทียบ PLS-R และวิธีการวิเคราะห์การถดถอยอื่นที่เป็นที่นิยมได้แก่ การวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยที่สุดสามัญ (Ordinary Least Square Method)

การถดถอยองค์ประกอบหลัก (Principal Component Regression) และการถดถอยแบบริดจ์ (Ridge Regression) โดยในงานวิจัยได้เสนอในมุมมองของทฤษฎี และได้นำไปประยุกต์กับข้อมูลจริง ซึ่งพบว่า PLS-R ให้ผลลัพธ์ที่ค่อนข้างดีกว่าในแง่ของความสามารถเชิงทำนายของตัวแบบ PLS-R มีความสามารถในการวิเคราะห์การถดถอยในกรณีที่มีตัวแปรตามมากกว่าหนึ่งตัว โดยสร้างเป็นตัวแบบเดียว ซึ่งทำให้เห็นภาพรวมได้ชัดเจนกว่าการแยกวิเคราะห์ทีละตัวแปรตาม แต่ถ้ากลุ่มตัวแปรตามนั้นวัดสิ่งที่ไม่เกี่ยวข้องกันซึ่งทำให้ตัวแปรตามไม่สัมพันธ์กัน การสร้างตัวแบบเดียวโดย PLS-R มีแนวโน้มที่จะได้องค์ประกอบจำนวนมากซึ่งทำให้ยากในการตีความ การสร้างตัวแบบแยกกันในแต่ละตัวแปรตามจะเหมาะสมกว่าเนื่องจากง่ายต่อการตีความ (Wold, Sjöström, & Eriksson, 2001)

PLS-R ใช้สำหรับการวิเคราะห์การถดถอย ซึ่งตัวแปรตามเป็นข้อมูลเชิงปริมาณ แต่สามารถประยุกต์กับงานการจำแนกประเภทข้อมูล ซึ่งตัวแปรตามเป็นข้อมูลเชิงคุณภาพได้ โดยมีขั้นตอนที่เพิ่มขึ้น 2 ส่วนได้แก่ ส่วนของการแปลงค่าเมทริกซ์ของตัวแปรตามให้เป็นเมทริกซ์ของตัวแปรดัมมี่ และการทำนายกลุ่มจากผลลัพธ์ที่ได้ ซึ่งจะกล่าวถึงในตอนที่ 3 หัวข้อการวิเคราะห์การจำแนกประเภทของวิธีกำลังสองน้อยที่สุดบางส่วน (Partial Least Squares Discriminant Analysis: PLS-DA)

ขั้นตอนการสร้างตัวแบบสำหรับ PLS-R

PLS-R เป็นวิธีการวิเคราะห์หลายตัวแปร โดยสร้างตัวแบบความสัมพันธ์เชิงเส้นในเชิงพยากรณ์ระหว่างกลุ่มตัวแปรสองกลุ่ม (กลุ่มตัวแปรตาม และกลุ่มตัวแปรทำนาย) โดยสร้างตัวแปรใหม่ซึ่งเป็นผลรวมเชิงเส้นของตัวแปรเดิม และเรียกตัวแปรใหม่นี้ว่าเป็นตัวแปรแฝงภายนอก หรือ X-scores หรือองค์ประกอบสำหรับตัวแปรทำนาย และเรียกว่าตัวแปรแฝงภายใน หรือ Y-scores หรือองค์ประกอบสำหรับตัวแปรตาม โดยองค์ประกอบที่สร้างขึ้นใหม่นี้มีจำนวนน้อยกว่าจำนวนตัวแปรเดิม และตั้งฉากกัน (Othogonal) แนวทางของ PLS-R มีหลายแนวทางซึ่งแตกต่างกันโดยขึ้นอยู่กับฟังก์ชันเป้าหมายสำหรับการหาค่าน้ำหนัก $\mathbf{W}$ อัลกอริทึมที่ใช้ในการคำนวณค่า $\mathbf{W}$ เป็นต้น (Boulesteix & Strimmer, 2006) PLS-R สำหรับกรณีที่มีตัวแปรตามหนึ่งตัวเรียก PLS1 ส่วนกรณีที่มีตัวแปรตามหลายตัวที่เป็นที่นิยมได้แก่ PLS2 และอัลกอริทึมที่เป็นที่รู้จักคือ วิธีกำลังสองน้อยที่สุดบางส่วนแบบทำซ้ำไม่เชิงเส้น (Nonlinear Iterative Partial Least Squares: NIPALS) และ วิธีกำลังสองน้อยที่สุดบางส่วนแบบเคอร์เนล (Kernel-Partial Least Squares: Kernel-PLS)

กำหนดให้ $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)'$ แทนเมทริกซ์ของข้อมูลขนาด $n\times p$ โดยที่ n แทนขนาดตัวอย่าง และ p แทนจำนวนตัวแปรทำนาย และ $\mathbf{Y}=(y_1,y_2,...,y_n)'$ แทนเมทริกซ์ของข้อมูลขนาด $n\times q$ โดยที่ q แทนจำนวนตัวแปรตาม และเวกเตอร์ $\mathbf{x}_i=(x_{i1},x_{i2},...,x_{ip})$ เป็นเวกเตอร์ที่แสดงค่าของตัวแปรทำนายของหน่วยตัวอย่างที่ i และเวกเตอร์ $\mathbf{y}_i=(y_{i1},y_{i2},...,y_{iq})$ แทน

เวกเตอร์ของตัวแปรตามของหน่วยตัวอย่างที่ i ผลรวมเชิงเส้นของตัวแปรทำนาย และตัวแปรตาม สำหรับองค์ประกอบที่ a (a=1,2,...,h) เขียนแทนด้วย $\mathbf{t}_a$ และ $\mathbf{u}_a$ ตามลำดับซึ่ง

$$\mathbf{T} = \mathbf{X}\mathbf{W}^* \tag{2.1}$$

$$\mathbf{U} = \mathbf{Y}\mathbf{C} \tag{2.2}$$

โดยที่ $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_h)$ และ $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2 ..., \mathbf{u}_h)$ เป็นเมทริกซ์ขนาด $n \times h$ ของเวกเตอร์องค์ประกอบจำนวน h องค์ประกอบที่ถูกสกัดจากตัวแปรทำนาย และตัวแปรตาม ตามลำดับ $\mathbf{W}^* = (\mathbf{w}_1^*, \mathbf{w}_2^*, ..., \mathbf{w}_h^*)$ และ $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_h)$ เป็นเมทริกซ์ขนาด $p \times h$ และ $q \times h$ ตามลำดับ ซึ่งสมาชิกประกอบด้วยเวกเตอร์น้ำหนัก (Weight Vector) ขององค์ประกอบที่ถูกสกัดจากตัวแปรทำนาย และตัวแปรตาม ตามลำดับ

PLS ได้แยกองค์ประกอบของ $\mathbf{X}$ และ $\mathbf{Y}$ ให้อยู่ในรูปแบบดังนี้

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \tag{2.3}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} \tag{2.4}$$

โดยที่ $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_h)$ และ $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_h)$ เป็นเมทริกซ์ขนาด $p \times h$ แทนค่าการให้น้ำหนัก (loading) สำหรับตัวแปรทำนาย และสำหรับตัวแปรตาม ตามลำดับ ซึ่งโดยทั่วไปแล้ว $\mathbf{P}$ และ $\mathbf{Q}$ ถูกประมาณค่าด้วยวิธีกำลังสองน้อยที่สุดสามัญ $\mathbf{E}$ เป็นเมทริกซ์ขนาด $n \times p$ และ $\mathbf{F}$ เป็นเมทริกซ์ขนาด $n \times q$ ซึ่งทั้งสองเป็นเมทริกซ์แทนค่าตกค้าง (Residual) สำหรับตัวแปรทำนาย และตัวแปรตาม ตามลำดับ

ความสัมพันธ์ทั้งสองถูกเชื่อมต่อด้วยความสัมพันธ์ภายใน (Inner Relationship) ดังนี้

$$\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{H} \tag{2.5}$$

โดยที่ $\mathbf{D}$ แทน เมทริกซ์ทแยงมุมขนาด $h \times h$

$\mathbf{H}$ แทน เมทริกซ์ค่าตกค้างขนาด $n \times h$

PLS โดยอัลกอริทึม NIPALS จะหาค่าพารามิเตอร์ของตัวแบบทีละองค์ประกอบ ซึ่งในองค์ประกอบที่ a ค่า $\mathbf{u}_a$, $\mathbf{w}_a$, $\mathbf{t}_a$ และ $\mathbf{c}_a$ จะได้รับการคำนวณตามลำดับ โดยในองค์ประกอบที่ 1 จะเริ่มจากสุ่มเวกเตอร์ $\mathbf{u}_1$ จากปริภูมิของ $\mathbf{Y}$ แต่ถ้าตัวแปรตามมีเพียงหนึ่งตัวแล้ว $\mathbf{u}_1 = \mathbf{y}$ จากนั้นจะทำการหาค่าน้ำหนัก $\mathbf{w}_1$ ที่ทำให้ $\max\left[\text{cov}(\mathbf{X}\mathbf{w}_1, \mathbf{Y}\mathbf{c}_1)\right]^2$ โดยที่ $\mathbf{w}_1'\mathbf{w}_1 = 1$, $\mathbf{c}_1'\mathbf{c}_1 = 1$ และ $\text{cov}(\mathbf{X}\mathbf{w}_1, \mathbf{Y}\mathbf{c}_1)$ แทนความแปรปรวนร่วมของตัวอย่างระหว่างองค์ประกอบ $\mathbf{X}\mathbf{w}_1$ และ $\mathbf{Y}\mathbf{c}_1$ องค์ประกอบของตัวแปรทำนายหาได้จากความสัมพันธ์ $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ และสุดท้ายเป็นการหาค่าน้ำหนัก $\mathbf{c}_1$ ซึ่งเป็นฟังก์ชันของ $\mathbf{t}_1$ ขั้นตอนเหล่านี้จะถูกทำซ้ำจนกระทั่ง $\mathbf{t}_1$ ลู่เข้า (Convergence) แต่หากตัวแปรตามมีเพียงหนึ่งตัวแล้วขั้นตอนเหล่านี้จะทำเพียงรอบเดียว

เมื่อค่า $\mathbf{t}_1$ ในองค์ประกอบที่ 1 ลู่เข้าแล้ว จะลดค่า (Deflate) $\mathbf{X}$ และ $\mathbf{Y}$ ลง โดยลบด้วยค่าประมาณของ $\mathbf{X}$ และ $\mathbf{Y}$ ตามลำดับ นั่นคือเมทริกซ์ $\mathbf{X}$ และ $\mathbf{Y}$ ลดค่าลงเป็นค่าตกค้าง $\mathbf{E}$ และ $\mathbf{F}$

ตามลำดับ รูปแบบการลดค่านี้มีหลายรูปแบบแต่ที่ใช้บ่อยจะเป็น PLS1 และ PLS2 (Rosipal, & Krämer, 2006) สำหรับตัวแปรตามหนึ่งตัว และตัวแปรตามสองตัว ตามลำดับ แนวทางการลดค่าลงนี้ทำให้เมทริกซ์องค์ประกอบ $\mathbf{T}$ เป็นเมทริกซ์ตั้งฉาก $\mathbf{E}$ และ $\mathbf{F}$ ซึ่งจะได้นำไปใช้ในการหาค่าพารามิเตอร์ขององค์ประกอบที่ 2 โดยใช้แทน $\mathbf{X}$ และ $\mathbf{Y}$ และดำเนินการในลักษณะเดียวกับการหาพารามิเตอร์ขององค์ประกอบที่ 1 และทำซ้ำในลักษณะเดียวกันนี้สำหรับองค์ประกอบอื่น ๆ

ขั้นตอนการสร้างตัวแบบ PLS-R (Höskuldsson, 1988) แสดงรายละเอียดได้ดังภาพที่ 4 ซึ่งก่อนที่จะเริ่มเข้าสู่ขั้นตอนจะต้องเตรียมข้อมูล โดยการทำให้ผลรวมของแต่ละตัวแปรเป็นศูนย์ ซึ่งทำได้โดยการลบค่าของแต่ละตัวแปรด้วยค่าเฉลี่ยของตัวแปรนั้น ๆ (Mean-Centering) หรือปรับขนาดของข้อมูล (Scaling) ในแต่ละตัวแปรให้มีความแปรปรวน 1 หน่วย ซึ่งการปรับขนาด $\mathbf{X}$ อาจไม่มีความจำเป็น แต่อย่างไรก็ตามองค์ประกอบที่หนึ่งของตัวแปรทำนายจะมีคุณสมบัติที่น่าสนใจต่อการคัดเลือกตัวแปร ถ้า $\mathbf{X}$ ถูกปรับขนาด (Boulesteix, 2004) การปรับขนาดให้แต่ละตัวแปรมีความแปรปรวน 1 หน่วย ทำได้โดยการหารแต่ละตัวแปรด้วยส่วนเบี่ยงเบนมาตรฐานของตัวแปรนั้น ๆ

ในกรณีเฉพาะที่จำนวนตัวแปรตามมีเพียงหนึ่งตัว อัลกอริทึม PLS2 มีลักษณะเดียวกับ PLS1 เนื่องจาก $\mathbf{c}_{p\times 1}$ ในขั้นตอนที่ 2.5 มีค่าเท่ากับเวกเตอร์ $\mathbf{1}_{p\times 1}$ และ $\mathbf{u}=\mathbf{y}$ ในขั้นตอนที่ 1 และขั้นตอนที่ 2.6

จากขั้นตอนการสร้างตัวแบบ PLS-R แต่ละองค์ประกอบ $\mathbf{t}_a$ หาได้จากผลคูณของค่าตกค้างกับค่าน้ำหนัก $\mathbf{w}_a$ ความสัมพันธ์ระหว่างค่าน้ำหนัก $\mathbf{w}_a$ และ $\mathbf{w}_a^*$ ซึ่งค่าหลังเป็นค่าน้ำหนักที่เกี่ยวข้องกับ $\mathbf{X}$ โดยตรง

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \tag{2.6}$$

1. กำหนดค่าเริ่มต้นของ $\mathbf{u}_1$ ซึ่งโดยมากกำหนดให้เป็นคอลัมน์ใดคอลัมน์หนึ่งของ $\mathbf{Y}$
2. สำหรับ a=1,..., h ทำซ้ำ

2.1 $\mathbf{w}_a = \mathbf{X}'\mathbf{u}_a / \mathbf{u}_a'\mathbf{u}_a$

2.2 $\mathbf{w}_a = \mathbf{w}_a / \|\mathbf{w}\|$ (ปรับให้ $\mathbf{w}$ มีขนาดหนึ่งหน่วย)

2.3 $\mathbf{t}_a = \mathbf{X}\mathbf{w}_a$

2.4 $\mathbf{c}_a = \mathbf{Y}'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a$

2.5 $\mathbf{c}_a = \mathbf{c}_a / \|\mathbf{c}\|$ (ปรับให้ $\mathbf{c}_a$ มีขนาดหนึ่งหน่วย)

2.6 $\mathbf{u}_a = \mathbf{Y}\mathbf{c}_a / \mathbf{c}_a'\mathbf{c}_a$

ทำซ้ำขั้นตอนที่ 2.1-2.6 จนกระทั่งเวกเตอร์ $\mathbf{t}_a$ ลู่เข้า มิฉะนั้นไปขั้นตอนที่ 2.7 ถ้ามีตัวแปรตามเพียงตัวเดียวไม่ต้องทำซ้ำให้ไปขั้นตอนที่ 2.7

2.7 $\mathbf{p}_a = \mathbf{X}'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a$

2.8 $\mathbf{q}_a = \mathbf{Y}'\mathbf{u}_a / \mathbf{u}_a'\mathbf{u}_a$

2.9 $d_a = \mathbf{u}_a'\mathbf{t}_a / \mathbf{t}_a'\mathbf{t}_a$

2.10 $\mathbf{X} = \mathbf{X} - \mathbf{t}_a\mathbf{p}_a'$

2.11 $\mathbf{Y} = \mathbf{Y} - d_a\mathbf{t}_a\mathbf{c}_a'$

ภาพที่ 4 ขั้นตอนการสร้างตัวแบบของ PLS-R

การกำหนดจำนวนองค์ประกอบ

แม้ว่าอัลกอริทึมของ PLS จะสามารถสกัดองค์ประกอบได้เท่ากับจำนวนลำดับที่ (Rank) ของ $\mathbf{X}$ ก็ตาม แต่องค์ประกอบที่สำคัญที่สกัดสารสนเทศจาก $\mathbf{X}$ และ $\mathbf{Y}$ มีเพียงองค์ประกอบแรก ๆ เท่านั้น ดังนั้นประเด็นที่สำคัญในการใช้ PLS คือการกำหนดจำนวนองค์ประกอบ ซึ่งมีวิธีในการกำหนดแตกต่างกันไปในแต่ละงานวิจัย โดยสรุปที่พบมีดังนี้

1. การกำหนดจำนวนองค์ประกอบที่แน่นอนเป็น 3-5 องค์ประกอบ (Li & Zeng, 2009) เช่น งานของ Nguyen and Rocke (2002a) และ Nguyen and Rocke (2002b) กำหนดจำนวนองค์ประกอบของยีนเป็น 3 องค์ประกอบ

2. การกำหนดจำนวนองค์ประกอบด้วยการเปรียบเทียบความสอดคล้องของตัวแบบเมื่อมีองค์ประกอบจำนวน a บนข้อมูลทดสอบ โดยการแบ่งส่วนข้อมูลสามารถทำได้ดังนี้

2.1 การใช้กระบวนการการตรวจสอบไขว้ (Cross-Validation: CV) ในการประเมินว่าองค์ประกอบที่ a สามารถเพิ่มความสามารถในการทำนายของตัวแบบได้หรือไม่ กระบวนการการตรวจสอบไขว้เป็นการแบ่งส่วนข้อมูลเริ่มต้นเป็นส่วนๆ อย่างสุ่ม เช่น แบ่งข้อมูลออกเป็น K ส่วนที่

เท่ากันอย่างสุ่มซึ่งแต่ละส่วนคือข้อมูลย่อย โดยที่ K เป็นค่าคงที่ใด ๆ ข้อมูล K–1 ส่วนจะถูกนำมาใช้ในการสร้างตัวแบบเรียกว่าเป็นข้อมูลฝึกฝน (Training Data) และข้อมูลย่อยส่วนที่เหลือซึ่งเรียกว่าข้อมูลทดสอบ (Test Data) จะถูกนำมาตรวจสอบความสอดคล้องของตัวแบบที่สามารถวัดความสอดคล้องด้วยค่าต่าง ๆ ขึ้นกับวัตถุประสงค์ การตรวจสอบไขว้ที่แบ่งข้อมูลออกเป็น K ส่วน เรียกว่า K-fold Cross-Validation ผลรวมของค่าวัดความสอดคล้องจากทุกข้อมูลย่อยจะถูกนำมาเป็นค่าที่ใช้ประเมินความสอดคล้องของตัวแบบ การทำ 5-fold Cross-Validation แสดงดังภาพที่ 5

ภาพที่ 5 การทำ 5-fold Cross-Validation

ค่ารากที่สองของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองของการตรวจสอบไขว้ (Root Mean Squared Error of Cross-Validation: RMSECV) เป็นค่าที่ได้นำมาใช้ในการวัดความสอดคล้องของตัวแบบที่มีองค์ประกอบจำนวน a องค์ประกอบ โดยจำนวนองค์ประกอบที่ทำให้ค่า RMSECV ต่ำสุดจะถูกเลือกเพื่อใช้เป็นพารามิเตอร์สำหรับการสร้างตัวแบบโดย PLS งานวิจัยที่กำหนดจำนวนองค์ประกอบด้วยแนวทางนี้ เช่น งานของ Teófilo, Martins, and Ferreira (2008) Gosselin, Rodrigue, and Duchesne (2010) เป็นต้น การคำนวณค่า RMSECV เป็นดังนี้

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \qquad (2.7)$$

โดยที่ y_i แทน ค่าสังเกตที่ i ของตัวแปรตาม

$\hat{y}_i$ แทน ค่าพยากรณ์ที่ i ของตัวแปรตาม

n แทน จำนวนตัวอย่าง

การวัดความสอดคล้องของตัวแบบสำหรับงานการจำแนกประเภทข้อมูล วัดด้วยประสิทธิภาพของการจำแนกประเภท โดยเลือกจำนวนองค์ประกอบที่ทำให้ประสิทธิภาพของการจำแนกประเภทสูงสุด คือมีความแม่นยำในการจำแนกประเภทสูงสุด หรือค่าคลาดเคลื่อนในการจำแนกประเภทต่ำสุด งานวิจัยที่ทำในลักษณะนี้เช่น งานของ Dai and Others (2006) เป็นต้น

2.2 วิธียกออก (Holdout) เป็นการแบ่งข้อมูลออกเป็นสองส่วนได้แก่ ข้อมูลฝึกฝนและข้อมูลทดสอบด้วยอัตราส่วนที่กำหนด เช่น 2:1 จากนั้นวัดความสอดคล้องของตัวแบบที่ใช้องค์ประกอบจำนวน *a* องค์ประกอบ จากการทำนายค่าตัวแปรตามในส่วนทดสอบด้วยค่ารากที่สองของค่าเฉลี่ยของค่าคลาดเคลื่อนกำลังสองของการพยากรณ์ (Root Mean Squared Error of Prediction: RMSEP) สำหรับงานการพยากรณ์ ซึ่งคำนวณได้ในลักษณะเดียวกับ RMSECV เพียงแต่ใช้เฉพาะข้อมูลในส่วนทดสอบ หรือวัดค่าความสอดคล้องด้วยประสิทธิภาพของการจำแนกประเภทข้อมูล สำหรับงานการจำแนกประเภทการแบ่งข้อมูลด้วยวิธีนี้สามารถทำซ้ำได้หลายรอบเพื่อปรับปรุงค่าประมาณของค่าวัดความสอดคล้องให้มีความเที่ยง งานวิจัยที่ใช้ลักษณะนี้ เช่น งานของ Boulesteix (2004) เป็นต้น

3. การใช้เกณฑ์ในการประเมินว่าองค์ประกอบที่เพิ่มเข้ามาในตัวแบบสามารถลดค่าคลาดเคลื่อนของตัวแบบ PLS ได้อย่างมีนัยสำคัญหรือไม่ เช่น สถิติเอฟ เกณฑ์สารสนเทศของ Akaike (Akaike Information Criterion) เกณฑ์ค่า R ของ Wold (Wold's R Criterion) สัมประสิทธิ์การตัดสินใจเชิงพหุ (Multiple Coefficient of Determination: R^2) สัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) (Li, Morris, & Martin, 2002; Zeng, Li, & Wu, n.d.; Lazraq et al., 2003; Gutkin et al., 2009)

อย่างไรก็ดี Boulesteix (2004) กล่าวว่าไม่มีกระบวนการใดกระบวนการหนึ่งที่ได้รับการยอมรับอย่างกว้างขวางในการกำหนดจำนวนองค์ประกอบของ PLS ได้อย่างถูกต้อง

ดัชนี VIP

ดัชนี VIP (Variable Influence on the Projection) เป็นดัชนีวัดความสำคัญของตัวแปรที่มีต่อการทำนายตัวแปรตาม โดยได้รวมความสัมพันธ์ระหว่างตัวแปรทำนายด้วยกัน ซึ่งอาศัยพารามิเตอร์ของ PLS-R ในการคำนวณค่า เผยแพร่ครั้งแรกโดย Wold and Others ในปี ค.ศ. 1993 (Chong & Jun, 2005) โดยอาศัยค่าน้ำหนัก w_{aj} ซึ่งวัดการมีส่วนร่วมของตัวแปรทำนายในองค์ประกอบ และถ่วงน้ำหนักด้วยความแปรผันของตัวแปรตามที่ถูกอธิบายได้ด้วยแต่ละองค์ประกอบ VIP สำหรับตัวแปรทำนายที่ *j* คำนวณได้ดังนี้

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{a=1}^{h} Red(\mathbf{Y},\mathbf{t}_a)\left(w_{aj} / \|\mathbf{w}_a\|\right)^2}{(1/p)\sum_{a=1}^{h} Red(\mathbf{Y},\mathbf{t}_a)}} \qquad (2.8)$$

โดยที่ p แทน จำนวนตัวแปรทำนาย

q แทน จำนวนตัวแปรตาม

h แทน จำนวนองค์ประกอบ

w_{aj} แทน น้ำหนักของตัวแปรทำนายที่ j ในองค์ประกอบ a

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

$$Red(\mathbf{Y},\mathbf{t}_a) = \sum_{k=1}^{q} R^2(\mathbf{y}_k,\mathbf{t}_a)$$

ดัชนี VIP ให้น้ำหนักกับการมีส่วนร่วมของแต่ละตัวแปรสอดคล้องกับความแปรผันที่อธิบายได้ด้วยแต่ละองค์ประกอบ ค่า $w_{aj} / \|\mathbf{w}_a\|$ แสดงถึงขนาดความสำคัญของตัวแปรทำนายที่ j ในองค์ประกอบที่ a และ $Red(\mathbf{Y},\mathbf{t}_a)$ แสดงถึงขนาดความสำคัญขององค์ประกอบที่ a ต่อการทำนายตัวแปรตาม ตัวแปรที่มีค่า VIP สูงกว่าแสดงถึงความสำคัญที่มากกว่า และเนื่องจากค่าเฉลี่ยของกำลังสองของ VIP เท่ากับ 1.0 ดังนั้นมักใช้เกณฑ์มากกว่า 1.0 เป็นเกณฑ์จุดตัดสำหรับการคัดเลือกตัวแปร (Chen, 2008, pp. 14-15; Cho et al., 2002; Chong & Jun, 2005; Lazraq et al., 2003; Sun et al., 2012) และถ้าตัวแปรใดที่มีค่า VIP น้อยกว่า 0.8 ตัวแปรนั้นไม่มีความสำคัญหรือไม่มีความเกี่ยวข้องต่อการทำนายค่า (Chen, 2008, pp. 14-15) การใช้เกณฑ์มากกว่า 1 เป็นเกณฑ์จุดตัดมีความหมายสำหรับเฉพาะข้อมูลที่ถูกทำให้เป็นมาตรฐาน แต่อย่างไรก็ดี Gosselin and Others (2010) ได้ใช้เกณฑ์ดังกล่าวกับข้อมูลที่ถูกทำให้ผลรวมเป็นศูนย์ซึ่งยังคงให้ผลที่ดี

ตอนที่ 3 การจำแนกประเภท

ความหมายของการจำแนกประเภท

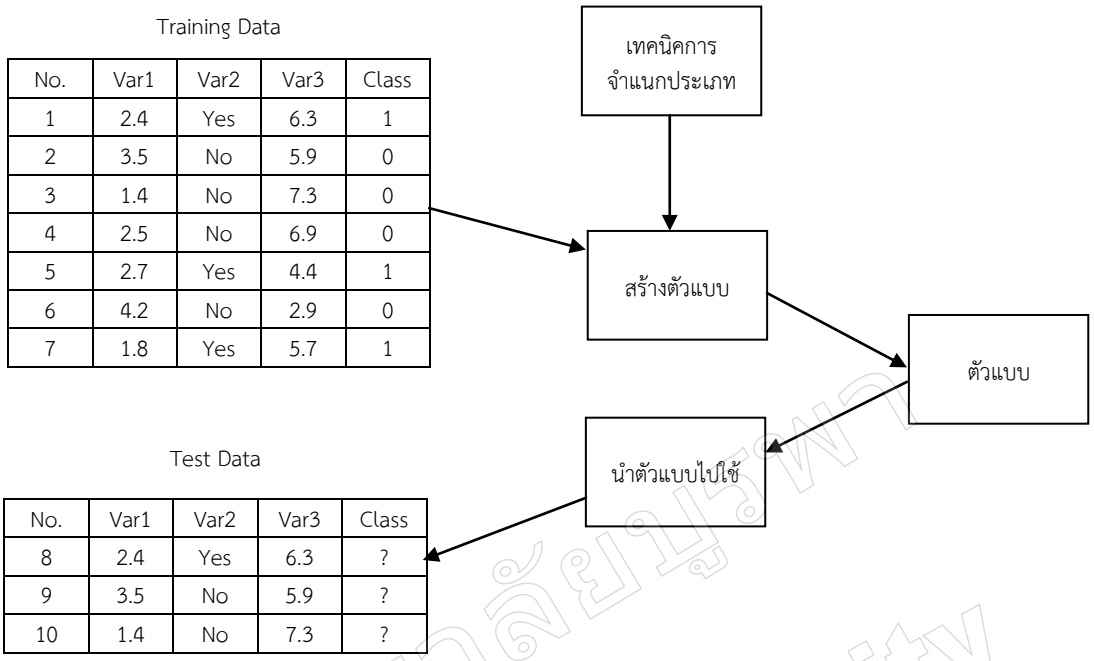
การจำแนกประเภท (Classification) เป็นการวิเคราะห์ข้อมูลสำหรับลักษณะของงานที่กำหนดกลุ่มให้กับหน่วยตัวอย่าง โดยอาศัยคุณลักษณะของหน่วยตัวอย่างซึ่งในที่นี้หมายถึงตัวแปรทำนาย เพื่อที่จะสามารถจำแนกกลุ่มได้จะต้องทราบล่วงหน้าว่ามีกลุ่มอะไรบ้าง (Predefined Categories) และมีข้อมูลของหน่วยตัวอย่างในแต่ละกลุ่มก่อน การจำแนกประเภทเป็นการทำเหมืองข้อมูลแบบที่นิยมใช้แพร่หลายซึ่งครอบคลุมการใช้งานที่หลากหลาย (Tan, Steinbach, & Kumar,

2006, p. 145) เช่น การทำนายการประสบความสำเร็จในการศึกษาระดับปริญญาตรีโดยอาศัยผลการทดสอบต่าง ๆ และผลการศึกษาในระดับชั้นมัธยมศึกษาตอนปลาย การคัดแยกอีเมลขยะจากอีเมลปกติโดยพิจารณาจากหัวเรื่องหรือเนื้อหาของอีเมล การจำแนกเซลที่เป็นหรือไม่เป็นอันตรายจากผลการตรวจด้วย MRI การจำแนกกาแล็กซีโดยอาศัยลักษณะรูปร่างของมัน การทำนายการเป็นมะเร็งโดยอาศัยข้อมูลระดับการแสดงออกของยีน เป็นต้น

แนวทางการดำเนินงานการจำแนกประเภท

เทคนิคการจำแนกประเภทเป็นวิธีการสร้างตัวแบบการจำแนกประเภท (Classification Model) อย่างมีระบบจากข้อมูลนำเข้า (Input Data) ตัวอย่างเทคนิคการจำแนกประเภทได้แก่ ตัวจำแนกต้นไม้การตัดสินใจ (Decision Tree Classifier) ตัวจำแนกด้วยกฎ (Rule-based Classifier) ข่ายงานประสาท (Neural Network) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) และตัวจำแนกนาอีฟเบส์ (Naïve Bayes Classifier: NB) เป็นต้น แต่ละเทคนิคใช้อัลกอริทึมการเรียนรู้เพื่อสร้างตัวแบบที่สอดคล้องกับข้อมูลที่สุด โดยแสดงความสัมพันธ์ระหว่างตัวแปรทำนาย และกลุ่มของข้อมูลนำเข้า ตัวแบบที่สร้างขึ้นนี้นอกจากที่จะมีความสอดคล้องกับข้อมูลนำเข้าแล้ว ยังควรมีความสามารถในการจำแนกประเภทข้อมูลใหม่ที่ไม่ได้ใช้ในการสร้างตัวแบบได้อย่างถูกต้องอีกด้วย

แนวทางการแก้ปัญหางานการจำแนกประเภทเริ่มจากการแบ่งข้อมูลเป็นสองส่วน ได้แก่ ข้อมูลฝึกฝน และข้อมูลทดสอบ แต่ละชุดข้อมูลมีข้อมูลของตัวแปรทำนาย และกลุ่มที่แต่ละหน่วยตัวอย่างเป็นสมาชิก ข้อมูลฝึกฝนจะถูกนำไปใช้ในการสร้างตัวแบบการจำแนกประเภท และตัวแบบนี้จะถูกนำไปใช้จำแนกประเภทข้อมูลของหน่วยตัวอย่างในข้อมูลทดสอบ ตัวแบบที่ได้รับการทดสอบว่ามีความถูกต้องสูงก็จะถูกนำไปใช้ทำนายข้อมูลใหม่ แนวทางการสร้างตัวแบบการจำแนกประเภทซึ่งดัดแปลงจาก Tan and others (2006, p. 148) แสดงดังภาพที่ 6



ภาพที่ 6 แนวทางการสร้างตัวแบบการจำแนกประเภท

การประเมินประสิทธิภาพของตัวแบบการจำแนกประเภท

ค่าวัดสำหรับการประเมินประสิทธิภาพ

การประเมินประสิทธิภาพของตัวแบบการจำแนกประเภท อาศัยผลการทำนายการเป็นสมาชิกกลุ่มข้อมูล เมื่อใช้ตัวแบบดังกล่าวกับข้อมูลทดสอบ โดยนับจำนวนตัวอย่างที่ได้รับการทำนายที่ถูกต้อง และไม่ถูกต้อง ค่าจำนวนตัวอย่างเหล่านี้ถูกนำมาสร้างเป็นตารางแสดงผลลัพธ์การจำแนกประเภทที่เรียกว่า เมทริกซ์ Confusion (Confusion Matrix) แสดงดังตารางที่ 3 ซึ่งใช้สำหรับกรณีปัญหาการจำแนกประเภทที่แบ่งข้อมูลเป็นสองกลุ่ม ตัวเลขที่แสดงในตารางเป็นจำนวนตัวอย่างในแต่ละสถานการณ์ ดังนี้

a_1 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่ม 1 และได้รับการทำนายว่าอยู่ในกลุ่ม 1

a_2 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่ม 1 และได้รับการทำนายว่าอยู่ในกลุ่ม 2

a_3 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่ม 2 และได้รับการทำนายว่าอยู่ในกลุ่ม 1

a_4 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่ม 2 และได้รับการทำนายว่าอยู่ในกลุ่ม 2

จำนวนหน่วยตัวอย่างที่ได้รับการทำนายได้อย่างถูกต้องมีจำนวน $a_1 + a_4$ ตัวอย่าง และจำนวนหน่วยตัวอย่างที่ได้รับการทำนายไม่ถูกต้องมีจำนวน $a_2 + a_3$ ตัวอย่าง

ตารางที่ 3 ผลการจำแนกประเภทของหน่วยตัวอย่าง

ประเภทของหน่วยตัวอย่างที่เป็นจริง	ประเภทของหน่วยตัวอย่างที่ถูกทำนาย	
	กลุ่ม 1	กลุ่ม 2
กลุ่ม 1	a_1	a_2
กลุ่ม 2	a_3	a_4

เมทริกซ์ Confusion จะจัดเตรียมสารสนเทศที่จำเป็นในการแสดงถึงประสิทธิภาพของตัวแบบ แต่การสรุปสารสนเทศเหล่านี้ด้วยตัวเลขเพียงตัวเดียวจะทำให้ง่ายต่อการเปรียบเทียบประสิทธิภาพของแต่ละตัวแบบ ซึ่งแนวทางหนึ่งที่สามารถทำได้โดยการหาค่าความแม่นยำ (Accuracy) ซึ่งมีค่าเท่ากับ $(a_1+ a_4)/(a_1+ a_2+ a_3+ a_4)$ หรือเทียบเท่ากับการหาอัตราค่าคลาดเคลื่อน (Error Rate) ซึ่งมีค่าเท่ากับ $(a_2+ a_3)/(a_1+ a_2+ a_3+ a_4)$ เทคนิคการจำแนกประเภทต่าง ๆ จะหาตัวแบบที่มีความแม่นยำสูงสุด หรือค่าคลาดเคลื่อนต่ำสุดเมื่อใช้ตัวแบบที่ได้กับข้อมูลทดสอบ

ค่าความแม่นยำให้ความสำคัญกับกลุ่มข้อมูลที่มีจำนวนมากกว่า ซึ่งถ้าตัวแบบใดจำแนกประเภทกลุ่มที่มีจำนวนมากกว่าได้ถูกต้องมาก ถึงแม้จะจำแนกกลุ่มที่มีจำนวนน้อยกว่าไม่ถูกต้องเลยก็จะให้ค่าความแม่นยำที่สูง หากกลุ่มที่มีจำนวนน้อยกว่านี้เป็นกลุ่มที่มีความสำคัญ การใช้ค่าความแม่นยำก็ไม่สามารถสะท้อนความสามารถในการจำแนกประเภทของตัวแบบได้อย่างถูกต้อง เพื่อแก้ปัญหาดังกล่าว BAC (Balanced Accuracy) ได้ถูกนำมาใช้ในการวัดประสิทธิภาพของตัวแบบการจำแนกประเภท ซึ่ง BAC เป็นค่าความแม่นยำที่ถ่วงสมดุลระหว่าง Sen (Sensitivity) และ Spe (Specificity) โดยที่ Sen คือสัดส่วนตัวอย่างในกลุ่ม 1 ที่ได้รับการทำนายว่าเป็นกลุ่ม 1 และ Spe คือสัดส่วนตัวอย่างในกลุ่ม 2 ที่ได้รับการทำนายว่าเป็นกลุ่ม 2 และ BAC=(Sen+Spe)/2

การแบ่งส่วนข้อมูลฝึกฝนและข้อมูลทดสอบ

ตัวแบบการจำแนกประเภทถูกสร้างจากข้อมูลฝึกฝน และวัดประสิทธิภาพของตัวแบบการจำแนกประเภทบนข้อมูลทดสอบ ซึ่งลักษณะการกำหนดข้อมูลฝึกฝน และข้อมูลทดสอบเพื่อประเมินประสิทธิภาพของตัวแบบมีหลักการเดียวกับการวัดความสอดคล้องของตัวแบบเมื่อใช้องค์ประกอบด้วยจำนวนต่าง ๆ ที่มีการแบ่งส่วนข้อมูล ดังที่กล่าวไว้แล้วในตอนที่ 2 หัวข้อการกำหนดจำนวนองค์ประกอบด้วยการเปรียบเทียบความเหมาะสมของตัวแบบ ซึ่งวิธีต่าง ๆ ได้แก่ การตรวจสอบไขว้ และวิธียกออก

วิธียกออกเป็นหนึ่งในวิธีที่ถูกใช้อย่างกว้างขวางในการศึกษาวิธีการจำแนกประเภทเชิงเปรียบเทียบ สำหรับข้อมูลไมโครอาร์เรย์ และเชื่อว่าเป็นวิธีที่มีความน่าเชื่อถือกว่า K-Fold Cross-

Validation (Boulesteix, 2004) โดย Boulesteix ได้กำหนดอัตราส่วนของข้อมูลฝึกฝน: ข้อมูลทดสอบ เป็น 7:3 และทำซ้ำจำนวน 200 ครั้งเพื่อหาค่าเฉลี่ยของอัตราค่าคลาดเคลื่อน

การวิเคราะห์การจำแนกประเภทของวิธีกำลังสองน้อยที่สุดบางส่วน

PLS-R ใช้สำหรับการพยากรณ์ค่าตัวแปรตามซึ่งเป็นข้อมูลเชิงปริมาณ แต่สามารถประยุกต์กับงานการจำแนกประเภทซึ่งตัวแปรตามเป็นข้อมูลเชิงคุณภาพได้ (Alsberg et al., 1998; Barker & Rayens, 2003; Cho et al., 2002; Gutkin et al., 2009; Huang & Pan, 2003; Man et al., 2004; Rajalahti & Kvalheim, 2011; Tan et al., 2004) โดยเรียกว่าการวิเคราะห์การจำแนกประเภทของวิธีกำลังสองน้อยที่สุดบางส่วน (PLS-DA) ซึ่งมีขั้นตอนหลักที่เพิ่มขึ้น 2 ส่วน ได้แก่ ส่วนของการแปลงค่าเมทริกซ์ของตัวแปรตามให้เป็นเมทริกซ์ของตัวแปรดัมมี (Dummy Variable) และการทำนายกลุ่มจากค่าพยากรณ์ $\hat{y}$ ที่ได้จากตัวแบบซึ่งเป็นตัวเลขที่มีค่าต่อเนื่อง

การแปลงค่าตัวแปรตามทำได้โดยการให้รหัสสมาชิกในกลุ่มด้วยค่าเชิงตัวเลขในลักษณะเดียวกับตัวแปรดัมมี่ ซึ่งค่าของตัวแปรดัมมี่นี้มีได้สองค่า อาจให้รหัสเป็น -1 และ 1 หรือ 1 และ 0 ในกรณีที่ข้อมูลแบ่งเป็นสองกลุ่ม (Binary Class) จะมีตัวแปรตามเพียงหนึ่งตัว แต่ถ้าข้อมูลจำแนกได้มากกว่าสองกลุ่ม (Multicategory Class) แนวทางหนึ่งที่สามารถทำได้โดยสร้างตัวแปรตามที่มีจำนวนเท่ากับจำนวนกลุ่ม (q) กำหนดให้ $\mathbf{Y}$ เป็นเมทริกซ์ขนาด $n \times q$ แทนตัวแปรเชิงกลุ่มที่สร้างขึ้นใหม่ตามลักษณะของตัวแปรดัมมี่ $\mathbf{g}$ เป็นเวกเตอร์ขนาด n แทนตัวแปรเชิงกลุ่มเดิมที่สมาชิกระบุกลุ่มของตัวอย่างที่ i การกำหนดค่าของสมาชิกในเมทริกซ์ $\mathbf{Y}$ ทำได้ดังนี้

$$y_{ik} = \begin{cases} 1 & ; g_i = k \\ 0 & ; g_i \neq k \end{cases}, i = 1,2,...,n, k = 1,2,...,q$$

PLS-DA ถือเป็นกรณีเฉพาะของ PLS-R เมื่อชุดของตัวแปรตามเป็นตัวแปรที่มีสองค่าซึ่งระบุกลุ่มของหน่วยตัวอย่าง โดยมีขั้นตอนการสร้างตัวแบบเช่นเดียวกับ PLS-R ดังภาพที่ 4 ซึ่งแสดงไว้ในตอนที่ 2

สำหรับขั้นตอนของการทำนายกลุ่มแสดงดังภาพที่ 7 (Alsberg et al., 1998; Gutkin et al., 2009) โดยอาศัยค่าพารามิเตอร์ที่ได้จากขั้นตอนการสร้างตัวแบบจากภาพที่ 4 ซึ่งต้องเตรียมข้อมูลโดยการลบด้วยค่าเฉลี่ยหรือหารส่วนเบี่ยงเบนมาตรฐานที่ได้จากข้อมูลในการสร้างตัวแบบ กำหนดให้ $\mathbf{z}$ เป็นเวกเตอร์ขนาด p ซึ่งแทนตัวอย่างใหม่ที่ต้องการทำนายค่า

1. สำหรับ $a = 1,2,...,h$ ทำซ้ำ
   1.1 $t_a = \mathbf{z}'\mathbf{w}_a$
   1.2 $\mathbf{z} = \mathbf{z} - t_a\mathbf{p}_a$
2. $\hat{y}_k = \overline{y}_k + s_{y_k}\sum_{a=1}^{h} t_a d_a c_{ka}$ โดยที่ $\overline{y}_k$ และ s_{y_k} แทนค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวแปรตามที่ k ตามลำดับ
3. ทำนายกลุ่มที่ได้ ($\hat{g}$) จากเกณฑ์ต่อไปนี้
   3.1 สำหรับกรณีตัวแปรตามหนึ่งตัว (k=1)[a]

$$\hat{g} = \begin{cases} +1 & ;\hat{y} \geq 0 \\ -1 & ;\hat{y} < 0 \end{cases}$$

   3.2 สำหรับกรณีตัวแปรตามมากกว่า 1 ตัว

$\hat{g} = k^*$ โดยที่ $k^* = \left\{k \middle| \forall j, |\hat{y}_j| \leq |\hat{y}_k|\right\}$ สำหรับ j=1, 2,..., q

[a]สำหรับกรณีให้รหัสกับตัวแปรตามเป็น +1/-1

ภาพที่ 7 ขั้นตอนการทำนายกลุ่มจากตัวแบบ PLS-DA

PLS-DA เป็นเทคนิคที่เหมาะกับการจัดการกับข้อมูลที่มีจำนวนตัวแปรทำนายมากกว่าขนาดตัวอย่าง และมีความสัมพันธ์กันเองระหว่างตัวแปรทำนาย ซึ่งเป็นปัญหาหลักสองปัญหาที่มักเผชิญกับการวิเคราะห์ข้อมูลไมโครอาร์เรย์ Barker and Rayens (2003) ได้แสดงให้เห็นความเชื่อมโยงของ PLS-DA กับ LDA ของ Fisher ส่วน Tan and Others (2004) ใช้วิธี PLS-DA เพื่อจำแนกประเภทชนิดของเนื้องอกโดยทดสอบกับข้อมูลไมโครอาร์เรย์ 4 ชุดข้อมูล ได้แก่ ข้อมูลมะเร็งเม็ดเลือดขาวชนิดเฉียบพลัน ข้อมูลโรคมะเร็งเต้านมกรรมพันธุ์ ข้อมูลเนื้องอกเซลล์สีฟ้าขนาดเล็กทรงกลม และข้อมูล NCI60 ส่วน Huang and Pan (2003) ได้เปรียบเทียบวิธี PLS-DA และ Penalized PLS เปรียบเทียบกับวิธีดั้งเดิม ได้แก่ แผนการลงคะแนนเสียงถ่วงน้ำหนัก (The Weighted Voting Scheme) วิธีตัวแปรร่วมประกอบ (The Compound Covariate Method) และวิธีเซนทรอยด์ที่เล็กลง (The Shrunken Centroids Method) ซึ่งได้ทดสอบบนข้อมูลมะเร็งเม็ดเลือดขาว และมะเร็งลำไส้ ปรากฏว่าวิธี PLS-DA และวิธีการ Penalized PLS ทำงานได้ดี โดยวัดผลด้วยค่าคลาดเคลื่อนในการทำนาย

ตอนที่ 4 งานวิจัยที่เกี่ยวข้อง

โดยมีเนื้อหาเกี่ยวกับดัชนีวัดความสำคัญของตัวแปร เกณฑ์จุดตัดสำหรับดัชนีวัดความสำคัญของตัวแปรในการคัดเลือกตัวแปรบนพื้นฐาน PLS และการจำแนกประเภทข้อมูลสำหรับข้อมูลไมโครอาร์เรย์ ดังนี้

1. ดัชนีวัดความสำคัญของตัวแปรบนพื้นฐาน PLS

Chong and Jun (2005) ได้เปรียบเทียบวิธีการคัดเลือกตัวแปรระหว่างวิธีที่ใช้ดัชนี VIP ที่มีเกณฑ์จุดตัด 1.0 (VIP-1) วิธีการลดลงสัมบูรณ์น้อยที่สุดและการคัดเลือกตัวดำเนินการ (Least Absolute Shrinkage and Selection Operator: Lasso) และการวิเคราะห์การถดถอยพหุคูณแบบทีละขั้น (Stepwise Multiple Regression) บนข้อมูลจำลองภายใต้ 4 พารามิเตอร์ ได้แก่ สัดส่วนของตัวแปรที่เกี่ยวข้อง ขนาดความสัมพันธ์ระหว่างตัวแปรทำนาย โครงสร้างสัมประสิทธิ์การถดถอย และขนาดของอัตราส่วนสัญญาณต่อสัญญาณรบกวน (Signal to Noise Ratio) ซึ่งพบว่าวิธี VIP-1 ให้ประสิทธิภาพที่ดีกว่าอีกสองวิธีในทุกสถานการณ์ นอกจากนี้แล้ว Chong and Jun ยังได้สำรวจถึงการใช้เกณฑ์จุดตัดที่เหมาะสมสำหรับข้อมูลลักษณะต่าง ๆ ซึ่งพบว่าถ้าสัดส่วนของตัวแปรที่เกี่ยวข้องมีค่าน้อย และขนาดความสัมพันธ์ระหว่างตัวแปรทำนายสูง หรือสัมประสิทธิ์การถดถอยของแต่ละตัวแปรเท่ากันแล้ว เกณฑ์จุดตัดสำหรับ VIP ควรมีค่ามากกว่า 1.0

Teófilo and Others (2008) ได้เรียงลำดับความสำคัญของตัวแปรโดยใช้ดัชนีที่ได้จากพารามิเตอร์ของ PLS-R จำนวน 7 ดัชนี ได้แก่ สัมประสิทธิ์การถดถอย สัมประสิทธิ์สหสัมพันธ์ ค่าตกค้างของตัวแปรทำนาย กระบวนการความแปรปรวนร่วมเกี่ยว (Covariance Procedure) VIP สัญญาณวิเคราะห์สุทธิ (Net Analyte Signal) และอัตราส่วนสัญญาณต่อสัญญาณรบกวน และใช้ดัชนีเหล่านี้ในอัลกอริทึมการคัดเลือกตัวแปรทำนายตามลำดับ (Ordered Predictors Selection: OPS-PLS) ในการลำดับความสำคัญของตัวแปร เพื่อกำหนดตัวแปรที่นำเข้าตัวแบบ โดยกำหนดเซตย่อยเริ่มต้นของตัวแปรในการสร้างตัวแบบและประเมินผล จากนั้นเพิ่มตัวแปรเข้าสู่ตัวแบบด้วยจำนวนที่คงที่และประเมินค่าตัวแบบ การเพิ่มนี้ทำจนกระทั่งมีตัวแปรทุกตัวอยู่ในตัวแบบหรือจนกระทั่งได้ครบกำหนดจำนวนที่ต้องการ ผลที่ได้พบว่าการใช้ดัชนีเพื่อคัดเลือกตัวแปรไม่ว่าจะเป็นดัชนีใดก็ตามให้ผลดีกว่าการไม่คัดเลือกตัวแปรออก และสัมประสิทธิ์การถดถอย และสัญญาณวิเคราะห์สุทธิเป็นดัชนีที่ดีที่สุดสำหรับการคัดเลือกตัวแปร

Chen (2008) ได้นำเสนอวิธีการเพื่อระบุยีนที่มีความเกี่ยวข้องซึ่งวิธีการอยู่บนพื้นฐานของค่าน้ำหนัก PLS-R โดยเรียกวิธีการนี้ว่า VIP แต่งเติมเชิงสุ่ม (Random Augmented Variance Influence on Projection: RA-VIP) วิธีดังกล่าวระบุยีนที่เกี่ยวข้องโดยการเปรียบเทียบค่าเฉลี่ยของ VIP ของแต่ละยีนกับการแจกแจงของ VIP เมื่อยีนทั้งหมดไม่สัมพันธ์กับตัวแปรกลุ่ม ซึ่งจะทำให้ได้ค่า p หรือ p-value ผู้วิจัยได้วัดประสิทธิภาพการคัดเลือกยีนของ RA-VIP ด้วย Sen และ Spe และ

นำมาเปรียบเทียบกับการคัดเลือกยีนเมื่อใช้ดัชนี VIP ที่มีเกณฑ์จุดตัด 1.0 และสัมประสิทธิ์การถดถอยของ PLS (B_{PLS}) โดยการใช้ข้อมูลจำลอง และข้อมูลไมโครอาร์เรย์ ผลที่ได้พบว่าวิธี B_{PLS} ให้ผลดีเมื่อวัดด้วยค่า Spe โดยที่ VIP ให้ผลดีเมื่อวัดด้วยค่า Sen ส่วน RA-VIP ให้ผลดีในภาพรวม

Gutkin and Others (2009) ได้พัฒนารูปแบบต่างๆ สำหรับการคัดเลือกตัวแปรบนพื้นฐานของ PLS-R โดยใช้ชื่อว่า SlimPLS ซึ่งได้รวมความสัมพันธ์ระหว่างตัวแปรเข้าไว้ด้วยกัน รูปแบบต่าง ๆ ของ SlimPLS ขึ้นอยู่กับ 1) วิธีการกำหนดจำนวนองค์ประกอบ โดยมีการกำหนดแบบคงที่ซึ่งผู้วิจัยกำหนดเป็น 2 และ 3 หรือกำหนดตามเกณฑ์ของค่า p-value 2) วิธีการเลือกตัวแปรจากแต่ละองค์ประกอบ โดยเลือกจากค่าน้ำหนัก w_{aj} ที่สูงสุดตามลำดับ หรือใช้กระบวนการค้นหาแบบปีนเขา (Hill Climbing Search) และเลือกเซตย่อยที่ทำให้ค่าประมาณค่าคลาดเคลื่อนมีค่าน้อยที่สุด 3) ผลลัพธ์ที่นำมาใช้ โดยนำตัวแปรที่ได้ไปใช้โดยตรง หรือใช้องค์ประกอบโดยมีเฉพาะตัวแปรที่คัดเลือก โดยนำไปทดสอบกับ ข้อมูลไมโครอาร์เรย์ 19 ชุด วิธีการที่ใช้ในการจำแนกประเภทได้แก่ SVM, KNN, NB และตัวจำแนกป่าแบบสุ่ม (Random Forest Classifier) วิธีการใหม่นี้ถูกเปรียบเทียบกับวิธีการคัดเลือกตัวแปรดั้งเดิมที่ใช้ดัชนีที่ไม่ได้นำความสัมพันธ์ระหว่างตัวแปรเข้ามามีส่วนในการคัดเลือกตัวแปร ได้แก่ สัมประสิทธิ์สหสัมพันธ์แบบเพียร์สัน (Pearson Correlation Coefficient) การทดสอบของเวลช์ (Welch Test) เกณฑ์ของโกลับ (Golub Criterion) และสารสนเทศร่วม (Mutual information) ผลที่ได้พบว่าวิธีการใหม่ที่นำเสนอให้ผลที่ดีกว่าวิธีการคัดเลือกตัวแปรแบบดั้งเดิม โดยรูปแบบที่ให้ผลที่ดีคือการร่วมกันระหว่างการกำหนดจำนวนองค์ประกอบตามเกณฑ์ค่า p-value และการเลือกเซตย่อยโดยใช้กระบวนการค้นหาแบบปีนเขา และเมื่อใช้องค์ประกอบจากแนวทางดังกล่าวจะให้ผลดีกว่าเมื่อใช้ตัวแปรโดยตรงเล็กน้อย SlimPLS ในรูปแบบการร่วมกันดังกล่าวให้ผลดีกว่าการคัดเลือกตัวแปรแบบดั้งเดิมอย่างชัดเจนเมื่อจำแนกประเภทโดยใช้วิธี KNN

Anzanello, Albin, and Chaovalitwongse (2009) ได้เสนอแนวทางในการวัดความสำคัญของตัวแปรในกระบวนการผลิตที่มีต่อการจำแนกประเภทของคุณภาพผลผลิต โดยนำเสนอดัชนีซึ่งอาศัยค่าพารามิเตอร์ที่ได้จาก PLS-R จำนวน 5 ดัชนี และนำดัชนีมาใช้ร่วมกับการกำจัดตัวแปรแบบย้อนกลับ โดยเลือกกำจัดตัวแปรที่มีค่าดัชนีต่ำที่สุด และจำแนกประเภทข้อมูลด้วยวิธี KNN, SVM และข่ายงานประสาทตามความน่าจะเป็น (Probabilistic Neural Network: PNN) กลุ่มตัวแปรที่ทำให้ค่าความแม่นยำสูงสุดจะได้รับเลือก วิธีการดังกล่าวได้ทดลองกับข้อมูล 5 ชุด ซึ่งเป็นข้อมูลที่เกี่ยวกับกระบวนการผลิตในอุตสาหกรรม โดยเปรียบเทียบวิธีการที่นำเสนอกับการวิเคราะห์การถดถอยพหุคูณแบบย้อนหลัง (Backward Multiple Regression: BMR) การวิเคราะห์การถดถอยพหุคูณแบบไปข้างหน้า (Forward Multiple Regression: FMR) การวิเคราะห์การถดถอยพหุคูณแบบทีละขั้น (Stepwise Multiple Regression: SMR) และ

Backward-Q^2_{cum} (BQ) ผลที่ได้พบว่า เมื่อใช้การจำแนกประเภทโดยวิธี KNN ร่วมกับดัชนีที่ใช้ค่าน้ำหนัก $\mathbf{w}_a$ และสัดส่วนความแปรผันในตัวแปรตามที่ถูกอธิบายด้วยองค์ประกอบ ($R^2(\mathbf{y},\mathbf{t}_a)$) ทำให้ความแม่นยำในการจำแนกประเภทมากกว่าวิธีอื่น ๆ โดยส่วนใหญ่และมากกว่าเมื่อใช้ตัวแปรทั้งหมดโดยเฉลี่ย 6% โดยใช้ตัวแปรเพียง 9% ของตัวแปรทั้งหมด

Ji, Yang, and You (2011) ได้เสนอดัชนีในการคัดเลือกยีนโดยอาศัย PLS-R ซึ่งดัชนีนี้เป็นการวัดความสามารถของยีน (ตัวแปรทำนาย) ในการอธิบายกลุ่ม (ตัวแปรตาม) ซึ่งพิจารณาถึงผลกระทบรวมของทุกยีน และความสัมพันธ์ระหว่างยีน ดัชนีที่นำเสนอได้แก่ อัตราการขยายของการอธิบายตัวแปรทำนาย (Independent-Variable Explanation Gain: IEG) และอัตราการขยายของการอธิบายตัวแปรตาม (Dependent-Variable Explanation Gain: DEG) ซึ่งดัชนีทั้งสองนี้วัดการเพิ่มขึ้นของความสามารถในการอธิบายตัวแปรทำนาย (สำหรับ IEG) หรือตัวแปรตาม (สำหรับ DEG) ของตัวแปรทำนายใด ๆ และได้นำมาเปรียบเทียบกับเมื่อใช้ดัชนี VIP ตัวแปรที่มีค่าวัดดังกล่าวสูงสุด k ตัวจะได้รับการคัดเลือก โดยกำหนดจำนวนตัวแปรสูงสุดที่เลือกเท่ากับ 100 เปรียบเทียบค่าความแม่นยำในการจำแนกประเภทข้อมูลเมื่อมีตัวแปรขนาดต่าง ๆ และเลือกขนาดตัวแปรที่ทำให้ความแม่นยำสูงสุด ผู้วิจัยได้ทดลองกับข้อมูลไมโครอาร์เรย์สองชุดข้อมูล ได้แก่ ข้อมูลมะเร็งเม็ดเลือดขาว และข้อมูลการเป็นเนื้องอก และใช้ SVM ในการจำแนกประเภท ผลที่ได้พบว่าวิธีที่นำเสนอให้ความแม่นยำในการจำแนกประเภท 100% สำหรับข้อมูลทั้งสองชุด ซึ่งเมื่อได้เปรียบเทียบกับงานวิจัยหลาย ชิ้นที่ใช้ข้อมูลสองชุดดังกล่าวนี้พบว่าวิธีที่นำเสนอให้ผลความแม่นยำที่เทียบเท่ากับหรือดีกว่าด้วยจำนวนยีนที่น้อยกว่า

สรุปทิศทางของงานวิจัย

PLS-R เป็นวิธีที่ได้รับการออกแบบมาให้เหมาะสมกับข้อมูลที่มีจำนวนมิติมาก ซึ่งข้อมูลลักษณะดังกล่าวมักมีความสัมพันธ์กันเอง พารามิเตอร์จากวิธี PLS-R จึงได้ถูกนำมาใช้วัดความสำคัญของตัวแปรทำนาย สำหรับข้อมูลในลักษณะดังกล่าว ดัชนีหนึ่งที่ได้รับความนิยมเพิ่มมากขึ้นคือ VIP (Chong & Jun, 2005) พารามิเตอร์อื่นที่ได้จากวิธี PLS-R ที่ได้นำมาใช้เพื่อวัดความสำคัญของตัวแปร เช่น เวกเตอร์น้ำหนัก สัมประสิทธิ์การถดถอย PLS (Gutkin et al., 2009; Teófilo et al., 2008) เป็นต้น นอกจากนี้ยังได้มีการนำพารามิเตอร์ของ PLS มาใช้ในการสร้างดัชนีวัดความสำคัญของตัวแปรขึ้นมาใหม่ เช่น งานของ Chen (2008) Anzanello and Others (2009) และ Ji and Others (2011) เป็นต้น

2. เกณฑ์จุดตัดสำหรับดัชนีวัดความสำคัญของตัวแปรบนพื้นฐาน PLS ในการคัดเลือกตัวแปร

Centner and Massart (1996) ได้เสนอวิธีการกำจัดตัวแปรที่ไม่ให้ข้อมูลโดยอาศัย PLS (Uninformative Variable Elimination by PLS: UVE-PLS) ซึ่งใช้ดัชนีวัดความสำคัญที่เป็น

ความเที่ยง (Reliability) c ของสัมประสิทธิ์การถดถอย PLS ถ้า c ของตัวแปรทำนายใดมีค่ามาก แสดงถึงความสำคัญของตัวแปรนั้นมีมากด้วย การสร้างเกณฑ์จุดตัดสำหรับคัดเลือกตัวแปรกำหนดได้ด้วยการสร้างตัวแปรรบกวนที่มีค่าน้อย ซึ่งมีจำนวนเท่ากับจำนวนตัวแปรเริ่มต้นเพิ่มเข้าไปในข้อมูลเริ่มต้น และคำนวณค่า c ของตัวแปรทำนายเริ่มต้นและตัวแปรรบกวนเหล่านี้ ค่า c ของตัวแปรทำนายเริ่มต้นใดที่ต่ำกว่าค่าสูงสุดของตัวแปรรบกวนเหล่านี้ จะถูกพิจารณาว่าตัวแปรทำนายนั้นเป็นตัวแปรที่ไม่สำคัญ และไม่มีความเกี่ยวข้องกับการทำนายค่าซึ่งควรถูกกำจัด นอกจากนี้ในงานเดียวกันของ Centner and Massart ยังได้มีการทดลองใช้เปอร์เซ็นไทล์ที่ 90, 95 และ 99 เป็นเกณฑ์จุดตัด ซึ่งพบว่าการใช้เกณฑ์ค่าสูงสุดให้ผลลัพธ์จำนวนตัวแปรที่เหลือด้วยจำนวนที่น้อยกว่าเกณฑ์เปอร์เซ็นไทล์ ในขณะที่ค่า RMSEP สูงกว่าเมื่อใช้เกณฑ์เปอร์เซ็นไทล์ที่ 95 เล็กน้อย Shao, Wang, Chen, and Su (2004) ได้พัฒนาวิธี UVE-PLS โดยกำหนดเกณฑ์จุดตัดสำหรับ c ต่างไปจาก Centner and Massart เล็กน้อยโดยคูณค่าคงที่ระหว่าง 0 ถึง 1 ที่ค่าสูงสุดของค่า c ของตัวแปรรบกวน ส่วน Moros, Kuligowski, Quintas, Garrigues, and Guardia (2008) ได้พัฒนาวิธี UVE-PLS โดยสร้างเกณฑ์จุดตัดสำหรับค่า c ซึ่งอาศัยการแจกแจงที (t Distribution) ผลลัพธ์ที่ได้พบว่าเมื่อใช้เกณฑ์ดังกล่าวทำให้ค่า RMSEP ใกล้เคียงกับเมื่อใช้เกณฑ์ค่าสูงสุดแต่จำนวนตัวแปรมากกว่าเล็กน้อย และเกณฑ์ดังกล่าวให้ผลที่ดีกว่าเกณฑ์เปอร์เซ็นต์ไทล์ที่ 99 ทั้งในด้านจำนวนตัวแปรที่คัดเลือกและค่า RMSEP แต่วิธีที่นำเสนอมีข้อดีในด้านของการทำซ้ำ เนื่องจากค่าเฉลี่ยของค่าที่ใช้วัดประสิทธิภาพมีความแปรปรวนต่ำกว่าของวิธีอื่น ๆ ทั้งนี้ทดลองกับข้อมูลเนียร์อินฟราเรดสเปกตรัม (Near Infrared Spectra) ของเฮโรอีน

Johansson and Others (2003) ได้คัดเลือกตัวแปรโดยใช้น้ำหนักขององค์ประกอบที่ 1 ($\mathbf{w}_1$) สำหรับเป็นดัชนีวัดความสำคัญของตัวแปร และสร้างเกณฑ์จุดตัด โดยการเพิ่มตัวแปรรบกวนซึ่งได้จากการเรียงสับเปลี่ยนตัวแปรเริ่มต้น จากนั้นหาค่าเปอร์เซ็นต์ไทล์ที่ α (กำหนด α=0.005, 0.01 และ 0.05) ของ $\mathbf{w}_1$ ของตัวแปรรบกวน และกำหนดเป็นเกณฑ์จุดตัด ตัวแปรที่มีค่า $\mathbf{w}_1$ มากกว่าเกณฑ์จุดตัดตัวแปรนั้นจะได้รับการคัดเลือก

Zeng and Others (2008) ได้นำเสนออัลกอริทึมในการคัดเลือกตัวแปรที่เกี่ยวข้อง และนำตัวแปรที่ได้ไปใช้ในการลดมิติของข้อมูลด้วย PLS-DR ซึ่งการคัดเลือกตัวแปรอาศัยค่าสถิติที (t Statisitc) และกำหนดเกณฑ์จุดตัดโดยการสร้างตัวแปรสุ่มจำนวนหนึ่งเพิ่มเข้าในข้อมูลเริ่มต้น ค่าสถิติทีของตัวแปรเริ่มต้นใดที่มีค่าสูงกว่าค่าเฉลี่ยของค่าสถิติทีของตัวแปรสุ่มเหล่านี้จะได้รับการคัดเลือก และนำไปสร้างองค์ประกอบด้วย PLS-DR อัลกอริทึมดังกล่าวนี้มีชื่อว่า PLSDR[9] วิธีการใหม่นี้นำไปทดลองกับข้อมูลไมโครอาร์เรย์ที่เป็นมาตรฐานจำนวน 6 ชุดข้อมูล จำแนกประเภทด้วยวิธี SVM, KNN และ C4.5 วัดความสามารถของวิธีการด้วยเวลาที่ใช้ในการประมวลผล Sen, Spe และ BAC ซึ่งจำแนกประเภทข้อมูลด้วยวิธี SVM, KNN และ C4.5 โดยเปรียบเทียบกับวิธีการลดมิติของ

ข้อมูลด้วย PLS-DR ที่ไม่ได้มีการคัดเลือกตัวแปร ผลการทดลองพบว่าวิธีการที่นำเสนอสามารถลดจำนวนตัวแปรลงได้ประมาณ 29% ซึ่งโดยส่วนใหญ่แล้วมีความแม่นยำสูงกว่าเมื่อไม่มีการคัดเลือกตัวแปร นอกจากนี้ถึงแม้ว่าจะมีขั้นตอนที่เพิ่มเข้ามาแต่เวลาที่ใช้ในการประมวลผลของวิธีการที่นำเสนอโดยเฉลี่ยแล้วน้อยกว่าเมื่อไม่ได้มีการคัดเลือกตัวแปร

สรุปทิศทางของงานวิจัย

พารามิเตอร์จากวิธี PLS-R ได้ถูกนำมาใช้ในการพัฒนาดัชนีวัดความสำคัญของตัวแปรทำนาย (Anzanello et al,, 2009; Ji et al., 2011; Teófilo et al., 2008) ดัชนีดังกล่าวมักนำไปใช้ร่วมกับการคัดเลือกตัวแปรอื่นโดยไม่ได้กำหนดเกณฑ์จุดตัดที่ชัดเจน ผู้ใช้อาจต้องกำหนดจำนวนตัวแปร หรือเกณฑ์จุดตัดด้วยตนเอง ดัชนี VIP เป็นดัชนีวัดความสำคัญของตัวแปรที่ได้รับความสนใจเพิ่มมากขึ้น โดยเฉพาะอย่างยิ่งเมื่อตัวแปรทำนายมีความสัมพันธ์กันเอง (Chong & Jun, 2005) ที่มักพบในข้อมูลที่มีจำนวนตัวแปรทำนายมาก ซึ่งเกณฑ์จุดตัดที่มักใช้สำหรับคัดเลือกตัวแปรเมื่อใช้ดัชนี VIP คือ VIP>1 (Chen, 2008, pp.14-15; Chong & Jun, 2005; Lazraq et al., 2003; Sun et al., 2012) แต่การกำหนดเกณฑ์จุดตัดที่เป็นค่าคงที่นี้ อาจไม่เหมาะกับโครงสร้างข้อมูลทุกลักษณะ (Chong & Jun, 2005) การสร้างตัวแปรรบกวนจำนวนหนึ่งรวมเข้ากับตัวแปรเริ่มต้น แล้วคำนวณค่าดัชนีวัดความสำคัญตามที่กำหนดทั้งตัวแปรเริ่มต้น และตัวแปรรบกวน จากนั้นสร้างเกณฑ์จุดตัดจากค่าดัชนีของตัวแปรรบกวน แนวทางนี้ถูกนำมาใช้ในการสร้างเกณฑ์จุดตัด (Centner & Massart, 1996; Johansson et al., 2003; Zeng et al., 2008)

3. การจำแนกประเภทข้อมูลสำหรับข้อมูลไมโครอาร์เรย์

งานวิจัยที่เกี่ยวข้องกับการจำแนกประเภทข้อมูลสำหรับข้อมูลไมโครอาร์เรย์ ในบางส่วนได้กล่าวถึงในหัวข้องานวิจัยก่อนหน้านี้ไปแล้ว นอกจากนี้แล้วยังมีงานวิจัยอื่น ๆ อีกดังนี้

Nguyen and Rocke (2002a) ได้ใช้ค่าสถิติที ในการเรียงลำดับความสำคัญของยีนเพื่อคัดเลือกยีนสำหรับการจำแนกประเภทข้อมูลไมโครอาร์เรย์ 5 ชุดข้อมูล ซึ่งเป็นงานการจำแนกประเภทข้อมูลที่แบ่งเป็นสองกลุ่ม โดยกำหนดเลือกยีนที่มีค่าสถิติทีสูงสุดจำนวน 50, 100, 500, 1,000 และ 1,500 ยีน แล้วนำเข้าสู่กระบวนการลดมิติของข้อมูลด้วยการสร้างองค์ประกอบ PLS และสร้างตัวแบบในการจำแนกประเภทด้วย LD และ QDA ต่อไป เช่นเดียวกับ Dai and Others (2006) ที่ใช้ค่าสถิติทีในการคัดเลือกยีนเช่นกัน แต่เลือกยีนที่มีค่าสถิติทีสูงสุดจำนวน 200, 500 และ 1,000

Cho and Others (2002) ได้ใช้ PLS-DA ในการแยกประเภทชนิดของมะเร็งเม็ดเลือดขาว โดยได้สกัดยีนที่เกี่ยวข้องในขั้นต้นด้วยสถิติทดสอบที ด้วยระดับนัยสำคัญ 0.05 ซึ่งจะนำยีนที่ได้ไปใช้เป็นตัวแปรสำหรับชุดข้อมูลเต็ม (Full Data Set) จากนั้นคัดเลือกยีนด้วยดัชนี VIP ซึ่งใช้เกณฑ์จุดตัด 1.0 ทำซ้ำกระบวนการ PLS-DA และการคัดเลือกยีนด้วยดัชนี VIP จนกว่าจะได้ขนาดจำนวนตัวแปรที่

ต้องการ ผลลัพธ์ที่ได้พบว่ายีนที่เกี่ยวข้องมีจำนวน 3 ยีนจาก 1,178 ยีน และจำแนกประเภทผิดพลาด 3 ตัวอย่างจาก 72 ตัวอย่าง

Huang and Pan (2003) ได้ชี้จุดที่เชื่อมโยงกันของวิธีการจำแนกประเภท 3 วิธีดั้งเดิม ได้แก่ แผนการลงคะแนนเสียงถ่วงน้ำหนัก วิธีตัวแปรร่วมประกอบ และวิธีเซนทรอยด์ที่เล็กลง ซึ่งถึงแม้ว่าทั้งสามก่อกำเนิดจากที่มาที่แตกต่างกัน แต่สามารถเชื่อมโยงกันด้วยตัวแบบการถดถอยเชิงเส้น และด้วยตัวแบบการถดถอยเชิงเส้นนี้จะนำไปสู่วิธีทั่วไปอื่น ๆ เช่น PLS-DA , Penalized PLS ซึ่งสามารถใช้เป็นวิธีทางเลือก ในงานวิจัยนี้ได้เปรียบเทียบวิธีทั้งห้า (แผนการลงคะแนนเสียงถ่วงน้ำหนัก วิธีตัวแปรร่วมประกอบ วิธีเซนทรอยด์ที่เล็กลง PLS-DA และ Penalized PLS) กับข้อมูลมะเร็งเม็ดเลือดขาว และมะเร็งลำไส้ ซึ่งพบว่าวิธี PLS-DA และวิธีการ Penalized PLS ทำงานได้ดีโดยวัดผลด้วยค่าคลาดเคลื่อนในการทำนาย

Tan and Others (2004) ได้ใช้วิธี PLS-DA ในการจำแนกประเภทชนิดเนื้องอก (ข้อมูลมะเร็งเม็ดเลือดขาวชนิดเฉียบพลัน ข้อมูลโรคมะเร็งเต้านมกรรมพันธุ์ ข้อมูลเนื้องอกเซลล์สีฟ้าขนาดเล็กทรงกลม ข้อมูล NCI60) ผลที่ได้พบว่าการใช้การตรวจสอบไขว้แบบแยกทีละครึ่ง (Leave-Half-Out Cross-Validation) ในการประเมินประสิทธิภาพตัวแบบและจำแนกประเภทด้วยวิธี PLS-DA ให้ค่าความสามารถในการทำนายที่สมเหตุสมผล

สรุปทิศทางของงานวิจัย

ดัชนีวัดความสำคัญของตัวแปรทำนายสำหรับข้อมูลไมโครอาร์เรย์ที่ใช้มากดัชนีหนึ่งคือ สถิติที (Dai et al., 2006; Nguyen & Rocke, 2002a; Saeys et al., 2007) แต่การใช้สถิติทีเป็นดัชนีวัดความสำคัญของตัวแปรนั้น พิจารณาแต่ละตัวแปรทำนายอย่างเป็นอิสระกัน โดยไม่ได้คำนึงถึงความสัมพันธ์ระหว่างตัวแปรทำนายด้วยกันให้มีผลต่อการคัดเลือกตัวแปร อาจทำให้ตัวแปรที่คัดเลือกมามีความแม่นยำน้อยในการจำแนกประเภทข้อมูล ดัชนี VIP เป็นดัชนีวัดความสำคัญของตัวแปรที่อาศัยพารามิเตอร์จาก PLS-R ซึ่งได้มีการพิจารณาถึงความสัมพันธ์ระหว่างตัวแปรทำนาย โดยถูกนำมาใช้ในการวัดความสำคัญของยีนสำหรับข้อมูลไมโครอาร์เรย์ (Chen, 2008; Ji et al., 2011) นอกจากการใช้ PLS-R ในการสร้างดัชนีความสำคัญของตัวแปรแล้ว PLS-DA ซึ่งเป็นกรณีเฉพาะของ PLS-R ได้นำมาใช้เป็นวิธีการจำแนกประเภทข้อมูล ซึ่งมีประสิทธิภาพดี (Cho et al., 2002; Huang & Pan, 2003; Tan et al., 2004)