การเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติสำหรับข้อมูลเข้าที่ใช้ในวิธีซัพพอร์ตเวกเตอร์แมชชีน : กรณีศึกษาการแจกแจงแบบเกาส์เซียน

นางสาวอารยา หลงชวน

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

COMPARISON THE EFFICIENCY OF DIMENSION REDUCTION TECHNIQUES FOR INPUT DATA IN SUPPORT VECTOR MACHINE: A CASE STUDY IN GAUSSIAN DISTRIBUTION

Miss Araya Longchuan

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Statistics
Department of Statistics
Faculty of Commerce and Accountancy
Chulalongkorn University
Academic Year 2013

หัวข้อวิทยานิพนธ์	การเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติสำหรับข้อมูลเข้าที่ใช้ในวิธีซัพพอร์ตเวกเตอร์แมชชีน : กรณีศึกษาการแจกแจงแบบเกาส์เซียน
โดย	นางสาวอารยา หลงชวน
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	อาจารย์ ดร.นัท กุลวานิช

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.พสุ เดชะรินทร์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ

(รองศาสตราจารย์ ดร.สุพล ดุรงค์วัฒนา)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(อาจารย์ ดร.นัท กุลวานิช)

.....กรรมการ

(อาจารย์ ดร.วิฐรา พึ่งพาพงศ์)

.....กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.อรุณี กำลัง)

อารยา หลงชวน : การเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติสำหรับข้อมูลเข้าที่ใช้ในวิธีซัพพอร์ตเวกเตอร์แมชชีน : กรณีศึกษาการแจกแจงแบบเกาส์เซียน. (COMPARISON THE EFFICIENCY OF DIMENSION REDUCTION TECHNIQUES FOR INPUT DATA IN SUPPORT VECTOR MACHINE: A CASE STUDY IN GAUSSIAN DISTRIBUTION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: อ. ดร.นัท กุลวานิช, 140 หน้า.

การวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติข้อมูลเข้า(input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS)และSliced Average Variance Estimator (SAVE) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน กรณีที่แบ่งตัวแปรตามออกเป็น 2 กลุ่ม ซึ่งดำเนินการภายใต้ขอบเขตของจำนวนตัวแปรอิสระที่ทำการศึกษาคือ 5, 10, 20 และ 40 ตัว โดยทำการจำลองข้อมูลและวิเคราะห์ผลด้วยโปรแกรม R 2.15.3 ทั้งนี้จะใช้ Receiver Operating Characteristic (ROC) เป็นเครื่องมือวัดประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูล โดยใช้พื้นที่ใต้โค้ง ROC (Area Under ROC Curve : AUC) และใช้อัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate : MCR)

การศึกษาภายใต้ขอบเขตดังกล่าวผลปรากฏว่ากรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 เมื่อขนาดตัวอย่างของทั้งสองกลุ่มเท่ากับ 30 และจำนวนตัวแปรอิสระเท่ากับ 10 เมื่อขนาดตัวอย่างทั้งหมดไม่เกิน 120 ให้ผลเหมือนกัน การลดข้อมูลด้วยวิธี PLS ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด แต่เมื่อเพิ่มขนาดตัวอย่างการลดข้อมูลด้วยวิธี SAVE จะให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีกว่าวิธี PLS และ PCA และเมื่อเพิ่มจำนวนตัวแปรอิสระเป็น 20 และ 40 การลดข้อมูลด้วยวิธี PLS ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด

ภาควิชา สถิติ ลายมือชื่อนิสิต .....

สาขาวิชา สถิติ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2556

5581636326 : MAJOR STATISTICS
KEYWORDS: PRINCIPLE COMPONENT ANALYSIS / PARTIAL LEAST SQUARES / SLICED AVERAGE VARIANCE ESTIMATOR / SUPPORT VECTOR MACHINE

ARAYA LONGCHUAN: COMPARISON THE EFFICIENCY OF DIMENSION REDUCTION TECHNIQUES FOR INPUT DATA IN SUPPORT VECTOR MACHINE: A CASE STUDY IN GAUSSIAN DISTRIBUTION. ADVISOR: NAT KULVANICH, Ph.D., 140 pp.

The purpose of this study is to compare the effectiveness of dimension reduction techniques between Principle Component Analysis (PCA), Partial Least Squares (PLS) and Sliced Average Variance Estimator (SAVE) for input data of Support Vector Machine. The datasets from four different number of independent variables (p=5, 10, 20 and 40) were simulated in this study. Simulating and analyzing data in this study use R 2.15.3.The area under ROC curve (AUC) and misclassification rate (MCR) were used to evaluate and compare the prediction performance in forecasting classification data.

In case of 5 independent variables on a sample size of both groups equal to 30 and 10 independent variables when sample size is less than 120. The results are the same, PLS is the most effective dimension reduction technique. On the other hand, when we increase the sample size, SAVE clearly performs better overall PLS than and PCA. And when we increase the number of independent variables , PLS is the most effective dimension reduction technique.

Department: Statistics Student's Signature .....

Field of Study: Statistics Advisor's Signature .....

Academic Year: 2013

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ลงได้ด้วยความอนุเคราะห์และความเอาใจใส่จากอาจารย์ที่ปรึกษาวิทยานิพนธ์ อาจารย์ ดร.นัท กุลวานิช ผู้วิจัยจะขอกราบขอบพระคุณท่านอาจารย์เป็นอย่างสูง ที่ให้ความกรุณาแก่ผู้วิจัยเป็นอย่างมาก ทั้งให้คำปรึกษา คำแนะนำเพื่อปรับปรุงแก้ไขวิทยานิพนธ์ และเป็นกำลังใจในการทำงาน ทั้งนี้ผู้วิจัยขอกราบขอบพระคุณท่านประธานกรรมการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร. สุพล ดุรงค์วัฒนา ท่านกรรมการ อาจารย์ ดร.วิฐรา พึ่งพาพงศ์ และท่านกรรมการภายนอกอาจารย์ ดร.อรุณี กำลัง เป็นอย่างสูงที่ท่านอาจารย์ทั้งสามท่านได้เสียสละเวลาเพื่อสอบและให้คำแนะนำที่ดีและมีประโยชน์ในการปรับปรุงแก้ไขงานของผู้วิจัยต่อไป

ขอกราบขอบพระคุณคณาจารย์ทุกท่านประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยที่ให้โอกาสทางการศึกษา และประสิทธิประสาทความรู้ให้แก่ผู้วิจัยจนกระทั่งสำเร็จการศึกษาในครั้งนี้

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณคุณพ่อ คุณแม่ และครอบครัว ที่ให้กำลังและให้ความห่วงใยมาตลอด และขอขอบคุณเพื่อน ๆ ทุกคน ที่คอยช่วยเหลือ ให้คำแนะนำและเป็นกำลังใจให้กับผู้วิจัยตลอดการทำวิทยานิพนธ์ฉบับนี้

สารบัญ

หน้า

หน้า

สารบัญตาราง

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

ตารางที่ หน้า

สารบัญภาพ

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

ภาพที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การลดมิติ (Dimension Reduction) เป็นเทคนิคหนึ่งของการเตรียมข้อมูล (data preprocessing)ก่อนที่จะทำการวิเคราะห์ข้อมูล เนื่องมาจากในปัจจุบันความก้าวหน้าของความสามารถในการเก็บรวบรวมข้อมูลมีประสิทธิภาพสูงขึ้นอย่างรวดเร็ว ทำให้นักวิจัยส่วนใหญ่จึงต้องเผชิญหน้ากับชุดข้อมูลที่มีจำนวนมิติหรือตัวแปรมาก และทำให้เกิดปัญหาของมิติข้อมูล (curse of dimensionality) ซึ่งส่งผลต่อเวลาในการคำนวณที่มากขึ้น และยังทำให้ยากต่อการวิเคราะห์และการอภิปรายผล จากข้อเสนอของ Cook(1998) กล่าวว่าการสร้างกราฟเป็นสิ่งสำคัญสำหรับการสรุป ดังนั้นข้อมูลมิติต่ำจะทำให้ง่ายต่อการอภิปรายผลด้วยกราฟ

ในเวลาต่อมาจึงมีงานวิจัยที่ได้ศึกษาเกี่ยวกับเทคนิคที่ใช้ในการลดมิติข้อมูลมากมายหลายวิธี และได้พัฒนาเทคนิคการลดมิติข้อมูล เพื่อประสิทธิภาพในการวิเคราะห์ที่ดีขึ้น อาทิเช่น

วิธีการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) นำเสนอโดย Karl Pearson(1901) ซึ่งเป็นการแปลงเชิงเส้นของตัวแปรกลายเป็นโครงสร้างใหม่ที่ไม่มีความสัมพันธ์กัน (uncorrelated) โดยการศึกษาโครงสร้างความสัมพันธ์ของตัวแปร และสร้างตัวแปรใหม่เรียกว่า องค์ประกอบ กล่าวคือ สร้างเซตของตัวแปรใหม่ให้เป็นฟังก์ชันเชิงเส้นของตัวแปรเดิม

วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS) นำเสนอโดย Herman Wold ในปี 1980 ใช้ในการสร้างสมการความถดถอยเพื่อการพยากรณ์ โดยใช้หลักการสร้างตัวแปรใหม่จากตัวแปรอิสระเดิมและตัวแปรตาม ซึ่งจะพิจารณาการสร้างตัวแปรใหม่อย่างเป็นลำดับ และตัวแปรอิสระใหม่จะสร้างมาจากผลบวกเชิงเส้นของตัวแปรอิสระเดิม

วิธีSliced Inverse Regression(SIR) นำเสนอโดย Ker-Chau Li (1991) ซึ่งมีประโยชน์ในการหาทิศทางในปริภูมิที่ศูนย์กลาง(central space) ภายใต้เงื่อนไขของค่าเฉลี่ย โดยแทนที่การวิเคราะห์ความถดถอยเมื่อ Y แปรผันตาม X ด้วยการวิเคราะห์ความถดถอยผกผันเมื่อ X แปรผันตาม Y จะได้ว่าการวิเคราะห์ความถดถอยของพิกัด X กับ Y เป็น 1 มิติกับ ปัญหาความถดถอย 1 มิติ ดังนั้นไม่มีปัญหามิติสูงอีก และวิธี Sliced Average Variance Estimator (SAVE) นำเสนอโดย

Cook and Weisberg (1991) ซึ่งได้พัฒนาวิธี Sliced Inverse Regression (SIR) เนื่องจากวิธี Sliced Inverse Regression (SIR) ไม่เหมาะสมกับข้อมูลที่มี $E(X \mid Y) = 0$ ได้แสดงว่าเป็นได้ที่จะพบทิศทางในปริภูมิที่ศูนย์กลางโดยเพิ่มการพิจารณาค่าความแปรปรวน แทนการพิจารณาด้วยค่าเฉลี่ยเพียงอย่างเดียว

สำหรับขั้นตอนในการวิเคราะห์ข้อมูลด้วยวิธีการพยากรณ์จำแนกประเภทของตัวแบบนับเป็นเทคนิคหนึ่งที่ได้รับความนิยมนำไปใช้ประโยชน์ในหลายสาขา เช่น ทางด้านวิทยาศาสตร์ สังคมศาสตร์ ธุรกิจ ฯลฯ ซึ่งงานวิจัยครั้งนี้จะศึกษาวิธีการพยากรณ์จำแนกประเภทข้อมูลของตัวแบบที่ไม่ใช้พารามิเตอร์ (Nonparametric) เนื่องจากวิธีการพยากรณ์จำแนกประเภทของตัวแบบที่ใช้พารามิเตอร์ (Parametric) จะมีประสิทธิภาพก็ต่อเมื่อชุดข้อมูลที่นำมาวิเคราะห์มีลักษณะที่สอดคล้องกับข้อสมมติของตัวแบบทางสถิตินั้น ๆ แต่ในความเป็นจริง ข้อมูลโดยส่วนใหญ่มีลักษณะที่ไม่สอดคล้องกับข้อสมมติของตัวแบบทางสถิติ จึงทำให้ประสิทธิภาพในการพยากรณ์จำแนกประเภทลดลง ซึ่งวิธีการพยากรณ์จำแนกประเภทข้อมูลของตัวแบบที่ไม่ใช้พารามิเตอร์ (Nonparametric) ที่จะกล่าวถึงมีดังนี้

วิธีโครงข่ายประสาทเทียม (Artificial Neural Networks : ANN) มีลักษณะการทำงานแบบขนาน เลียนแบบการทำงานของมนุษย์ โดยอาศัยการนำเข้าข้อมูล เพื่อสร้างตัวแบบการจำลอง เพื่อใช้ในการ-พยากรณ์ข้อมูลในอนาคต แล้วทำการปรับปรุงข้อมูลให้มีเหมาะสมกับเงื่อนไขของข้อมูลที่มีการเปลี่ยน- แปลง โดยแนวความคิดนี้ จะพยายามลดจำนวนของการพยากรณ์เพื่อจำแนกประเภทให้ผิดพลาดต่ำที่สุด

วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) นำเสนอโดย Vapnik(1995) มีแนวความคิดที่มีการพัฒนาจากวิธีโครงข่ายประสาทเทียมแบบชั้นเดียว (Single – Layer Neural Networks) ซึ่งเป็นเทคนิคในการจัดประเภทข้อมูลมากกว่า 2 กลุ่ม ด้วยระนาบหลายมิติ โดยเป้าหมายของวิธีการซัพพอร์ตเวคเตอร์แมชชีน คือ การสร้างตัวแยกประเภทข้อมูล (Classifier) ที่มีความเป็นทั่วไป (Generalize) สูง ซึ่งตัวแยกประเภทข้อมูลที่ดีควรมีโครงสร้างแบบเชิงเส้น (Linear Classifier) ที่สร้างระยะห่างมากที่สุดระหว่างตัวแยกประเภทข้อมูลเองกับค่าที่ใกล้ที่สุดของแต่ละกลุ่มข้อมูล ซึ่งเส้นที่เหมาะสมดังกล่าวถูกเรียกว่า “The Optimal Separating Hyper plane” แต่ในความเป็นจริงแล้วข้อมูลที่ป้อนเข้าสู่เทคนิคมักเป็นข้อมูลแบบไม่เชิงเส้น ซึ่งสามารถแก้ปัญหานี้ได้โดยการนำเคอร์เนลฟังก์ชันมาใช้

นันทณัฐ พันธุ์สีดา (2013) ศึกษาการจำลองข้อมูลเพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ระหว่างวิธีโครงข่ายประสาทเทียมกับวิธีซัพพอร์ตเวกเตอร์แมชชีน จะได้ว่าวิธีซัพพอร์ตเวกเตอร์แมชชีนด้วย Laplacian Kernel ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนก

ประเภทของข้อมูลได้ดีที่สุดในกรณีที่ข้อมูลมีการแจกแจงแบบชี้กำลังและข้อมูลที่มีการแจกแจงแบบปกติ ส่วนกรณีที่ข้อมูลมีการแจงแจกแบบปัวส์ซงนั้น วิธีโครงข่ายประสาทเทียมแบบย้อนกลับ ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด

L.J. Caoa;* , K.S. Chuab , W.K. Chongc , H.P. Leea , Q.M. Gud(2003) ได้ศึกษาเปรียบเทียบเทคนิค PCA, KPCA และ ICA สำหรับการลดมิติข้อมูลในซัพพอร์ตเวกเตอร์แมชชีน โดยใช้ข้อมูล 3 ชุด คือ Sunspot data , Santa Fe data และ Financial data sets ในงานวิจัยครั้งนี้ปรากฏว่าวิธีการลดมิติข้อมูลเข้าก่อนการวิเคราะห์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพภาพมากกว่าการนำข้อมูลเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน และพบว่าการลดมิติข้อมูลด้วยวิธี KPCA ดีที่สุด และวิธี ICA รองลงมา และตามด้วยวิธี PCA

ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียร และพยุง มีสัจ (2009)ได้ศึกษาเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสม สำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น(BSFS) และการวิเคราะห์องค์ประกอบหลัก(PCA) เพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง จากประสิทธิภาพของเทคนิค BSFS แสดงให้เห็นว่ามีความเหมาะสมเมื่อนำมาเป็นอินพุตให้กับโครงข่ายประสาทเทียมมากกว่าเทคนิคการวิเคราะห์องค์ประกอบหลัก

จากงานวิจัยของนันทณัฐ พันธุ์สีดา (2013) ผู้วิจัยจะนำมาศึกษาต่อในการพยากรณ์จำแนกประเภทของข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนด้วย Laplacian Kernel สำหรับข้อมูลที่มีการแจกแจงแบบปกติหรือเกาส์เซียน และจากงานวิจัยของ ภัทราวุฒิ แสงศิริ และคณะ (2009) Hyunsoo Kim และคณะ(2005) ผู้วิจัยจึงเห็นความสำคัญของการลดมิติข้อมูลเข้าก่อนการพยากรณ์จำแนกประเภท เพื่อประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของตัวแบบที่ไม่ใช้พารามิเตอร์ของข้อมูล ดังนั้นในการวิจัยครั้งนี้จึงศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติสำหรับข้อมูลเข้าที่ใช้ในวิธีซัพพอร์ตเวกเตอร์แมชชีน : กรณีศึกษาการแจกแจงแบบเกาส์เซียน

1.2 วัตถุประสงค์

เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติข้อมูลเข้า(input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS)และ Sliced Average Variance Estimator(SAVE) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ในกรณีศึกษาการแจกแจงแบบเกาส์เซียน

1.3 ข้อตกลงเบื้องต้น

ศึกษาตัวแปรอิสระที่มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) โดยเวกเตอร์ตัวแปรอิสระ p มิติคือ $X' = (x_1, x_2, ..., x_p)$ ที่มีค่าเวกเตอร์ค่าเฉลี่ย p มิติเป็น μ และเมตริกซ์ความแปรปรวนร่วมขนาด p x p เป็น $\sum$ โดยฟังก์ชันความหนาแน่น เป็นดังนี้

$$f(X) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu)'\sum^{-1}(X-\mu)\right\}$$

ในการศึกษาครั้งนี้ได้ทำการศึกษาตัวแปรอิสระที่มีการแจกแจงแบบปกติหลายตัวแปรที่มีพารามิเตอร์ค่าเฉลี่ยคือ $\mu = (0,...,0)'$ และมีเมตริกซ์ความแปรปรวนร่วมตามระดับความสัมพันธ์ของตัวแปรอิสระที่กำหนดไว้

1.4 ขอบเขตของการวิจัย

การศึกษาครั้งนี้ได้ทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาเหตุการณ์เกิดขึ้นสองเหตุการณ์ (dichotomous) คือ เหตุการณ์ที่สนใจ (Group1) กับเหตุการณ์ที่ไม่สนใจ (Group2) เนื่องมาจากโดยหลักการพื้นฐานของวิธีซัพพอร์ตเวกเตอร์แมชชีน กำหนดให้ตัวแปรตาม (Y) เป็นข้อมูลเชิงกลุ่มที่อยู่ในระดับนามบัญญัติ (Nominal Scale) โดยตัวแปรตาม (Y) แบ่งเป็น 2 กลุ่ม คือ

$$Y = \begin{cases} 1 \; ; \; Group\,1 \\ -1 \; ; \; Group\,2 \end{cases}$$

2. จำนวนตัวแปรอิสระ (p) ที่ใช้ในการวิจัยมี 4 ระดับ คือ 5, 10, 20 และ 40 ตัวแปร ซึ่งมีการแจกแจงแบบปกติหลายตัวแปร ที่มีพารามิเตอร์ค่าเฉลี่ยคือ $\mu = (0,...,0)'$ และมีเมตริกซ์ความแปรปรวนร่วมตามระดับความสัมพันธ์ของตัวแปรอิสระดังนี้

เมตริกซ์ความแปรปรวนร่วม $(\sum)$ =

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{bmatrix}_{p\times p}$$

โดยที่มีค่าความแปรปรวนของตัวแปรอิสระแต่ละตัวเป็น 1 และมีความสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และ j หรือ $\rho_{ij} = \rho^{|i-j|}$ เมื่อ $i, j = 1, 2, ..., p$

กำหนดให้ ρ มีค่าเป็น 0.85, 0.9 และ 0.95 จะได้ว่า

<u>กรณีตัวแปรอิสระ 5 ตัว (p=5)</u>

- $\rho = 0.85$ จะได้ว่า เมตริกซ์ความแปรปรวนร่วม คือ

$$\begin{bmatrix} 1 & 0.85^{|1-2|} & 0.85^{|1-3|} & 0.85^{|1-4|} & 0.85^{|1-5|} \\ 0.85^{|2-1|} & 1 & 0.85^{|2-3|} & 0.85^{|2-4|} & 0.85^{|2-5|} \\ 0.85^{|3-1|} & 0.85^{|3-2|} & 1 & 0.85^{|3-4|} & 0.85^{|3-5|} \\ 0.85^{|4-1|} & 0.85^{|4-2|} & 0.85^{|4-3|} & 1 & 0.85^{|4-5|} \\ 0.85^{|5-1|} & 0.85^{|5-2|} & 0.85^{|5-3|} & 0.85^{|5-4|} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.85 & 0.72 & 0.61 & 0.52 \\ 0.85 & 1 & 0.85 & 0.72 & 0.61 \\ 0.72 & 0.85 & 1 & 0.85 & 0.72 \\ 0.61 & 0.72 & 0.85 & 1 & 0.85 \\ 0.52 & 0.61 & 0.72 & 0.85 & 1 \end{bmatrix}$$

ดังนั้น $\rho_{ij} \in [0.52, 0.85]$

- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.65, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.81, 0.95]$

<u>กรณีตัวแปรอิสระ 10 ตัว (p=10)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.23, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.38, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.63, 0.95]$

<u>กรณีตัวแปรอิสระ 20 ตัว (p=20)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.04, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.13, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.37, 0.95]$

<u>กรณีตัวแปรอิสระ 40 ตัว (p=40)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.001, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.016, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.135, 0.95]$

3. ศึกษาภายใต้ขนาดตัวอย่าง ซึ่งกำหนดตัวอย่างของกลุ่มที่หนึ่งเป็นกลุ่มตัวอย่างที่สนใจ (n_1) และกลุ่มที่สองเป็นกลุ่มตัวอย่างที่ไม่สนใจ (n_2) ซึ่งสามารถกำหนดขนาดของจำนวนตัวอย่างดังนี้ $n_s = 30, 60, 120 \;\; ; \;\; s = 1, 2$

4. ศึกษาลักษณะของการแจกแจงข้อมูล โดยมีการศึกษาตัวอย่างที่มีตัวแปรอิสระ 5, 10, 20 และ 40 ตัวแปร ซึ่งมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) โดยตัวแปรตาม (Y) แบ่งเป็น 2 กลุ่ม คือ

กลุ่มที่สนใจ $(Y = 1)$ ซึ่งเป็นกลุ่มที่มีเกณฑ์กำหนดที่ชัดเจน โดยที่ค่าของพารามิเตอร์ของการแจกแจงข้อมูลเป็นค่าคงที่

กลุ่มที่ไม่สนใจ $(Y=-1)$ ซึ่งเป็นกลุ่มที่มีลักษณะการแจกแจงของข้อมูลแตกต่างกับกลุ่มที่สนใจ โดยที่ค่าของพารามิเตอร์ของการแจกแจงข้อมูลมีการเปลี่ยนแปลงตามค่าของ d

ทำการวิเคราะห์โดยเปลี่ยนค่าของ d ไปเรื่อย ๆ ตามค่าที่กำหนดจนครบสำหรับการแจกแจงที่ทำการศึกษา ซึ่งสามารถอธิบายได้ว่า เมื่อค่าของ d มีค่าเพิ่มมากขึ้น จะทำให้กลุ่มตัวอย่างของกลุ่มที่สนใจกับกลุ่มที่ไม่สนใจมีลักษณะการแจกแจงของข้อมูลแตกต่างกันเพิ่มมากขึ้นหรือสามารถอธิบายความแตกต่างระหว่างข้อมูลทั้งสองกลุ่มได้ชัดเจนเพิ่มมากขึ้น ในทางตรงข้ามเมื่อ d มีค่าลดน้อยลง จะทำให้ลักษณะการแจกแจงของทั้งสองกลุ่มใกล้เคียงกันมากขึ้นหรืออธิบายความแตกต่างได้น้อยลง

5. ในการศึกษาครั้งนี้ทำการจำลองข้อมูลให้มีสถานการณ์ที่แตกต่างกัน ตามข้อกำหนดข้างต้นโดยใช้เทคนิคมอนติคาร์โล (Monte Carlo Simulation Technique) โดยทำการจำลองในแต่ละสถานการณ์จะกระทำซ้ำ 500 รอบ

1.5 คำจำกัดความที่ใช้ในการศึกษา

1. วิธีการวิเคราะห์องค์ประกอบหลัก (Principle Component Analysis, PCA) ซึ่งเป็นการแปลงเชิงเส้นของตัวแปรกลายเป็นโครงสร้างใหม่ที่ไม่มีความสัมพันธ์กัน (uncorrelated) โดยการศึกษาโครงสร้างความสัมพันธ์ของตัวแปร และสร้างตัวแปรใหม่เรียกว่า องค์ประกอบ

2. วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS) โดยหลักการสร้างตัวแปรใหม่จากตัวแปรอิสระเดิมและตัวแปรตาม ซึ่งจะพิจารณาการสร้างตัวแปรใหม่อย่างเป็นลำดับ และตัวแปรอิสระใหม่จะสร้างมาจากผลบวกเชิงเส้นของตัวแปรอิสระเดิม

3. วิธี Sliced Average Variance Estimator (SAVE) ซึ่งมีประโยชน์ในการหาทิศทางในปริภูมิที่ศูนย์กลาง (central space) โดยแทนที่การวิเคราะห์ความถดถอยเมื่อ Y แปรผันตาม X ด้วยการวิเคราะห์ความถดถอยผกผันเมื่อ X แปรผันตาม Y จะได้ว่าการวิเคราะห์ความถดถอยของพิกัด X กับ Y เป็น 1 มิติกับปัญหาความถดถอย 1 มิติ

4. วิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) เป็นเทคนิคในการจัดประเภทข้อมูลมากกว่า 2 กลุ่ม ด้วยระนาบหลายมิติ โดยเป้าหมายของวิธีการซัพพอร์ตเวคเตอร์แมชชีน คือ การสร้างเส้นแบ่งแยกประเภทของข้อมูลที่ดีที่สุด (Optimal Separating Hyper plane)

5. ค่าพื้นที่ใต้โค้งอาร์โอซี (Area Under ROC Curve) คือ ค่าที่อธิบายความสามารถในการจำแนกประเภทของข้อมูลหรือความเชื่อถือได้ของตัวแบบกรณีที่มีเหตุการณ์เกิดขึ้น 2 เหตุการณ์

1.6 วิธีดำเนินการวิจัย

1. กำหนดเงื่อนไขและขอบเขตของการวิจัย
   - จำนวนตัวแปรอิสระ (p)
   - กำหนดค่าพารามิเตอร์ตามการแจกแจงที่กำหนดในขอบเขตของการศึกษา (d)
   - กำหนดขนาดตัวอย่าง (n)
   - ค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ (ρ)
2. จำลองข้อมูลตามการแจกแจงและขอบเขตที่ต้องการศึกษา

   • จำลองค่า X ตามการแจกแจงของข้อตกลงเบื้องต้น และจำลองค่า y เป็นข้อมูลเชิงกลุ่มที่อยู่ในระดับนามบัญญัติ (Nominal Scale) คือ

$$y = \begin{cases} 1 & ; X \sim N\left(\mu = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \Sigma\right) \\ -1 & ; X \sim N\left(\mu = \begin{bmatrix} d \\ \vdots \\ d \end{bmatrix}, \Sigma\right) \end{cases}$$

3. ลดตัวแปรข้อมูลหรือลดมิติข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลัก(PCA), วิธีกำลังสองน้อยสุดเชิงส่วน (PLS) และ Sliced Average Variance Estimator(SAVE) โดยมีเกณฑ์ในการเลือกมิติข้อมูลดังนี้

   • **การวิเคราะห์องค์ประกอบหลัก(PCA)** จะเลือกมิติของตัวแปรอิสระใหม่หรือจำนวนองค์ประกอบหลัก โดยพิจารณาจากร้อยละสัดส่วนของความแปรปรวนสะสม ถ้าร้อยละความแปรปรวนสะสมของตัวประกอบหลัก k ตัวแรก เป็นอย่างต่ำร้อยละ 80 ก็ควรให้จำนวนตัวประกอบหลักเท่ากับ k โดยที่ k < p

   • **วิธีกำลังสองน้อยสุดเชิงส่วน(PLS)** จะเลือกมิติของตัวแปรอิสระใหม่หรือจำนวนองค์ประกอบหลัก พิจารณาจากการประเมินความน่าเชื่อถือได้ของการวิเคราะห์ความถดถอย โดยพิจารณาค่าสถิติ Adjusted Cross Validation (adjCV) ของวิธี cross-validation หรือ Leave-One-Out cross-Validation ดังนั้นการเปรียบเทียบค่า adjCV ของการวิเคราะห์ที่มีจำนวนองค์ประกอบต่างกัน ถ้าค่า adjCV ของการวิเคราะห์ที่มีจำนวนองค์ประกอบใดๆมีค่าน้อยสุดจะเลือกจำนวนองค์ประกอบดังกล่าวเป็นจำนวนองค์ประกอบที่เหมาะสม

   • **Sliced Average Variance Estimator (SAVE)** จะเลือกมิติของตัวแปรอิสระใหม่จำนวน k มิติ โดยใช้หลักการของ Marginal dimension hypothesis สำหรับการทดสอบสมมติฐานว่ามิติ $k = q$ โดยที่ $k < p$ และ $q = 0,1,2,...,p$ ซึ่งมีรายละเอียดดังนี้

กำหนด : $\hat{A}_i = \sqrt{\frac{n_i}{n}}\left(I_p - \mathrm{var}(Z \mid Y \in I_i)\right)$ เมื่อเมตริกซ์ความแปรปรวนร่วมของน้ำหนัก $\hat{M} = \sum_{i=1}^{H} \hat{A}_i^2$ โดยที่ H เป็นจำนวนส่วน(slice)แบ่งลำดับของ Y และ $\hat{m}$ เป็นเมตริกซ์ขนาด $p \times (p-m)$ ประกอบด้วยหลักที่เป็นเวกเตอร์เจาะจง(eigenvector) ของ $\hat{M}$ ซึ่งสอดคล้องกับค่าเฉพาะ(eigenvalue) $p-m$ ค่า

ผลลัพธ์ : $k = q$

1) เริ่มต้นให้ $q = 0$

2) สมมติฐาน $\mathrm{H}_0 : k = q$ vs. $\mathrm{H}_0 : k > q$

3) สถิติทดสอบ $T_n(\hat{m}) = \frac{n}{2}\sum_{i=1}^{H} tr\left\{\left(\hat{m}^T \hat{A}_i \hat{m}\right)^2\right\}$ มีการแจกแจงแบบไคสแควร์ที่มีองศาอิสระเท่ากับ $(H-1)(p-q)(p-q+1)/2$

4) ให้ $\alpha = 0.05$ จะปฏิเสธ H_0 เมื่อ $T_n(\hat{m}) > \chi^2_{df=(H-1)(p-q)(p-q+1)/2}$ จะกลับไปทำข้อ 1) โดยให้ $k = q+1$ และเมื่อผลการทดสอบยอมรับ H_0 จะได้จำนวนมิติที่เหมาะสม q มิติ

4. ทำการประมาณค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนเพื่อสร้างตัวแบบสำหรับนำมาพยากรณ์
5. นำตัวแบบที่ใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนด้วย Laplacian Kernel แล้วนำตัวแบบที่ได้ไปพยากรณ์ข้อมูลต่อไป เพื่อตรวจสอบผลของการพยากรณ์เทียบกับเหตุการณ์ที่เกิดขึ้นจริง และนำผลลัพธ์ของการพยากรณ์ที่ได้ไปสร้างตารางการแบ่งกลุ่ม
6. นำข้อมูลที่ได้ไปคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC (Area Under ROC Curve : AUC) และคำนวณค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate : MCR)
7. วิเคราะห์และสรุปผลการเปรียบเทียบวิธีการที่ใช้ในการวิจัย

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถใช้เทคนิคการลดมิติข้อมูลนำเข้า(input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก วิธีกำลังสองน้อยสุดเชิงส่วน และ Sliced Average Variance Estimator สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน เพื่อให้มีประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด สำหรับข้อมูลที่มีการแจกแจงแบบเกาส์เซียน

2. เป็นแนวทางเพื่อเลือกใช้เทคนิคการลดมิติข้อมูลนำเข้าระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก วิธีกำลังสองน้อยสุดเชิงส่วน และ Sliced Average Variance Estimator สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน เพื่อให้เหมาะสมกับข้อมูลจริง

บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

2.1 การวิเคราะห์องค์ประกอบหลัก(PCA)

การวิเคราะห์องค์ประกอบหลัก เป็นเทคนิคหนึ่งที่ใช้ในการลดตัวแปร p ตัว โดยสร้างตัวแปรใหม่ที่เรียกว่า องค์ประกอบหลัก(ปัจจัย) k ตัว ที่มีความสัมพันธ์เชิงเส้นของตัวแปรเดิม p ตัว ดังนั้นจำนวนองค์ประกอบหลักจะต้องมีจำนวนไม่เกินจำนวนตัวแปร กล่าวคือ $k \le p$ โดยอ้างอิงจาก กัลยา วานิชย์บัญชา (2551)

ขั้นตอนการคำนวณของวิธีการวิเคราะห์องค์ประกอบหลัก

ให้ $X' = (x_1, x_2, ..., x_p)$ ซึ่งมีเมตริกซ์ความแปรปรวนร่วม $\sum$ ที่มีค่าเจาะจง (eigenvalue) และเวกเตอร์เจาะจง(eigenvector) เป็น $(\lambda_1, w_1), (\lambda_2, w_2), ..., (\lambda_p, w_p)$ โดยที่ค่าเจาะจง $\lambda_1 \ge \lambda_2 \ge ... \ge \lambda_p \ge 0$ และเวกเตอร์เจาะจง $W' = (w_1, w_2, ..., w_p)$ สร้างสมการที่ใช้ประมาณค่าองค์ประกอบหลักหรือปัจจัยตัวที่ i (PC_i) เมื่อ $i = 1, ..., p$ ดังนี้

$$\begin{aligned} PC_1 &= w_1' X = w_{11}x_1 + w_{12}x_2 + ... + w_{1p}x_p \\ PC_2 &= w_2' X = w_{21}x_1 + w_{22}x_2 + ... + w_{2p}x_p \\ &\vdots \quad\quad \vdots \quad\quad \vdots \\ PC_p &= w_p' X = w_{p1}x_1 + w_{p2}x_2 + ... + w_{pp}x_p \end{aligned} \qquad (2.1.1)$$

โดยเริ่มจาก (PC_1) ที่มีค่าความแปรปรวนสูงสุดไปจนถึง (PC_p) ที่มีค่าความแปรปรวนต่ำสุด หรือ $Var(PC_1) \ge Var(PC_2) \ge ... \ge Var(PC_p)$ จะได้ว่า

$$\sum_{i=1}^{p} Var(x_i) = \sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} Var(PC_i)$$

$\therefore PC_k$ มีสัดส่วนของค่าแปรปรวน $= \dfrac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$; $k = 1, 2, ..., p$

2.2 วิธีกำลังสองน้อยสุดเชิงส่วน (PLS)

วิธีกำลังสองน้อยสุดเชิงส่วน อ้างอิงจาก ปราณี คำแก้ว (2552)ใช้ในการสร้างสมการความถดถอยเพื่อการพยากรณ์ โดยใช้หลักการสร้างตัวแปรใหม่จากตัวแปรอิสระเดิมและตัวแปรตาม ซึ่งจะพิจารณาการสร้างตัวแปรใหม่อย่างเป็นลำดับ และตัวแปรอิสระใหม่จะสร้างมาจากผลบวกเชิงเส้นของตัวแปรอิสระเดิม โดยไม่ได้นำสหสัมพันธ์ระหว่างตัวแปรอิสระเดิมมาพิจารณา และเป็นวิธีที่สามารถแก้ไขปัญหาการประมาณค่าพารามิเตอร์เมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ แต่ก็ยังเป็นประมาณที่มีความเอนเอียงอยู่ ซึ่งข้อดีของวิธีกำลังสองน้อยสุดเชิงส่วน คือ ทำให้สหพันธ์ของตัวแปรอิสระใหม่แต่ละคู่มีค่าเป็นศูนย์ เนื่องจากตัวแปรอิสระใหม่จากส่วนเหลือของตัวแปรเดิม โดยสามารถเขียนตัวแบบความถดถอยของวิธีกำลังสองน้อยสุดเชิงส่วนได้ดังนี้

$$\hat{Y} = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + ... + \beta_p P_p \qquad (2.2.1)$$

โดยที่ P_j เป็นตัวแปรอิสระตัวใหม่ที่เป็นผลบวกเชิงเส้นของ X_j เมื่อ j= 1, ..., p และค่าสหสัมพันธ์พันธ์ตัวอย่างสำหรับแต่ละคู่ของตัวแปรอิสระใหม่มีค่าเท่ากับศูนย์

การวิเคราะห์วิธี PLS จะพิจารณาตัวแบบ ดังนี้

$$X = T \cdot P^T + E$$

$$y = T \cdot Q^T + F$$

โดยที่ $X = \left(x_{ij}\right) \in \mathbb{R}^{n \times p}, i = 1,...,n, j = 1,...,p$ คือ เมตริกซ์ของตัวแปรอิสระ

$T = [t_1,...,t_K] \in \mathbb{R}^{n \times K}$ คือ เมตริกซ์ Score ของตัวแปรอิสระ และตัวแปรตาม

$P = [p_1,...,p_K] \in \mathbb{R}^{p \times K}$ คือ เมตริกซ์ loading ของตัวแปรอิสระ

$E = [e_1,...,e_K] \in \mathbb{R}^{n \times p}$ คือ เมตริกซ์ของความคลาดเคลื่อนที่เกิดจากตัวแปรอิสระ

$y = \left(y_i\right) \in \mathbb{R}^{n \times 1}, i = 1,...,n$ คือ เวกเตอร์ของตัวแปรตาม

$Q = [q_1,...,q_K] \in \mathbb{R}^{1 \times K}$ คือ เวกเตอร์ loading ของตัวแปรตาม

$F = [f_1,...,f_n] \in \mathbb{R}^{n \times 1}$ คือ เวกเตอร์ของความคลาดเคลื่อนที่เกิดจากตัวแปรตาม

กระบวนการสำหรับการวิเคราะห์ของวิธีกำลังสองน้อยสุดเชิงส่วน

ขั้นตอนในการหาองค์ประกอบหลักของ PLS ตัวที่หนึ่ง

1. คำนวณเวกเตอร์ $t_1 = Xw_1$ เมื่อ w_1 เป็นเวกเตอร์เจาะจง(eigenvector) ที่สอดคล้องกับค่าเจาะจง(eigenvalue)ของ $X'YY'X$
2. คำนวณเมตริกซ์ loading ของตัวแปรอิสระ จากสูตร $p_1' = (t_1't_1)^{-1} t_1'X$
3. คำนวณเวกเตอร์ loading ของตัวแปรตาม จากสูตร $q_1' = (t_1't_1)^{-1} t_1'Y$

เมื่อสิ้นสุดขั้นตอนที่ 3. จะได้ตัวประกอบของ PLS ตัวที่หนึ่ง P_1

ขั้นตอนในการระบุค่าความคลาดเคลื่อน

4. คำนวณค่าเมตริกซ์ของความคลาดเคลื่อนที่เกิดจากตัวแปรอิสระ จากสูตร $E = X - T \cdot P'$
5. คำนวณค่าเวกเตอร์ของความคลาดเคลื่อนที่เกิดจากตัวแปรตาม จากสูตร $F = Y - T \cdot Q'$

ดังนั้นถ้าต้องการวิเคราะห์ตัวประกอบหลักตัวถัดไปให้ย้อนกลับไปเริ่มทำในขั้นตอนที่ 1 ใหม่

เกณฑ์ในการพิจารณาจำนวนตัวประกอบหลักที่เหมาะสมมีดังนี้

พิจารณาจากค่าการตรวจสอบความถูกต้อง Cross-validation หรือบางครั้งเรียกว่า rotation estimation (Devijver, 1982) เป็นการแบ่งตัวอย่างของข้อมูลออกเป็นกลุ่มย่อย เริ่มต้นการทำงานจากกลุ่มย่อยเดี่ยว ขณะที่กลุ่มย่อยอื่นๆถูกเก็บไว้สำหรับใช้ต่อมา เพื่อตรวจสอบความถูกต้องในการวิเคราะห์ กลุ่มย่อยเริ่มต้นของข้อมูลเรียกว่า กลุ่มสร้างตัวแบบ ส่วนกลุ่มย่อยอื่นๆ เรียกว่ากลุ่มตรวจสอบความถูกต้อง (Devijver, 1982และ Tutorial, 2006) หรือกลุ่มทดสอบ ชนิดของการตรวจสอบความถูกต้องที่ใช้ในงานวิจัยครั้งนี้คือ

- **Leave-one-out cross-validation (LOOCV)**

Leave-one-out cross-validation เป็นวิธีการที่ข้อมูลดั้งเดิมถูกแบ่งเป็น N กลุ่มย่อย (N เป็นจำนวนตัวอย่าง) จาก N กลุ่มย่อย กลุ่มย่อยหนึ่งกลุ่มนำไปใช้เป็นกลุ่มตรวจสอบความถูกต้องเพื่อทดสอบโมเดล และเก็บ N-1 กลุ่มย่อย ใช้เป็นกลุ่มสร้างตัวแบบ วิธีการนี้เป็นการทำซ้ำโดยที่จะเปลี่ยนกลุ่มตรวจสอบความถูกต้องไปเรื่อยๆ จนครบทั้งหมด N ครั้ง ข้อดีคือข้อมูลสามารถใช้เพื่อสร้างตัวแบบโดยมีกลุ่มตรวจสอบหลากหลาย และสามารถทำการพยากรณ์ LOO ซึ่งง่ายกว่าการพยากรณ์ปกติ

2.3 Sliced Inverse Regression(SIR)

Sliced Inverse Regression ได้เสนอโดย Li (1991)ซึ่งมีประโยชน์ในการหาทิศทางในปริภูมิที่ศูนย์กลาง(central space) โดยแทนที่การวิเคราะห์ความถดถอยเมื่อ Y แปรผันตาม X ด้วยการวิเคราะห์ความถดถอยผกผันเมื่อ X แปรผันตาม Y จะได้ว่าการวิเคราะห์ความถดถอยของพิกัด X กับ Y เป็น 1 มิติกับ ปัญหาความถดถอย 1 มิติ ดังนั้นไม่มีปัญหามิติสูงอีก

สมมติฐาน 2.3.1 ให้ $(\beta_1,...,\beta_k)$ เป็นฐานหลัก(basis) ของ $S_{Y|X}$ สมมติว่า $E\left(X \mid \beta_1^T X,...,\beta_k^T X\right)$ เป็นเชิงเส้นใน $X \mid \beta_1^T X,...,\beta_k^T X$

ผลลัพธ์ที่ศูนย์กลางภายใต้สมมติฐาน 2.3.1 ค่าเฉลี่ยแบบมีเงื่อนไขเชิงเส้นและการวิเคราะห์ความถดถอยผกผัน แนวโค้ง $\sum^{-1}\left[E\left(X \mid Y\right)-E\left(X\right)\right]$ เป็นของ $S_{Y|X}$ เมื่อ $\sum$ เป็นเมตริกซ์ความแปรปรวนของ X

ให้ Z เป็นค่าที่ได้ปรับมาตรฐาน(standardized)ของ X

$$Z=\sum^{-1/2}\left[X-E\left(X\right)\right] \tag{2.3.1}$$

ทฤษฎีบท 2.3.2 ให้ X ได้ปรับค่ามาตรฐานเป็น Z ดังนั้นภายใต้สมมติฐาน 2.3.1ค่าเฉลี่ยแบบมีเงื่อนไขเชิงเส้นและการวิเคราะห์ความถดถอยผกผันที่ศูนย์กลาง เส้นโค้ง $E\left(Z \mid Y\right)$ เป็นของ $S_{Y|Z}$

กระบวนการสำหรับ Sliced Inverse Regression

ให้ $\left(X_1,Y_1\right),...,\left(X_n,Y_n\right)$ เป็นข้อมูลตัวอย่างที่เป็นอิสระกันจากการสุ่ม $\left(X,Y\right)$

1. คำนวณค่าเฉลี่ย(mean) และเมตริกซ์ความแปรปรวนของค่าพยากรณ์ X

$$\hat{\mu}=n^{-1}\sum_{i=1}^{n}X_i,\hat{\Sigma}=n^{-1}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)\left(X_i-\hat{\mu}\right)^T \tag{2.3.2}$$

2. ปรับค่ามาตรฐาน X_i เป็น Z_i โดย $Z_i=\sum^{-1/2}\left[X_i-\hat{\mu}\right]$
3. แบ่งลำดับของ Y เป็น H ส่วน(slice) กล่าวคือ $I_1,...,I_H$ และคำนวณค่าเฉลี่ยตัวอย่างของ Z ของแต่ละส่วน $\hat{\mu}_i=\frac{1}{n_i}\sum_{j\in I_i}Z_j$

   เมื่อ n_i คือจำนวนของ Y_i ที่อยู่ในส่วนที่ I_i

4. โครงสร้างเมตริกซ์ความแปรปรวนร่วมของน้ำหนัก (the weighted covariance matrix)

$$\hat{V} = \sum_{i=1}^{H} \frac{n_i}{n} \hat{\mu}_i \hat{\mu}_i^T$$

5. ให้ $v_1,...,v_p$ เป็นเวกเตอร์เจาะจง(eigenvector) ของ V ซึ่งสอดคล้องกับค่าเจาะจง (eigenvalue) $\lambda_1, \lambda_2,..., \lambda_p$

เนื่องจากการวิจัยครั้งนี้ศึกษาเหตุการณ์ที่เกิดขึ้นสองเหตุการณ์ (dichotomous) คือ เหตุการณ์ที่สนใจ (Group1) กับเหตุการณ์ที่ไม่สนใจ (Group2) ดังนั้นแบ่งข้อมูลได้เป็น 2 ส่วน(slice) จากการศึกษาไม่สามารถหาจำนวนมิติใหม่ที่เหมาะสมได้

2.4 Sliced Average Variance Estimator(SAVE)

เนื่องจากวิธี Sliced Inverse Regression (SIR) ไม่เหมาะสมกับข้อมูลที่มี $E(X \mid Y) = 0$ ดังนั้น Cook and Weisberg (1991) ได้แสดงว่าเป็นได้ที่จะพบทิศทางโดยการใช้ค่าสถิติมากกว่า 1 ค่าในการพิจารณาด้วยค่าเฉลี่ยเพียงอย่างเดียว

สมมติฐาน 2.4.1 ให้ $(\beta_1,...,\beta_k)$ เป็นฐานหลัก(basis) ของ $S_{Y|X}$ สมมติว่า $\mathrm{var}(X \mid \beta_1 X,...,\beta_k X)$ เป็นเมตริกซ์คงที่

ทฤษฎีบท 2.4.2 ให้ I_p เป็นเมตริกซ์เอกลักษณ์ของ p มิติ ถ้าสมมติฐาน 2.3.1ค่าเฉลี่ยสมมติฐานเชิงเส้นและสมมติฐาน 2.4.1 ความแปรปรวนแบบมีเงื่อนไขคงที่ และ Z เป็นค่าที่ได้ปรับมาตรฐาน (standardized)ของ X ดังนั้นหลักของเมตริกซ์ $I_p - \mathrm{var}(X \mid Y)$ เป็นส่วนของ $S_{Y|X}$ จะได้ว่าหลักของเมตริกซ์ $E\left[I_p - \mathrm{var}(X \mid Y)\right]^2$ เป็นส่วนของ $S_{Y|X}$ ด้วย

กระบวนการสำหรับ Sliced Average Variance Estimator(SAVE)

ขั้นที่1 -3 เหมือนวิธี Sliced Inverse Regression(SIR)

4.โครงสร้างเมตริกซ์ความแปรปรวนร่วมของน้ำหนัก (weighted covariance matrix)

$$\hat{M} = \sum_{i=1}^{H} \frac{n_i}{n} \left(I_p - \mathrm{var}(Z \mid Y \in I_i)\right)^2 \qquad (2.4.1)$$

5. การประมาณค่ามิติสำหรับวิธีการลดมิติ SAVE โดยงานวิจัยของ Shao, Cook and Weisberg (2007) ได้เสนอวิธี Marginal dimension tests

2.5 Marginal dimension hypothesis

การทดสอบสำหรับ Marginal dimension hypothesis สมมติฐานการทดสอบว่ามิติ $k = q$ โดยที่ $k < p$ ให้ $\hat{m}$ เป็นเมตริกซ์ซึ่งประกอบด้วยหลักที่เป็นเวกเตอร์เจาะจง (eigenvector) ของ $\hat{M}$ โดยที่ $\hat{M} = \sum_{i=1}^{H} \frac{n_i}{n}\left(I_p - \text{var}(Z \mid Y \in I_i)\right)^2 = \sum_{i=1}^{H} \hat{A}_i^2$ เมื่อ $\hat{A}_i = \sqrt{\frac{n_i}{n}}\left(I_p - \text{var}(Z \mid Y \in I_i)\right)$ ซึ่งสอดคล้องกับค่าเจาะจง (eigenvalue) โดย Tyler (1981)ให้นิยามการทดสอบทางสถิติ คือ $T_n(\hat{m}) = \frac{n}{2}\sum_{i=1}^{H} tr\left\{\left(\hat{m}^T \hat{A}_i \hat{m}\right)^2\right\}$

สมมติฐาน 2.5.1 สำหรับทุก $\beta \neq 0$ และ $\beta \in S_{Y|X}$ แล้ว $\text{var}\left\{E\left(\beta^T X \mid Y\right)\right\} > 0$ หรือ $\text{var}\left\{\text{var}\left(\beta^T X \mid Y\right)\right\} > 0$

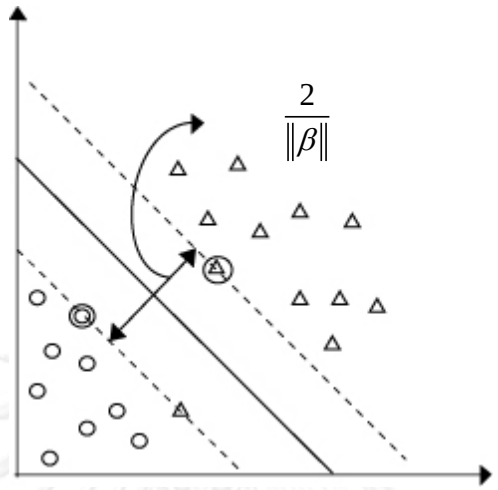
ทฤษฎีบท 2.5.2 สมมติว่า สมมติฐาน 2.3.1, 2.3.2 และ 2.4.1 เป็นจริง และ $\text{var}\left(m^T Z \otimes m^T Z \mid \gamma^T Z\right)$ ไม่เป็นอย่างสุ่ม (เช่น คงที่) ถ้า X มีการแจกแจงแบบปกติแล้วภายใต้สมมติฐาน $d = q$ ค่าสถิติทดสอบคือ $T_n(\hat{m}) \to \chi^2_{(H-1)(p-q)(p-m+1)/2}$

2.6 ซัพพอร์ตเวกเตอร์แมชชีน (SVM)

เป็นวิธีการที่สามารถนำมาใช้ในการจำแนกรูปแบบหรือกลุ่มของข้อมูลได้ โดยจะสร้างเส้นแบ่ง (Plane) ที่เป็นเส้นตรงขึ้นมา ในการแบ่งเขตของข้อมูลออกเป็นสองฝั่ง โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด (Optimal Separating Hyper plane) อ้างอิงจาก นันทณัฐ พันธุ์สีดา (2556) ซึ่งต้องมีคุณสมบัติของเงื่อนไขดังนี้

1. ค่าความผิดพลาดในการปฏิบัติเป็นศูนย์ (Zero Training Error)
2. ระยะระหว่างซัพพอร์ตเวกเตอร์ของทั้ง 2 ชนิดห่างกันมากที่สุด (Maximum Margin)

<u>ภาพที่ 2.6.1</u> แสดงลักษณะหลักการหาระนาบเส้นแบ่งแยกประเภทของข้อมูลที่ดีที่สุดด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน



กำหนดให้ลักษณะของข้อมูลเป็น $(\vec{x}_i, y_i)$ ซึ่ง $\vec{x}_i \in R^p$ โดยที่ $i = 1, 2, \ldots, p$ และ $y_i \in \{-1, 1\}$ สร้างเส้นตรงบนไฮเปอร์เพลน (Hyper – plane) ซึ่งแบ่งกลุ่มข้อมูลที่มีลักษณะเชิงเส้นสองกลุ่มออกจากกัน วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือ การเพิ่มเส้นขอบ (Margin)ให้กับเส้นแบ่งทั้งสองข้าง เส้นขอบของทั้งสองเส้นที่เพิ่มขึ้นมานี้ถูกแทนด้วยสมการ

$$\beta' X + \beta_0 \geq +1 \quad for \quad y_i = 1 \tag{2.6.1}$$

และ
$$\beta' X + \beta_0 \leq -1 \quad for \quad y_i = -1 \tag{2.6.2}$$

เมื่อ β คือ ค่าความชัน

X คือ เวกเตอร์ข้อมูล

β_0 คือ ค่าคงที่ (ค่าตัดแกน y)

โดยสามารถนำสมการมาเขียนรวมกัน ได้ดังนี้

$$y_i(\beta' x_i + \beta_0) \geq 1 \quad ; \quad \forall i \tag{2.6.3}$$

การหาระนาบสำหรับการแบ่งกลุ่มที่เหมาะสมที่สุด ทำได้โดยการหาค่าระยะ $(\text{Distance}; d(\beta, \beta_0; x))$ จากตำแหน่งของซัพพอร์ตเวกเตอร์ x ถึงระนาบ (β, β_0) ได้ดังนี้

$$d(\beta, \beta_0; x) = \frac{|\langle \beta, x_i \rangle + \beta_0|}{\|\beta\|} \tag{2.6.4}$$

โดยวิเคราะห์จากค่าของระยะขอบที่มากที่สุด (Maximum margin) ซึ่งสามารถหาได้ดังนี้

$$
\begin{aligned}
p(\beta,\beta_0) &= \min_{x_i,y_i=-1} d(\beta,\beta_0;x_i) + \min_{x_i,y_i=+1} d(\beta,\beta_0;x_i) \\
&= \min_{x_i,y_i=-1} \frac{\left|\langle \beta, x_i \rangle + \beta_0\right|}{\|\beta\|} + \min_{x_i,y_i=+1} \frac{\left|\langle \beta, x_i \rangle + \beta_0\right|}{\|\beta\|} \\
&= \frac{1}{\|\beta\|}\left(\min_{x_i,y_i=-1} \left|\langle \beta, x_i \rangle + \beta_0\right| + \min_{x_i,y_i=+1} \left|\langle \beta, x_i \rangle + \beta_0\right| \right) \\
&= \frac{2}{\|\beta\|}
\end{aligned}
$$

สำหรับการหาสัมประสิทธิ์ที่ดีที่สุด ในการลดระยะทางดังกล่าวได้ดังนี้

$$\text{Minimize } \Phi(\beta,\beta_0) = \frac{1}{2}\|\beta^2\| \qquad (2.6.5)$$

$$\textit{Subject to} \quad y_i(\beta' x_i + \beta_0) \geq 1 \quad ; \quad \forall i$$

กรณีที่ไม่สามารถแยกข้อมูลได้ด้วยไฮเปอร์เพลน (Hyper – plane)

เราสามารถหาได้เพียงไฮเปอร์เพลนที่สามารถแยกจุดตัวอย่างออกจากกันให้ได้มากที่สุด และยอมให้มีจุดตัวอย่างส่วนน้อยเพียงบางจุดที่ผิดพลาด โดยที่ข้อผิดพลาดที่เกิดขึ้นในกรณีที่ไม่สามารถทำการแบ่งกลุ่มของข้อมูลได้ กำหนดให้ $\xi = (\xi_1, \xi_2, \ldots, \xi_n)^T$ แทน เวกเตอร์ของตัวแปรที่ทำให้เกิดความผิดพลาดในการแบ่งข้อมูล (Slack Variables)

นำมาเขียนเป็นสมการตามเงื่อนไขใหม่ได้ดังนี้

$$\text{Minimize } \Phi(\beta,\beta_0) = \frac{1}{2}\|\beta^2\| + \frac{1}{2}C\sum_{i=1}^{n}\xi_i \qquad (2.6.6)$$

$$\textit{Subject to} \quad y_i(\beta' x_i + \beta_0) \geq 1 - \xi_i \quad ; \quad \xi_i \geq 0 \quad ; \quad \forall i$$

เมื่อ C คือ ค่าควบคุมการ Trade – off ระหว่างขอบเขตกับค่าผิดพลาด

นำมาเขียนสมการให้อยู่ในรูปของ Lagrangian [x] ได้ดังนี้

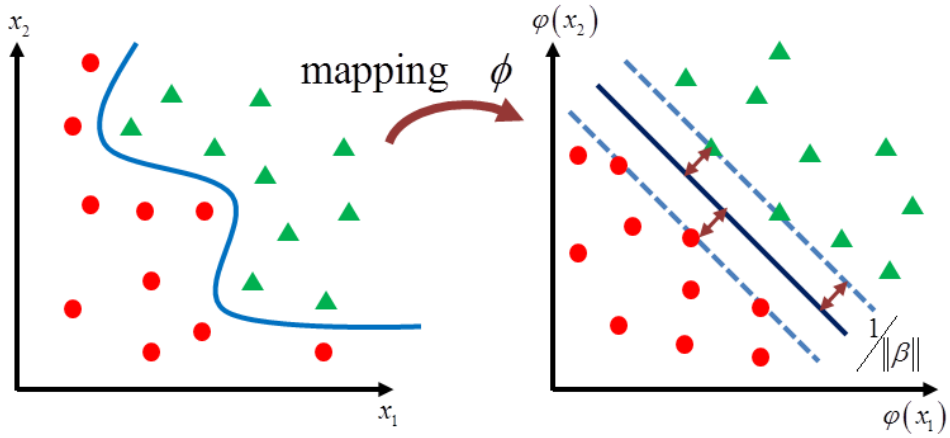
$$L_P(\beta,\beta_0,\xi,\alpha,\gamma) = \sum_{i=1}^{n}\xi_i + \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n}\alpha_i\left[1 - y_i(\beta' x_i + \beta_0) - \xi_i\right] - \sum_{i=1}^{n}\gamma_i\xi_i \qquad (2.6.7)$$

$$\textit{Subject to} \quad \alpha_i \geq 0 \text{ และ } \gamma_i \geq 0 \quad ; \quad \forall i$$

กรณีการแบ่งกลุ่มโดยการใช้ระนาบแบบไม่เป็นเส้นตรง

ซัพพอร์ตเวกเตอร์แมชชีน จะอาศัยหลักการแปลงข้อมูลจากปริภูมิขาเข้า (Input space) ให้เป็นปริภูมิลักษณะ (Feature space) ที่มีมิติสูงขึ้น โดยใช้ฟังก์ชันเคอร์เนล (Kernel Function)

ภาพที่ 2.6.2 แสดงหลักการแปลงข้อมูลจากปริภูมิขาเข้าให้เป็นปริภูมิที่มีมิติสูงขึ้น



คุณสมบัติตามทฤษฎีของ Mercer (Mercer's Theorem) ดังนี้

$$\text{Kernel Function} \quad : \quad k(x_i, x_j) \quad = \quad \Phi(\vec{x}_i)' \, \Phi(\vec{x}_j) \tag{2.6.8}$$

นำมาจัดให้อยู่ในรูปแบบของปัญหาคู่ (Dual Problem) โดยการแทนค่า β ในสมการของลา กรานจ์ จะได้

$$\text{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j \Phi(x_i)' \, \Phi(x_j) \tag{2.6.9}$$

$$\textit{With respect to} \quad \alpha \quad = \quad (\alpha_1, \alpha_2, \ldots \alpha_n)'$$

$$\textit{Subject to} \quad 0 \le \alpha_i \le C \quad \textit{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad ; \quad \forall i$$

สมการที่แสดงการจำแนกข้อมูลบนไฮเปอร์เพลน (Hyper – plane) ได้ดังนี้

$$\textit{Hyperplane} \, ; h(\vec{x}) \quad = \quad \text{Sgn}\left(\sum_{x_i=1}^{n} \alpha_i y_i \, K(x_i, x) + \beta_0 \right) \tag{2.6.10}$$

เมื่อ Sgn คือ Signum Function

กำหนดเคอร์เนลฟังก์ชันดังนี้

$$\text{Laplacian Kernel} : k(x_i, x_j) \quad = \quad \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2} \right) \tag{2.6.11}$$

ขั้นตอนการคำนวณของวิธีซัพพอร์ตเวกเตอร์แมชชีน

1. นำเข้าเวกเตอร์ของข้อมูล และกำหนดขอบเขตของความผิดพลาดที่ยอมรับได้ $(C = \{1:10\})$ และกำหนดพารามิเตอร์ของเคอร์เนลฟังก์ชัน $(\sigma^2 = \{2^{-5} : 2^5\})$ ทำการคัดเลือกค่าพารามิเตอร์ต่างๆ ให้เหมาะสมกับข้อมูลในแต่ละชุด โดยการกระทำซ้ำขั้นตอนที่ 2 – 6 เพื่อปรับเปลี่ยนค่าพารามิเตอร์ที่กำหนดไว้ให้ครบตามที่กำหนด

2. คำนวณตามหลักการของการแปลงข้อมูลจากปริภูมิขาเข้า (Input space) ให้เป็นปริภูมิลักษณะ

(Feature space) ที่มีมิติสูงขึ้น คุณสมบัติตามทฤษฎีของ Mercer (Mercer'sTheorem) ดังนี้

$$\text{Kernel Function} \quad : \quad k\left(x_i, x_j\right) \quad = \quad \Phi\left(\vec{x}_i\right)' \Phi\left(\vec{x}_j\right) \qquad (2.6.12)$$

3. เขียนสมการให้อยู่ในรูปของ Lagrangian [x] ได้ดังนี้

$$\text{Maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j \Phi\left(x_i\right)' \Phi\left(x_j\right) \qquad (2.6.13)$$

$$\textit{With respect to} \quad \alpha \quad = \quad \left(\alpha_1, \alpha_2, \ldots \alpha_n\right)'$$

$$\textit{Subject to} \quad 0 \le \alpha_i \le C \quad \textit{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \quad ; \quad \forall i$$

และใช้วิธีการ Quadratic Programming Problem เพื่อหาผลลัพธ์ของตัวคูณลากรานจ์ (Lagrange Multipliers)

4. ทำการหาขนาดของขอบเขตที่ได้จากผลลัพธ์ของตัวคูณลากรานจ์ (Lagrange Multipliers) เพื่อหาตำแหน่งของข้อมูลที่เป็นซัพพอร์ตเวกเตอร์ของข้อมูลที่มีค่าตัวคูณลากรานจ์ ไม่เท่ากับศูนย์

5. การหาค่าพารามิเตอร์ β, β_0 และ ξ_i โดยการ Differential ซึ่งกำหนดให้ผลลัพธ์มีค่าเป็นศูนย์ ตามสมการ (2.6.8) ดังนี้

$$L_P\left(\beta, \beta_0, \xi, \alpha, \gamma\right) \quad = \quad \sum_{i=1}^{n} \xi_i + \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{n} \alpha_i \left[1 - y_i\left(\beta' x_i + \beta_0\right) - \xi_i\right] - \sum_{i=1}^{n} \gamma_i \xi_i$$

$$\textit{Subject to} \quad \alpha_i \ge 0 \quad \textit{and} \quad \gamma_i \ge 0 \quad ; \quad \forall i$$

$$\frac{\partial L_P}{\partial \beta} \quad = \quad \beta - \sum_{i=1}^{n} \alpha_i y_i x_i \quad = \quad 0 \qquad \text{จะได้} \qquad \beta \quad = \quad \sum_{i=1}^{n} \alpha_i y_i x_i \qquad (2.6.14)$$

$$\frac{\partial L_P}{\partial \beta_0} \quad = \quad \sum_{i=1}^{n} \alpha_i y_i \quad = \quad 0 \qquad \text{จะได้} \qquad \sum_{i=1}^{n} \alpha_i y_i \quad = \quad 0 \qquad (2.6.15)$$

$$\frac{\partial L_P}{\partial \xi_i} \quad = \quad C - \alpha_i - \gamma_i \quad = \quad 0 \qquad \text{จะได้} \qquad \gamma_i \quad = \quad C - \alpha_i \qquad (2.6.16)$$

6. คำนวณหาสมการที่แสดงการแบ่งกลุ่มข้อมูลบนไฮเปอร์เพลน (Hyper – plane) ได้ดังนี้

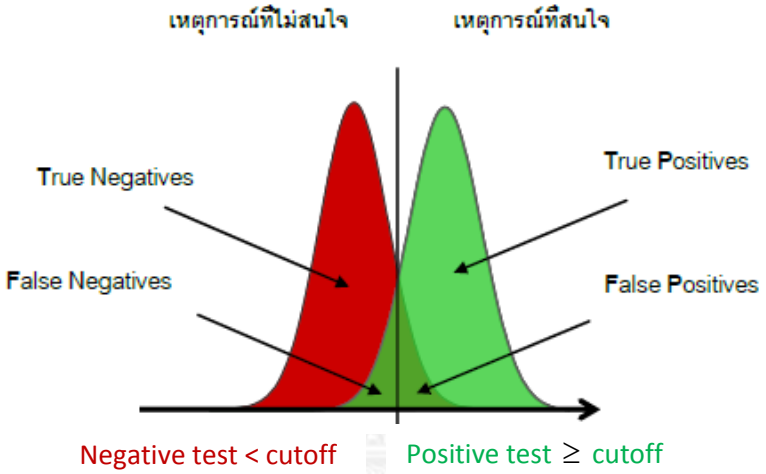
$$\textit{Hyperplane}\,; h\left(\vec{x}\right) \quad = \quad \text{Sgn}\left(\sum_{x_i=1}^{n} \alpha_i y_i \; K\left(x_i, x\right) + \beta_0\right) \qquad (2.6.17)$$

2.7 เครื่องมือวัดความมีประสิทธิภาพของการพยากรณ์จำแนกประเภท

Receiver Operating Characteristic (ROC) ถูกนำมาใช้ในการประเมินความถูกต้องของการพยากรณ์ในการจำแนกกลุ่มกรณีแบ่งข้อมูลเป็น 2 กลุ่ม ได้แก่ กลุ่มที่เกิดเหตุการณ์ที่สนใจหรือผล

การทดสอบเป็นบวก(positive) และกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจหรือผลการทดสอบเป็นลบ (negative) ณ จุดตัด(cut point) ของตัวแปรตอบสนอง ซึ่งจุดตัดที่ดีที่สุดควรอยู่ที่จุดวกกลับของเส้นโค้ง เพราะจะให้พื้นที่ใต้เส้นโค้งด้านซ้ายของเส้นทแยงมุมหรือการพยากรณ์ถูกต้องเพิ่มขึ้น โดยอยู่ในรูปของกราฟที่พล็อตระหว่าง ค่า Sensitivity และค่า 1 – Specificity ซึ่งกราฟอยู่ในช่วง $[0,1]$ ดังนี้

ภาพที่ 2.7.1 แสดงผลการพยากรณ์จำแนกประชากรออกเป็นกลุ่มเหตุการณ์ที่สนใจและกลุ่มเหตุการณ์ที่ไม่สนใจ



จากภาพที่ 2.7.1 หากพิจารณาจุดตัด ซึ่งเป็นตำแหน่งบนเส้นตรงเป็นเกณฑ์ในการจำแนกเหตุการณ์ออกเป็นกลุ่มของเหตุการณ์ที่ไม่สนใจ และกลุ่มของเหตุการณ์ที่สนใจ จะได้ว่า

TP (True Positive) คือ จำนวนตัวอย่างที่พยากรณ์**ถูกต้อง**ของการ**เกิด**เหตุการณ์

FP (False Positive) คือ จำนวนตัวอย่างที่พยากรณ์**ผิด**ของการ**ไม่เกิด**เหตุการณ์

TN (True Negative) คือ จำนวนตัวอย่างที่พยากรณ์**ถูกต้อง**ของการ**ไม่เกิด**เหตุการณ์

FN (False Negative) คือ จำนวนตัวอย่างที่พยากรณ์**ผิด**ของการ**เกิด**เหตุการณ์

Sensitivity (True Positive Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ได้**ถูกต้อง**ของการ**เกิด**เหตุการณ์ที่สนใจ

Specificity (True Negative Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ได้**ถูกต้อง**ของการ**ไม่เกิด**เหตุการณ์ที่สนใจ

1 – Specificity (False – Positive Rate) เป็นความน่าจะเป็นหรืออัตราส่วนของการพยากรณ์เหตุการณ์ได้**ผิด**ของการ**ไม่เกิด**เหตุการณ์ที่สนใจ

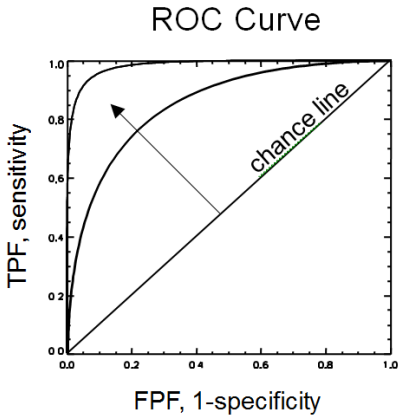
ดังนั้นสามารถคำนวณหาสัดส่วนของการพยากรณ์เหตุการณ์ เพื่อนำค่าที่ได้ไปทำการพล็อตโค้ง ROC และคำนวณหาพื้นที่ใต้โค้ง ซึ่งสูตรการคำนวณ Sensitivity และ 1 – Specificity ได้ดังนี้

$$Sensitivity = \frac{TP}{Total\ acttual\ Positive} = \frac{TP}{TP + FN} \qquad (2.7.1)$$

$$Specificity = \frac{TN}{Total\ acttual\ Negative} = \frac{TN}{FP + TN} \tag{2.7.2}$$

$$1 - Specificity = \frac{FP}{Total\ acttual\ Negative} = \frac{FP}{FP + TN} \tag{2.7.3}$$

ภาพที่ 2.7.2 แสดงพื้นที่ใต้โค้ง ROC



จากกราฟที่ได้จะนำมาหาค่าประมาณพื้นที่ใต้โค้ง ROC (Area under the Curve หรือ $\hat{AUC}$) โดย $\hat{AUC}$ จะใช้เทคนิคการประมาณค่าเกี่ยวกับการคำนวณอินทริกรัลจำกัดเขต ซึ่งแสดงได้ดังนี้

$$\int_a^b f(x)dx \quad ; \quad a \le x \le b \tag{2.7.4}$$

ฟังก์ชัน $f(x); a \le x \le b$ ซึ่งจะคำนวณโดยอินทิกรัลจำกัดเขต

2.8 อัตราความผิดพลาดในการจำแนกประเภทข้อมูล (MCR)

อัตราความผิดพลาดในการจำแนกประเภทข้อมูล โดยการกำหนดจุดตัดเป็นค่ากึ่งกลางของการจำแนกข้อมูล $(p_i = 0.5)$ แล้วนำมาใช้ในการประเมินความความผิดพลาดของการพยากรณ์เหตุการณ์ในการจำแนกประเภทกรณีแบ่งข้อมูลเป็น 2 กลุ่ม

$$\text{MCR} = \frac{FP + FN}{n} \quad \text{เมื่อ } n \text{ คือ จำนวนข้อมูลทั้งหมด} \tag{2.8.1}$$

บทที่ 3

วิธีการดำเนินการศึกษา

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติข้อมูลเข้า (input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS) และ Sliced Average Variance Estimator(SAVE) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ในกรณีศึกษาการแจกแจงแบบเกาส์เซียน โดยในการเปรียบเทียบความแม่นยำในการพยากรณ์จำแนกประเภทจะพิจารณาจากพื้นที่ใต้โค้งอาร์โอซี (Receiver Operating Characteristic : ROC) และอัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate : MCR) ซึ่งจะทำการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล (Monte Carlo Method) โดยใช้โปรแกรม R เวอร์ชัน 2.15.3 ในการทำการศึกษาภายใต้ขอบเขตดังต่อไปนี้

3.1 ขอบเขตของการวิจัย

ในการวิจัยครั้งนี้จะทำการศึกษาภายใต้ขอบเขตดังนี้

1. ศึกษาเหตุการณ์เกิดขึ้นสองเหตุการณ์ (dichotomous) คือ เหตุการณ์ที่สนใจ (Group1) กับเหตุการณ์ที่ไม่สนใจ (Group2) เนื่องมาจากโดยหลักการพื้นฐานของวิธีซัพพอร์ตเวกเตอร์แมชชีน กำหนดให้ตัวแปรตาม (Y) เป็นข้อมูลเชิงกลุ่มที่อยู่ในระดับนามบัญญัติ (Nominal Scale) โดยตัวแปรตาม (Y) แบ่งเป็น 2 กลุ่ม คือ

$$Y = \begin{cases} 1 & ; \quad Group\,1 \\ -1 & ; \quad Group\,2 \end{cases}$$

2. จำนวนตัวแปรอิสระ (p) ที่ใช้ในการวิจัยมี 4 ระดับ คือ 5, 10, 20 และ 40 ตัวแปร ซึ่งมีการแจกแจงแบบปกติหลายตัวแปร ที่มีพารามิเตอร์ค่าเฉลี่ยคือ $\mu = (0,...,0)'$ และมีเมตริกซ์ความแปรปรวนร่วมตามระดับความสัมพันธ์ของตัวแปรอิสระ

เมตริกซ์ความแปรปรวนร่วม $\left(\sum\right) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2p} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & \rho_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \rho_{p3} & \cdots & 1 \end{bmatrix}_{p\times p}$

โดยที่มีค่าความแปรปรวนของตัวแปรอิสระแต่ละตัวเป็น 1 และมีความสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และ j หรือ $\rho_{ij} = \rho^{|i-j|}$ เมื่อ $i, j = 1, 2, ..., p$

กำหนดให้ ρ มีค่าเป็น 0.85, 0.9 และ 0.95 จะได้ว่า

<u>กรณีตัวแปรอิสระ 5 ตัว (p=5)</u>

- $\rho = 0.85$ จะได้ว่า เมตริกซ์ความแปรปรวนร่วม คือ

$$\begin{bmatrix} 1 & 0.85^{|1-2|} & 0.85^{|1-3|} & 0.85^{|1-4|} & 0.85^{|1-5|} \\ 0.85^{|2-1|} & 1 & 0.85^{|2-3|} & 0.85^{|2-4|} & 0.85^{|2-5|} \\ 0.85^{|3-1|} & 0.85^{|3-2|} & 1 & 0.85^{|3-4|} & 0.85^{|3-5|} \\ 0.85^{|4-1|} & 0.85^{|4-2|} & 0.85^{|4-3|} & 1 & 0.85^{|4-5|} \\ 0.85^{|5-1|} & 0.85^{|5-2|} & 0.85^{|5-3|} & 0.85^{|5-4|} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.85 & 0.72 & 0.61 & 0.52 \\ 0.85 & 1 & 0.85 & 0.72 & 0.61 \\ 0.72 & 0.85 & 1 & 0.85 & 0.72 \\ 0.61 & 0.72 & 0.85 & 1 & 0.85 \\ 0.52 & 0.61 & 0.72 & 0.85 & 1 \end{bmatrix}$$

ดังนั้น $\rho_{ij} \in [0.52, 0.85]$

- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.65, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.81, 0.95]$

<u>กรณีตัวแปรอิสระ 10 ตัว (p=10)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.23, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.38, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.63, 0.95]$

<u>กรณีตัวแปรอิสระ 20 ตัว (p=20)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.04, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.13, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.37, 0.95]$

<u>กรณีตัวแปรอิสระ 40 ตัว (p=40)</u>

- $\rho = 0.85$ จะได้ว่า $\rho_{ij} \in [0.001, 0.85]$
- $\rho = 0.90$ จะได้ว่า $\rho_{ij} \in [0.016, 0.90]$
- $\rho = 0.95$ จะได้ว่า $\rho_{ij} \in [0.135, 0.95]$

3. ศึกษาภายใต้ขนาดตัวอย่าง ซึ่งกำหนดตัวอย่างของกลุ่มที่หนึ่งเป็นกลุ่มตัวอย่างที่สนใจ (n_1) และกลุ่มที่สองเป็นกลุ่มตัวอย่างที่ไม่สนใจ (n_2) ซึ่งสามารถกำหนดขนาดของจำนวนตัวอย่างดังนี้ $n_s = 30, 60, 120 \;\; ; \;\; s = 1, 2$

3.1 กรณีที่ขนาดตัวอย่างของแต่ละกลุ่มเท่ากัน ดังนี้ $n_1 = n_2$ จะสามารถ จัดกลุ่มของขนาดตัวอย่างได้เท่ากับ $\binom{3}{1}$ = 3 กรณี

3.2 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจ ดังนี้ $n_1 > n_2$ จะสามารถจัดกลุ่มของขนาดตัวอย่างได้เท่ากับ $\binom{3}{2} = 3$ กรณี

3.3 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจ ดังนี้ $n_1 < n_2$ จะสามารถจัดกลุ่มของขนาดตัวอย่างได้เท่ากับ $\binom{3}{2} = 3$ กรณี

4. ศึกษาลักษณะของการแจกแจงข้อมูล โดยวัดค่าการซ้อนทับ(overlap) ได้เสนอโดย(Melnykov and Maitra. (2010)) ในงานวิจัยครั้งนี้จะวัดการซ้อนทับระหว่างองค์ประกอบ(component)ที่ i กับ j ในกรณีที่เมตริกซ์ความแปรปรวนร่วมระหว่างองค์ประกอบดังกล่าวมีค่าเท่ากัน $\left(\sum_i = \sum_j = \sum\right)$ คือ

$$\omega_{ij} = 2\Phi\left(-\frac{1}{2}\sqrt{\left(\mu_j - \mu_i\right)' \sum^{-1} \left(\mu_j - \mu_i\right)}\right)$$

เมื่อ $\Phi\left(X; \mu_k, \sum_k\right)$ เป็นฟังก์ชันความหนาแน่นของการแจกแจงแบบปกติหลายตัวแปร (multivariate normal density) ขององค์ประกอบที่ k ที่มีพารามิเตอร์ค่าเฉลี่ยคือ μ_k แลเมตริกซ์ความแปรปรวนร่วมเป็น $\sum_k$

โดยมีการศึกษาตัวอย่างที่มีตัวแปรอิสระ 5, 10, 20 และ 40 ตัวแปร โดยตัวแปรตาม (Y) แบ่งเป็น 2 กลุ่ม คือ

กลุ่มที่สนใจ $(Y = 1)$ ซึ่งเป็นกลุ่มที่มีเกณฑ์กำหนดที่ชัดเจน โดยที่ค่าของพารามิเตอร์ของการแจกแจงข้อมูลเป็นค่าคงที่

กลุ่มที่ไม่สนใจ $(Y = -1)$ ซึ่งเป็นกลุ่มที่มีลักษณะการแจกแจงของข้อมูลแตกต่างกับกลุ่มที่สนใจ โดยที่ค่าของพารามิเตอร์ของการแจกแจงข้อมูลมีการเปลี่ยนแปลงตามค่าของ d

ทำการวิเคราะห์โดยเปลี่ยนค่าของ d ไปเรื่อย ๆ ตามค่าที่กำหนดจนครบสำหรับการแจกแจงที่ทำการศึกษา ซึ่งสามารถวัดค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap) จากการจำลองข้อมูลซ้ำจำนวน 100 รอบ ได้ดังนี้

ตารางที่ 3.1.1 แสดงค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap) กรณีจำนวนตัวแปรอิสระเท่ากับ 5 ตัว

n	ρ	d			
		1	1.5	2	2.5
$n_1 = n_2 = 30$	0.85	21%	14%	10%	6%
	0.9	22%	16%	10%	7%
	0.95	21%	17%	11%	8%
$n_1 = n_2 = 60$	0.85	25%	17%	11%	7%
	0.9	25%	18%	12%	7%
	0.95	26%	19%	13%	8%
$n_1 = n_2 = 120$	0.85	26%	18%	12%	7%
	0.9	27%	19%	13%	8%
	0.95	28%	20%	14%	9%
$n_1 = 60, n_2 = 30$ หรือ $n_1 = 30, n_2 = 60$	0.85	25%	17%	11%	7%
	0.9	25%	19%	12%	7%
	0.95	27%	19%	13%	8%
$n_1 = 120, n_2 = 60$ หรือ $n_1 = 60, n_2 = 120$	0.85	32%	22%	14%	8%
	0.9	33%	23%	16%	9%
	0.95	34%	25%	17%	11%
$n_1 = 120, n_2 = 30$ หรือ $n_1 = 30, n_2 = 120$	0.85	28%	19%	12%	7%
	0.9	29%	20%	13%	8%
	0.95	30%	22%	14%	9%

ตารางที่ 3.1.2 แสดงค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap)กรณีจำนวนตัวแปรอิสระเท่ากับ 10 ตัว

n	ρ	d			
		1	1.5	2	2.5
$n_1 = n_2 = 30$	0.85	10%	6%	4%	2%
	0.9	10%	7%	5%	3%
	0.95	11%	8%	5%	3%
$n_1 = n_2 = 60$	0.85	16%	11%	6%	3%
	0.9	17%	12%	7%	4%
	0.95	17%	13%	9%	6%
$n_1 = n_2 = 120$	0.85	20%	13%	8%	4%
	0.9	22%	15%	9%	5%
	0.95	23%	16%	11%	7%
$n_1 = 60, n_2 = 30$	0.85	13%	9%	5%	3%
หรือ	0.9	14%	10%	6%	3%
$n_1 = 30, n_2 = 60$	0.95	15%	11%	7%	5%
$n_1 = 120, n_2 = 60$	0.85	19%	12%	7%	4%
หรือ	0.9	20%	14%	8%	5%
$n_1 = 60, n_2 = 120$	0.95	20%	15%	10%	6%
$n_1 = 120, n_2 = 30$	0.85	19%	13%	7%	4%
หรือ	0.9	21%	14%	9%	5%
$n_1 = 30, n_2 = 120$	0.95	22%	16%	11%	6%

ตารางที่ 3.1.3 แสดงค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap)กรณีจำนวนตัวแปรอิสระเท่ากับ 20 ตัว

n	ρ	d			
		1	1.5	2	2.5
$n_1 = n_2 = 30$	0.85	0%	0%	0%	0%
	0.9	1%	0%	0%	0%
	0.95	1%	0%	0%	0%
$n_1 = n_2 = 60$	0.85	4%	3%	1%	0%
	0.9	5%	3%	2%	1%
	0.95	5%	4%	2%	1%
$n_1 = n_2 = 120$	0.85	10%	6%	3%	1%
	0.9	11%	7%	4%	2%
	0.95	12%	8%	5%	3%
$n_1 = 60, n_2 = 30$ หรือ $n_1 = 30, n_2 = 60$	0.85	1%	1%	0%	0%
	0.9	1%	1%	0%	0%
	0.95	2%	1%	1%	0%
$n_1 = 120, n_2 = 60$ หรือ $n_1 = 60, n_2 = 120$	0.85	2%	1%	1%	0%
	0.9	2%	2%	1%	0%
	0.95	3%	2%	1%	1%
$n_1 = 120, n_2 = 30$ หรือ $n_1 = 30, n_2 = 120$	0.85	7%	4%	2%	1%
	0.9	8%	5%	3%	1%
	0.95	8%	6%	4%	2%

<u>ตารางที่ 3.1.4</u> แสดงค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap)กรณีจำนวนตัวแปรอิสระเท่ากับ 40 ตัว

n	ρ	d			
		1	1.5	2	2.5
$n_1 = n_2 = 30$	0.85	0%	0%	0%	0%
	0.9	0%	0%	0%	0%
	0.95	0%	0%	0%	0%
$n_1 = n_2 = 60$	0.85	0%	0%	0%	0%
	0.9	0%	0%	0%	0%
	0.95	0%	0%	0%	0%
$n_1 = n_2 = 120$	0.85	1%	0%	0%	0%
	0.9	1%	1%	0%	0%
	0.95	1%	1%	1%	0%
$n_1 = 60, n_2 = 30$ หรือ $n_1 = 30, n_2 = 60$	0.85	0%	0%	0%	0%
	0.9	0%	0%	0%	0%
	0.95	0%	0%	0%	0%
$n_1 = 120, n_2 = 60$ หรือ $n_1 = 60, n_2 = 120$	0.85	0%	0%	0%	0%
	0.9	0%	0%	0%	0%
	0.95	0%	0%	0%	0%
$n_1 = 120, n_2 = 30$ หรือ $n_1 = 30, n_2 = 120$	0.85	0%	0%	0%	0%
	0.9	0%	0%	0%	0%
	0.95	0%	0%	0%	0%

จากตารางที่ 3.1.1-3.1.4 จะได้ว่าค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap) มีค่าลดลง เมื่อค่าพารามิเตอร์ของค่าเฉลี่ยของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ (d) มีค่าเพิ่มขึ้น นั่นคือ จะทำให้กลุ่มตัวอย่างของกลุ่มที่สนใจกับกลุ่มที่ไม่สนใจมีลักษณะการแจกแจงของข้อมูลแตกต่างกันเพิ่มมากขึ้นหรือสามารถอธิบายความแตกต่างระหว่างข้อมูลทั้งสองกลุ่มได้ชัดเจนเพิ่มมากขึ้น ในทางตรงข้ามเมื่อ d มีค่าลดน้อยลง ค่าเฉลี่ยเปอร์เซ็นต์การซ้อนทับ(overlap) มีค่าเพิ่มขึ้นนั่นคือ จะทำให้ลักษณะการแจกแจงของทั้งสองกลุ่มใกล้เคียงกันมากขึ้นหรืออธิบายความแตกต่างได้น้อยลง

5. ในการศึกษาครั้งนี้ทำการจำลองข้อมูลให้มีสถานการณ์ที่แตกต่างกัน ตามข้อกำหนดข้างต้นโดยใช้เทคนิคมอนติคาร์โล (Monte Carlo Simulation Technique) โดยทำการจำลองในแต่ละสถานการณ์จะกระทำซ้ำ 500 รอบ

3.2 ขั้นตอนในการดำเนินการศึกษา

1. กำหนดเงื่อนไขและขอบเขตของการวิจัย

- กำหนดจำนวนตัวแปรอิสระ $(p = 5,10,20,40)$
- กำหนดค่าพารามิเตอร์ตามการแจกแจงที่กำหนดในขอบเขตของการศึกษา $(d = 1,1.5,2,2.5)$
- กำหนดขนาดตัวอย่าง $(n_s = 30,60,120 \quad ; \quad s = 1,2)$
- ค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ $(\rho =, 0.85, 0.9, 0.95)$

2. จำลองข้อมูลตามการแจกแจงและขอบเขตที่ต้องการศึกษา

- จำลองค่า X ตามการแจกแจงของข้อตกลงเบื้องต้น และจำลองค่า y เป็นข้อมูลเชิงกลุ่มที่อยู่ในระดับนามบัญญัติ (Nominal Scale) คือ

$$y = \begin{cases} 1 \quad ; X \sim N\left(\mu = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \Sigma\right) \\ -1 \quad ; X \sim N\left(\mu = \begin{bmatrix} d \\ \vdots \\ d \end{bmatrix}, \Sigma\right) \end{cases}$$

3. ลดตัวแปรข้อมูลหรือลดมิติข้อมูลด้วยวิธีการวิเคราะห์องค์ประกอบหลัก(PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (PLS)และ Sliced Average Variance Estimator(SAVE) โดยมีเกณฑ์ในการเลือกมิติข้อมูลดังนี้

- **การวิเคราะห์องค์ประกอบหลัก(PCA)** จะเลือกมิติข้อมูลหรือจำนวนองค์ประกอบหลัก โดยพิจารณาจากร้อยละสัดส่วนของความแปรปรวนสะสม ถ้าร้อยละความแปรปรวนสะสมของตัวประกอบหลัก k ตัวแรก เป็นอย่างต่ำร้อยละ 80 ก็ควรให้จำนวนตัวประกอบหลักเท่ากับ k โดยที่ k < p
- **วิธีกำลังสองน้อยสุดเชิงส่วน(PLS)** จะเลือกมิติข้อมูลหรือจำนวนองค์ประกอบหลัก โดยพิจารณาจากการประเมินความน่าเชื่อถือได้ของการวิเคราะห์ความถดถอย โดยพิจารณาค่าสถิติ Adjusted Cross Validation (adjCV) ของวิธี cross-validation หรือ Leave-One-Out cross-Validation ดังนั้นการเปรียบเทียบค่า adjCV ของการวิเคราะห์ที่มีจำนวนองค์ประกอบต่างกัน ถ้าค่า adjCV ของการวิเคราะห์ที่มีจำนวนองค์ประกอบใดๆมีค่าน้อยสุดจะเลือกจำนวนองค์ประกอบดังกล่าวเป็นจำนวนองค์ประกอบที่เหมาะสม
- **Sliced Average Variance Estimator (SAVE)** จะเลือกมิติข้อมูลจำนวน q มิติ เมื่อ $q = 0,1,2,...,p$ สำหรับการทดสอบสมมติฐานว่ามิติ $k = q$ โดยที่ $k < p$ จาก Marginal dimension hypothesis ซึ่งมีรายละเอียดดังนี้

ผลลัพธ์ : $k = q$

1) เริ่มต้นให้ $q = 0$

2) สมมติฐาน $\mathrm{H}_0 : k = q$ vs. $\mathrm{H}_0 : k > q$

3) สถิติทดสอบ $T_n(\hat{m}) = \frac{n}{2}\sum_{i=1}^{H} tr\left\{\left(\hat{m}^T \hat{A}_i \hat{m}\right)^2\right\}$ มีการแจกแจงแบบไคสแควร์ที่มีองศาอิสระเท่ากับ $(H-1)(p-q)(p-q+1)/2$

4) ให้ $\alpha = 0.05$ จะปฏิเสธ H_0 เมื่อ $T_n(\hat{m}) > \chi^2_{df=(H-1)(p-q)(p-q+1)/2}$ จะกลับไปทำข้อ 1) โดยให้ $k = q+1$ และเมื่อผลการทดสอบยอมรับ H_0 จะได้จำนวนมิติที่น้อยที่สุด คือ q มิติ

4. ทำการประมาณค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน เพื่อสร้างตัวแบบสำหรับนำมาพยากรณ์

กำหนดค่าที่ใช้ในวิธีการซัพพอร์ตเวกเตอร์แมชชีน ดังนี้

- กำหนดช่วงของค่าควบคุมการ Trade – off ระหว่างขอบเขตกับค่าความผิดพลาดที่ดีที่สุด เป็นค่าตั้งแต่ 1 ถึง 50
- กำหนดช่วงของ Laplacian Kernel เป็นค่าตั้งแต่ 2^{-5} ถึง 2^5

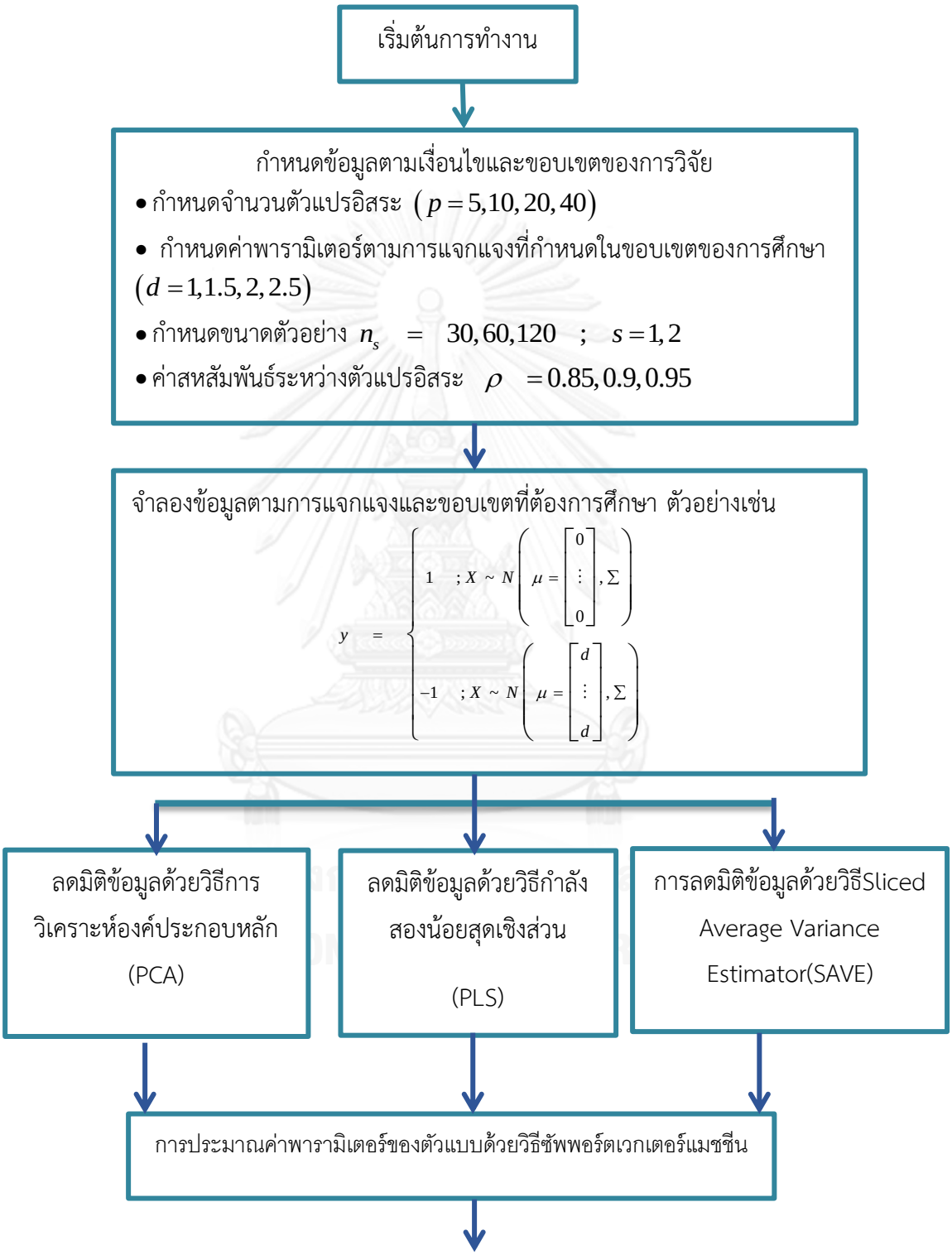
ในการการประมาณค่าด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีนนั้น ผู้วิจัยได้มีการคัดเลือกค่าพารามิเตอร์ต่าง ๆ ให้เหมาะสมกับข้อมูลในแต่ละชุดก่อน โดยพิจารณาค่าตามที่กำหนดไว้ข้างต้น แล้วจึงทำการประมาณค่าสัมประสิทธิ์การถดถอยของพารามิเตอร์นั้นๆ

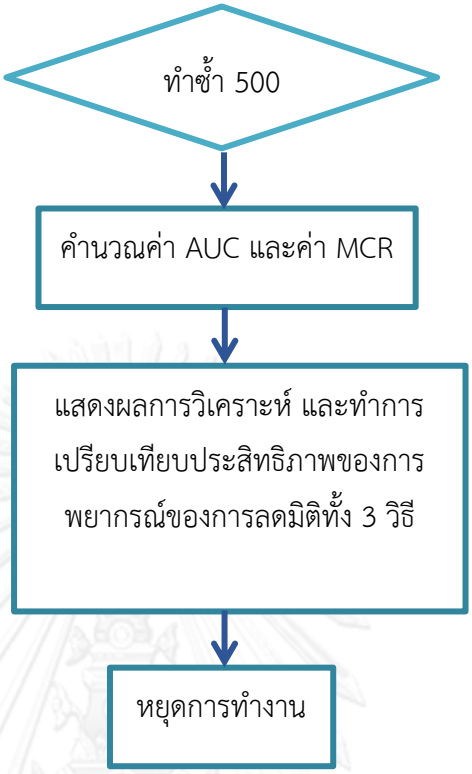
5. นำตัวแบบที่ใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนด้วยฟังก์ชันเคอร์เนลต่าง ๆ แล้วนำตัวแบบที่ได้ไปพยากรณ์ข้อมูลต่อไป เพื่อตรวจสอบผลของการพยากรณ์เทียบกับเหตุการณ์ที่เกิดขึ้นจริง และนำผลลัพธ์ของการพยากรณ์ที่ได้ไปสร้าง ตารางการแบ่งกลุ่ม

6. นำข้อมูลที่ได้ไปคำนวณหาค่าประมาณพื้นที่ใต้โค้ง ROC (Area Under ROC Curve : AUC) และคำนวณค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate : MCR)

7. วิเคราะห์และสรุปผลการเปรียบเทียบวิธีการที่ใช้ในการวิจัย

3.3 ขั้นตอนการทำงานของโปรแกรม





บทที่ 4

ผลการวิเคราะห์ข้อมูล

การศึกษางานวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติข้อมูลเข้า(input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS) และ Sliced Average Variance Estimator(SAVE) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ในกรณีศึกษาการแจกแจงแบบเกาส์เซียน โดยทำการจำลองข้อมูลเพื่อศึกษาผลกระทบจากระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล (d), ค่าระดับความสัมพันธ์ของตัวแปรอิสระ (ρ) และขนาดของกลุ่มตัวอย่าง (n_1, n_2) ทำการพิจารณาผลการศึกษาด้วย Receiver Operating Characteristic (ROC) ใช้เป็นเครื่องมือวัดประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูล โดยใช้พื้นที่ใต้โค้ง ROC และใช้อัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate : MCR) เพื่อศึกษาว่าวิธีการใดมีความผิดพลาดในการจำแนกประเภท ซึ่งในงานวิจัยนี้ทำการศึกษาผลของเหตุการณ์เกิดขึ้นสองเหตุการณ์ (dichotomous)

ในการนำเสนอผลการวิจัยจะแสดงในรูปแบบของตาราง โดยมีสัญลักษณ์ที่ใช้แทนความหมายต่างๆ ดังนี้

n_1 แทน ขนาดของตัวอย่างของกลุ่มที่ 1

n_2 แทน ขนาดของตัวอย่างของกลุ่มที่ 2

d แทน ค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจที่มีการเปลี่ยนแปลงไปเรื่อย ๆ เมื่อตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) กล่าวคือ $N\left(\mu = \begin{bmatrix} d \\ \vdots \\ d \end{bmatrix}, \Sigma\right)$

ρ แทน ความสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และ j หรือ $\rho_{ij} = \rho^{|i-j|}$

AUC แทน ค่าของพื้นที่ใต้โค้ง ROC (Area Under ROC Curve)

MCR แทน ค่าของอัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate)

การพิจารณาผลการศึกษาการจำลองข้อมูลเพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติสำหรับข้อมูลเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน ด้วยค่าจำนวนมิติของตัวแปรอิสระที่ใช้สำหรับข้อมูลเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน(dim) ค่า Receiver Operating Characteristic (ROC) ใช้เป็นเครื่องมือวัดประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูล โดยใช้พื้นที่ใต้โค้ง ROC(AUC) และใช้อัตราความผิดพลาดในการจำแนกประเภทข้อมูล (Misclassification Rate :

MCR) เพื่อศึกษาว่าวิธีการใดมีความผิดพลาดในการจำแนกประเภท ซึ่งในงานวิจัยนี้ทำการศึกษาผลของเหตุการณ์เกิดขึ้นสองเหตุการณ์ (dichotomous) กรณีศึกษาการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution)

4.1 ตัวแปรอิสระ 5 ตัวแปร

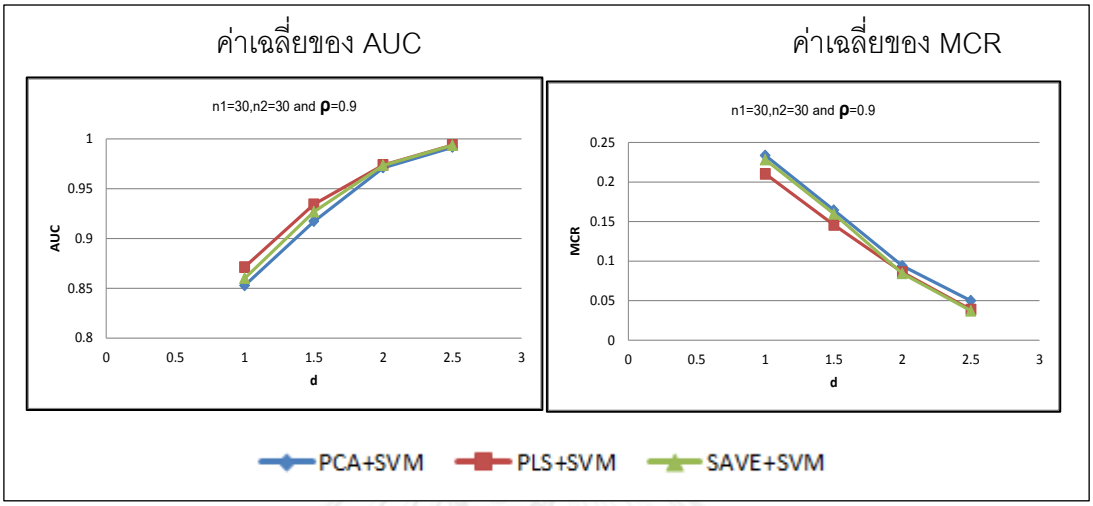
4.1.1 กรณีศึกษาขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจและไม่สนใจมีขนาดเท่ากัน

4.1.1.1 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 30 $(n_1 = n_2 = 30)$

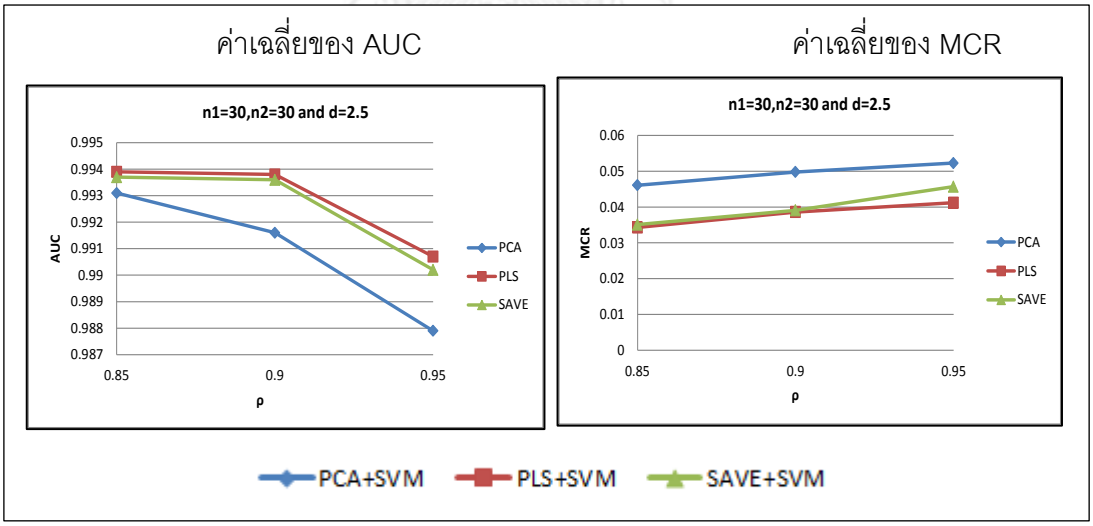
ตารางที่ 4.1.1 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8682	0.8786	0.8723	0.2205	0.2012	0.2138
	1.5	0.9322	0.9408	0.9344	0.1492	0.1354	0.1452
	2	0.9739	0.9801	0.9784	0.0911	0.0720	0.0773
	2.5	0.9931	0.9939	0.9937	0.0461	0.0343	0.0351
0.9	1	0.8527	0.8712	0.8699	0.2336	0.2102	0.2283
	1.5	0.9171	0.9343	0.9268	0.1644	0.1456	0.1596
	2	0.9710	0.9738	0.9734	0.0939	0.0860	0.0843
	2.5	0.9916	0.9938	0.9936	0.0498	0.0386	0.0391
0.95	1	0.8520	0.8688	0.8666	0.2398	0.2110	0.2293
	1.5	0.9158	0.9267	0.9185	0.1675	0.1496	0.1607
	2	0.9659	0.9722	0.9663	0.1033	0.0927	0.0995
	2.5	0.9879	0.9907	0.9902	0.0523	0.0412	0.0457

ภาพที่ 4.1.1 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.2 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี SAVE และวิธี PCA ตามลำดับ

2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด SAVE และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

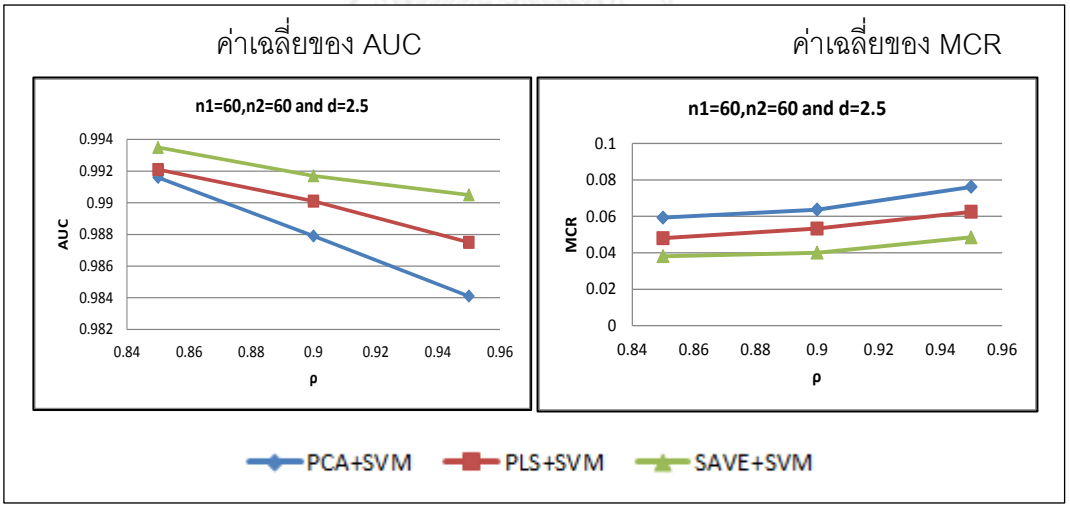
4.1.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 60 $\left(n_1 = n_2 = 60\right)$

ตารางที่ 4.1.2 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8371	0.8402	0.8487	0.2670	0.2491	0.2382
	1.5	0.9239	0.9313	0.9364	0.1724	0.1583	0.1485
	2	0.9730	0.9771	0.9812	0.1020	0.0849	0.0716
	2.5	0.9916	0.9921	0.9935	0.0593	0.0480	0.0381
0.9	1	0.8210	0.8216	0.8365	0.2687	0.2572	0.2520
	1.5	0.9193	0.9237	0.9247	0.1788	0.1683	0.1614
	2	0.9705	0.9707	0.9755	0.1060	0.1003	0.0836
	2.5	0.9879	0.9901	0.9917	0.0637	0.0533	0.0400
0.95	1	0.8039	0.8112	0.8280	0.2846	0.2773	0.2542
	1.5	0.9079	0.9090	0.9136	0.1872	0.1807	0.1787
	2	0.9631	0.9667	0.9705	0.1214	0.1051	0.0948
	2.5	0.9841	0.9875	0.9905	0.0762	0.0625	0.0485

ภาพที่ 4.1.3 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5

ภาพที่ 4.1.4 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธีSAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ

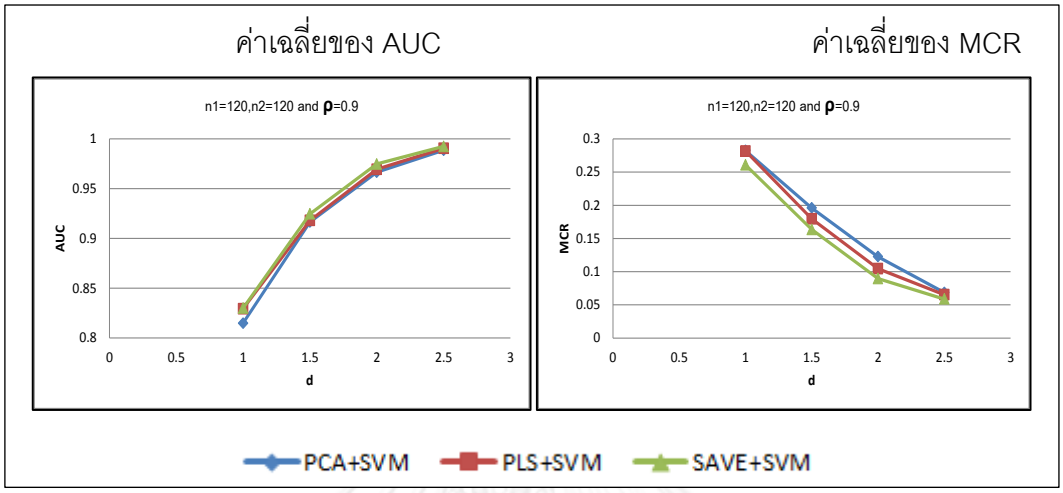
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.1.3 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 120 $\left(n_1 = n_2 = 120\right)$

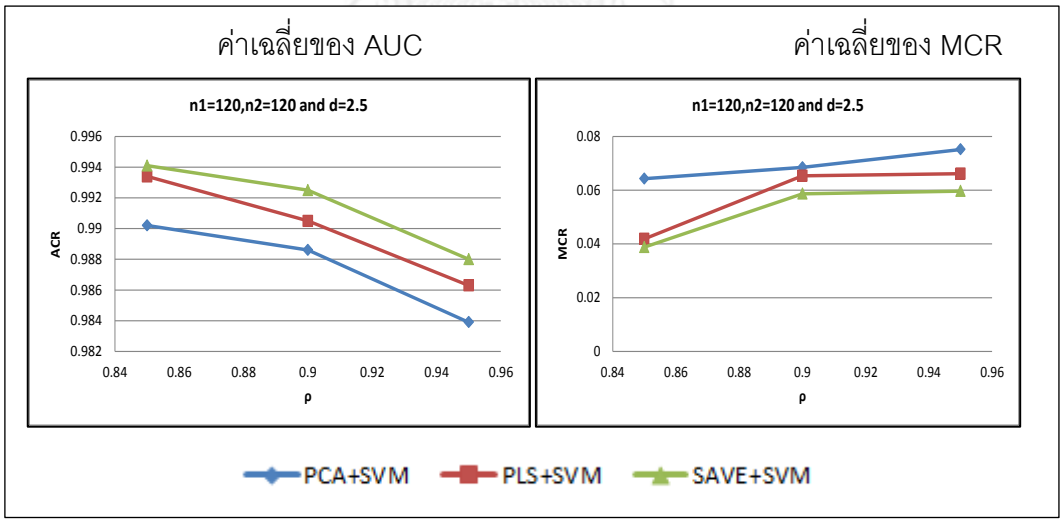
ตารางที่ 4.1.3 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8179	0.8301	0.8303	0.2799	0.2615	0.2719
	1.5	0.9217	0.9304	0.9305	0.1856	0.1617	0.1588
	2	0.9724	0.9758	0.9784	0.1119	0.0895	0.0832
	2.5	0.9902	0.9934	0.9941	0.0643	0.0419	0.0388
0.9	1	0.8148	0.8293	0.8299	0.2825	0.2618	0.2838
	1.5	0.9165	0.9181	0.9247	0.1958	0.1795	0.1634
	2	0.9665	0.9695	0.9748	0.1224	0.1045	0.0895
	2.5	0.9886	0.9905	0.9925	0.0685	0.0654	0.0587
0.95	1	0.8053	0.8163	0.8001	0.2888	0.2828	0.2813
	1.5	0.9047	0.9094	0.9126	0.2036	0.1914	0.1821
	2	0.9587	0.9598	0.9664	0.1361	0.1257	0.1054
	2.5	0.9839	0.9863	0.9880	0.0752	0.0662	0.0597

ภาพที่ 4.1.5 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.6 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ

2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.2 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจมีความแตกต่างกับกลุ่มตัวอย่างที่ไม่สนใจ

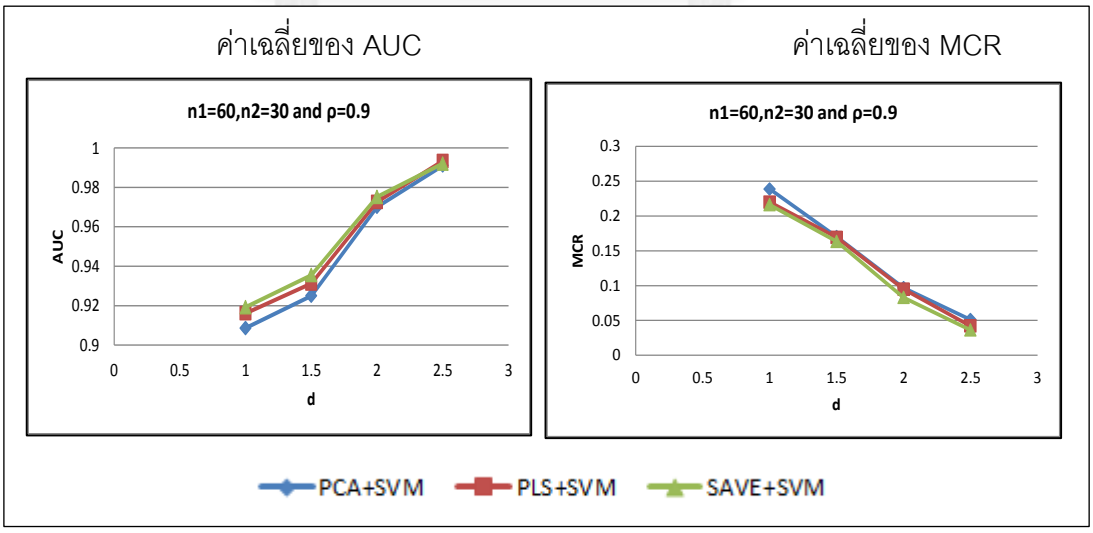
4.1.2.1 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างมากกว่ากลุ่มที่ไม่สนใจ

4.1.2.1.1 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $\left(n_1 = 60, n_2 = 30\right)$

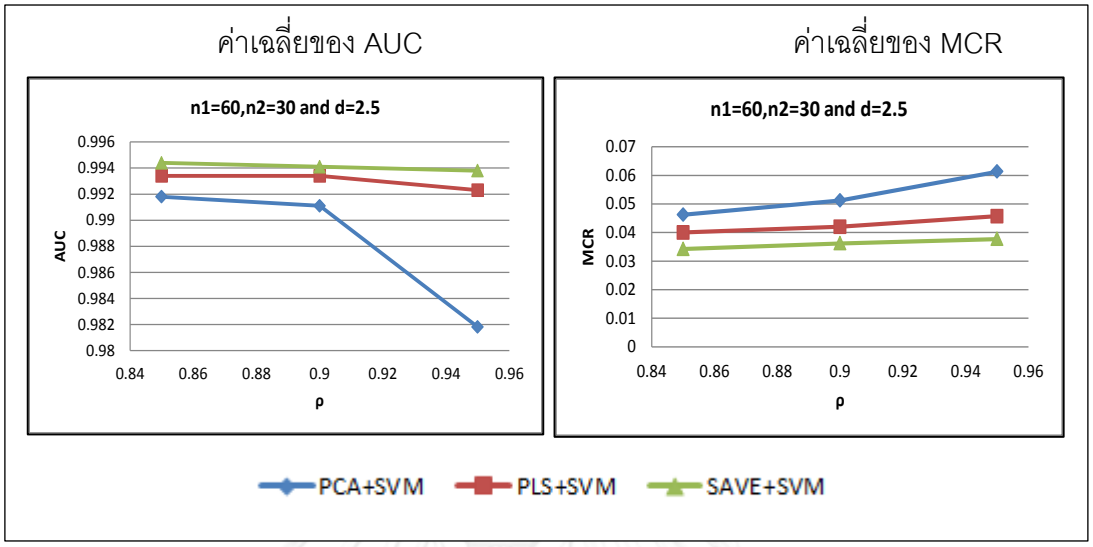
ตารางที่ 4.1.4 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9091	0.9132	0.9227	0.2182	0.2143	0.2109
	1.5	0.9410	0.9470	0.9488	0.1489	0.1473	0.1402
	2	0.9769	0.9795	0.9823	0.0857	0.0748	0.0664
	2.5	0.9918	0.9934	0.9944	0.0462	0.0400	0.0342
0.9	1	0.9086	0.9161	0.9192	0.2387	0.2199	0.2162
	1.5	0.9250	0.9312	0.9355	0.1703	0.1693	0.1636
	2	0.9701	0.9726	0.9752	0.0970	0.0949	0.0830
	2.5	0.9911	0.9934	0.9941	0.0512	0.0420	0.0362
0.95	1	0.9007	0.9150	0.9224	0.2529	0.2247	0.2189
	1.5	0.9174	0.9224	0.9235	0.1714	0.1699	0.1639
	2	0.9666	0.9709	0.9742	0.1034	0.0970	0.0846
	2.5	0.9818	0.9923	0.9938	0.0613	0.0457	0.0377

ภาพที่ 4.1.7 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.8 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

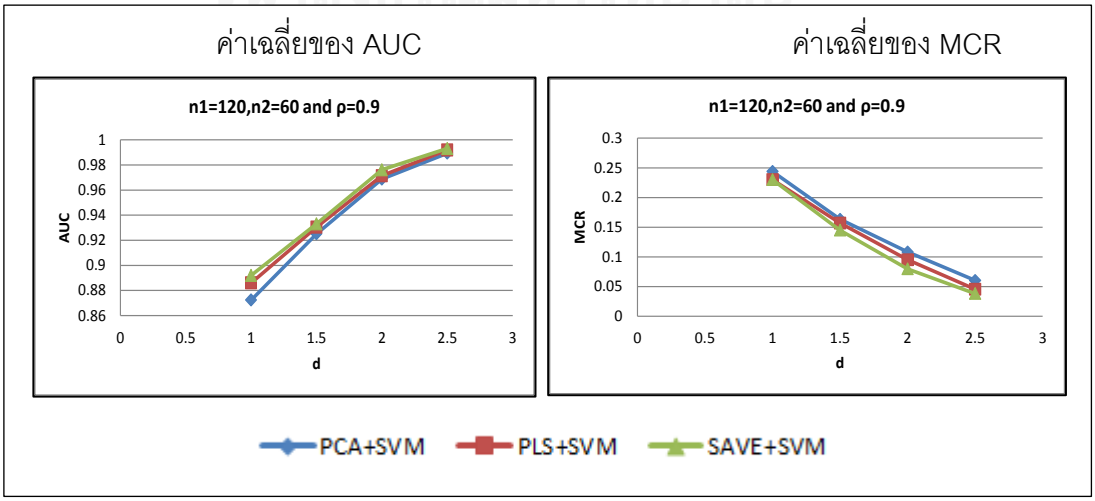
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.2.1.2 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 120, n_2 = 60)$

ตารางที่ 4.1.5 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8769	0.8906	0.8922	0.2315	0.2137	0.2128
	1.5	0.9319	0.9399	0.9410	0.1581	0.1444	0.1379
	2	0.9747	0.9770	0.9802	0.0983	0.0836	0.0710
	2.5	0.9914	0.9935	0.9949	0.0562	0.0386	0.0315
0.9	1	0.8723	0.8860	0.8920	0.2439	0.2301	0.2300
	1.5	0.9251	0.9305	0.9331	0.1633	0.1569	0.1450
	2	0.9688	0.9714	0.9762	0.1084	0.0952	0.0799
	2.5	0.9894	0.9920	0.9931	0.0605	0.0455	0.0384
0.95	1	0.8714	0.8848	0.8913	0.2477	0.2321	0.2318
	1.5	0.9203	0.9232	0.9237	0.1716	0.1668	0.1614
	2	0.9640	0.9653	0.9697	0.1154	0.1075	0.0913
	2.5	0.9877	0.9889	0.9909	0.0671	0.0580	0.0456

ภาพที่ 4.1.9 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.10 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5

ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

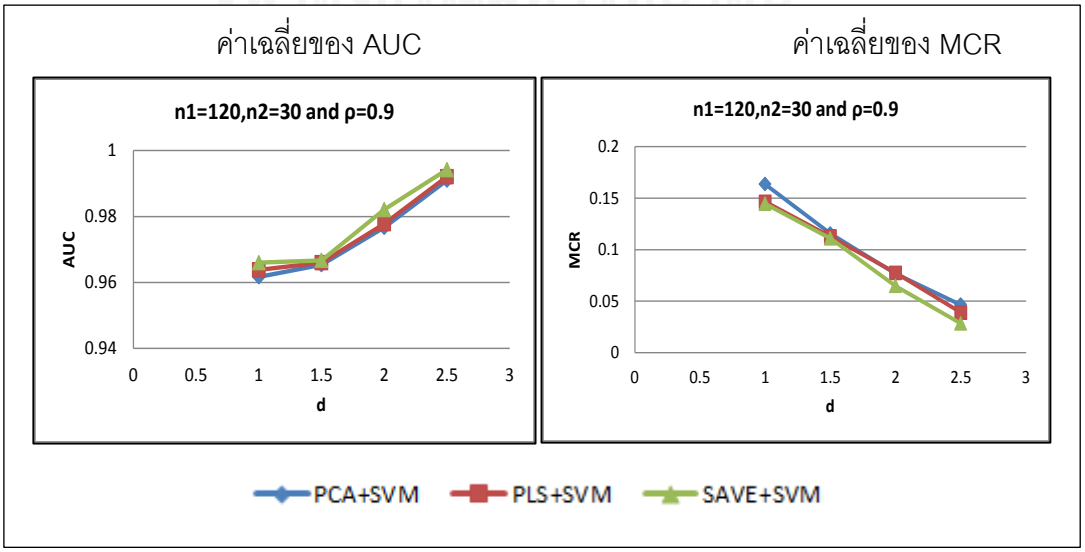
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.2.1.3 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $\left(n_1 = 120, n_2 = 30\right)$

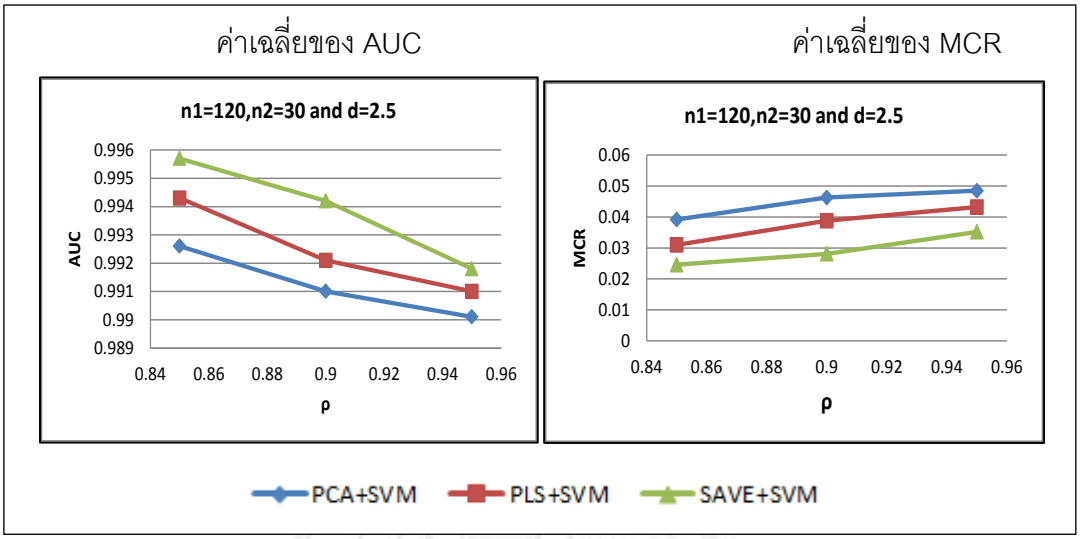
ตารางที่ 4.1.6 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9653	0.9659	0.9660	0.1466	0.1463	0.1434
	1.5	0.9655	0.9662	0.9665	0.1133	0.1093	0.1046
	2	0.9785	0.9818	0.9839	0.0744	0.0659	0.0576
	2.5	0.9926	0.9943	0.9957	0.0392	0.0310	0.0246
0.9	1	0.9617	0.9638	0.9660	0.1634	0.1464	0.1444
	1.5	0.9654	0.9660	0.9667	0.1154	0.1127	0.1107
	2	0.9767	0.9778	0.9821	0.0766	0.0774	0.0646
	2.5	0.9910	0.9921	0.9942	0.0463	0.0388	0.0281
0.95	1	0.9613	0.9631	0.9654	0.1663	0.1481	0.1451
	1.5	0.9639	0.9655	0.9656	0.1324	0.1225	0.1205
	2	0.9737	0.9768	0.9779	0.0840	0.0784	0.0747
	2.5	0.9901	0.9910	0.9918	0.0485	0.0432	0.0352

ภาพที่ 4.1.11 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.1.12</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางจะได้ว่า

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

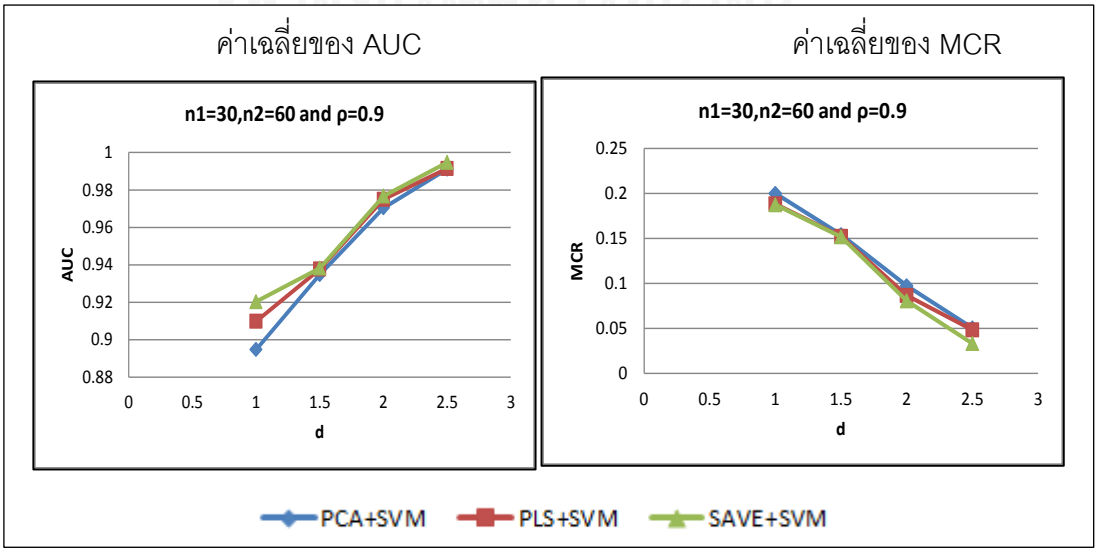
4.1.2.2 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างน้อยกว่ากลุ่มที่ไม่สนใจ

4.1.2.2.1 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $(n_1 = 30, n_2 = 60)$

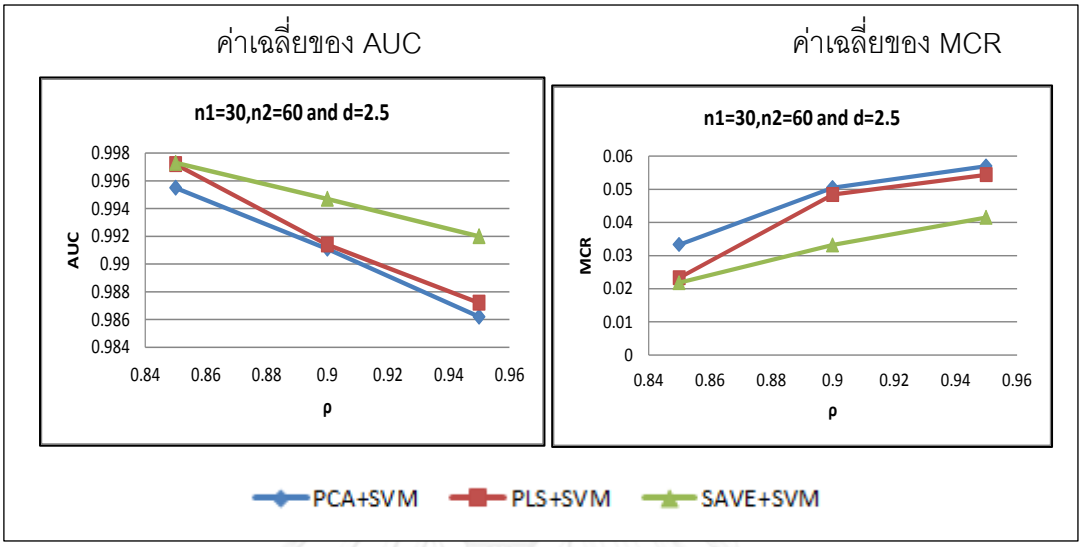
ตารางที่ 4.1.7 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9105	0.9152	0.9173	0.187	0.1805	0.1803
	1.5	0.9374	0.9457	0.9470	0.1476	0.1360	0.1359
	2	0.9757	0.9802	0.9828	0.0861	0.0745	0.0667
	2.5	0.9955	0.9972	0.9973	0.0333	0.0233	0.0218
0.9	1	0.8947	0.9098	0.9203	0.2000	0.1884	0.1875
	1.5	0.9348	0.9377	0.9382	0.1542	0.1523	0.1518
	2	0.9704	0.9750	0.9767	0.0974	0.0866	0.0806
	2.5	0.9911	0.9914	0.9947	0.0505	0.0484	0.0332
0.95	1	0.8995	0.9099	0.9188	0.2051	0.1930	0.1910
	1.5	0.9292	0.9351	0.9360	0.1608	0.1557	0.1527
	2	0.9645	0.9692	0.9724	0.1057	0.0983	0.0872
	2.5	0.9862	0.9872	0.9920	0.0570	0.0544	0.0415

ภาพที่ 4.1.13 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.14 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

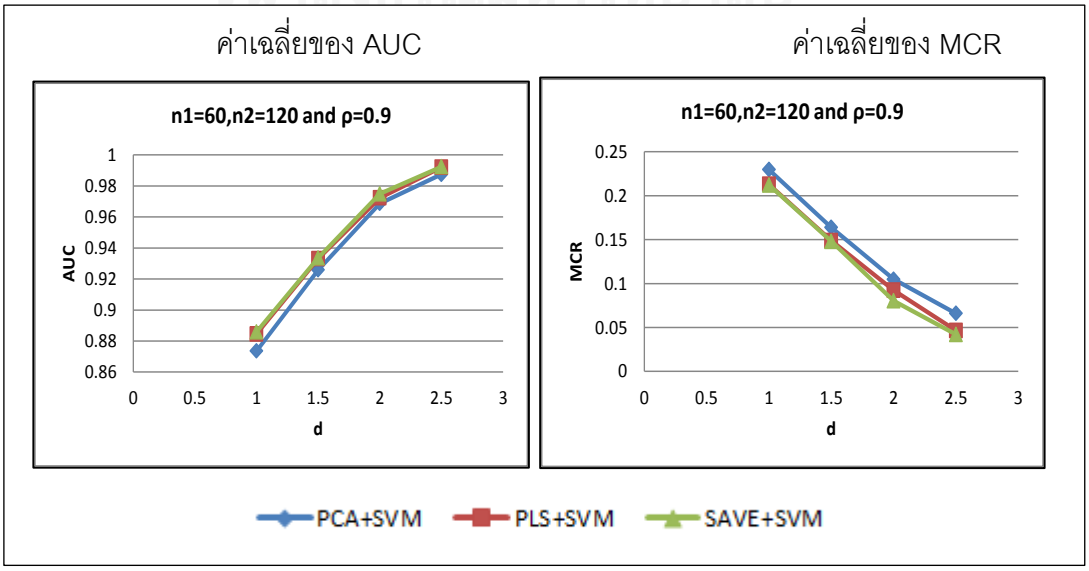
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.2.2.2 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 60, n_2 = 120)$

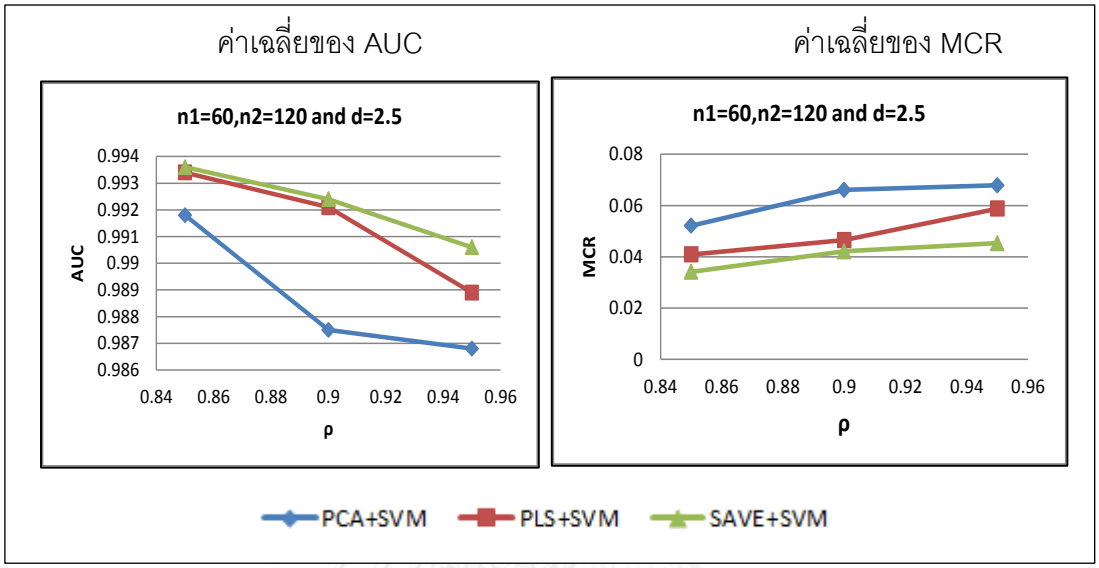
ตารางที่ 4.1.8 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8800	0.8854	0.8867	0.2178	0.2127	0.2119
	1.5	0.9283	0.9345	0.9414	0.1611	0.1483	0.1361
	2	0.9743	0.9762	0.9800	0.0968	0.0824	0.0714
	2.5	0.9918	0.9934	0.9936	0.0521	0.0409	0.0341
0.9	1	0.8736	0.8845	0.8859	0.2298	0.2129	0.2122
	1.5	0.9259	0.9331	0.9336	0.1638	0.1489	0.1481
	2	0.9687	0.9723	0.9749	0.1049	0.0924	0.0804
	2.5	0.9875	0.9921	0.9924	0.0661	0.0465	0.0421
0.95	1	0.8701	0.8842	0.8853	0.2304	0.2163	0.2160
	1.5	0.9145	0.9225	0.9242	0.1753	0.1641	0.1605
	2	0.9651	0.9654	0.9705	0.1122	0.1072	0.0900
	2.5	0.9868	0.9889	0.9906	0.0679	0.0588	0.0453

ภาพที่ 4.1.15 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.16 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

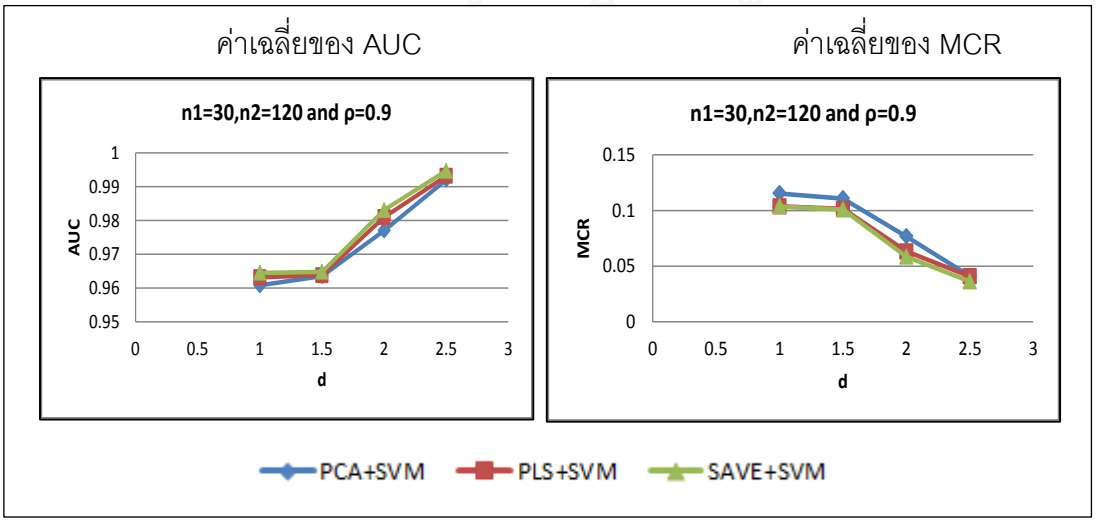
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.1.2.2.3 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 30, n_2 = 120)$

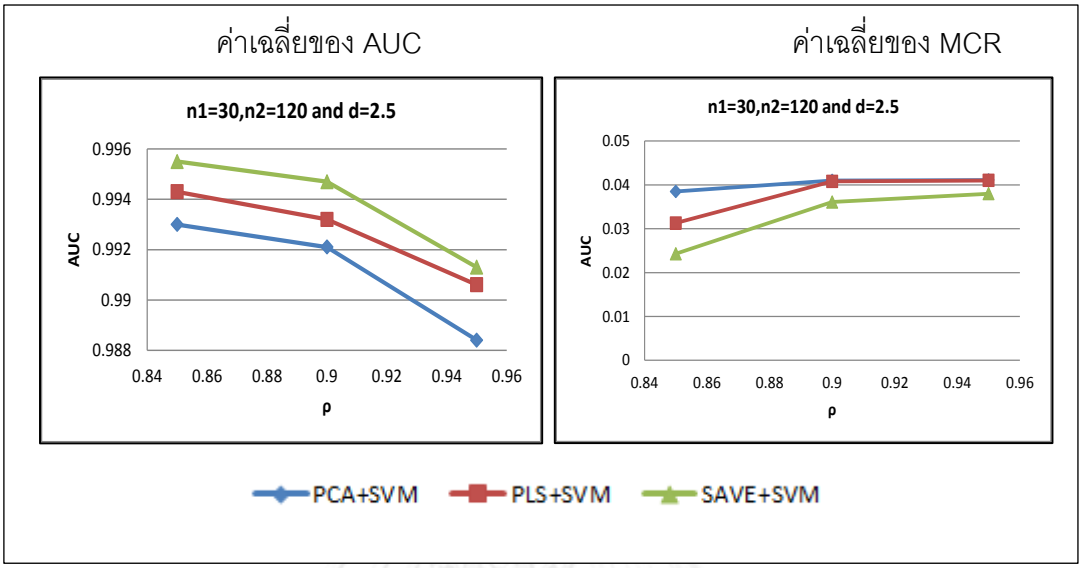
ตารางที่ 4.1.9 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9660	0.9662	0.9672	0.1073	0.1060	0.1021
	1.5	0.9662	0.9664	0.9674	0.1047	0.0966	0.0925
	2	0.9802	0.9832	0.9843	0.0698	0.0612	0.0528
	2.5	0.9930	0.9943	0.9955	0.0385	0.0313	0.0243
0.9	1	0.9608	0.9632	0.9645	0.1153	0.1037	0.1036
	1.5	0.9635	0.9638	0.9648	0.1107	0.1015	0.1008
	2	0.9770	0.9810	0.9831	0.0766	0.0632	0.0584
	2.5	0.9921	0.9932	0.9947	0.0410	0.0408	0.0361
0.95	1	0.9605	0.9630	0.9640	0.1207	0.1068	0.1057
	1.5	0.9606	0.9629	0.9683	0.1112	0.1066	0.1013
	2	0.9730	0.9760	0.9785	0.0818	0.0749	0.0671
	2.5	0.9884	0.9906	0.9913	0.0411	0.0410	0.0380

ภาพที่ 4.1.17 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.1.18 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด PLS และรองลงมาเป็นวิธี และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2 ตัวแปรอิสระ 10 ตัวแปร

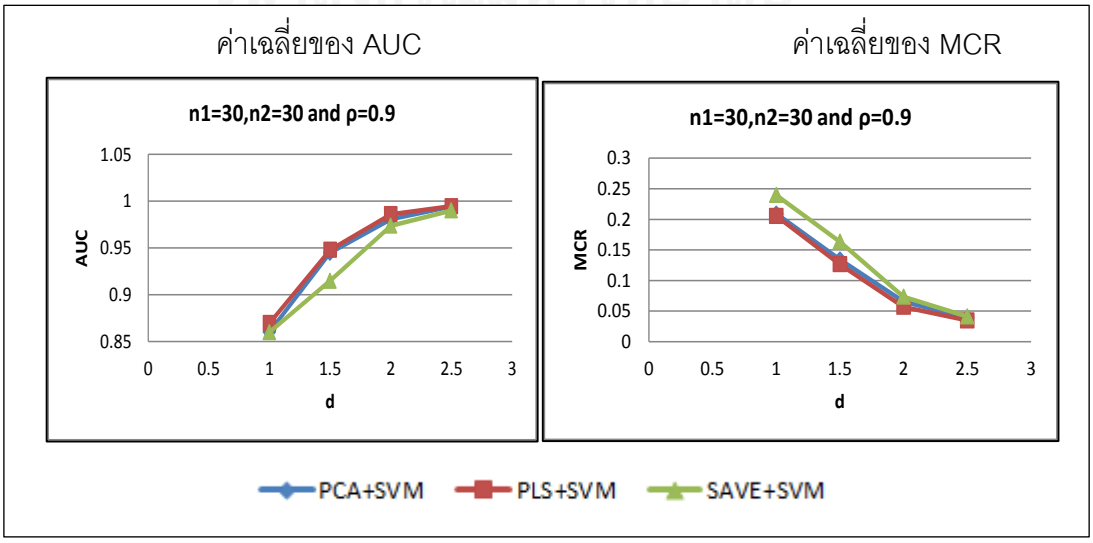
4.1.1 กรณีที่ขนาดตัวอย่างของแต่ละกลุ่มเท่ากัน

4.1.1.1 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 30 $(n_1 = n_2 = 30)$

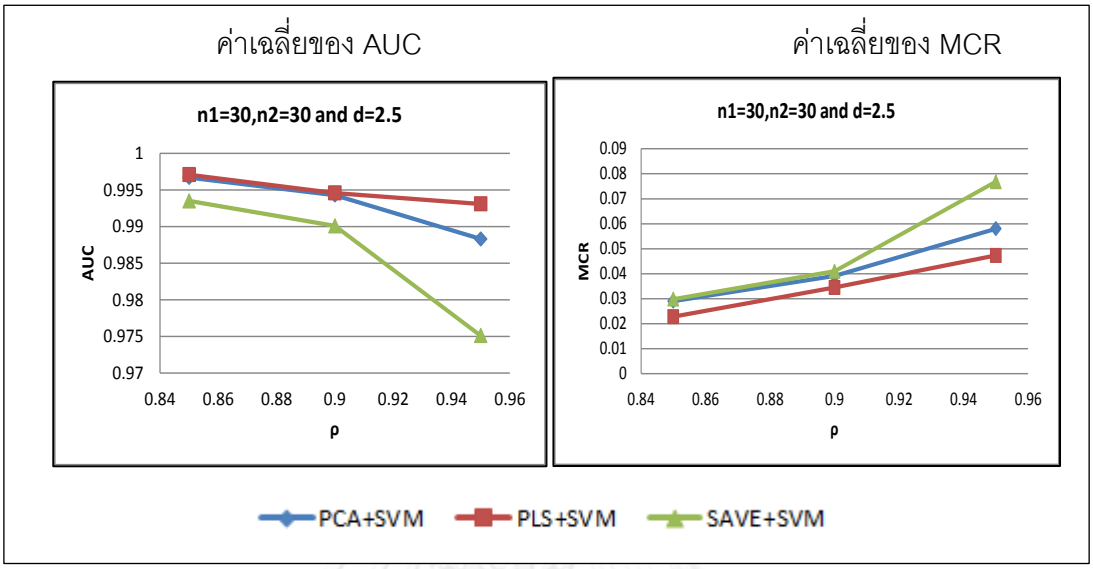
ตารางที่ 4.2.1 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8997	0.9016	0.8991	0.1985	0.1936	0.1994
	1.5	0.9419	0.9596	0.9316	0.1253	0.1035	0.1515
	2	0.9852	0.9880	0.9784	0.0578	0.0552	0.0704
	2.5	0.9967	0.9971	0.9935	0.0290	0.0228	0.0298
0.9	1	0.8603	0.8697	0.8600	0.2092	0.2055	0.2397
	1.5	0.9449	0.9480	0.9150	0.1342	0.1264	0.1633
	2	0.9807	0.9860	0.9738	0.0653	0.0567	0.0733
	2.5	0.9943	0.9946	0.9901	0.0391	0.0345	0.0410
0.95	1	0.8901	0.8923	0.8842	0.2148	0.2137	0.2167
	1.5	0.9260	0.9374	0.9119	0.1576	0.1399	0.1772
	2	0.9686	0.9717	0.941	0.1008	0.0884	0.1261
	2.5	0.9883	0.9931	0.9751	0.0580	0.0473	0.0768

ภาพที่ 4.2.1 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.2.2</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

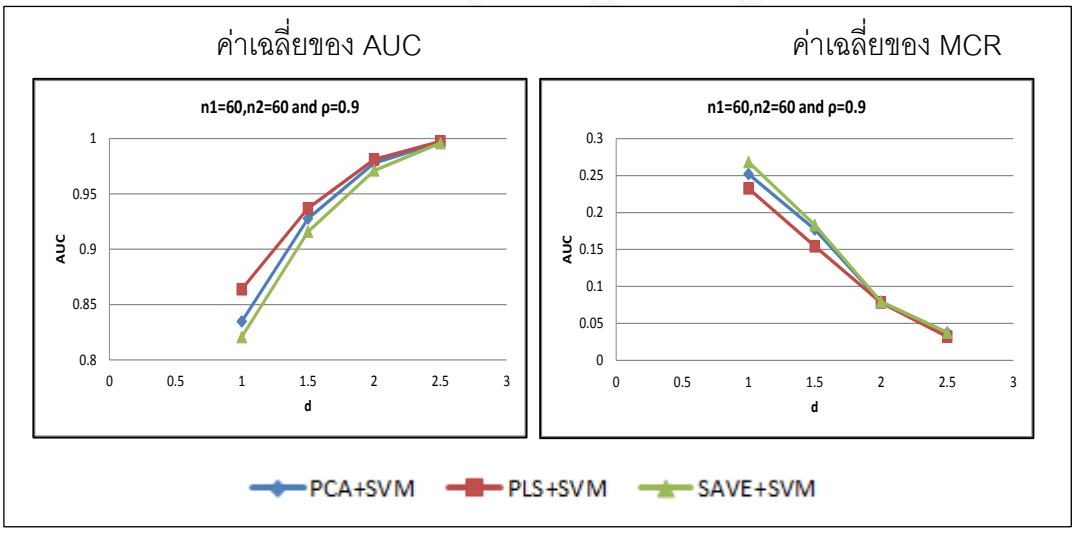
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. เมื่อค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 60 $\left(n_1 = n_2 = 60\right)$

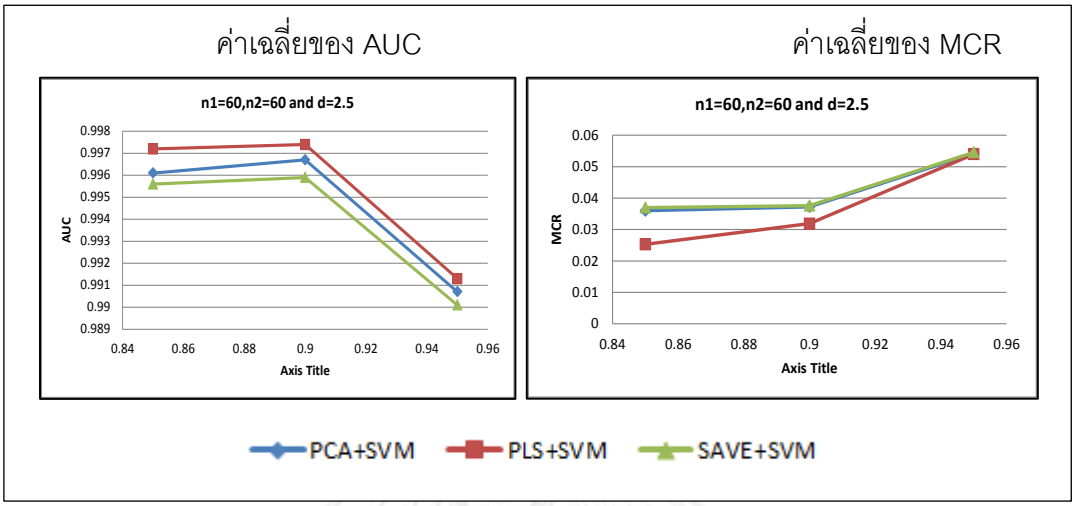
ตารางที่ 4.2.2 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8611	0.8714	0.8281	0.2256	0.2240	0.2643
	1.5	0.9307	0.9458	0.9254	0.1657	0.1542	0.1753
	2	0.9757	0.9802	0.9728	0.0539	0.0536	0.0582
	2	0.9961	0.9972	0.9956	0.0360	0.0253	0.0370
0.9	1	0.8346	0.8638	0.8206	0.2521	0.2327	0.2685
	1.5	0.9277	0.9370	0.9158	0.1771	0.1543	0.1830
	2	0.9780	0.9811	0.9712	0.0789	0.0782	0.0792
	2.5	0.9967	0.9974	0.9959	0.0372	0.0319	0.0376
0.95	1	0.8297	0.8475	0.8281	0.2608	0.2542	0.2620
	1.5	0.9120	0.9140	0.8965	0.1926	0.1842	0.2088
	2	0.9709	0.9741	0.9700	0.0911	0.0899	0.0969
	2.5	0.9907	0.9913	0.9901	0.0542	0.0540	0.0546

ภาพที่ 4.2.3 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.4 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

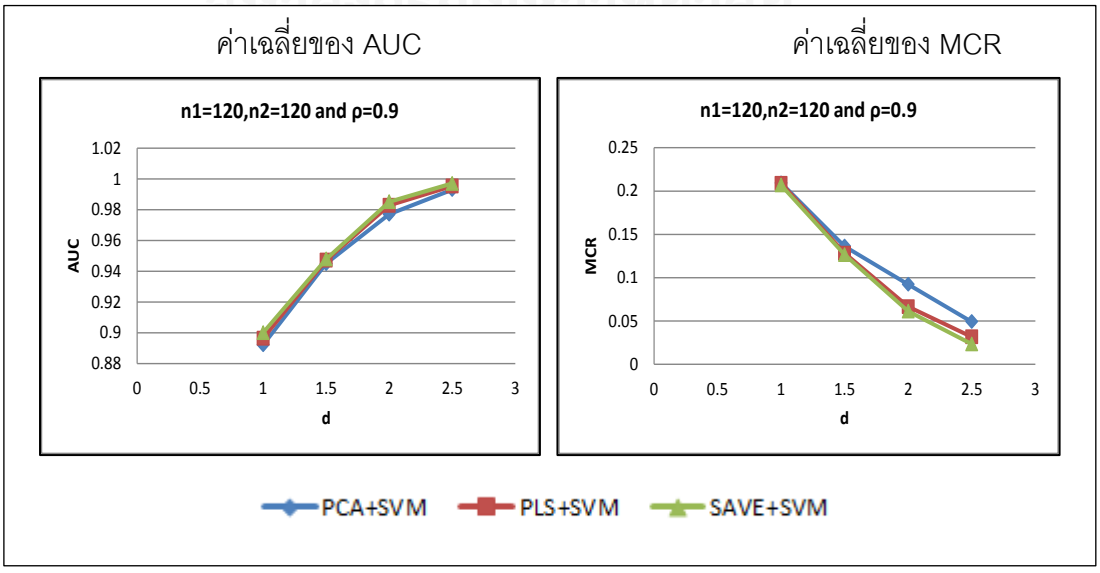
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.1.3 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ 120 $\left(n_1 = n_2 = 120\right)$

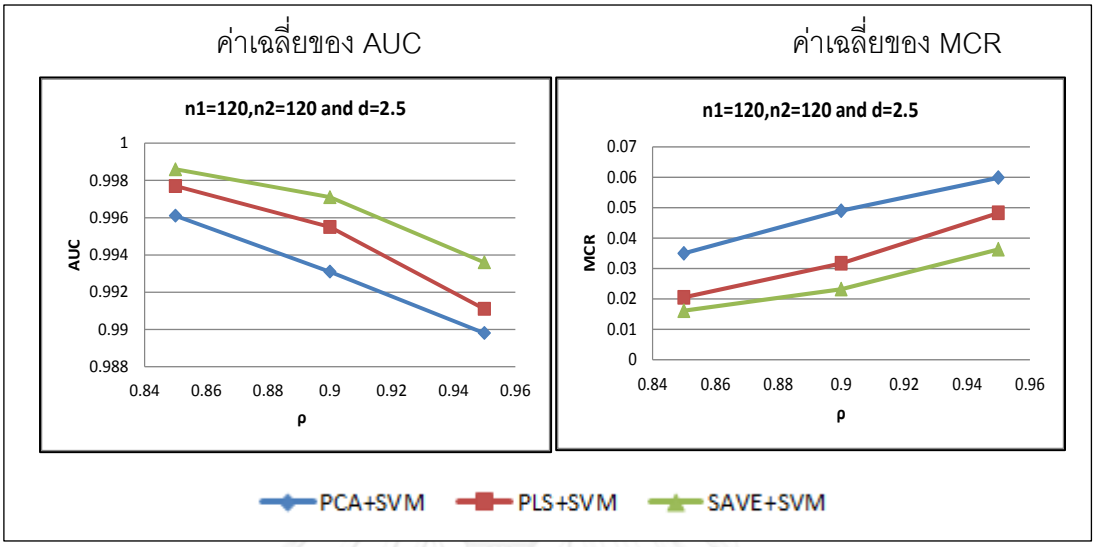
ตารางที่ 4.2.3 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8932	0.9054	0.9092	0.2065	0.1764	0.1727
	1.5	0.9511	0.9590	0.9597	0.1188	0.1122	0.1119
	2	0.9861	0.9888	0.9918	0.0673	0.0529	0.0444
	2.5	0.9961	0.9977	0.9986	0.0350	0.0205	0.0161
0.9	1	0.8921	0.8964	0.9001	0.2092	0.2089	0.2066
	1.5	0.9450	0.9472	0.9479	0.1361	0.1282	0.1263
	2	0.9771	0.9829	0.9852	0.0921	0.0665	0.0611
	2.5	0.9931	0.9955	0.9971	0.0490	0.0317	0.0232
0.95	1	0.8767	0.8951	0.9052	0.2228	0.2191	0.2143
	1.5	0.9265	0.9340	0.9360	0.1622	0.1685	0.1638
	2	0.9697	0.9729	0.9769	0.1053	0.0894	0.0814
	2.5	0.9898	0.9911	0.9936	0.0599	0.0483	0.0363

ภาพที่ 4.2.5 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.6 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.2 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจมีความแตกต่างกับกลุ่มตัวอย่างที่ไม่สนใจ

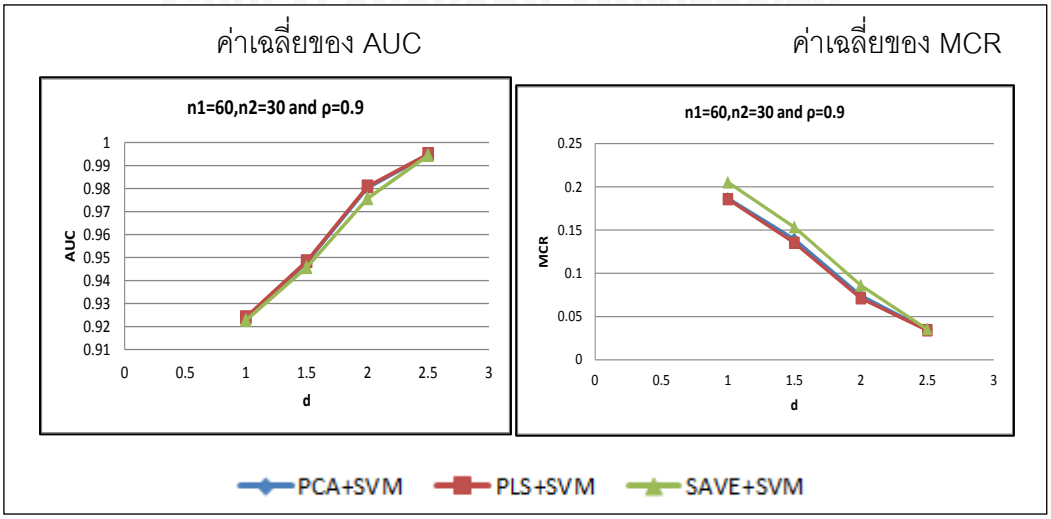
4.2.2.1 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างมากกว่ากลุ่มที่ไม่สนใจ

4.2.2.1.1 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $(n_1 = 60, n_2 = 30)$

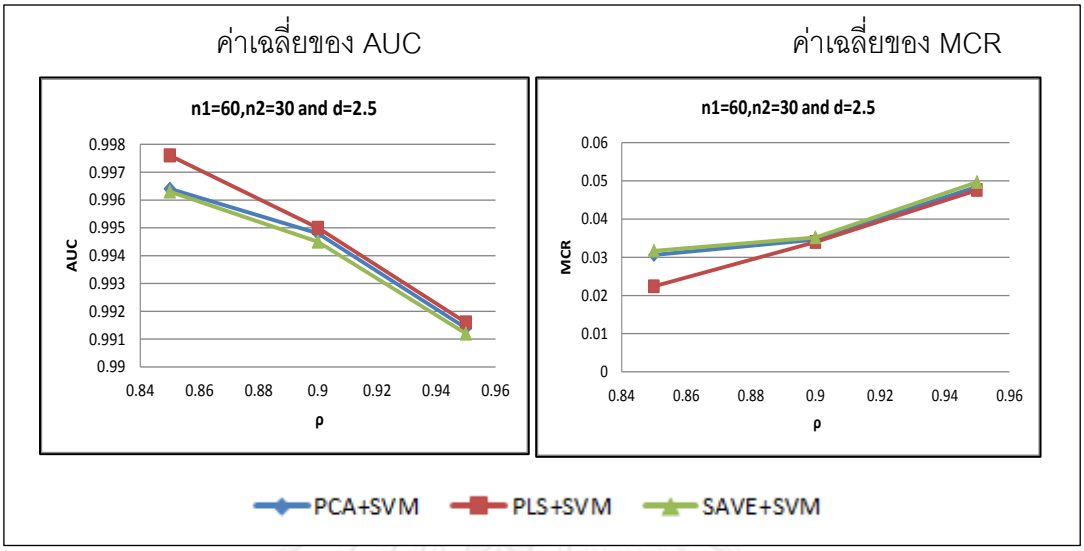
ตารางที่ 4.2.4 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9294	0.9299	0.9291	0.1761	0.1709	0.1972
	1.5	0.9604	0.9601	0.9529	0.1120	0.1109	0.1395
	2	0.9864	0.9877	0.9854	0.0579	0.0540	0.0649
	2.5	0.9964	0.9976	0.9963	0.0306	0.0224	0.0317
0.9	1	0.9235	0.9241	0.9228	0.1867	0.1856	0.2052
	1.5	0.9480	0.9484	0.9458	0.1395	0.1352	0.1533
	2	0.9803	0.9809	0.9757	0.0743	0.0710	0.0860
	2.5	0.9948	0.9950	0.9945	0.0346	0.0340	0.0352
0.95	1	0.9059	0.9085	0.9048	0.2305	0.2264	0.2204
	1.5	0.9380	0.9399	0.9351	0.1560	0.1536	0.1740
	2	0.9729	0.9734	0.9670	0.0909	0.0902	0.1063
	2.5	0.9914	0.9916	0.9912	0.0485	0.0476	0.0497

ภาพที่ 4.2.7 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.8 กราฟแสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

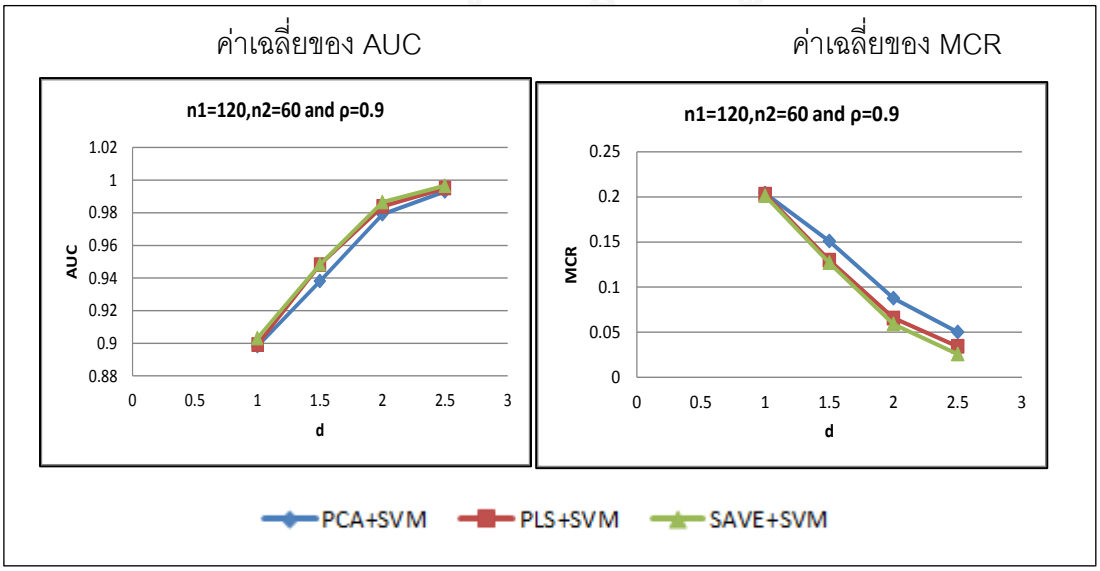
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.2.1.2 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 120, n_2 = 60)$

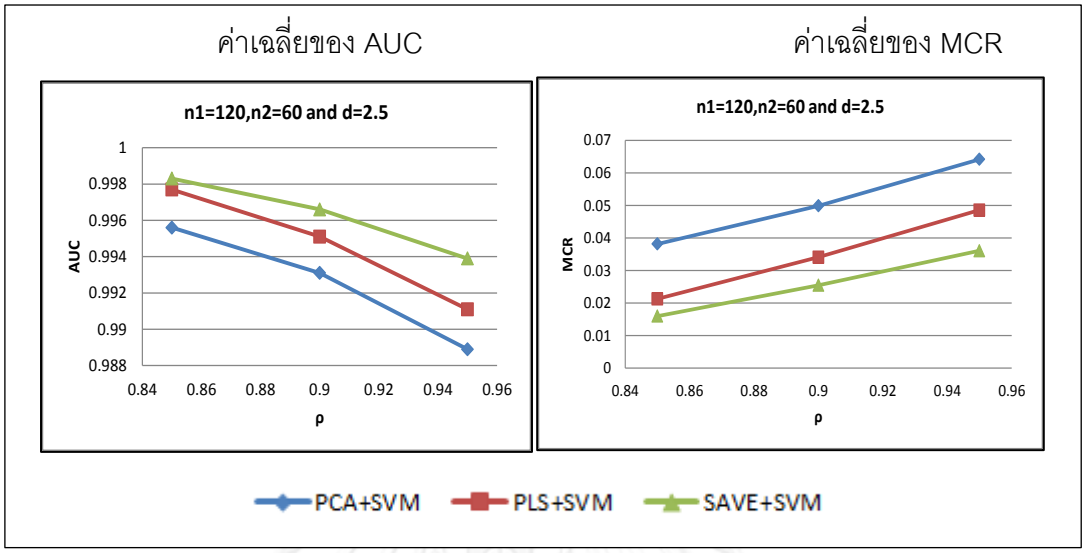
ตารางที่ 4.2.5 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8994	0.9035	0.9039	0.2032	0.1964	0.1933
	1.5	0.9503	0.9563	0.9575	0.1209	0.1163	0.1161
	2	0.9854	0.9886	0.9917	0.0690	0.0528	0.0445
	2.5	0.9956	0.9977	0.9983	0.0382	0.0213	0.0160
0.9	1	0.8982	0.8992	0.9032	0.2040	0.2031	0.2011
	1.5	0.9381	0.9482	0.9484	0.1509	0.1298	0.1269
	2	0.9791	0.9839	0.9866	0.0874	0.0656	0.0588
	2.5	0.9931	0.9951	0.9966	0.0499	0.0341	0.0255
0.95	1	0.8747	0.8903	0.9023	0.2401	0.2264	0.2213
	1.5	0.9252	0.9288	0.9286	0.1681	0.1578	0.1505
	2	0.9711	0.9750	0.9775	0.1036	0.0893	0.0810
	2.5	0.9889	0.9911	0.9939	0.0642	0.0486	0.0361

ภาพที่ 4.2.9 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.2.10</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

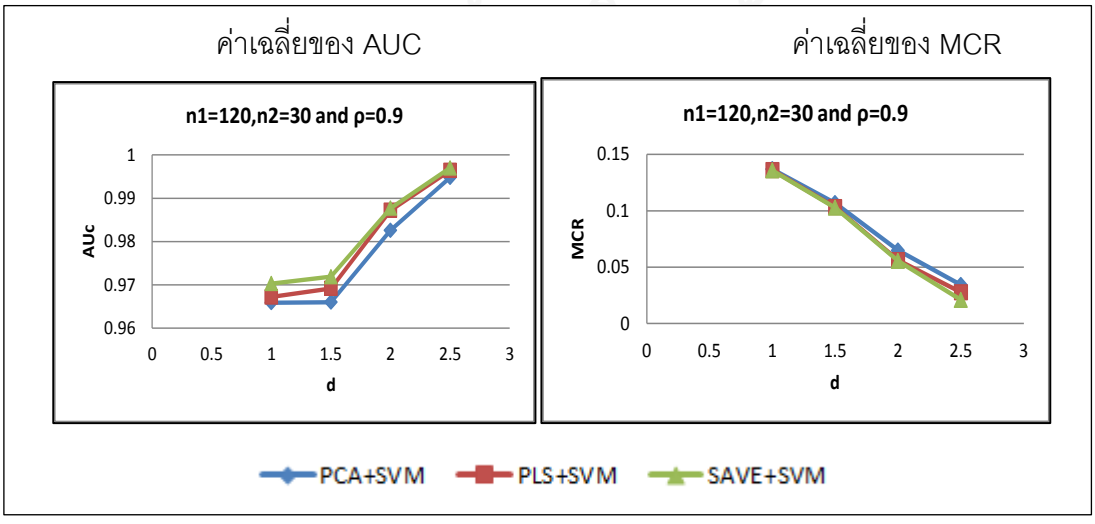
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. เมื่อพิจารณาค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.2.1.3 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 120, n_2 = 30)$

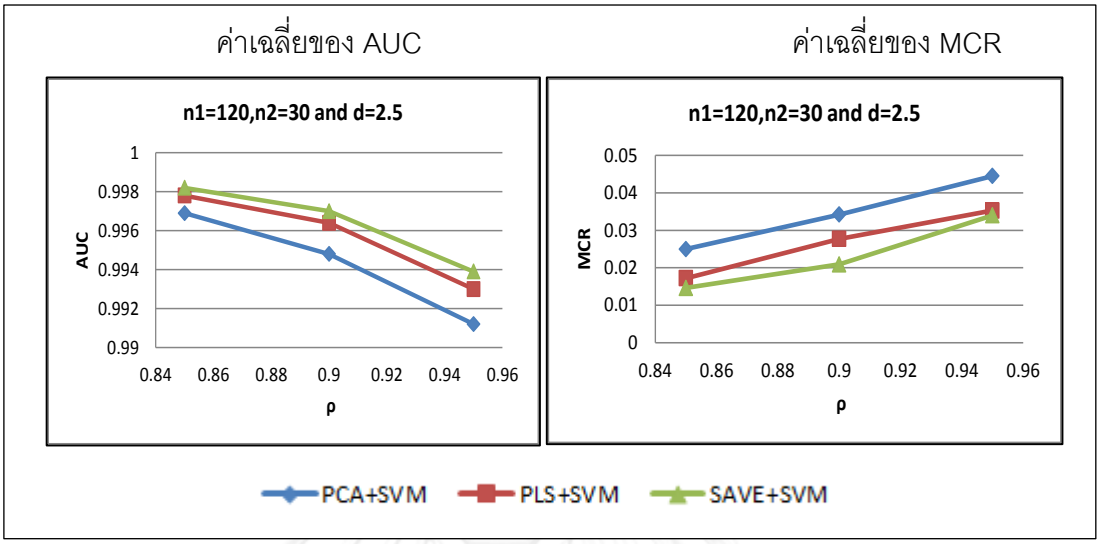
ตารางที่ 4.2.6 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9669	0.9681	0.9744	0.1159	0.1321	0.1355
	1.5	0.9738	0.9750	0.9751	0.0824	0.0871	0.0998
	2	0.9879	0.9898	0.9905	0.0487	0.0423	0.0416
	2.5	0.9969	0.9978	0.9982	0.0250	0.0172	0.0146
0.9	1	0.9659	0.9672	0.9703	0.1369	0.1361	0.1358
	1.5	0.9660	0.9692	0.9719	0.1067	0.1032	0.1027
	2	0.9826	0.9872	0.9877	0.0648	0.0562	0.0555
	2.5	0.9948	0.9964	0.9970	0.0342	0.0277	0.0209
0.95	1	0.9625	0.9701	0.9736	0.1570	0.1376	0.1375
	1.5	0.9599	0.9678	0.9706	0.1193	0.1071	0.1062
	2	0.9767	0.9805	0.9808	0.0791	0.0783	0.0781
	2.5	0.9912	0.9930	0.9939	0.0445	0.0353	0.0340

ภาพที่ 4.2.11 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.2.12</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

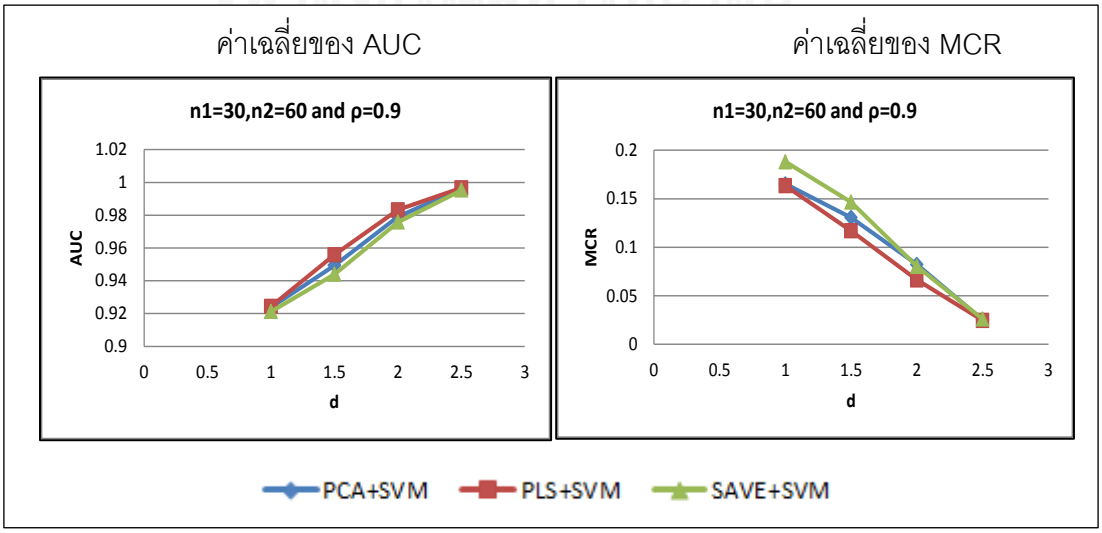
4.2.2.2 เมื่อขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างน้อยกว่ากลุ่มที่ไม่สนใจ

4.2.2.2.1 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $(n_1 = 30, n_2 = 60)$

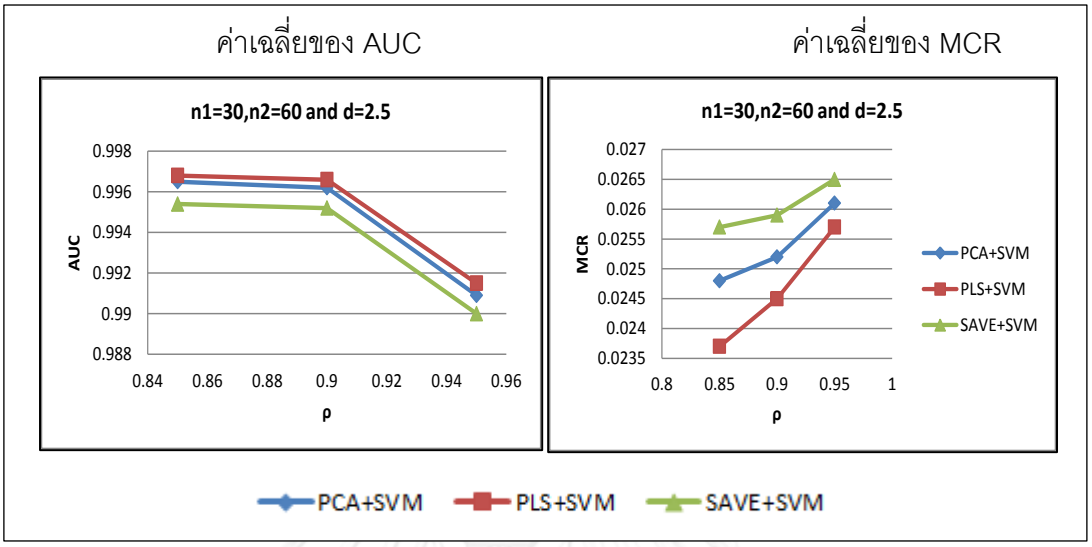
ตารางที่ 4.2.7 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9264	0.9316	0.9214	0.1510	0.1508	0.1828
	1.5	0.9606	0.9633	0.9509	0.1093	0.1072	0.1355
	2	0.9868	0.9890	0.9866	0.0602	0.0517	0.0579
	2.5	0.9965	0.9968	0.9954	0.0248	0.0237	0.0257
0.9	1	0.9239	0.9244	0.9212	0.1653	0.1634	0.1881
	1.5	0.9493	0.9559	0.9439	0.1305	0.1169	0.1465
	2	0.9791	0.9833	0.9757	0.0822	0.0663	0.0802
	2.5	0.9962	0.9966	0.9952	0.0252	0.0245	0.0259
0.95	1	0.9141	0.9233	0.9010	0.1878	0.1828	0.2007
	1.5	0.9352	0.9436	0.9244	0.1405	0.1380	0.1810
	2	0.9734	0.9744	0.9663	0.0910	0.0862	0.1032
	2.5	0.9909	0.9915	0.9900	0.0261	0.0257	0.0265

ภาพที่ 4.2.13 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.14 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

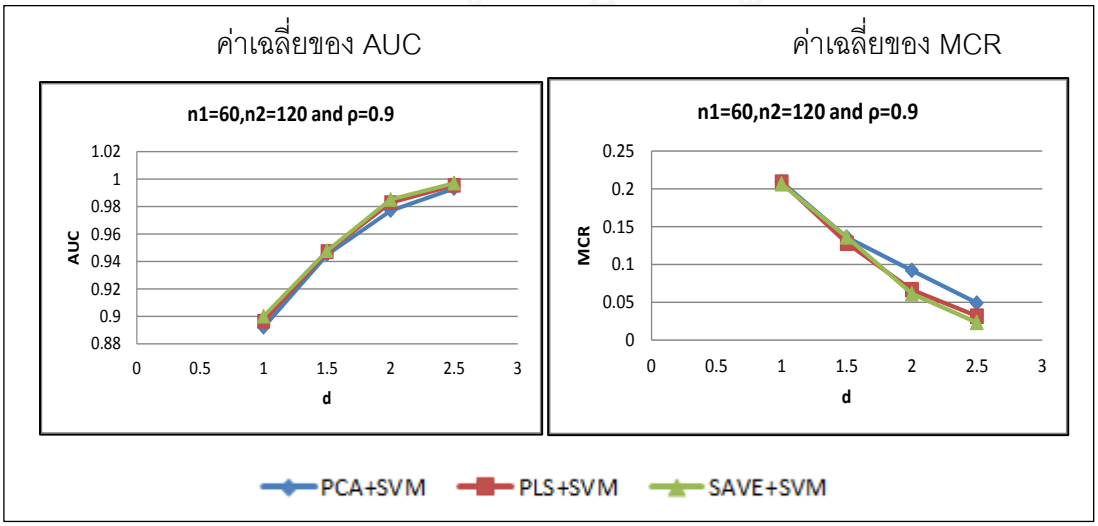
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.2.2.2 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 60, n_2 = 120)$

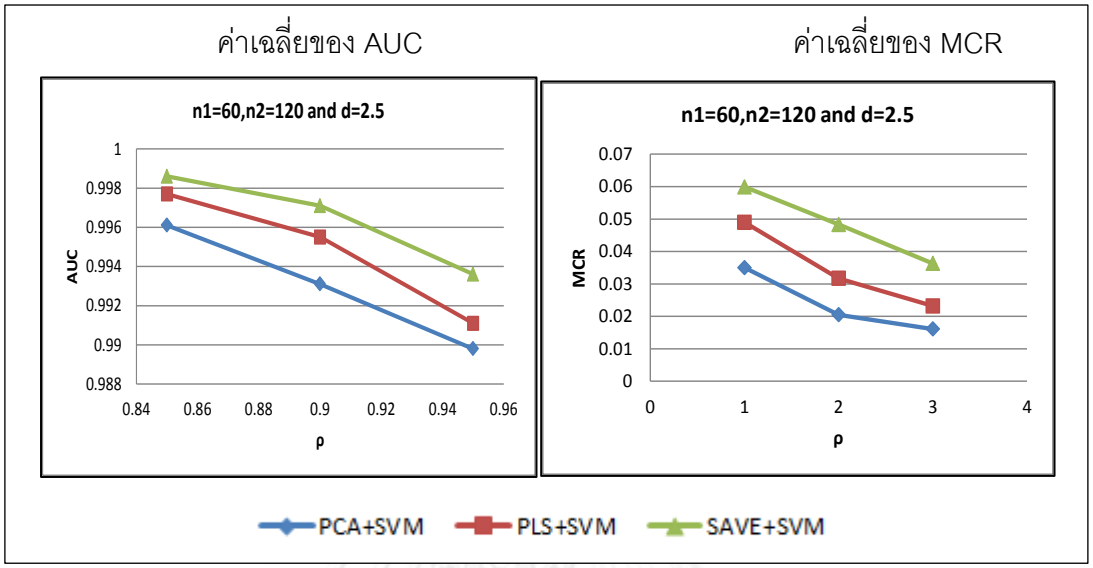
ตารางที่ 4.2.8 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.8931	0.9154	0.9192	0.2065	0.1864	0.1827
	1.5	0.9511	0.9590	0.9597	0.1188	0.1122	0.1139
	2	0.9861	0.9888	0.9918	0.0673	0.0529	0.0444
	2.5	0.9961	0.9977	0.9986	0.0350	0.0205	0.0161
0.9	1	0.8921	0.8964	0.9001	0.2092	0.2089	0.2066
	1.5	0.9450	0.9472	0.9479	0.1361	0.1282	0.1363
	2	0.9771	0.9829	0.9852	0.0921	0.0665	0.0611
	2.5	0.9931	0.9955	0.9971	0.0490	0.0317	0.0232
0.95	1	0.8717	0.8951	0.8992	0.2228	0.2191	0.2143
	1.5	0.9265	0.9340	0.9360	0.1692	0.1685	0.1638
	2	0.9697	0.9729	0.9769	0.1053	0.0894	0.0814
	2.5	0.9898	0.9911	0.9936	0.0599	0.0483	0.0363

ภาพที่ 4.2.15 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.16 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

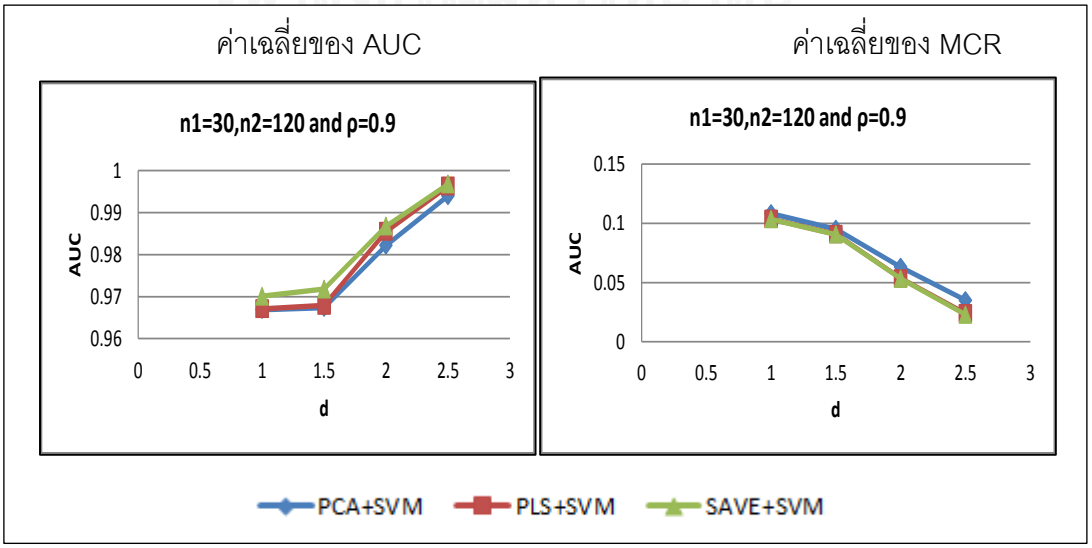
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.2.2.2.3 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 30, n_2 = 120)$

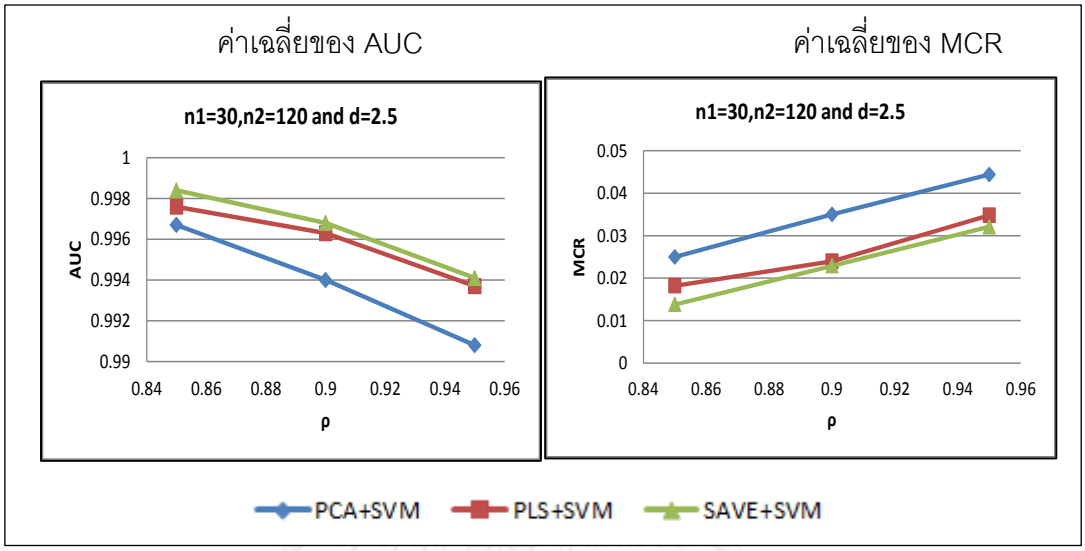
ตารางที่ 4.2.9 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9720	0.9729	0.9733	0.1011	0.1009	0.1001
	1.5	0.9734	0.9742	0.9744	0.0784	0.0784	0.0716
	2	0.9883	0.9892	0.9907	0.0450	0.0433	0.0414
	2.5	0.9967	0.9976	0.9984	0.0250	0.0182	0.0138
0.9	1	0.9669	0.9671	0.9701	0.1083	0.1037	0.1035
	1.5	0.9674	0.9679	0.9718	0.0952	0.0908	0.0906
	2	0.9822	0.9855	0.9868	0.0631	0.0535	0.0533
	2.5	0.9940	0.9963	0.9968	0.0350	0.0240	0.0229
0.95	1	0.9620	0.9686	0.9692	0.1232	0.1145	0.1134
	1.5	0.9664	0.9673	0.9699	0.1060	0.0989	0.0983
	2	0.9765	0.9810	0.9812	0.0750	0.0624	0.0616
	2.5	0.9908	0.9937	0.9941	0.0444	0.0348	0.0321

ภาพที่ 4.2.17 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.2.18 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี SAVE สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PLS และวิธี PCA ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้นทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3 ตัวแปรอิสระ 20 ตัวแปร

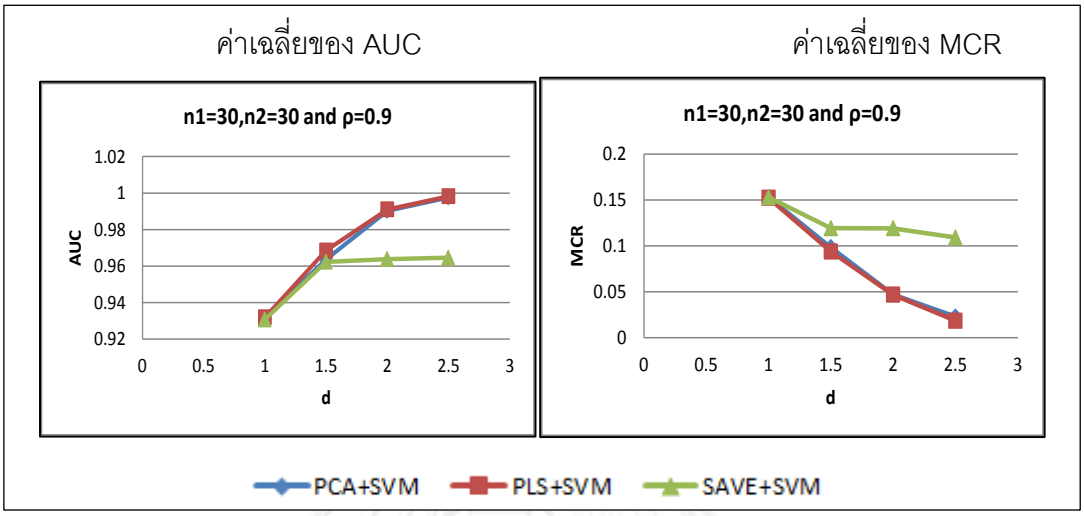
4.3.1 กรณีที่ขนาดตัวอย่างของแต่ละกลุ่มเท่ากัน

4.3.1.1 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $(n_1 = n_2 = 30)$

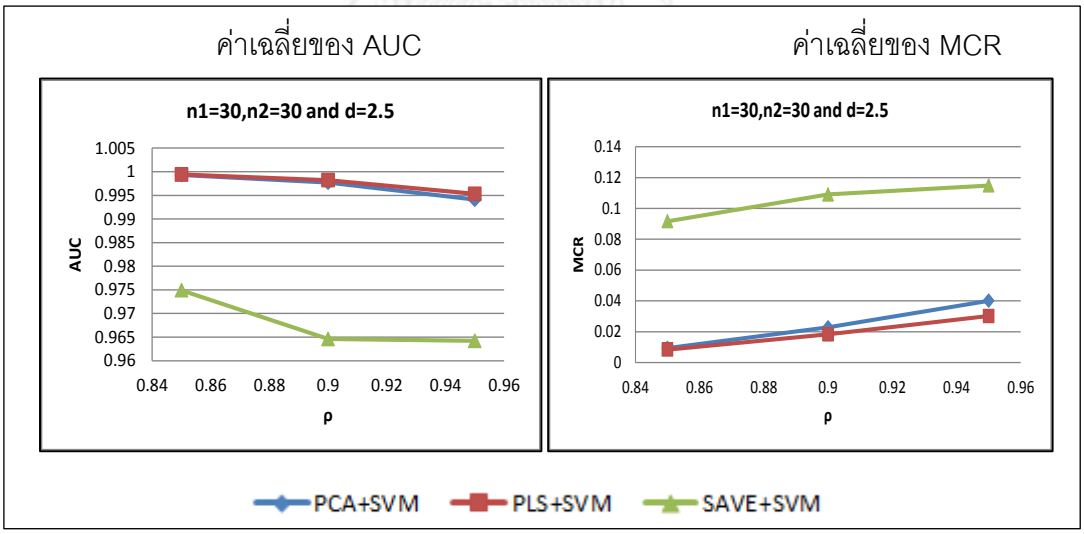
ตารางที่ 4.3.1 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9328	0.9331	0.9327	0.1374	0.1370	0.1392
	1.5	0.9795	0.9796	0.9539	0.0672	0.0651	0.1190
	2	0.9956	0.9957	0.9680	0.0278	0.0276	0.1180
	2.5	0.9993	0.9994	0.9749	0.0094	0.0084	0.0917
0.9	1	0.9317	0.9319	0.9307	0.1530	0.1523	0.1531
	1.5	0.9636	0.9685	0.9623	0.0984	0.0938	0.1193
	2	0.9903	0.9910	0.9638	0.0469	0.0467	0.1192
	2.5	0.9977	0.9982	0.9646	0.0228	0.0183	0.1091
0.95	1	0.9259	0.9282	0.9246	0.1752	0.1750	0.1759
	1.5	0.9601	0.9603	0.9586	0.1701	0.1700	0.1713
	2	0.9823	0.9839	0.9590	0.0766	0.0659	0.1199
	2.5	0.9941	0.9953	0.9642	0.0401	0.0302	0.1149

ภาพที่ 4.3.1 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.2 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

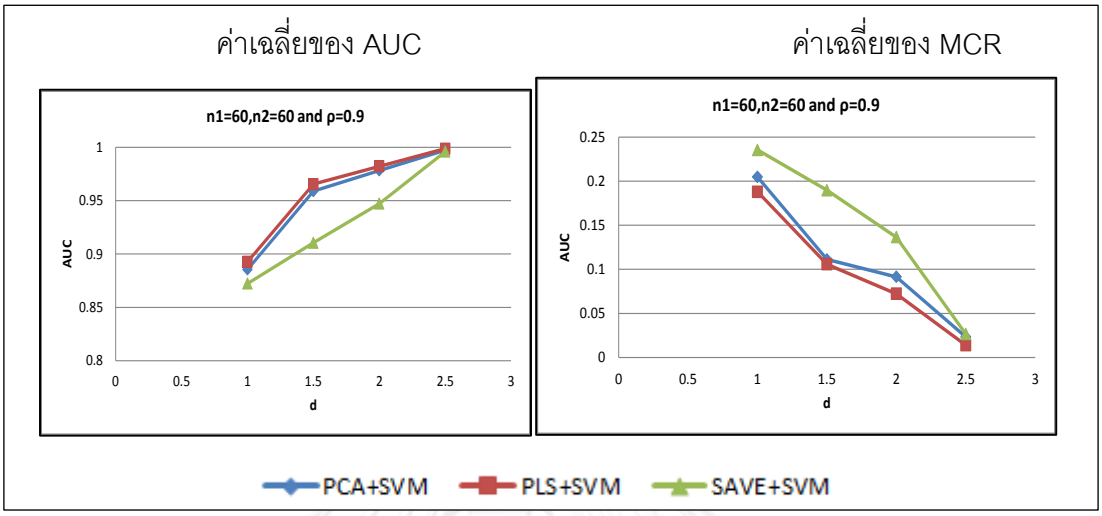
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $(n_1 = n_2 = 60)$

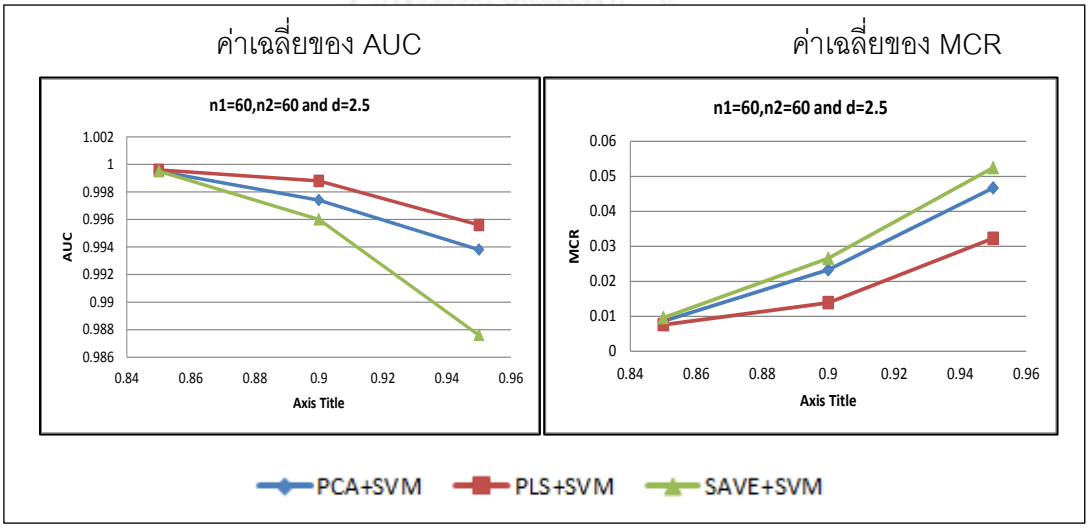
ตารางที่ 4.3.2 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9085	0.9193	0.8750	0.1783	0.1568	0.2340
	1.5	0.9788	0.9793	0.9385	0.0781	0.0718	0.1479
	2	0.9958	0.9964	0.9921	0.0298	0.0290	0.0407
	2.5	0.9995	0.9996	0.9995	0.0086	0.0076	0.0096
0.9	1	0.8855	0.8926	0.8723	0.2048	0.1880	0.2355
	1.5	0.9590	0.9656	0.9106	0.1112	0.1058	0.1898
	2	0.9783	0.9823	0.9473	0.0915	0.0723	0.1365
	2.5	0.9974	0.9988	0.9960	0.0233	0.0139	0.0266
0.95	1	0.8544	0.8580	0.8542	0.2280	0.2225	0.2388
	1.5	0.9320	0.9439	0.8902	0.1637	0.1389	0.2156
	2	0.9765	0.9818	0.9463	0.0938	0.0724	0.1399
	2.5	0.9938	0.9956	0.9876	0.0467	0.0323	0.0525

ภาพที่ 4.3.3 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.4 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

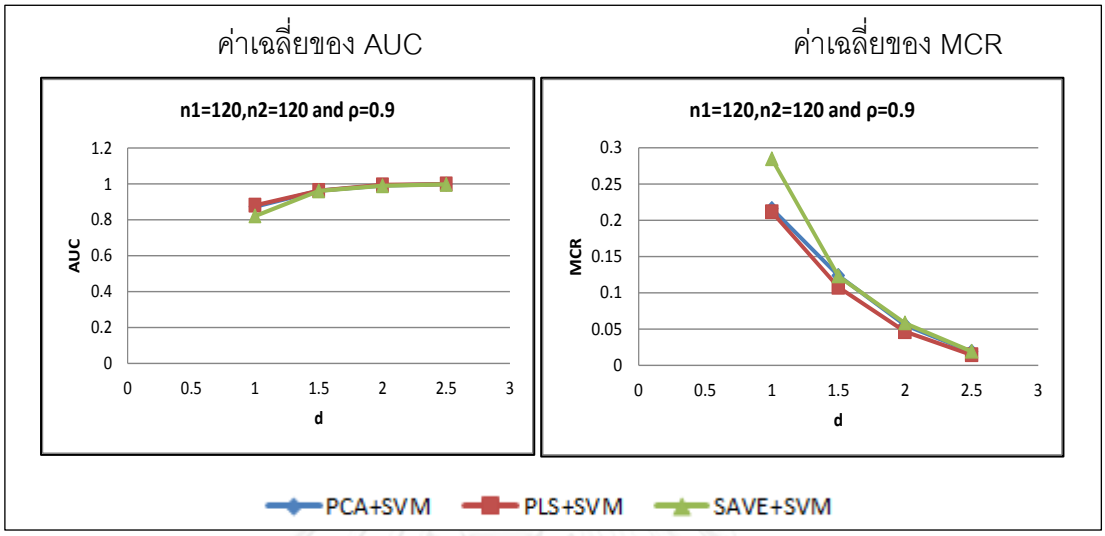
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $(n_1 = n_2 = 120)$

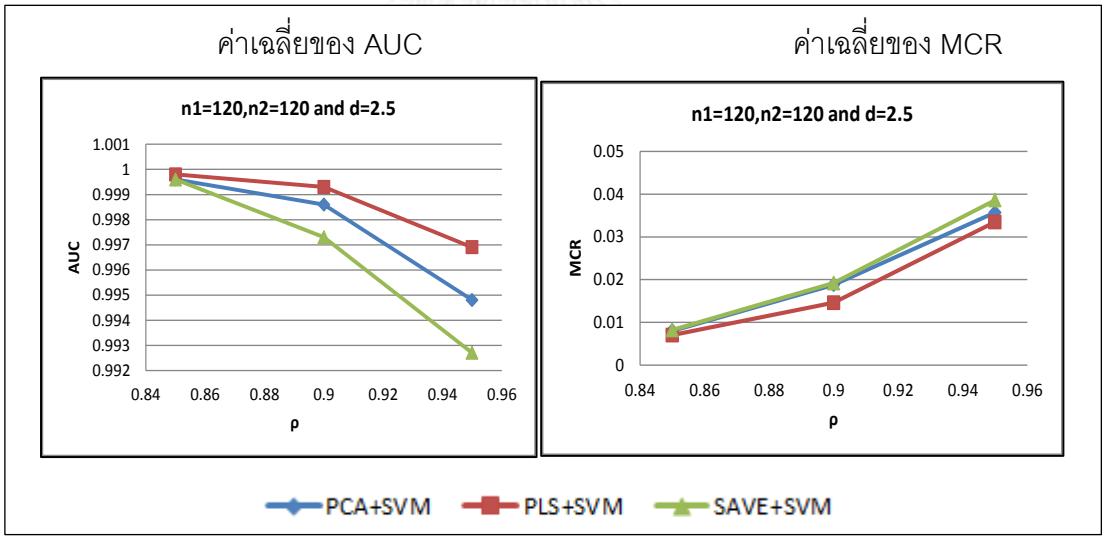
ตารางที่ 4.3.3 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9081	0.9113	0.8442	0.1738	0.1822	0.2606
	1.5	0.9780	0.9795	0.9704	0.0762	0.0784	0.0847
	2	0.9968	0.9980	0.9957	0.0220	0.0214	0.0226
	2.5	0.9996	0.9998	0.9996	0.0080	0.0070	0.0082
0.9	1	0.8734	0.8811	0.8216	0.2161	0.2117	0.2846
	1.5	0.9622	0.9629	0.9614	0.1237	0.1077	0.1229
	2	0.9908	0.9947	0.9897	0.0555	0.0464	0.0582
	2.5	0.9986	0.9993	0.9973	0.0188	0.0146	0.0192
0.95	1	0.8330	0.8403	0.7822	0.2555	0.2483	0.3192
	1.5	0.9327	0.9397	0.9121	0.1745	0.1426	0.1894
	2	0.9795	0.9845	0.9750	0.0768	0.0764	0.0775
	2.5	0.9948	0.9969	0.9927	0.0357	0.0335	0.0386

ภาพที่ 4.3.5 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.6 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจมีความแตกต่างกับกลุ่มตัวอย่างที่ไม่สนใจ

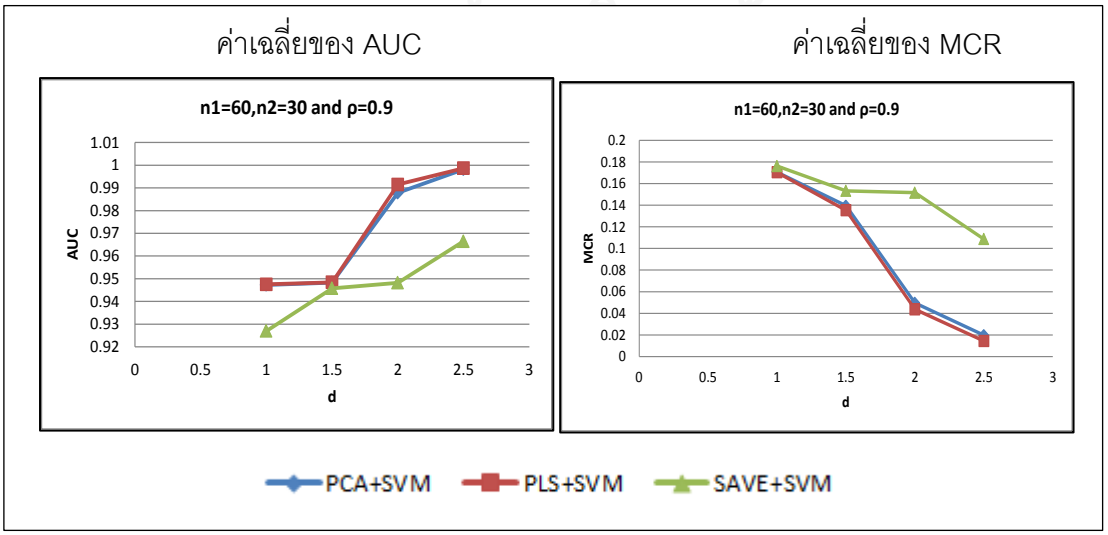
4.3.2.1 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างมากกว่ากลุ่มที่ไม่สนใจ

4.3.2.1.1 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $\left(n_1 = 60, n_2 = 30\right)$

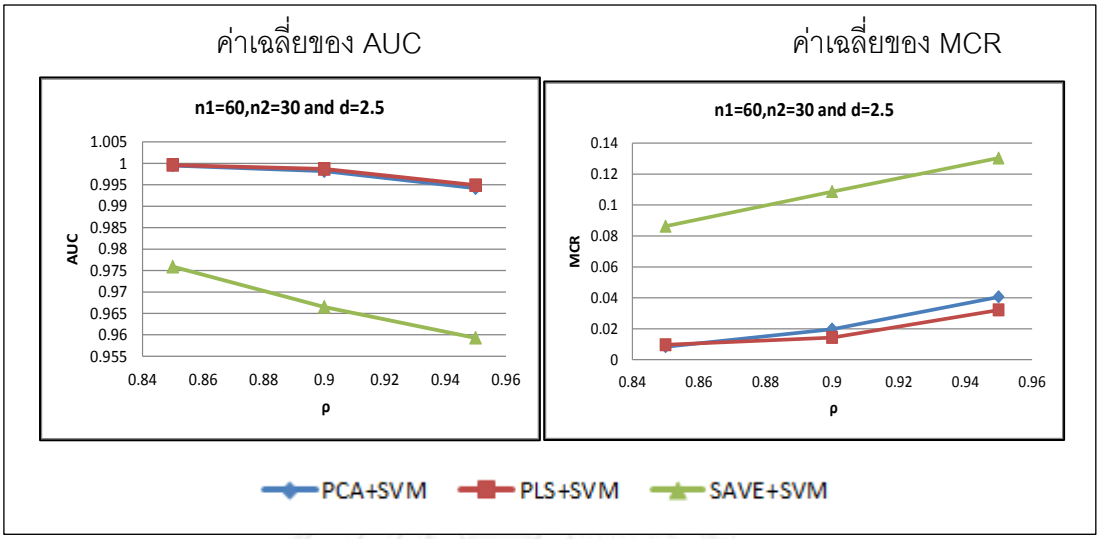
ตารางที่ 4.3.4 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9555	0.9566	0.9443	0.1536	0.1217	0.1545
	1.5	0.9782	0.9792	0.9450	0.0845	0.0737	0.1535
	2	0.9967	0.9968	0.9610	0.0216	0.0266	0.1178
	2.5	0.9995	0.9996	0.9759	0.0084	0.0096	0.0863
0.9	1	0.9473	0.9475	0.9270	0.1707	0.1705	0.1763
	1.5	0.9483	0.9484	0.9458	0.1395	0.1353	0.1533
	2	0.9879	0.9915	0.9482	0.0496	0.0436	0.1516
	2.5	0.9982	0.9987	0.9665	0.0197	0.0143	0.1087
0.95	1	0.9223	0.9225	0.9215	0.1983	0.1982	0.1988
	1.5	0.9380	0.9385	0.9351	0.1560	0.1536	0.1740
	2	0.9804	0.9831	0.9511	0.0758	0.0699	0.1551
	2.5	0.9942	0.9949	0.9593	0.0406	0.0321	0.1304

ภาพที่ 4.3.7 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.3.8</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

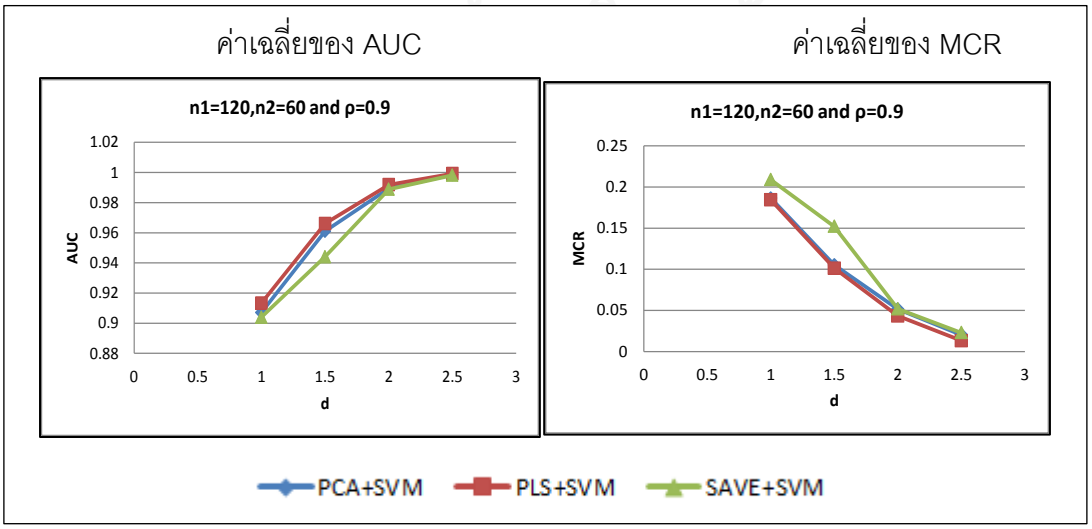
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.1.2 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 120, n_2 = 60)$

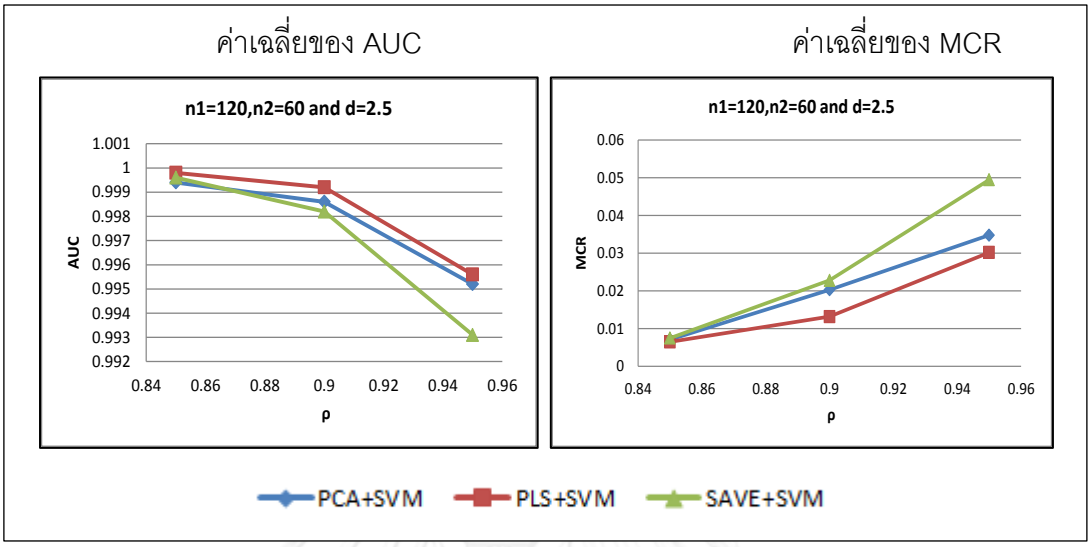
ตารางที่ 4.3.5 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9218	0.9222	0.9194	0.1631	0.1617	0.1984
	1.5	0.9761	0.9814	0.9673	0.0759	0.0741	0.1041
	2	0.9966	0.9968	0.9957	0.0278	0.0251	0.0293
	2.5	0.9994	0.9998	0.9996	0.0071	0.0065	0.0075
0.9	1	0.9071	0.9133	0.9040	0.1864	0.1844	0.2088
	1.5	0.9610	0.9661	0.9441	0.1052	0.1013	0.1520
	2	0.9894	0.9919	0.9889	0.0515	0.0432	0.0517
	2.5	0.9986	0.9992	0.9982	0.0263	0.0132	0.0228
0.95	1	0.8974	0.9096	0.8959	0.2124	0.2119	0.2179
	1.5	0.9383	0.9447	0.9243	0.1491	0.1336	0.1872
	2	0.9775	0.9823	0.9707	0.0910	0.0677	0.1036
	2.5	0.9952	0.9956	0.9931	0.0348	0.0302	0.0495

ภาพที่ 4.3.9 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.10 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

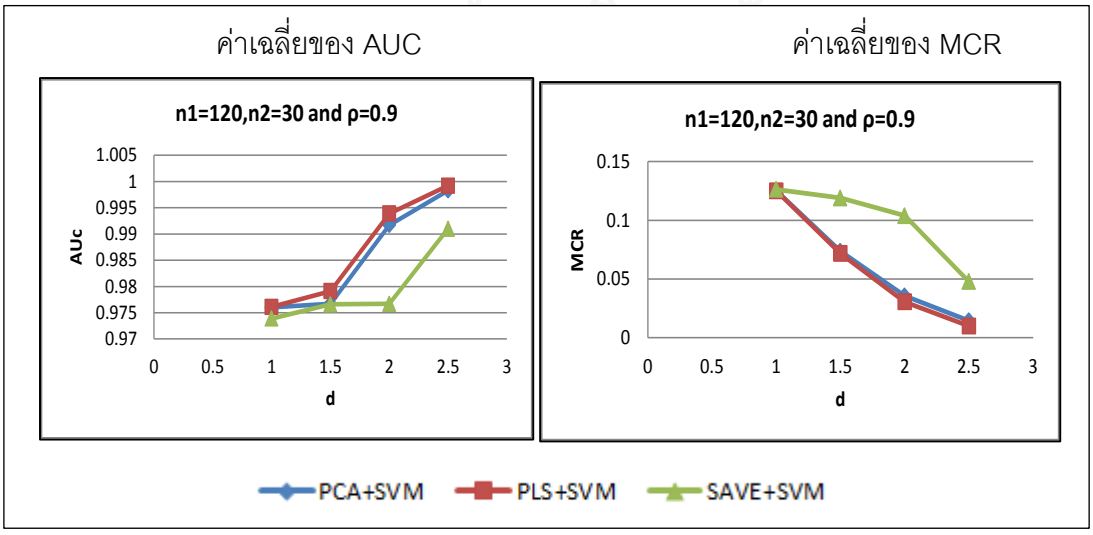
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.1.3 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $\left(n_1 = 120, n_2 = 30\right)$

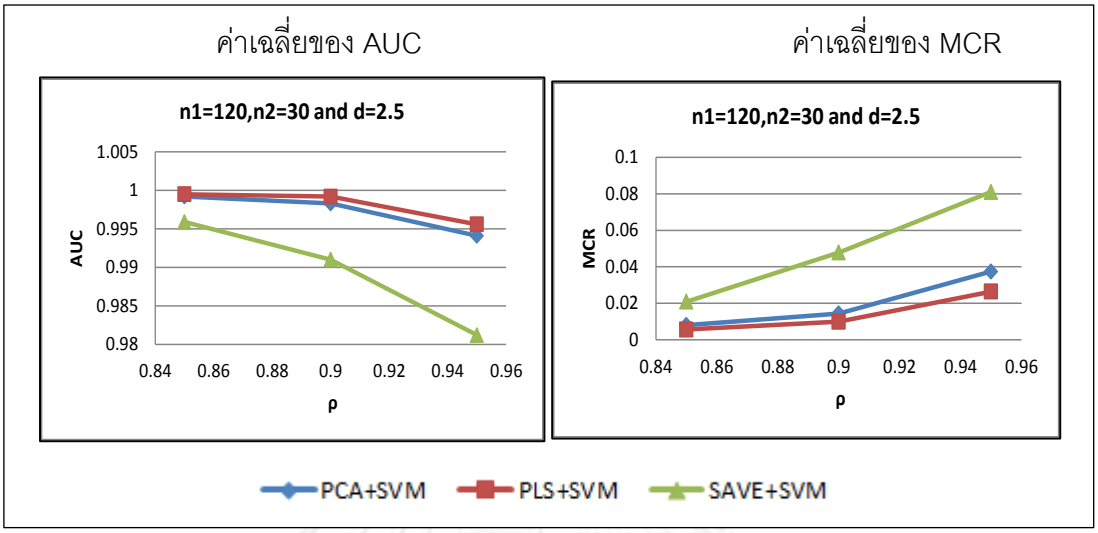
ตารางที่ 4.3.6 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9763	0.9769	0.9740	0.1119	0.1259	0.1240
	1.5	0.9852	0.9854	0.9754	0.0548	0.0528	0.1081
	2	0.9967	0.9970	0.9848	0.0202	0.0209	0.0769
	2.5	0.9992	0.9995	0.9959	0.0080	0.0056	0.0209
0.9	1	0.9760	0.9761	0.9739	0.1254	0.1251	0.1263
	1.5	0.9767	0.9791	0.9766	0.0736	0.0717	0.1190
	2	0.9917	0.9939	0.9767	0.0356	0.0305	0.1039
	2.5	0.9983	0.9992	0.9910	0.0144	0.0099	0.0478
0.95	1	0.9754	0.9759	0.9737	0.1253	0.1201	0.1355
	1.5	0.9788	0.9790	0.9747	0.1108	0.0956	0.1220
	2	0.9815	0.9866	0.9740	0.0666	0.0517	0.1168
	2.5	0.9941	0.9956	0.9812	0.0374	0.0264	0.0811

ภาพที่ 4.3.11 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.3.12</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

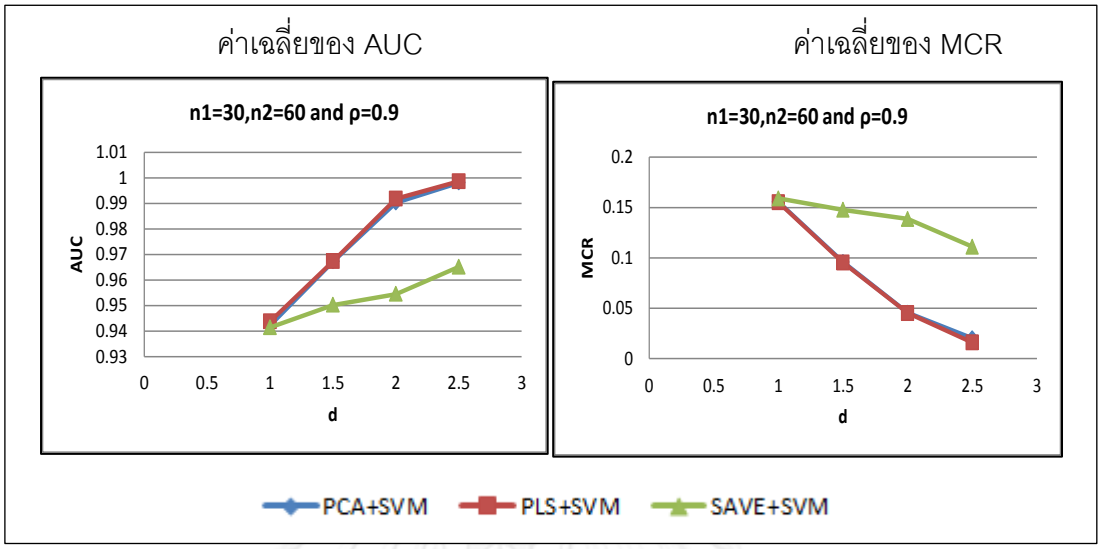
4.3.2.2 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างน้อยกว่ากลุ่มที่ไม่สนใจ

4.3.2.2.1 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $(n_1 = 30, n_2 = 60)$

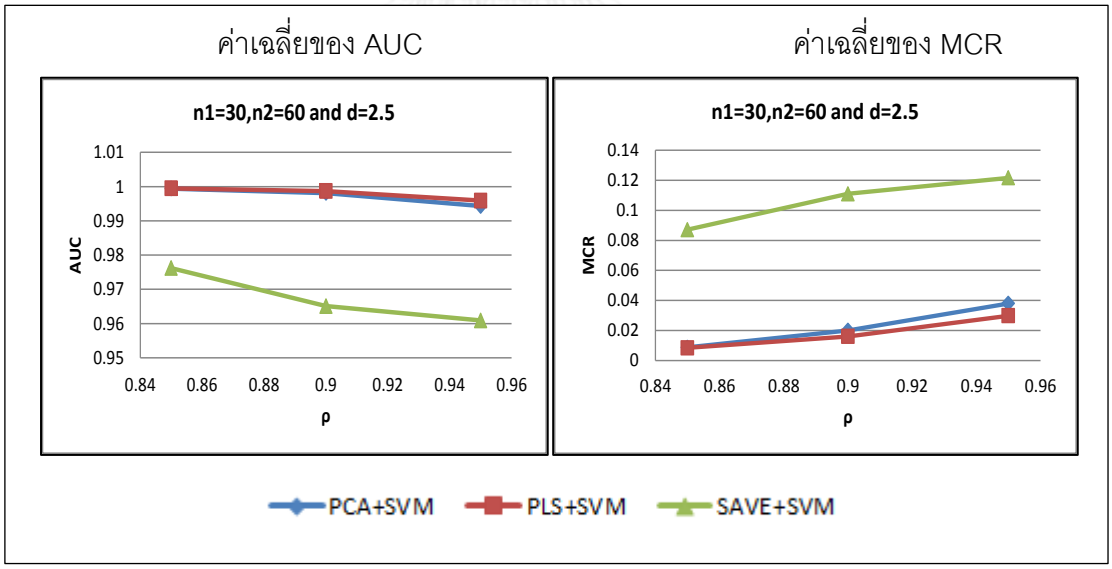
ตารางที่ 4.3.7 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9429	0.9450	0.9353	0.1485	0.1415	0.1588
	1.5	0.9827	0.9855	0.9548	0.0733	0.0727	0.1433
	2	0.9964	0.9969	0.9622	0.0238	0.0233	0.1197
	2.5	0.9994	0.9995	0.9762	0.0087	0.0084	0.0871
0.9	1	0.9422	0.9438	0.9414	0.1557	0.1552	0.1590
	1.5	0.9672	0.9674	0.9502	0.0959	0.0953	0.1477
	2	0.9903	0.9919	0.9545	0.0454	0.0452	0.1387
	2.5	0.9981	0.9987	0.9651	0.0200	0.0160	0.1110
0.95	1	0.9362	0.9370	0.9336	0.1626	0.1622	0.1690
	1.5	0.9461	0.9542	0.9424	0.1394	0.1224	0.1463
	2	0.9779	0.9827	0.9495	0.0804	0.0709	0.1391
	2.5	0.9943	0.9959	0.9609	0.0380	0.0298	0.1217

ภาพที่ 4.3.13 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.14 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

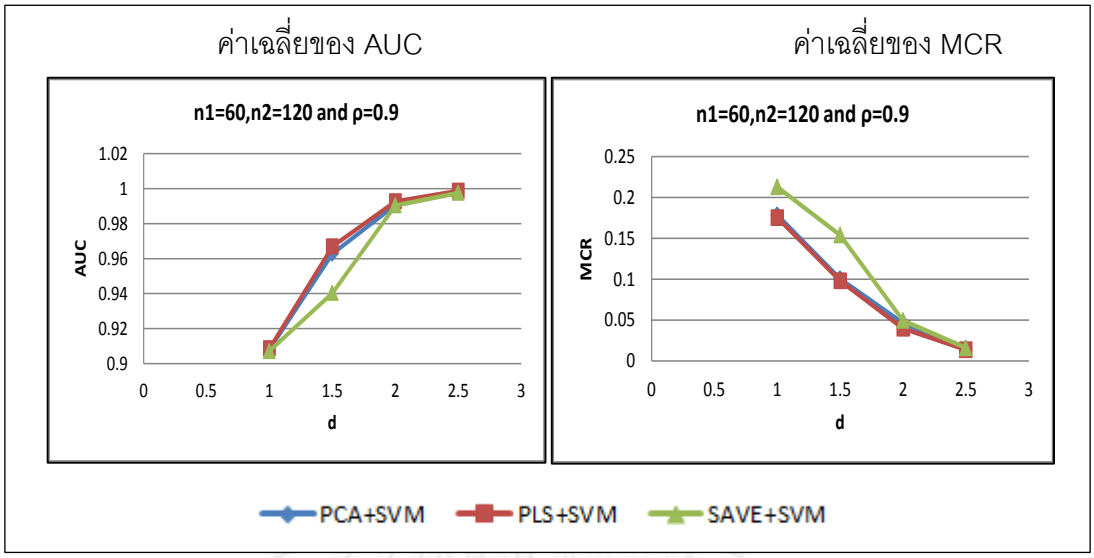
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.2.2 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 60, n_2 = 120)$

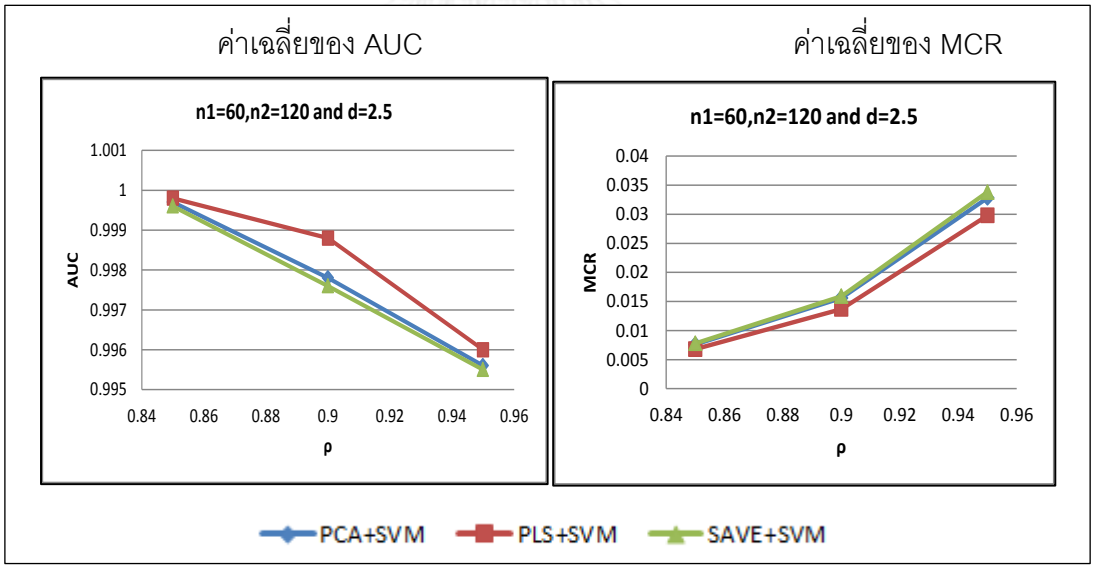
ตารางที่ 4.3.8 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9205	0.9249	0.9116	0.1638	0.1591	0.2045
	1.5	0.9786	0.9808	0.9674	0.0686	0.0690	0.1072
	2	0.9968	0.9971	0.9961	0.0259	0.0254	0.0262
	2.5	0.9997	0.9998	0.9996	0.0076	0.0068	0.0078
0.9	1	0.9079	0.9087	0.9072	0.1787	0.1754	0.2130
	1.5	0.9628	0.9670	0.9404	0.1005	0.0980	0.1540
	2	0.9906	0.9927	0.9904	0.0463	0.0398	0.0495
	2.5	0.9978	0.9988	0.9976	0.0156	0.0137	0.0159
0.95	1	0.8959	0.8973	0.8958	0.1977	0.1974	0.2105
	1.5	0.9374	0.9469	0.9220	0.1518	0.1309	0.1885
	2	0.9772	0.9823	0.9716	0.0901	0.0686	0.0986
	2.5	0.9956	0.9960	0.9955	0.0328	0.0298	0.0338

ภาพที่ 4.3.15 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.16 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางจะได้ว่า

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

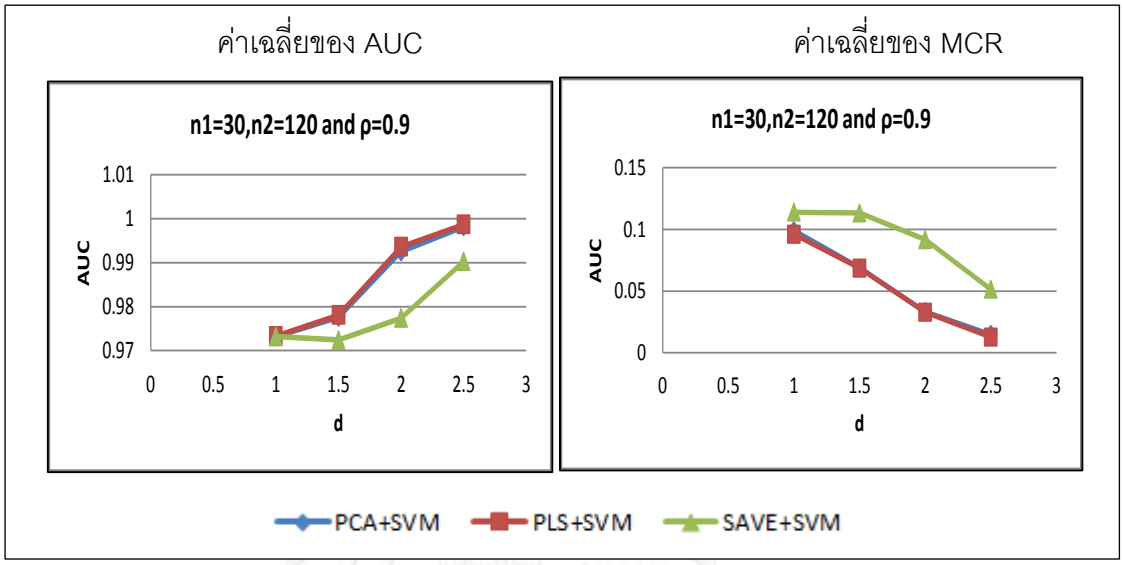
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.2.3 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 30, n_2 = 120)$

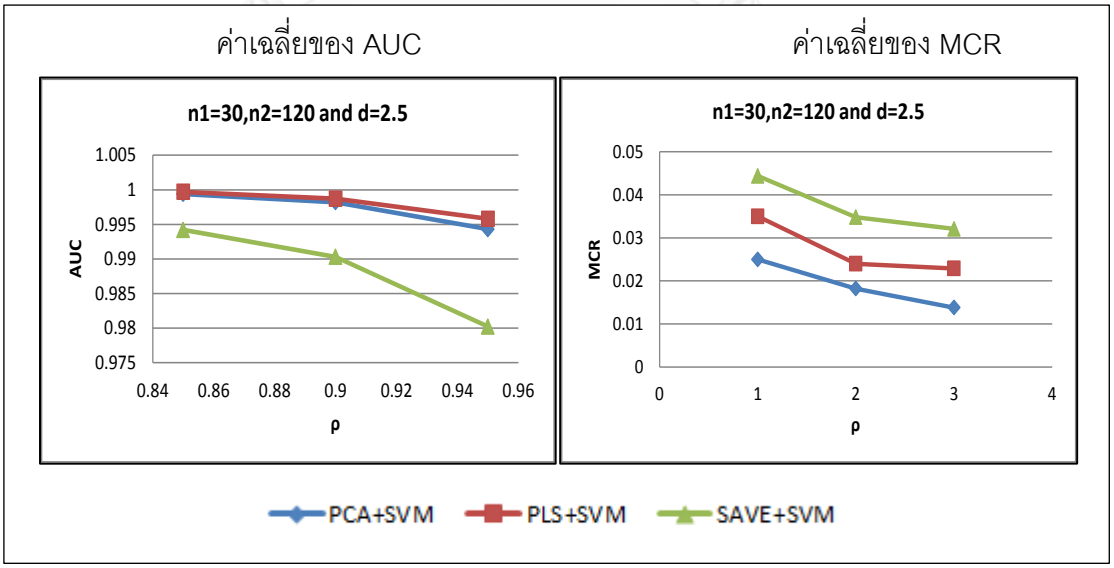
ตารางที่ 4.3.9 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9741	0.9742	0.9728	0.0906	0.0891	0.1134
	1.5	0.9847	0.9851	0.9732	0.0518	0.0538	0.1088
	2	0.9968	0.9968	0.9842	0.0195	0.0215	0.0720
	2.5	0.9994	0.9997	0.9942	0.0078	0.0062	0.0304
0.9	1	0.9732	0.9733	0.9732	0.0990	0.0961	0.1139
	1.5	0.9776	0.9781	0.9724	0.0689	0.0686	0.1133
	2	0.9925	0.9936	0.9774	0.0332	0.0328	0.0917
	2.5	0.9982	0.9987	0.9903	0.0148	0.0127	0.0515
0.95	1	0.9724	0.9731	0.9714	0.0956	0.0955	0.1154
	1.5	0.9752	0.9772	0.9715	0.0952	0.0889	0.1139
	2	0.9815	0.9854	0.9723	0.0659	0.0528	0.1094
	2.5	0.9943	0.9958	0.9802	0.0344	0.0259	0.0820

ภาพที่ 4.3.17 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.3.18 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.4 ตัวแปรอิสระ 40 ตัวแปร

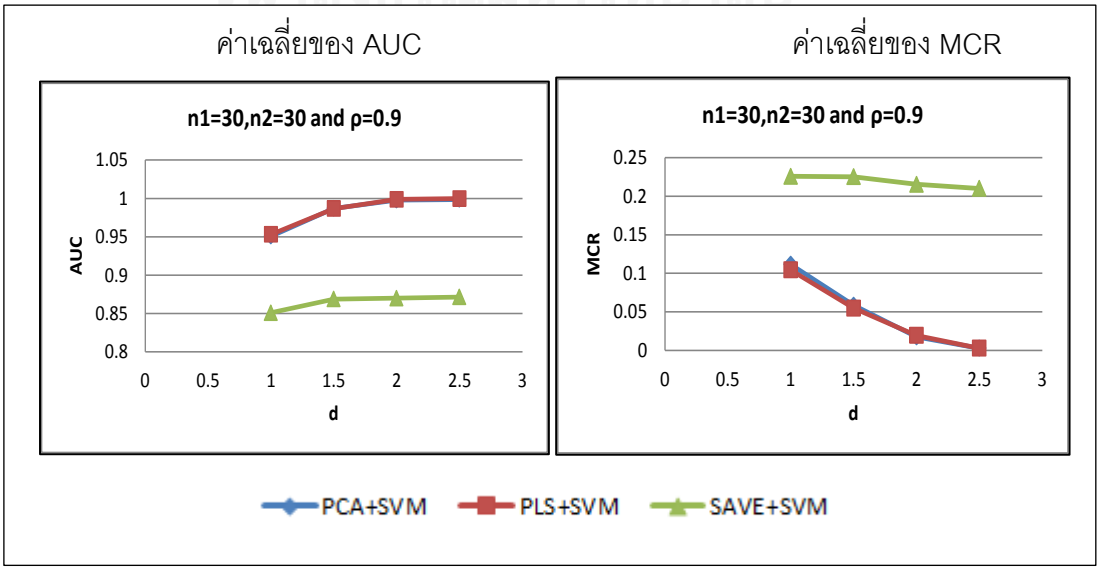
4.4.1 กรณีที่ขนาดตัวอย่างของแต่ละกลุ่มเท่ากัน

4.4.1.1 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $\left(n_1 = n_2 = 30\right)$

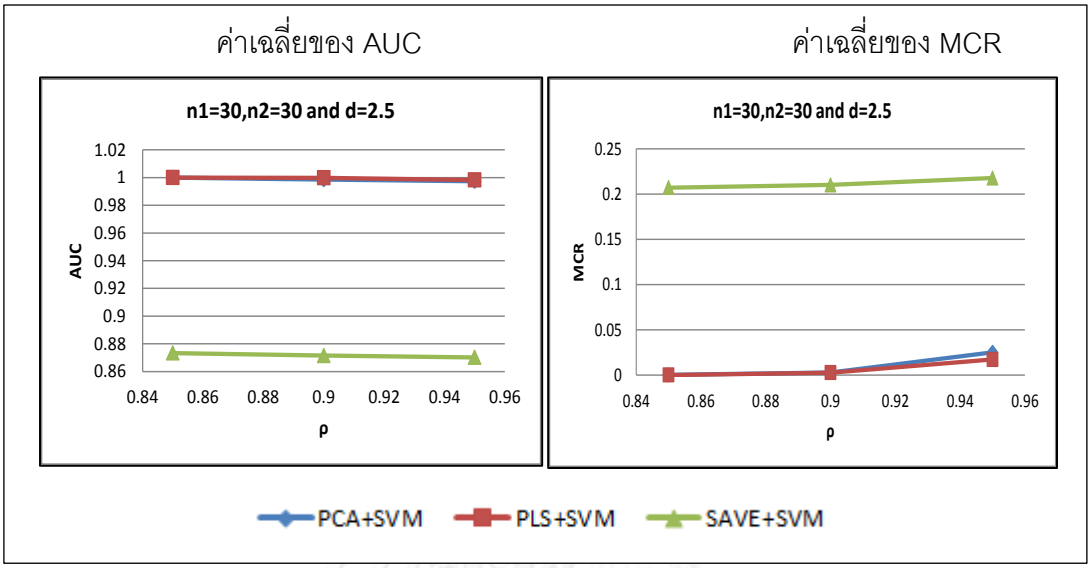
ตารางที่ 4.4.1 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9709	0.9730	0.8518	0.0927	0.0922	0.2253
	1.5	0.9948	0.9953	0.8589	0.0312	0.0310	0.2245
	2	0.9987	0.9997	0.8702	0.0055	0.0053	0.2140
	2.5	1.0000	1.0000	0.8733	0.0002	0.0001	0.2072
0.9	1	0.9506	0.9531	0.8509	0.1117	0.1043	0.2258
	1.5	0.9870	0.9867	0.8587	0.0585	0.0548	0.2253
	2	0.9978	0.9988	0.8700	0.0175	0.0193	0.2155
	2.5	0.9985	0.9998	0.8715	0.0028	0.0027	0.2102
0.95	1	0.9149	0.9154	0.8432	0.1508	0.1507	0.2268
	1.5	0.9674	0.9675	0.8548	0.0830	0.0783	0.2259
	2	0.9896	0.9927	0.8659	0.0463	0.0423	0.2180
	2.5	0.9973	0.9983	0.8702	0.0250	0.0172	0.2178

ภาพที่ 4.4.1 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.2 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

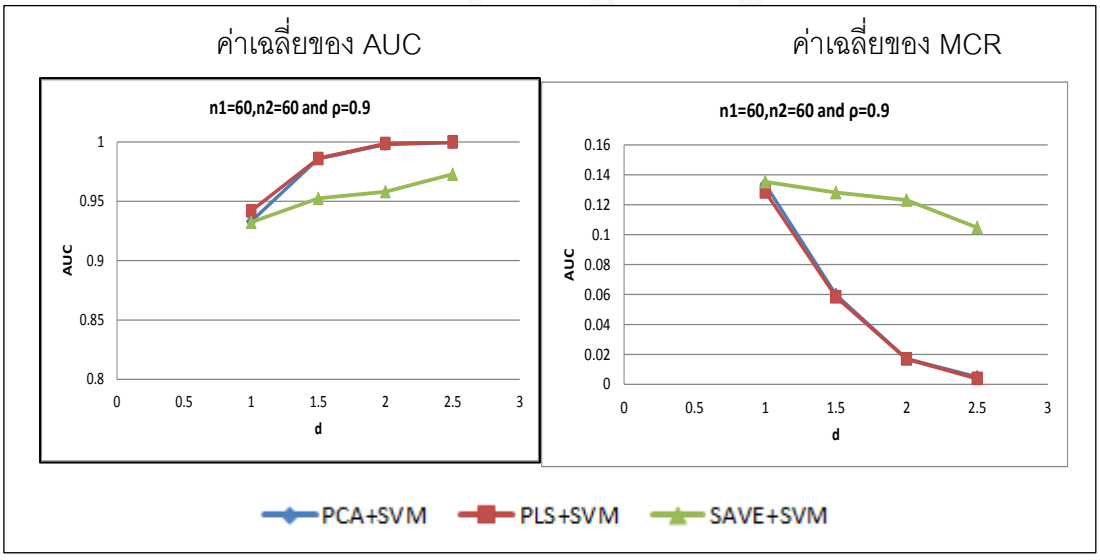
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.4.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $(n_1 = n_2 = 60)$

ตารางที่ 4.4.2 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9544	0.9565	0.9534	0.1128	0.1119	0.1269
	1.5	0.9941	0.9958	0.9620	0.0303	0.0302	0.1234
	2	0.9996	0.9997	0.9672	0.0053	0.0046	0.1224
	2.5	1.0000	1.0000	0.9861	0.0011	0.0010	0.0646
0.9	1	0.9331	0.9420	0.9324	0.1337	0.1286	0.1354
	1.5	0.9862	0.9863	0.9525	0.0598	0.0584	0.1282
	2	0.9982	0.9987	0.9582	0.0170	0.0169	0.1232
	2.5	0.9997	0.9999	0.9729	0.0048	0.0038	0.1048
0.95	1	0.9256	0.9273	0.9223	0.1310	0.1291	0.1358
	1.5	0.9592	0.9658	0.9522	0.1061	0.1029	0.1340
	2	0.9900	0.9910	0.9540	0.0504	0.0493	0.1413
	2.5	0.9974	0.9984	0.9610	0.0274	0.0198	0.1310

ภาพที่ 4.4.3 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.4 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5

ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

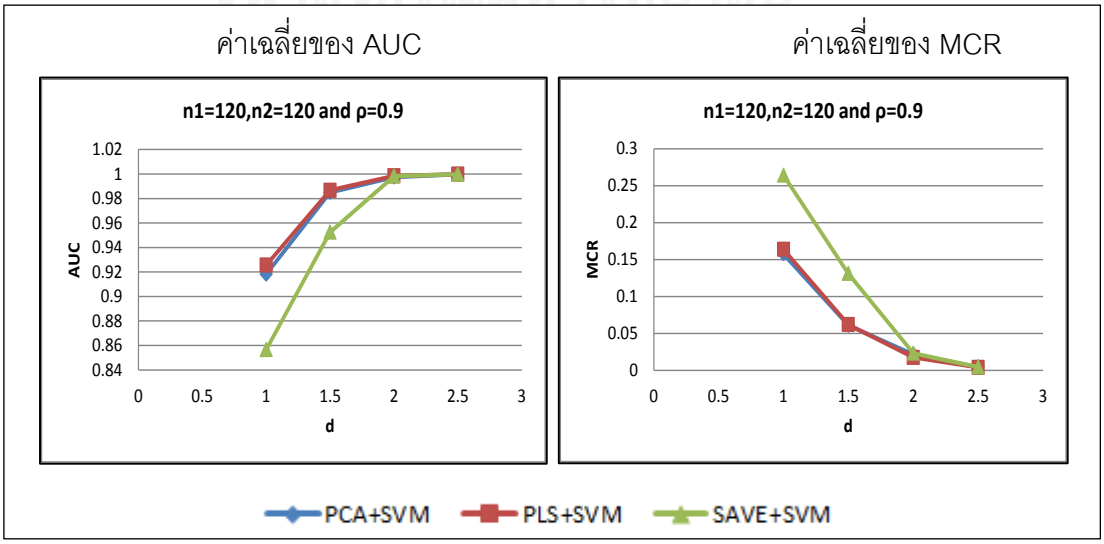
1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.4.1.2 ขนาดตัวอย่างของแต่ละกลุ่มเท่ากับ $(n_1 = n_2 = 120)$

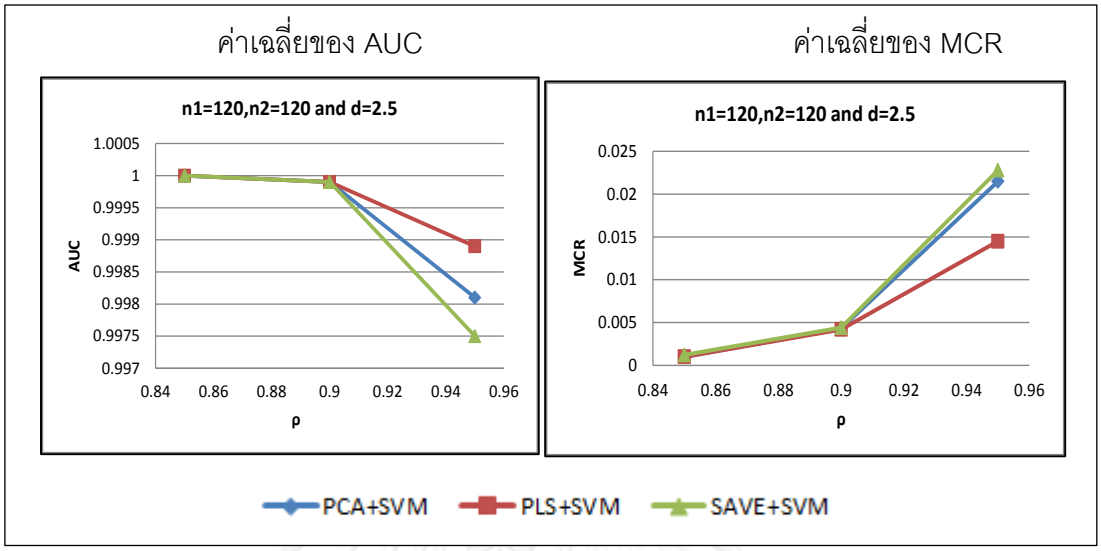
ตารางที่ 4.4.3 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9522	0.9528	0.8647	0.1153	0.1143	0.2572
	1.5	0.9941	0.9948	0.9917	0.0329	0.0381	0.0501
	2	0.9997	0.9997	0.9997	0.0063	0.0059	0.0080
	2.5	1.0000	1.0000	1.0000	0.0010	0.0010	0.0012
0.9	1	0.9182	0.9256	0.8566	0.1575	0.1639	0.2642
	1.5	0.9850	0.9867	0.9525	0.0609	0.0620	0.1309
	2	0.9975	0.9985	0.9983	0.0216	0.0176	0.0231
	2.5	0.9999	0.9999	0.9999	0.0043	0.0042	0.0044
0.95	1	0.8618	0.8757	0.8542	0.2244	0.2171	0.2688
	1.5	0.9554	0.9615	0.8803	0.1273	0.1066	0.2375
	2	0.9902	0.9924	0.9790	0.0546	0.0425	0.0848
	2.5	0.9981	0.9989	0.9975	0.0215	0.0145	0.0228

ภาพที่ 4.4.5 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



<u>ภาพที่ 4.4.6</u> แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2 กรณีที่ขนาดตัวอย่างของกลุ่มตัวอย่างที่สนใจมีความแตกต่างกับกลุ่มตัวอย่างที่ไม่สนใจ

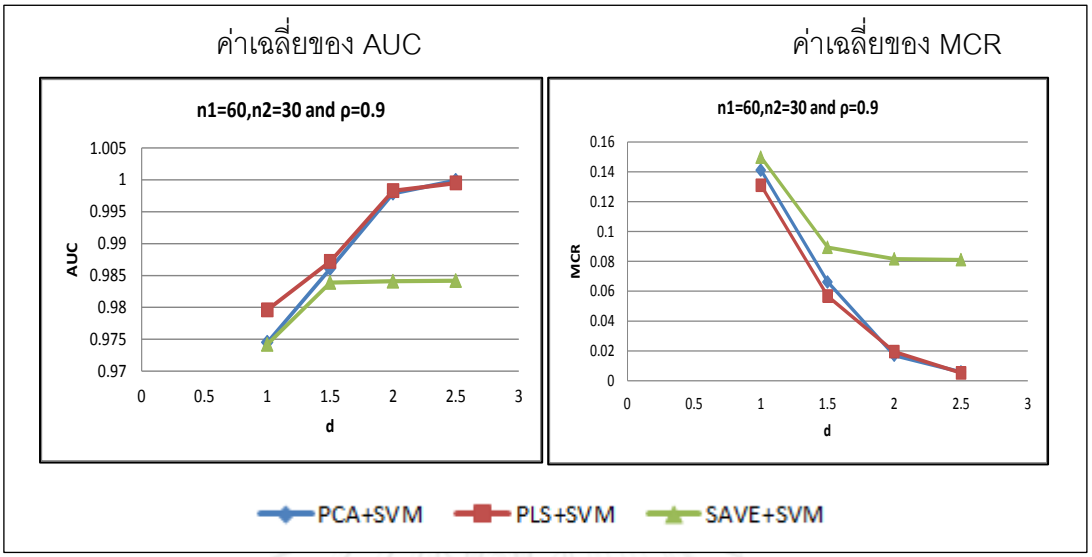
4.3.2.1 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างมากกว่ากลุ่มที่ไม่สนใจ

4.3.2.1.1 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $(n_1 = 60, n_2 = 30)$

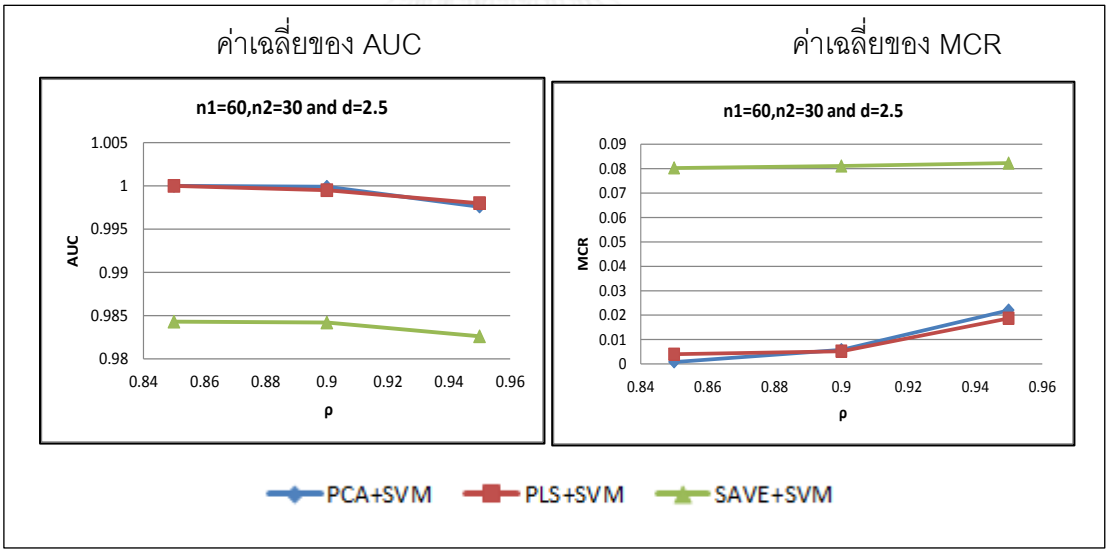
ตารางที่ 4.4.4 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9751	0.9798	0.9748	0.1683	0.1561	0.1693
	1.5	0.9946	0.9949	0.9868	0.0538	0.0344	0.0790
	2	0.9993	0.9997	0.9859	0.0119	0.0083	0.0800
	2.5	1.0000	1.0000	0.9843	0.0008	0.0040	0.0803
0.9	1	0.9745	0.9796	0.9742	0.1412	0.1310	0.1499
	1.5	0.9861	0.9872	0.9839	0.0662	0.0567	0.0894
	2	0.9979	0.9983	0.9841	0.0170	0.0194	0.0817
	2.5	0.9999	0.9995	0.9842	0.0028	0.0032	0.0811
0.95	1	0.9646	0.9686	0.9630	0.1832	0.1816	0.1849
	1.5	0.9660	0.9655	0.9817	0.0917	0.0909	0.0928
	2	0.9913	0.9907	0.9838	0.0523	0.0509	0.0836
	2.5	0.9976	0.9980	0.9826	0.0220	0.0187	0.0823

ภาพที่ 4.4.7 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.8 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

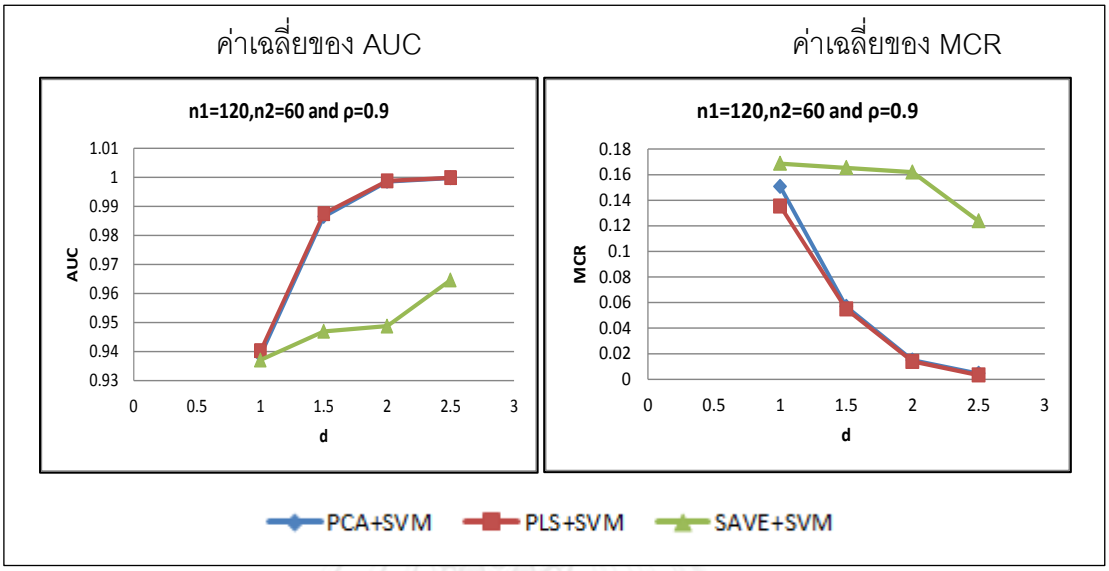
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.1.2 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 120, n_2 = 60)$

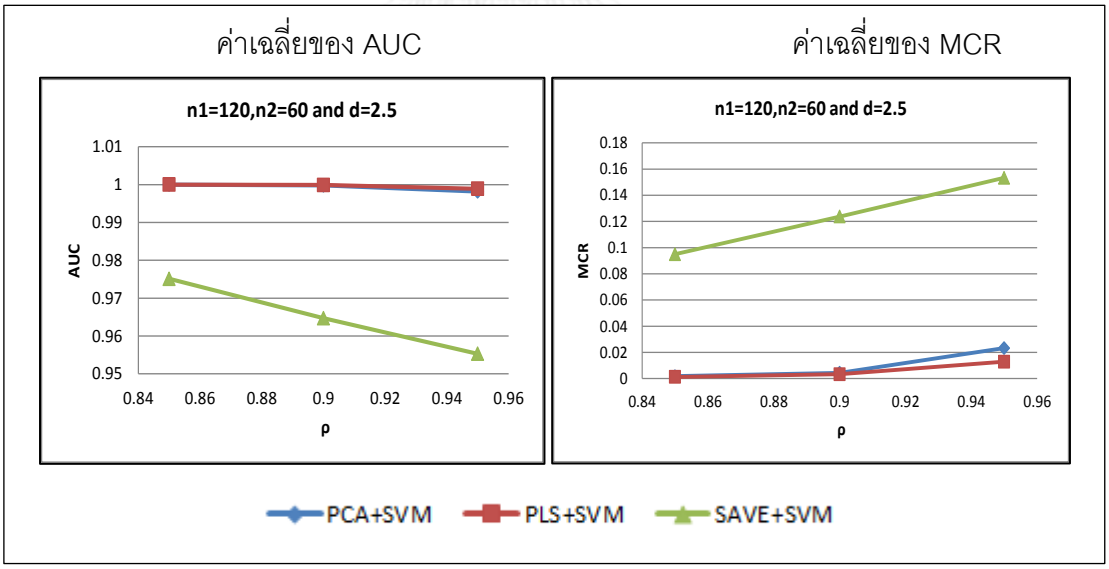
ตารางที่ 4.4.5 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9597	0.9599	0.9503	0.1353	0.1180	0.1607
	1.5	0.9949	0.9956	0.9572	0.0320	0.0316	0.1468
	2	0.9997	0.9998	0.9632	0.0048	0.0044	0.1302
	2.5	1.0000	1.0000	0.9751	0.0018	0.0013	0.0950
0.9	1	0.9386	0.9403	0.9371	0.1506	0.1355	0.1687
	1.5	0.9865	0.9875	0.9470	0.0568	0.0549	0.1653
	2	0.9986	0.9988	0.9488	0.0149	0.0139	0.1619
	2.5	0.9998	0.9999	0.9647	0.0044	0.0033	0.1237
0.95	1	0.9069	0.9050	0.9369	0.1825	0.1817	0.1851
	1.5	0.9618	0.9670	0.9446	0.1031	0.0988	0.1697
	2	0.9905	0.9930	0.9486	0.0482	0.0377	0.1606
	2.5	0.9982	0.9989	0.9553	0.0233	0.0129	0.1533

ภาพที่ 4.4.9 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.10 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

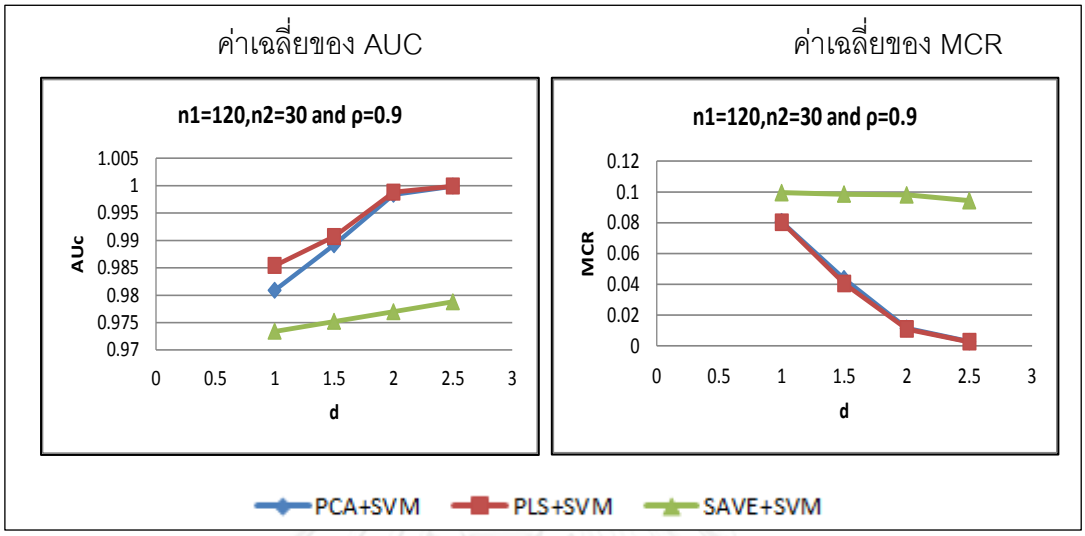
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.1.3 ขนาดของกลุ่มตัวอย่างที่สนใจมากกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 120, n_2 = 30)$

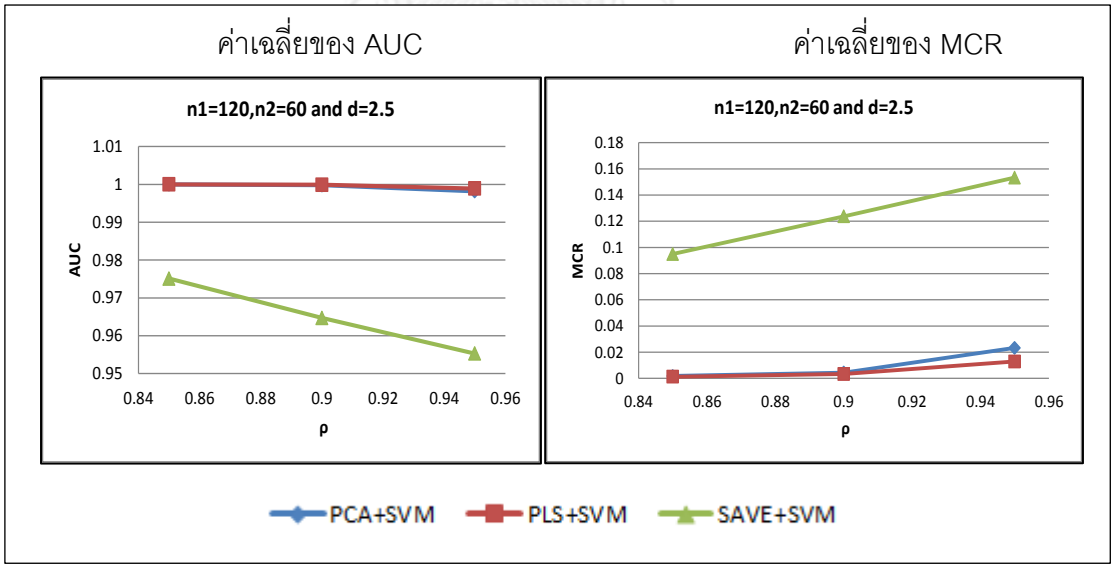
ตารางที่ 4.4.6 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9830	0.9862	0.9736	0.0919	0.0851	0.0990
	1.5	0.9943	0.9949	0.9759	0.0241	0.0239	0.0983
	2	0.9996	0.9997	0.9772	0.0047	0.0042	0.0979
	2.5	1.0000	1.0000	0.9793	0.0009	0.0007	0.0903
0.9	1	0.9809	0.9854	0.9734	0.0805	0.0803	0.0995
	1.5	0.9892	0.9907	0.9752	0.0435	0.0404	0.0985
	2	0.9984	0.9988	0.9770	0.0114	0.0109	0.0981
	2.5	0.9999	0.9999	0.9788	0.0027	0.0026	0.0943
0.95	1	0.9788	0.9789	0.9732	0.1038	0.1034	0.1056
	1.5	0.9780	0.9784	0.9748	0.0713	0.0709	0.0995
	2	0.9919	0.9925	0.9767	0.0356	0.0352	0.0990
	2.5	0.9985	0.9988	0.9773	0.0125	0.0122	0.0984

ภาพที่ 4.4.11 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.12 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 120, n_2 = 30$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้น ทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

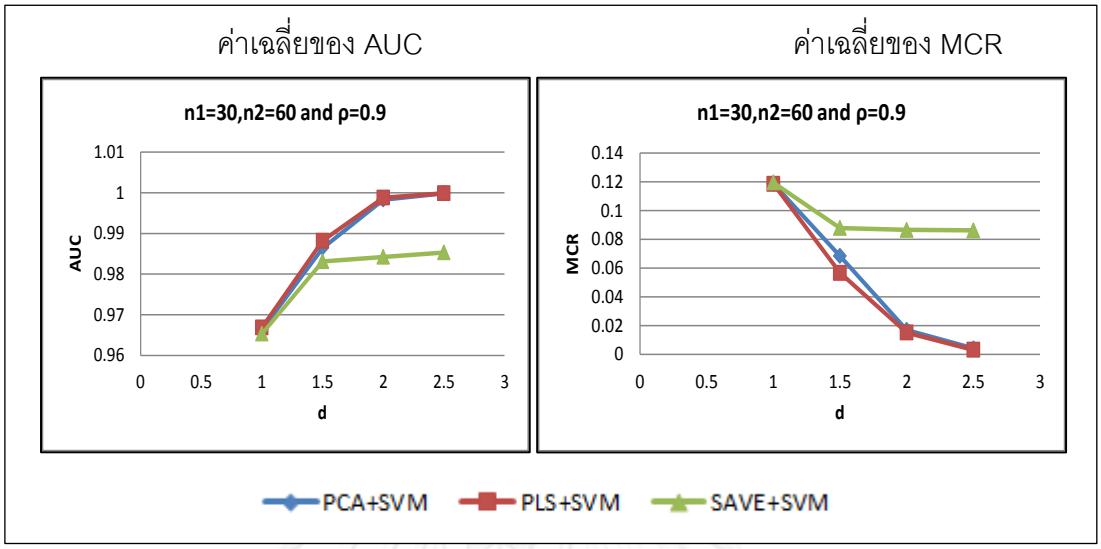
4.3.2.2 ขนาดตัวอย่างของกลุ่มที่สนใจมีขนาดตัวอย่างน้อยกว่ากลุ่มที่ไม่สนใจ

4.3.2.2.1 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 30 $\left(n_1 = 30, n_2 = 60\right)$

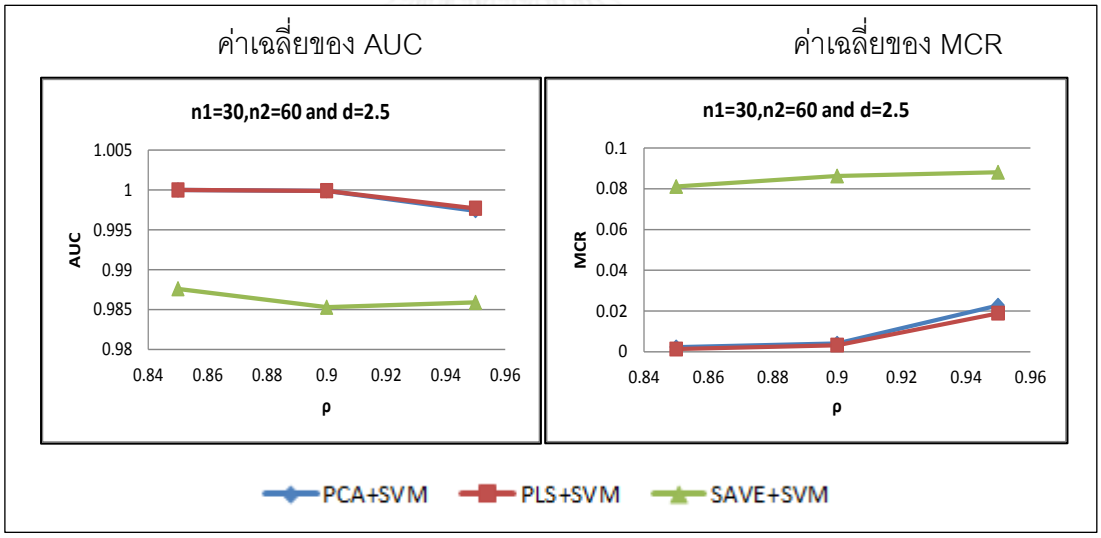
ตารางที่ 4.4.7 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9778	0.9779	0.9769	0.1049	0.1041	0.1052
	1.5	0.9952	0.9961	0.9862	0.0568	0.0306	0.0750
	2	0.9997	0.9998	0.9867	0.0103	0.0072	0.0746
	2.5	1.0000	1.0000	0.9876	0.0023	0.0013	0.0811
0.9	1	0.9663	0.9669	0.9653	0.1188	0.1186	0.1196
	1.5	0.9864	0.9882	0.9831	0.0684	0.0567	0.0879
	2	0.9983	0.9988	0.9842	0.0169	0.0152	0.0866
	2.5	0.9999	0.9999	0.9853	0.0041	0.0032	0.0863
0.95	1	0.9528	0.9539	0.9508	0.1261	0.1257	0.1282
	1.5	0.9725	0.9733	0.9703	0.0956	0.0898	0.0979
	2	0.9880	0.9908	0.9861	0.0501	0.0479	0.0897
	2.5	0.9974	0.9977	0.9859	0.0228	0.0189	0.0881

ภาพที่ 4.4.13 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.14 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 60$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

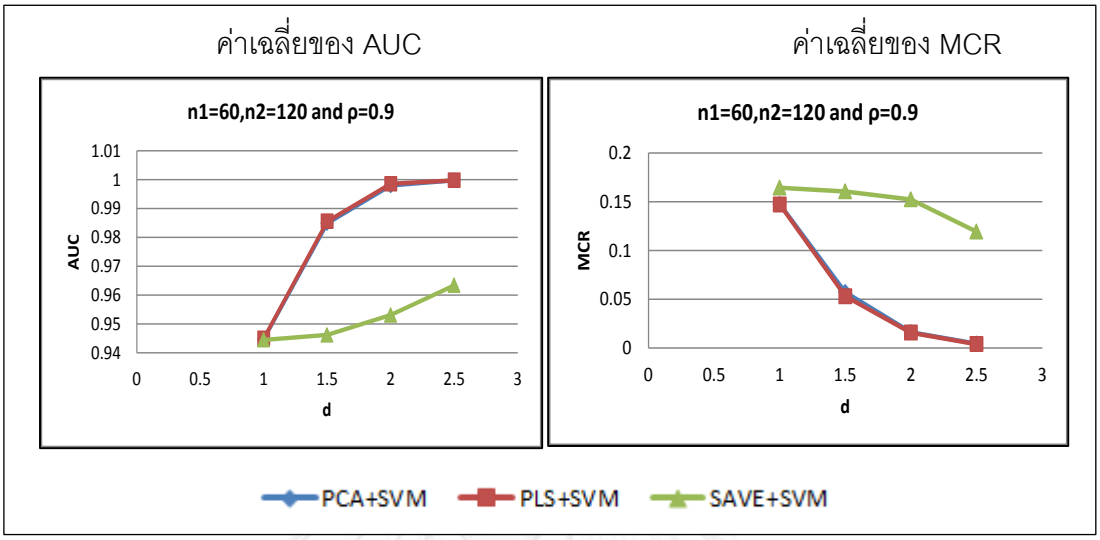
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.2.2 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 60 $(n_1 = 60, n_2 = 120)$

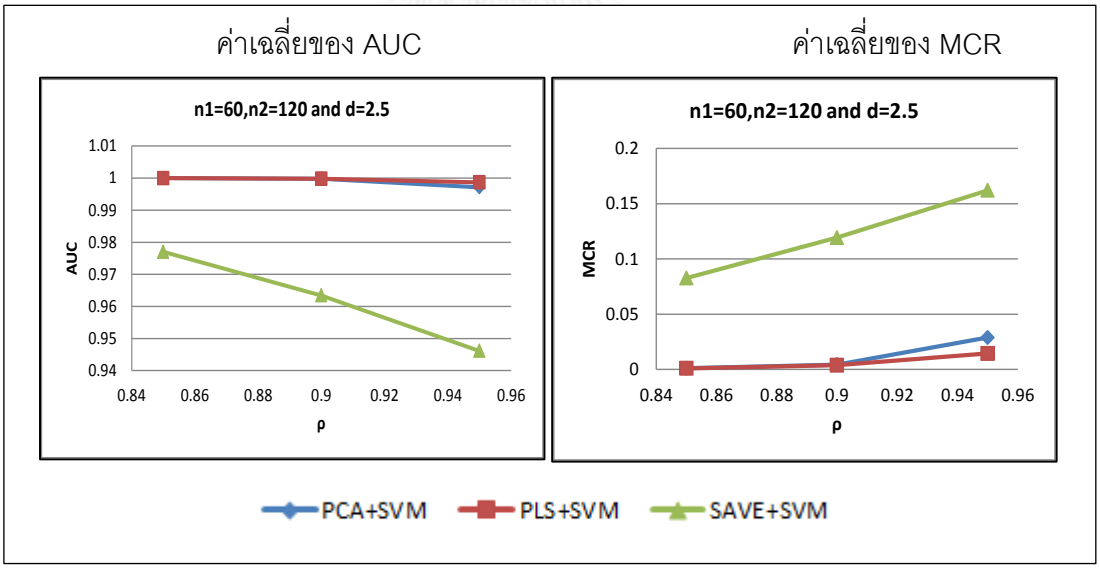
ตารางที่ 4.4.8 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9594	0.9596	0.9407	0.1312	0.1203	0.1643
	1.5	0.9942	0.9947	0.9437	0.0354	0.0352	0.1542
	2	0.9998	0.9998	0.9636	0.0049	0.0042	0.1251
	2.5	1.0000	1.0000	0.9770	0.0011	0.0010	0.0827
0.9	1	0.9447	0.9450	0.9445	0.1477	0.1472	0.1645
	1.5	0.9849	0.9856	0.9462	0.0572	0.0530	0.1606
	2	0.9980	0.9986	0.9531	0.0162	0.0157	0.1524
	2.5	0.9998	0.9998	0.9634	0.0044	0.0039	0.1194
0.95	1	0.9375	0.9389	0.9345	0.1708	0.1701	0.1738
	1.5	0.9585	0.9639	0.9401	0.1105	0.1032	0.1666
	2	0.9902	0.9921	0.9418	0.0482	0.0413	0.1661
	2.5	0.9971	0.9987	0.9461	0.0289	0.0145	0.1621

ภาพที่ 4.4.15 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.16 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 60, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

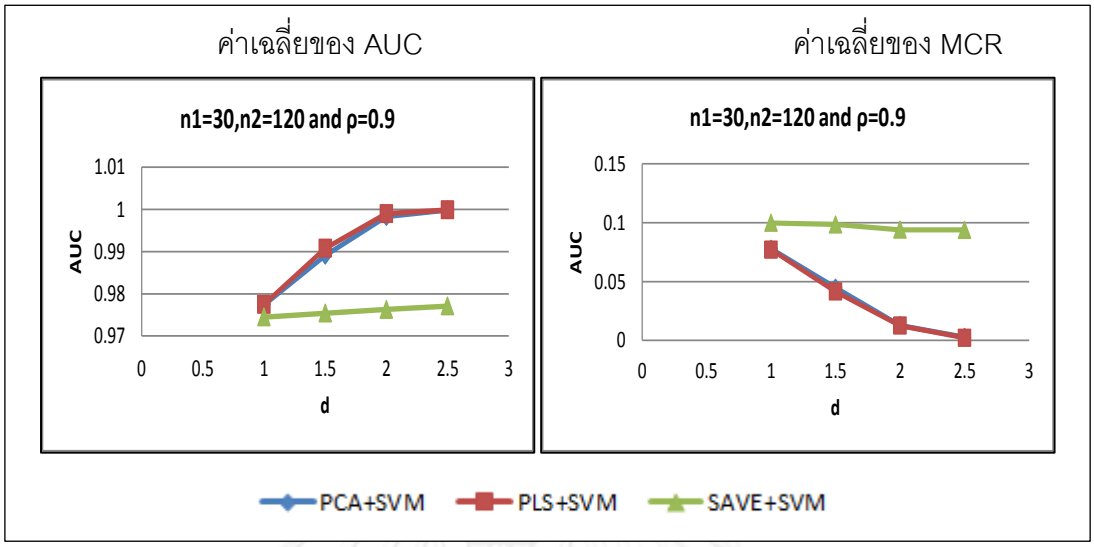
2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.3.2.2.3 ขนาดของกลุ่มตัวอย่างที่สนใจน้อยกว่ากลุ่มตัวอย่างที่ไม่สนใจอยู่ 90 $(n_1 = 30, n_2 = 120)$

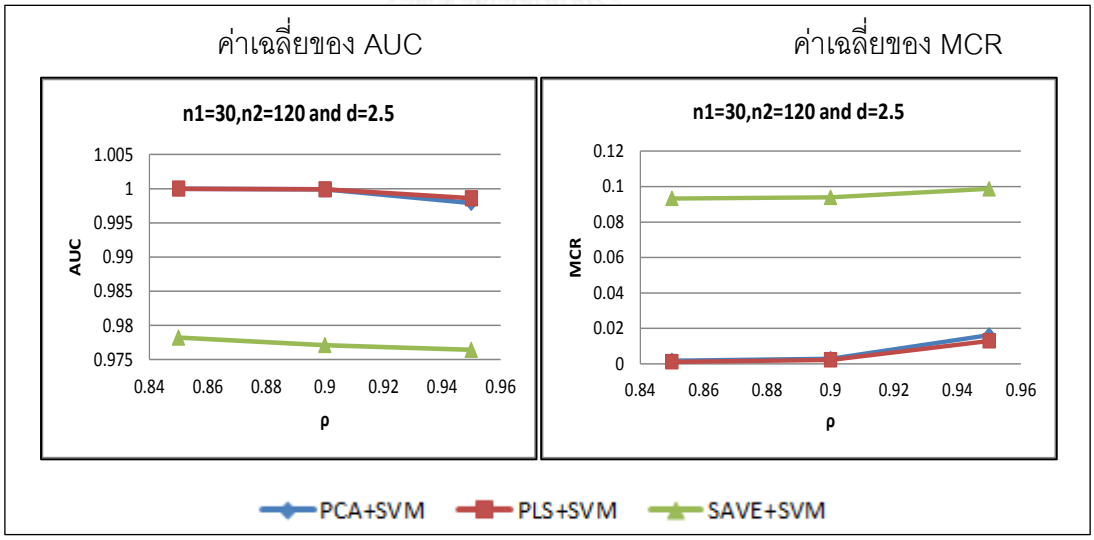
ตารางที่ 4.4.9 แสดงค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC(AUC) และค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภท(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$

ρ	d	AUC			MCR		
		PCA+SVM	PLS+SVM	SAVE+SVM	PCA+SVM	PLS+SVM	SAVE+SVM
0.85	1	0.9735	0.9764	0.9720	0.0731	0.0728	0.0986
	1.5	0.9962	0.9970	0.9752	0.0296	0.0219	0.0984
	2	0.9996	0.9996	0.9772	0.0059	0.0053	0.0932
	2.5	1.0000	1.0000	0.9782	0.0018	0.0011	0.0933
0.9	1	0.9774	0.9775	0.9745	0.0778	0.0771	0.0998
	1.5	0.9891	0.9908	0.9754	0.0448	0.0415	0.0985
	2	0.9983	0.9990	0.9763	0.0128	0.0126	0.0940
	2.5	0.9999	0.9999	0.9771	0.0028	0.0022	0.0939
0.95	1	0.9764	0.9770	0.9734	0.0839	0.0806	0.0978
	1.5	0.9840	0.9844	0.9756	0.0758	0.0756	0.0989
	2	0.9906	0.9926	0.9757	0.0397	0.0362	0.0941
	2.5	0.9979	0.9986	0.9764	0.0161	0.0129	0.0987

ภาพที่ 4.4.17 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.9 และ d =1,1.5,2,2.5



ภาพที่ 4.4.18 แสดงการเปรียบเทียบค่าเฉลี่ยของ AUC และค่าเฉลี่ยของ MCR กรณีที่มีจำนวนข้อมูล $n_1 = 30, n_2 = 120$ และ ρ = 0.85,0.9,0.95 และ d =2.5



ผลจากตารางและกราฟด้านบนพบประเด็นที่น่าสนใจ ดังนี้

1. ค่าเฉลี่ยของ AUC ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการ

พยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ

2. ค่าเฉลี่ยของ MCR ในทุกการจำลองของข้อมูลพบว่าวิธีการลดข้อมูลด้วยวิธี PLS สำหรับการพยากรณ์ด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด และรองลงมาเป็นวิธี PCA และวิธี SAVE ตามลำดับ
3. เมื่อระดับความสัมพันธ์ของตัวแปรอิสระเพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC ลดลงเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าเพิ่มขึ้น ทุกกรณีของ d ที่เปลี่ยนแปลงไป
4. เมื่อค่าของพารามิเตอร์ของการแจกแจงของข้อมูลในตัวอย่างกลุ่มที่ไม่สนใจ(d)เพิ่มขึ้นทุกวิธีการที่ทำการศึกษาจะมีค่าเฉลี่ยของ AUC เพิ่มขึ้นเล็กน้อย ในทางตรงกันข้ามค่าเฉลี่ยของ MCR จะมีค่าลดลง ทุกกรณีของ ρ ที่เปลี่ยนแปลงไป

4.5 กรณีศึกษาข้อมูล crab data

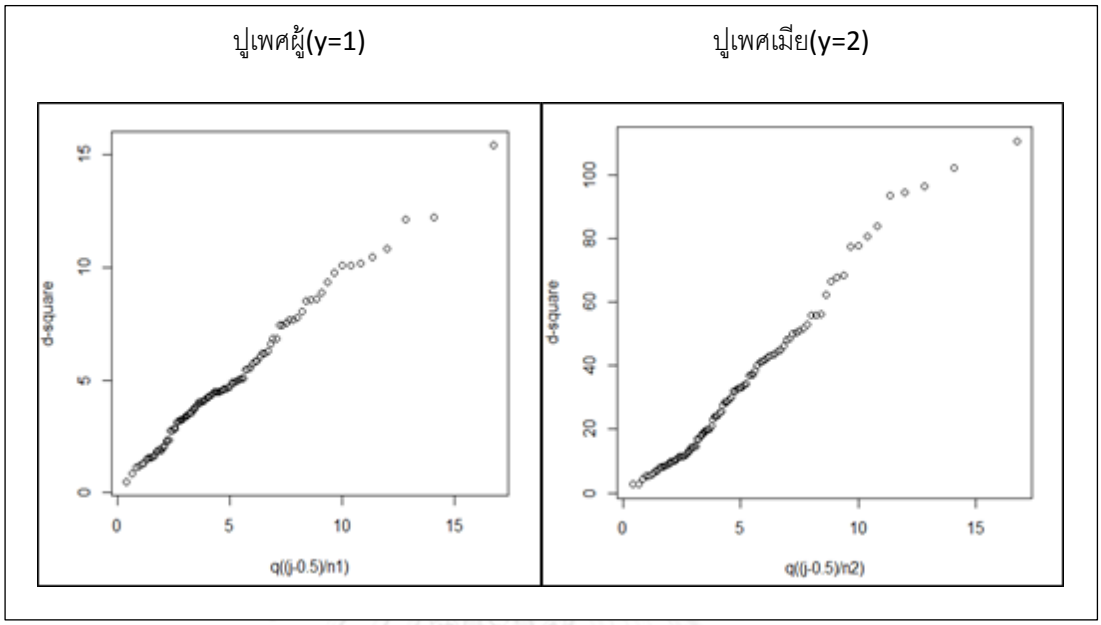
จากฐานข้อมูล crab data ได้ทำการศึกษาปูจำนวน 200 ตัว ซึ่งในงานวิจัยครั้งนี้ได้คัดเลือกข้อมูลมาบางส่วนซึ่งประกอบด้วย ตัวแปรเชิงกลุ่ม 1 ตัวแปรคือ ปูเพศผู้จำนวนข้อมูล 100 ตัว และปูเพศเมียจำนวนข้อมูล 100 ตัว และตัวแปรเชิงปริมาณ 5 ตัวแปรซึ่งวัดค่าจากรูปพรรณสัณฐานของปู ดังนี้

(1) y แทน เพศซึ่งประกอบด้วย 1=เพศผู้ และ 2= เพศเมีย
(2) X_1 แทน ขนาดซีกด้านหน้า(เซนติเมตร)
(3) X_2 แทน ความกว้างส่วนหลัง(เซนติเมตร)
(4) X_3 แทน ความยาวของcarapace(เซนติเมตร)
(5) X_4 แทน ความกว้างของcarapace(เซนติเมตร)
(6) X_5 แทน ความลึกของร่างกาย(เซนติเมตร)

วัตถุประสงค์ : เพื่อแยกกลุ่มเพศของปู โดยการสร้างแบบจำลองการพยากรณ์จำแนกประเภทด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน เริ่มจากขั้นตอนการเตรียมข้อมูล โดยใช้เทคนิคการลดมิติข้อมูลด้วยเทคนิคการวิเคราะห์องค์ประกอบหลัก(PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (PLS)และ Sliced Average Variance Estimator(SAVE) สามารถสรุปผลได้ดังนี้

จากการวัดค่าการซ้อนทับ(overlap) ของข้อมูล 2 กลุ่ม มีค่าเท่ากับ 3.24%

ภาพที่ 4.5.1 แสดงกราฟไคสแควร์ สำหรับข้อมูล crab data กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 100$



สรุปผลจากภาพที่ 4.5.1 ข้อมูลชุดนี้มีตัวแปรเชิงกลุ่มซึ่งทั้ง 2 กลุ่มมีการแจกแจงแบบปกติหลายตัวแปร

ตารางที่ 4.5.1 แสดงเมตริกซ์สหสัมพันธ์(correlation matrix) สำหรับข้อมูล crab data กรณีที่มีจำนวนข้อมูล $n_1 = n_2 = 100$

	x1	x2	x3	x4	x5
x1	1.0000	0.9070	0.9788	0.9650	0.9876
x2	0.9070	1.0000	0.8927	0.9004	0.8892
x3	0.9788	0.8927	1.0000	0.9950	0.9832
x4	0.9650	0.9004	0.9950	1.0000	0.9678
x5	0.9876	0.8892	0.9832	0.9678	1.0000

สรุปผลจากตารางที่ 4.5.1 แสดงความสัมพันธ์สูงระหว่างตัวแปรอิสระตัวที่ i และ j เมื่อ i, j = 1, 2, ..., 5

ตารางที่ 4.5.2 แสดงค่าเฉลี่ยพื้นที่ใต้โค้ง ROC(AUC)และค่าเฉลี่ยอัตราความผิดพลาด(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีข้อมูล crab data

	AUC	MCR
PCA+SVM	0.8093	0.2646
PLS+SVM	0.9984	0.0141
SAVE+SVM	0.9995	0.0128

สรุปผลจากตารางที่ 4.5.2 ปรากฏว่าวิธีการลดข้อมูลด้วยวิธี Sliced Average Variance(SAVE) มีความเหมาะสมเมื่อนำมาเป็นข้อมูลนำเข้าสำหรับวิธีซัพพอร์ตเวกเตอร์แมชชีน เนื่องจากมีความแม่นยำ 99.95% ซึ่งใกล้เคียงกับวิธี PLS ที่มีค่าความแม่นยำ 99.84% และต่างจากวิธี PCA ที่มีค่าความแม่นยำ 80.93% ในทำนองเดียวกันเมื่อพิจารณาค่าเฉลี่ยอัตราความผิดพลาดวิธี Sliced Average Variance (SAVE) มีค่าเฉลี่ยอัตราความผิดพลาดเท่ากับ 1.28% ซึ่งน้อยกว่าเทคนิค PLS และ PCA ที่มีค่าเฉลี่ยอัตราความผิดพลาดเท่ากับ 1.41% และ 26.46% ตามลำดับ ซึ่งผลลัพธ์ที่ได้สอดคล้องกับการทดลองกรณีที่ตัวแปรอิสระ 5 ตัวแปร

4.6 กรณีศึกษาข้อมูล wine data

จากฐานข้อมูล wine data ซึ่งศึกษาผลการวิเคราะห์ทางเคมีของการเจริญเติบโตขององุ่นที่ปลูกในพื้นที่เดียวกันในอิตาลีแต่ใช้วัฒนธรรมการปลูกที่แตกต่างกันซึ่งในงานวิจัยนี้จะคัดเลือกมา 2 วัฒนธรรม คือ ไวน์ประเภทที่ 1 มีจำนวนข้อมูล 51 ข้อมูล และไวน์ประเภทที่ 2 มีจำนวนข้อมูล 48 ข้อมูล เป็นตัวแปรเชิงกลุ่ม และตัวแปรเชิงคุณภาพ 13 ตัวแปรซึ่งมาจากการวัดผลการวิเคราะห์ทางเคมีของไวน์ ดังนี้

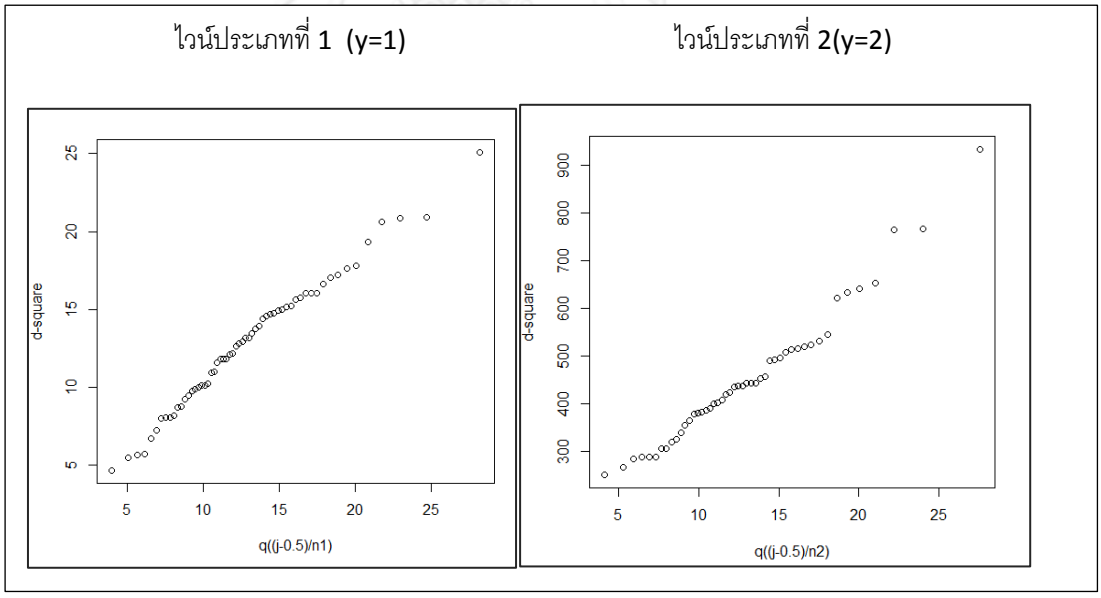
(1) y แทน ประเภทของไวน์ซึ่งประกอบด้วย 1=ไวน์ประเภทที่ 1 และ 2= ไวน์ประเภทที่ 2
(2) X_1 แทน แอลกอฮอล์ (Alcohol)
(3) X_2 แทน กรดมาลิก (Malic acid)
(4) X_3 แทน เถ้าถ่าน (Ash)
(5) X_4 แทน สภาพด่างของเถ้าถ่าน (Alcalinity of ash)
(6) X_5 แทน แมกนีเซียม (Magnesium)
(7) X_6 แทน ปริมาณฟีนอลโดยรวม (Total phenols)
(8) X_7 แทน ฟลาโวนอยด์ (Flavanoids)
(9) X_8 แทน นอนฟลาโวนอยด์ ฟีนอล (Nonflavanoid phenols)

(10) X_9 แทน โปรแอนโทไซยานิน (Proanthocyanins)
(11) X_{10} แทน ความเข้มของสี (Color intensity)
(12) X_{11} แทน สี (Hue)
(13) X_{12} แทน อัตราส่วน OD280 ต่อ OD315 ของไวน์ที่เจือจาง (OD280/OD315 of diluted wines)
(14) X_{13} แทน โพรลีน(Proline)

วัตถุประสงค์ : เพื่อแยกประเภทของไวน์ โดยการสร้างแบบจำลองการพยากรณ์จำแนกประเภทด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน เริ่มจากขั้นตอนการเตรียมข้อมูล โดยใช้เทคนิคการลดมิติข้อมูลด้วยเทคนิคการวิเคราะห์องค์ประกอบหลัก(PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (PLS)และ Sliced Average Variance Estimator(SAVE) สามารถสรุปผลได้ดังนี้

จากการวัดค่าการซ้อนทับ(overlap) ของข้อมูล 2 กลุ่ม มีค่าเท่ากับ 0%

<u>ภาพที่ 4.6.1</u> แสดงกราฟไคสแควร์ สำหรับข้อมูล wine data กรณีที่มีจำนวนข้อมูล $n_1 = 59, n_2 = 48$



สรุปผลจากภาพที่ 4.6.1 ข้อมูลชุดนี้มีตัวแปรเชิงกลุ่มซึ่งทั้ง 2 กลุ่มมีการแจกแจงแบบปกติหลายตัวแปร

ตารางที่ 4.6.1 แสดงเมตริกซ์สหสัมพันธ์(correlation matrix) สำหรับข้อมูล wine data กรณีที่มีจำนวนข้อมูล $n_1 = 59, n_2 = 48$

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
x1	1.0000	-0.2747	0.0405	-0.3963	0.1940	0.5813	0.5613	-0.3040	0.5622	0.0368	0.4516	0.5044	0.5155
x2	-0.2747	1.0000	-0.0098	0.4439	-0.2367	-0.5654	-0.6233	0.4203	-0.4957	0.1416	-0.5687	-0.5213	-0.5722
x3	0.0405	-0.0098	1.0000	0.4331	0.3087	0.1373	0.0555	0.1195	0.0216	-0.0097	0.1495	0.0491	0.0007
x4	-0.3963	0.4439	0.4331	1.0000	-0.0675	-0.5663	-0.6611	0.4892	-0.4484	0.3118	-0.5494	-0.6345	-0.5889
x5	0.1940	-0.2367	0.3087	-0.0675	1.0000	0.3397	0.3829	-0.3450	0.2378	-0.0354	0.2392	0.2827	0.2272
x6	0.5813	-0.5654	0.1373	-0.5663	0.3397	1.0000	0.9083	-0.4664	0.7628	-0.1976	0.6991	0.8104	0.7521
x7	0.5613	-0.6233	0.0555	-0.6611	0.3829	0.9083	1.0000	-0.6780	0.7525	-0.3071	0.7977	0.8494	0.8053
x8	-0.3040	0.4203	0.1195	0.4892	-0.3450	-0.4664	-0.6780	1.0000	-0.4002	0.2748	-0.4359	-0.5742	-0.4731
x9	0.5622	-0.4957	0.0216	-0.4484	0.2378	0.7628	0.7525	-0.4002	1.0000	0.0517	0.5269	0.6042	0.6070
x10	0.0368	0.1416	-0.0097	0.3118	-0.0354	-0.1976	-0.3071	0.2748	0.0517	1.0000	-0.5391	-0.4694	-0.2028
x11	0.4516	-0.5687	0.1495	-0.5494	0.2392	0.6991	0.7977	-0.4359	0.5269	-0.5391	1.0000	0.7725	0.7600
x12	0.5044	-0.5213	0.0491	-0.6345	0.2827	0.8104	0.8494	-0.5742	0.6042	-0.4694	0.7725	1.0000	0.6846
x13	0.5155	-0.5722	0.0007	-0.5889	0.2272	0.7521	0.8053	-0.4731	0.6070	-0.2028	0.7600	0.6846	1.0000

สรุปผลจากภาพที่ 4.6.1 แสดงความสัมพันธ์สูงระหว่างตัวแปร (X6,X7), (X6,X12), (X7,X12), และ (X7,X13) และแสดงความสัมพันธ์ค่อนข้างสูงระหว่างตัวแปร (X6,X9), (X6,X13), (X7,X9), (X7,X11), (X11,X12) และ (X11,X13)

ตารางที่ 4.6.2 แสดงค่าเฉลี่ยพื้นที่ใต้โค้ง ROC(AUC)และค่าเฉลี่ยอัตราความผิดพลาด(MCR) เมื่อใช้เทคนิคการลดมิติข้อมูลสำหรับเป็นข้อมูลนำเข้าวิธีซัพพอร์ตเวกเตอร์แมชชีน กรณีข้อมูล wine data

	AUC	MCR
PCA+SVM	0.9990	0.0069
PLS+SVM	1.0000	0.0000
SAVE+SVM	0.9985	0.0087

สรุปผลจากตารางที่ 4.6.2 ปรากฏว่าวิธีการลดข้อมูลด้วยวิธีกำลังสองน้อยสุดเชิงส่วน (PLS) มีความเหมาะสมเมื่อนำมาเป็นข้อมูลนำเข้าสำหรับวิธีซัพพอร์ตเวกเตอร์แมชชีน เนื่องจากมีความแม่นยำ 100% ซึ่งใกล้เคียงกับเทคนิค PCA และ SAVE ที่มีค่าความแม่นยำ 99.90% และ 99.85% ตามลำดับ เมื่อพิจารณาค่าเฉลี่ยอัตราความผิดพลาดเทคนิค PLS มีค่าเฉลี่ยอัตราความผิดพลาด 0% ซึ่งน้อยกว่าเทคนิค PCA และ SAVE ซึ่งผลลัพธ์ที่ได้สอดคล้องกับการทดลองกรณีที่ตัวแปรอิสระ 10 ตัวแปร

บทที่ 5

สรุปผลการวิจัย และข้อเสนอแนะ

การศึกษางานวิจัยในครั้งนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติข้อมูลเข้า(input data) ระหว่างเทคนิคการวิเคราะห์องค์ประกอบหลัก(Principle Component Analysis, PCA) วิธีกำลังสองน้อยสุดเชิงส่วน (Partial Least Squares, PLS) และ Sliced Average Variance Estimator(SAVE) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ด้วย Laplacian Kernel โดยทำการจำลองข้อมูลเพื่อศึกษาผลกระทบจากระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล (d), ค่าระดับความสัมพันธ์ของตัวแปรอิสระ (ρ) และขนาดของกลุ่มตัวอย่าง (n_1, n_2) ทำการพิจารณาผลการศึกษาด้วย Receiver Operating Characteristic (ROC) ใช้เป็นเครื่องมือวัดประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูล โดยใช้พื้นที่ใต้โค้ง ROC และใช้อัตราความผิดพลาดในการจำแนก ประเภทข้อมูล (Misclassification Rate : MCR) เพื่อศึกษาว่าวิธีการใดมีความผิดพลาดในการ จำแนกประเภท ซึ่งในงานวิจัยนี้ทำการศึกษาผลของ เหตุการณ์เกิดขึ้นสองเหตุการณ์ (dichotomous) สามารถสรุปผลการศึกษาในกรณีต่าง ๆ ได้ดังนี้

5.1 สรุปผลการศึกษา

5.1.1 ผลกระทบจากขนาดตัวอย่าง เมื่อกำหนดจำนวนตัวแปรอิสระ ดังนี้

ผลการศึกษาจากการจำลองข้อมูลตามขอบเขตที่กำหนด สามารถสรุปวิธีที่มีประสิทธิภาพการพยากรณ์จำแนกประเภทที่ดีที่สุดจากผลการศึกษาข้อมูลได้ดังนี้

5.1.1.1 กรณีที่มีตัวแปรอิสระ 5 ตัวแปร

ตารางที่ 5.1.1 แสดงเทคนิคการลดมิติข้อมูลนำเข้า(input data) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ที่ให้ค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC มากที่สุดและค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภทข้อมูล(MCR)ที่มีค่าน้อยที่สุด โดยจำแนกตามขนาดของกลุ่มตัวอย่างที่สนใจและไม่สนใจ เมื่อ $\rho = 0.85, 0.9, 0.95$ และ $d = 1, 1.5, 2, 2.5$ กรณีจำนวนตัวแปรอิสระเท่ากับ 5 ตัว

n	AUC	MCR
$n_1 = n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 60$	SAVE+SVM	SAVE+SVM
$n_1 = n_2 = 120$	SAVE+SVM	SAVE+SVM
$n_1 = 60, n_2 = 30$	SAVE+SVM	SAVE+SVM
$n_1 = 120, n_2 = 60$	SAVE+SVM	SAVE+SVM
$n_1 = 120, n_2 = 30$	SAVE+SVM	SAVE+SVM
$n_1 = 30, n_2 = 60$	SAVE+SVM	SAVE+SVM
$n_1 = 60, n_2 = 120$	SAVE+SVM	SAVE+SVM
$n_1 = 30, n_2 = 120$	SAVE+SVM	SAVE+SVM

ผลจากตารางจะได้ว่า ในกรณีที่ขนาดของกลุ่มตัวอย่างแต่ละกลุ่มเป็น 30 วิธีการลดมิติข้อมูลด้วยเทคนิควิธีกำลังสองน้อยสุดเชิงส่วน (PLS) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด แต่ในกรณีอื่นๆที่ทำการศึกษาพบว่าวิธีการลดมิติข้อมูลด้วยเทคนิค Sliced Average Variance Estimator (SAVE) ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และเมื่อพิจารณาค่าเฉลี่ยของ MCR จะได้ว่าวิธีที่ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุดจะให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุดเช่นเดียวกัน

5.1.1.2 กรณีที่มีตัวแปรอิสระ 10 ตัวแปร

ตารางที่ 5.1.2 แสดงเทคนิคการลดมิติข้อมูลนำเข้า(input data) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีนที่ให้ค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC มากที่สุดและค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภทข้อมูลที่มีค่าน้อยที่สุด โดยจำแนกตามขนาดของกลุ่มตัวอย่างที่สนใจและไม่สนใจ เมื่อ $\rho = 0.85, 0.9, 0.95$ และ $d = 1, 1.5, 2, 2.5$ กรณีจำนวนตัวแปรอิสระเท่ากับ 10 ตัว

n	AUC	MCR
$n_1 = n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 120$	SAVE+SVM	SAVE+SVM
$n_1 = 60, n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = 120, n_2 = 60$	SAVE+SVM	SAVE+SVM
$n_1 = 120, n_2 = 30$	SAVE+SVM	SAVE+SVM
$n_1 = 30, n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = 60, n_2 = 120$	SAVE+SVM	SAVE+SVM
$n_1 = 30, n_2 = 120$	SAVE+SVM	SAVE+SVM

ผลจากตารางจะได้ว่า ในกรณีที่ขนาดของกลุ่มตัวอย่างทั้งหมดมีขนาดไม่เกิน 120 ตัวอย่าง คือ $n_1 = n_2 = 30$, $n_1 = n_2 = 60$, $n_1 = 60, n_2 = 30$ และ $n_1 = 30, n_2 = 60$ วิธีการลดมิติข้อมูลด้วยเทคนิควิธีกำลังสองน้อยสุดเชิงส่วน (PLS) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด ในกรณีที่ขนาดของกลุ่มตัวอย่างทั้งหมดมีขนาดใหญ่กว่า 120 ตัวอย่าง วิธีการลดมิติข้อมูลด้วยเทคนิค Sliced Average Variance Estimator(SAVE) ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และเมื่อพิจารณาค่าเฉลี่ยของ MCR จะได้ว่าวิธีที่ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด จะให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด

5.1.1.3 กรณีที่มีตัวแปรอิสระ 20 ตัวแปร

ตารางที่ 5.1.3 แสดงเทคนิคการลดมิติข้อมูลนำเข้า(input data) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีนด้วย Laplacian Kernel ที่ให้ค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC มากที่สุดและค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภทข้อมูลที่มีค่าน้อยที่สุด โดยจำแนกตามขนาดของกลุ่มตัวอย่างที่สนใจและไม่สนใจ เมื่อ $\rho = 0.85, 0.9, 0.95$ และ $d = 1, 1.5, 2, 2.5$ กรณีจำนวนตัวแปรอิสระเท่ากับ 20 ตัว

n	AUC	MCR
$n_1 = n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 120$	PLS+SVM	PLS+SVM
$n_1 = 60, n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = 120, n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = 120, n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = 30, n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = 60, n_2 = 120$	PLS+SVM	PLS+SVM
$n_1 = 30, n_2 = 120$	PLS+SVM	PLS+SVM

ผลจากตารางจะได้ว่า ในกรณีที่ขนาดของกลุ่มตัวอย่างเท่ากันทุกกรณีที่ทำการศึกษา วิธีการลดมิติข้อมูลด้วยเทคนิควิธีกำลังสองน้อยสุดเชิงส่วน (PLS) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และเมื่อพิจารณาค่าเฉลี่ยของ MCR จะให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด

5.1.1.4 กรณีที่มีตัวแปรอิสระ 40 ตัวแปร

ตารางที่ 5.1.4 แสดงเทคนิคการลดมิติข้อมูลนำเข้า(input data) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ที่ให้ค่าเฉลี่ยของพื้นที่ใต้โค้ง ROC มากที่สุดและค่าเฉลี่ยอัตราความผิดพลาดในการจำแนกประเภทข้อมูลที่มีค่าน้อยที่สุด โดยจำแนกตามขนาดของกลุ่มตัวอย่างที่สนใจและไม่สนใจ เมื่อ $\rho = 0.85, 0.9, 0.95$ และ $d = 1, 1.5, 2, 2.5$ กรณีจำนวนตัวแปรอิสระเท่ากับ 40 ตัว

n	AUC	MCR
$n_1 = n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = n_2 = 120$	PLS+SVM	PLS+SVM
$n_1 = 60, n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = 120, n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = 120, n_2 = 30$	PLS+SVM	PLS+SVM
$n_1 = 30, n_2 = 60$	PLS+SVM	PLS+SVM
$n_1 = 60, n_2 = 120$	PLS+SVM	PLS+SVM
$n_1 = 30, n_2 = 120$	PLS+SVM	PLS+SVM

ผลจากตารางจะได้ว่า ในกรณีที่ขนาดของกลุ่มตัวอย่างเท่ากันทุกกรณีที่ทำการศึกษา วิธีการลดมิติข้อมูลด้วยวิธีกำลังสองน้อยสุดเชิงส่วน(PLS) สำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด ให้ประสิทธิภาพความแม่นยำในการพยากรณ์จำแนกประเภทของข้อมูลดีที่สุด และเมื่อพิจารณาค่าเฉลี่ยของ MCR จะให้ค่าอัตราความผิดพลาดในการจำแนกประเภทข้อมูลได้ต่ำที่สุด

5.1.2 ผลกระทบจากระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล (d)

พิจารณาจากผลกระทบที่ได้รับจากค่าพารามิเตอร์ของการแจกแจงข้อมูล (d) จะได้ว่าเมื่อระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล (d) มีค่าเพิ่มมากขึ้น จะได้ว่าค่าเปอร์เซ็นต์การซ้อนทับ(overlap) มีค่าลดลง นั่นคือ จะทำให้กลุ่มตัวอย่างของกลุ่มที่สนใจกับกลุ่มที่ไม่สนใจมีลักษณะการแจกแจงของข้อมูลแตกต่างกันเพิ่มมากขึ้นหรือสามารถอธิบายความแตกต่างระหว่างข้อมูลทั้งสองกลุ่มได้ชัดเจนเพิ่มมากขึ้น ดังนั้นประสิทธิภาพของการพยากรณ์จำแนกประเภทจะเพิ่มขึ้นในทุกวิธีการที่ทำการศึกษา

5.1.3 ผลกระทบจากค่าระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (ρ)

พิจารณาจากผลกระทบที่ได้รับจากค่าระดับความสัมพันธ์ระหว่างตัวแปรอิสระ เมื่อกำหนดระดับค่าพารามิเตอร์ของการแจกแจงข้อมูลและขนาดตัวอย่าง จะได้ว่าเมื่อระดับค่าความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าเพิ่มมากขึ้น ประสิทธิภาพของการพยากรณ์จำแนกประเภทจะลดลงเล็กน้อยในทุกวิธีการที่ทำการศึกษา

5.2 สรุปผลจากฐานข้อมูล crab data และ wine data

จากการจัดการข้อมูล crab data และ wine data กล่าวคือเทคนิคการลดมิติข้อมูลสำหรับการพยากรณ์ด้วยซัพพอร์ตเวกเตอร์แมชชีนที่เหมาะสมสำหรับฐานข้อมูล crab data คือ เทคนิค Sliced Average Variance(SAVE) เมื่อคัดเลือกจำนวนข้อมูล 200 ข้อมูล และเทคนิคลดมิติข้อมูลที่เหมาะสมสำหรับ wine data คือวิธีกำลังสองน้อยสุดเชิงส่วน(PLS) เมื่อคัดเลือกจำนวนข้อมูล 107 ข้อมูล

5.3 แนวทางการศึกษาต่อ

การศึกษาครั้งนี้ได้ทำการศึกษาภายใต้ขอบเขตของตัวแปรตามเป็นข้อมูลเชิงกลุ่มที่อยู่ในระดับนามบัญญัติ (Nominal Scale) มีลักษณะของการเกิดเหตุการณ์เพียงสองเหตุการณ์ ขนาดตัวอย่าง ตัวแปรอิสระมีลักษณะการแจกแจงแบบปกติซึ่งมีระดับความสัมพันธ์และระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล แต่ได้ทำการศึกษาเพียงบางกรณีของเงื่อนไขเหล่านี้ ดังนั้นเพื่อประโยชน์ในการใช้งานต่อไปในอนาคตจึงควรมีการศึกษาเพิ่มเติมโดยขยายขอบเขตให้กว้างขึ้น และครอบคลุมกรณีต่างๆมากขึ้น เพราะฉะนั้นทางผู้ศึกษาจึงได้เสนอแนะนำในการศึกษาต่อในกรณีต่างๆ ดังนี้

5.2.1 ในงานวิจัยต่อไปอาจทำการศึกษาในกรณีที่ตัวแปรอิสระมาจากการแจกแจงแบบอื่นๆ และตัวแปรตามเป็นเป็นข้อมูลเชิงกลุ่มอยู่ในมาตรเรียงอันดับ (Ordinal Scale) ที่สามารถเกิดเหตุการณ์ได้หลากหลายมากยิ่งขึ้น

5.2.2 ศึกษาขนาดตัวอย่างที่สูงขึ้น เพื่อพิจารณาความสามารถในการจำแนกประเภทของข้อมูลหรือความเชื่อถือได้ของตัวแบบ รวมทั้งหาแนวโน้มเมื่อขนาดตัวอย่างเพิ่มขึ้น

5.2.3 จากงานวิจัยนี้ศึกษาผลกระทบระดับค่าพารามิเตอร์ของการแจกแจงข้อมูล (d) ซึ่งทำการศึกษาลักษณะการกระจายของข้อมูลแตกต่างกันของข้อมูลเพียงสองกลุ่ม โดยวัดค่าการซ้อนทับของข้อมูลทั้งสองกลุ่ม ดังนั้นในงานวิจัยต่อไปอาจทำการศึกษาในกรณีที่มีการแจกแจงข้อมูลแบบผสมกันระหว่างการแจกแจงที่เหมือนกันและการแจกแจงที่ต่างกัน

5.2.4 จากงานวิจัยนี้ศึกษาผลกระทบระดับความสัมพันธ์ของการตัวแปรอิสระที่ค่อนข้างสูง ดังนั้นในงานวิจัยต่อไปอาจทำการศึกษาในกรณีที่มีระดับความสัมพันธ์ของตัวแปรอิสระระดับต่างๆ เพื่อเปรียบจำนวนองค์ประกอบหลักหรือตัวแปรอิสระที่ลดได้

5.2.5 ในงานวิจัยนี้ศึกษาตัวแปรอิสระเพียง 5, 10, 20และ 40 ตัวแปร และตัวแปรอิสระนั้นมีลักษณะการแจกแจงพารามิเตอร์ของข้อมูลเหมือนกัน ดังนั้นในงานวิจัยต่อไปจึงน่าจะกำหนดพารามิเตอร์ของการแจกแจงข้อมูลในตัวแปรอิสระที่แตกต่างกัน

5.2.6 ในงานวิจัยนี้ศึกษาฟังก์ชันเคอร์เนลเพียง Laplacian Kernel ซึ่งอาจมีฟังก์ชันเคอร์เนลอื่นๆที่มีประสิทธิภาพในการสร้างระนาบเส้นแบ่งประเภทข้อมูลได้ดีกว่า ดังนั้นในงานวิจัยต่อไปอาจทำการศึกษาฟังก์ชันเคอร์เนลเพิ่มเติม หรือมีการผสมผสานฟังก์ชันเคอร์เนลเพื่อให้มีประสิทธิภาพในการพยากรณ์จำแนกประเภทเพิ่มมากขึ้น

5.2.7 ในงานวิจัยนี้ศึกษาเทคนิคการลดมิติข้อมูลเพียง 3 วิธี ซึ่งอาจมีเทคนิคการลดมิติข้อมูลอื่นๆที่มีประสิทธิภาพในการสร้างระนาบเส้นแบ่งประเภทข้อมูลได้ดีกว่า ดังนั้นในงานวิจัยถัดไปอาจทำการศึกษาเทคนิคการลดมิติข้อมูลเพิ่มเติม เพื่อให้มีประสิทธิภาพในการพยากรณ์จำแนกประเภทเพิ่มมากขึ้น

5.2.8 .ในงานวิจัยครั้งนี้ศึกษากรณีที่จำนวนตัวแปรอิสระน้อยกว่าจำนวนขนาดตัวอย่าง $(p<n)$ ดังนั้นในงานวิจัยถัดไปอาจทำการศึกษากรณีที่จำนวนตัวแปรอิสระมากกว่าจำนวนขนาดตัวอย่าง $(p>n)$ เพื่อประสิทธิภาพในการลดมิติข้อมูลสำหรับการพยากรณ์จำแนกประเภทด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนเพิ่มมากขึ้น

รายการอ้างอิง

Cook, R. D. and Weisberg (1991). "Discussion of Sliced inverse regression for dimension reduction"." Journal of the American Statistical Association 86: 316-342.

Li, K. C. (1991). "Sliced inverse regression for dimension reduction (with discussion)." Journal of the American Statistical Association 86: 316-342.

Melnykov and Maitra. (2010). "Finite Mixture Model and Model-Based Clustering." statistics surveys: 80-116.

กัลยา วานิชย์บัญชา (2551). การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพฯ บริษัทธรรมสาร.

นันทณัฐ พันธุ์สีดา (2556). การจำลองข้อมูลเพื่อเปรียบเทียบความแม่นยำในการพยากรณ์ระหว่างวิธีโครงข่ายประสาทเทียมกับวิธีซัพพอร์ตเวกเตอร์แมชชีน. คณะพาณิชยศาสตร์และการบัญชี. กรุงเทพฯ , จุฬาลงกรณ์มหาวิทยาลัย. วิทยานิพนธ์ปริญญามหาบัณฑิต.

ปราณี คำแก้ว (2552). การวิเคราะห์โครงสร้างยา โดยวิธีการถดถอยกำลังสองน้อยที่สุดบางส่วน. คณะวิทยาศาสตร์ เชียงใหม่ มหาวิทยาลัยเชียงใหม่. วิทยานิพนธ์ปริญญามหาบัณฑิต.

บรรณานุกรม

กิตติ วารีรัตน์.การแก้ไขปัญหาการประมาณค่าพารามิเตอร์ของตัวแบบความถดถอยพหุเชิงเส้นที่เกิดปัญหาพหุสัมพันธ์ด้วยวิธีการวิเคราะห์ตัวประกอบหลัก, วิทยานิพนธ์ปริญญามหาบัณฑิต, คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2555.

พิมพลอย ภูววิเชียรฉาย.การวิเคราะห์ปริมาตรแบบ HYPER ใต้โค้ง ROC สาหรับตัวแบบโลจิสติกพหุนาม,วิทยานิพนธ์ปริญญามหาบัณฑิต, คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย, 2553.

ภัทราวุฒิ แสงศิริ ศจีมาจ ณ วิเชียร และพยุง มีสัจ. การเปรียบเทียบประสิทธิภาพการลดตัวแปรข้อมูลเข้าที่เหมาะสม สำหรับโครงข่ายประสาทเทียมระหว่างเทคนิคการเลือกตัวแปรแบบถอยหลังทีละขั้น และการวิเคราะห์องค์ประกอบเพื่อพยากรณ์กลุ่มข้อมูลโรคมะเร็ง. The 5 th National Conference on Computing and Information Technology, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2552.

Efstathia Bura. Estimating the structural dimension of regressions via parametric Inverse regression.J.R.Statist. Soc. B (2001)63,Part 2,pp.393-410

L.J. Caoa;* , K.S.Chuab , W.K. Chongc , H.P. Leea , Q.M. Gud. A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. Neurocomputing 55 (2003) 321 – 336

Sanford Weisberg. Dimension Reduction Regression in R. National Science Foundation Grant DUE 0109756,2002

Sanford Weisberg. The dr package. School of Statistics, University of Minnesota, Minneapolis,MN55455. 2011

Zhihui Tang. THREE TOPICS ON DIMENSION REDUCTION. Thesis in Statistics, The Pennsylvania State University, 2007

Yuan – chin Ivan Chang. Boosting SVM Classifiers with Logistic Regression. Institute of Statistical Science, Academia Sinica, Taipei, Taiwan.

Yongwu Shao, R. Dennis Cook and Sanford Weisberg. Marginal tests with sliced average variance estimation. Biometrika (2007), pp. 1–12

ภาคผนวก

สั่งที่ใช้ในการวิเคราะห์ข้อมูลจากการจำลองด้วยโปรแกรม R

ตัวอย่างกรณีที่มีตัวแปรอิสระ 5 ตัวแปรที่มีการแจกแจงแบบปกติหลายตัวแปร (The Multivariate Normal Distribution) และตัวแปรตาม 2 กลุ่ม ซึ่งกำหนดระดับความสัมพันธ์ของตัวแปรอิสระเป็น 0.85 $(\rho = 0.85)$ จำนวนขนาดของตัวอย่างคือ $n_1 = 30$ และ $n_2 = 30$ โดยที่ค่าเฉลี่ยของแต่ละกลุ่มคือ $\mu_1 = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$ และ $\mu_2 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$

```
#######################################

###       Multivariate Normal Distribution    ###

###            n1=30  +  n2=30                ###

#######################################

library(class)

library(e1071)

library(kernlab)

library(grid)

library(MASS)

library(mvtnorm)

library(dr)

library(pls)

library(gtools)

library(gdata)

library(caTools)

library(KernSmooth)
```

```
library(gplots)

library(ROCR)

library(plyr)

library(utils)

library(pROC)

###################################

###      Collect  the data set          ###

###################################

Result_AUC_PCA<-c()

Result_AUC_PLS<-c()

Result_AUC_SAVE<-c()

Result_MCR_PCA<-c()

Result_MCR_PLS<-c()

Result_MCR_SAVE<-c()

dim_PCA<-c()

dim_PLS<-c()

dim_SAVE<-c()

###################################

###      Replication (i in 1:500)          ###

###################################

for (r in 1:500)
```

```
        {
################################################
###     Define the parameters of the distribution      ###
################################################
npos<-30                ## The number of samples positive group
nneg<-30                ## The number of samples negative group
n<-npos+nneg            ## The number of samples
d<-0.5          ## Mean of the negative group
p<-5            ## The number of independence variable
rho<-0.85
########################################
###         Simulation of the data                ###
###       Multivariate Normal Distribution        ###
########################################
mu_X_gp1 <- matrix(0, p, 1)                 ## Mean vector of the positive group
mu_X_gp2 <- matrix(d, p, 1)                 ## Mean  vector of the negative group
corr_X <- matrix(, p, p)
for(j in 1:p){
  for(i in 1:p){
      corr_X[i,j] = rho^abs(i-j)
  }
```

```
}

X_gp1<- rmvnorm (n_gp1, mu_X_gp1, corr_X)

X_gp2<- rmvnorm (n_gp2, mu_X_gp2, corr_X)

X<-rbind(X_gp1,X_gp2)                                    ## Independent
variables

colnames(X)<-c("X1","X2","X3","X4","X5")

y<-matrix(c(rep(1,n_gp1),rep(-1,n_gp2)))                 ## Dependent variable

Data<-data.frame(X,y)                                    ## Data set

#####################################################################

####                  Principle Component Analysis                 ####

#####################################################################

pca = princomp(X, scores = TRUE,cor=TRUE)

summary(pca)

Corr<-cor(X)            ## Correlation of the Independent Variables

e<-eigen(Corr)          ##eigenvalue of  correlation matrix

lamda<-c(e$values)   ##eigenvalue of  correlation matrix

sumlamda<-sum(lamda)  ## summation of  eigenvalue

PC=lamda[1]/sumlamda     ## the first component

for(q_PCA in 1:p) {if(PC<0.8){PC=PC+(lamda[q_PCA+1]/sumlamda)

}

else{break}

}                                        ##find number of component
```

```
q_PCA                                   ##number of component
w_PCA<-matrix(e$vector,p,p)             ##eigenvector  of  correlation matrix
comp_PCA=matrix(, nrow=n, ncol= q_PCA)              ##suitable  component
for(i in 1: q_PCA){
  PC=0
     for(j in 1:p){
     PC=w_PCA [j,i]*X[,j]+PC
}
comp_PCA [,i]=PC
}
################################################
####        Input data for support vector machines     ####
################################################
if(q_PCA ==1){formula_PCA=(y~PCA1)                    ## formular for q
component
colnames(comp_PCA)<-"PCA1"
}else if(q_PCA ==2){formula_PCA =(y~PCA1+PCA2)
colnames(comp_PCA)<-c("PCA1","PCA2")
}else if(q_PCA ==3){formula_PCA =(y~PCA1+PCA2+PCA3)
colnames(comp_PCA)<-c("PCA1","PCA2","PCA3")
}else if(q_PCA ==4){formula_PCA =(y~PCA1+PCA2+PCA3+PCA4)
colnames(comp_PCA)<-c("PCA1","PCA2","PCA3","PCA4")
```

```
}else {formula_PCA =(y~PCA1+PCA2+PCA3+PCA4+PCA5)

colnames(comp_PCA)<-c("PCA1","PCA2","PCA3","PCA4","PCA5")}

formula_PCA

Input_PCA<-data.frame(comp_PCA,y)

#################################################################

###     Selection of the parameters of support vector machines is best      ###

#################################################################

##### Support vector machines with Laplacian kernel        #####

SVM_simple_PCA<- tune(svm, formula_PCA, data =Input_PCA,ranges = list(gamma =
2^(-5:5), cost = (1:50)),tunecontrol = tune.control(sampling = "fix"))          ##
Simulation data with support vector machines

best_para_PCA<-SVM_simple_PCA$best.parameters            ## The  best parameters

best_gamma_PCA  <-best_para_PCA$gamma          ## The best gamma

best_cost_PCA<-best_para_PCA$cost                    ## The  best C

################################################

##### Support vector machines with Laplacian Kernel     #####

################################################

##### Parameter estimation  #####

SVM_PCA<-ksvm(formula_PCA, data =Input_PCA,type="C-
svc",kernel="laplacedot",kpar=list(sigma= best_gamma_PCA  ),C=
best_cost_PCA,prob.model=TRUE)

beta_PCA<-b(SVM_PCA)

##### Predictive of classification       #####
```

```
ypred_PCA<-predict(SVM_PCA, comp_PCA)                    ## Prediction for y

table_ypred_PCA<-table(y,ypred_PCA)                      ## Results of predictions

MCR_PCA<-(1-(sum(diag(table_ypred_PCA))/sum(table_ypred_PCA)))   ##
Misclassification Rate

##### Graphing and finding the area under the ROC curve            ######

ypredscore_PCA<-predict(SVM_PCA, comp_PCA,type="decision")

table_ypredscore_PCA<-table(ypredscore_PCA>0,ypred_PCA)

pred_PCA<-prediction(ypredscore_PCA,y)

perf_PCA<-performance(pred_PCA,measure="tpr",x.measure="fpr")  ## Graphing the
ROC curve

AUC_PCA<-auc(y,ypredscore_PCA)

#####################################################################

####                          Partial Least Squares                          ####

#####################################################################

PLS <- plsr(y ~ X1+X2+X3+X4+X5, ncomp = p, data = Data, validation = "LOO")

RMSEP<-RMSEP(PLS)                                        ##Cross-validated

Int_adjCV<-c()

for(i in 1:6){

  Int_adjCV<-c(Int_adjCV,RMSEP$val[2,1,r])

}

adjCV<-Int_adjCV[-1]                                     ##no intercept term
```

```
q_PLS<-which(adjCV==min(adjCV))       ##the least  dimension of Independent
Variables

w_PLS<-matrix(PLS$loadings,nrow=p,ncol=p)          ##eigenvector

comp_PLS =matrix(, nrow=n, ncol= q_PLS)              ##suitable  component

for(i in 1:q_PLS){

  PLS=0

     for(j in 1:p){

     PLS=w_PLS[j,i]*X[,j]+PLS

}

comp_PLS[,i]=PLS

}

comp_PLS

#####################################################################

##  Input data of Sliced Average Variance Estimator for  support vector machines  ##

#####################################################################

if(q_PLS==1){formula_PLS=(y~PLS1)

colnames(comp_PLS)<-"PLS1"

}else if(q_PLS==2){formula_PLS =(y~PLS1+PLS2)

colnames(comp_PLS)<-c("PLS1","PLS2")

}else if(q_PLS==3){formula_PLS =(y~PLS1+PLS2+PLS3)

colnames(comp_PLS)<-c("PLS1","PLS2","PLS3")

}else if(q_PLS==4){formula_PLS=(y~PLS1+PLS2+PLS3+PLS4)
```

```
colnames(comp_PLS)<-c("PLS1","PLS2","PLS3","PLS4")

}else{formula_PLS=(y~ PLS1+PLS2+PLS3+PLS4+PLS5)

colnames(comp_PLS)<-c("PLS1","PLS2","PLS3","PLS4","PLS5")}

formula_PLS

Input_PLS<-data.frame(comp_PLS,y)

#####################################################################

###          Selection of the parameters of support vector machines is best          ###

#####################################################################

##### Support vector machines with Laplacian kernel       #####

SVM_simple_PLS<- tune(svm, formula_PLS,data =Input_PLS,ranges = list(gamma =
2^(-5:5), cost = (1:50)),tunecontrol = tune.control(sampling = "fix"))    ## Simulation
data with support vector machines

best_para_PLS<-SVM_simple_PLS$best.parameters                  ## The best
parameters

best_gamma_PLS<-best_para_PLS$gamma                            ## The best gamma

best_cost_PLS<-best_para_PLS$cost                              ## The best C

#####################################################################

#####              Support vector machines with Laplacian Kernel                #####

#####################################################################

##### Parameter estimation  #####
```

```
SVM_PLS<-ksvm(formula_PLS, data =Input_PLS,type="C-
svc",kernel="laplacedot",kpar=list(sigma= best_gamma_PLS),C=
best_cost_PLS,prob.model=TRUE)

beta_PLS<-b(SVM_PLS)

##### Predictive of classification      #####

ypred_PLS<-predict(SVM_PLS,comp_PLS)                    ## Prediction for y

table_ypred_PLS<-table(y,ypred_PLS)                     ## Results of
predictions

MCR_PLS<-(1-(sum(diag(table_ypred_PLS))/sum(table_ypred_PLS)))     ##
Misclassification Rate

##### Graphing and finding the area under the ROC curve          ######

ypredscore_PLS<-predict(SVM_PLS,comp_PLS,type="decision")

table_ypredscore_PLS<-table(ypredscore_PLS>0,ypred_PLS)

pred_PLS<-prediction(ypredscore_PLS,y)

perf_PLS<-performance(pred_PLS,measure="tpr",x.measure="fpr")      ## Graphing the
ROC curve

AUC_PLS<-auc(y,ypredscore_PLS)

#####################################################################

####                    Sliced Average Variance Estimator                    ####

#####################################################################

SAVE<- dr(y~ X1+X2+X3+X4+X5, method="save", nslices=2)          ## SAVE method

sumSAVE<-summary(SAVE)                  ##basic summay statistics

SAVEtest<-sumSAVE$test                  ##SAVE test
```

```
pvalue<-c(SAVEtest$"p.value(Nor)")              ## p-value
for(q_SAVE  in 1:length(pvalue)){if(pvalue[q_SAVE]<0.05)
q_SAVE = q_SAVE+1
else{break}
}
q_SAVE<-q_SAVE-1                    ##the least  dimension of Independent Variables
w_SAVE<-matrix(SAVE$evectors,p,p)       ##Eigenvector  of  weighted covariance
matrix
comp_SAVE=matrix(, nrow=n, ncol=q_SAVE)          ##suitable  component
for(i in 1:q_SAVE){
   SAVE=0
     for(j in 1:p){
        SAVE =w_SAVE[j,i]*X[,j]+ PC
}
comp_SAVE [,i]= SAVE
}
comp_SAVE
############################################################
####              Input data for support vector machines                  ####
############################################################
if(q_SAVE==1){formula_SAVE=(y~SAVE1)
colnames(comp_SAVE)<-"SAVE1"
```

```
}else if(q_SAVE==2){formula_SAVE =(y~SAVE1+SAVE2)

colnames(comp_SAVE)<-c("SAVE1","SAVE2")

}else if(q_SAVE==3){formula_SAVE =(y~SAVE1+SAVE2+SAVE3)

colnames(comp_SAVE)<-c("SAVE1","SAVE2","SAVE3")

}else if(q_SAVE==4){formula_SAVE =(y~SAVE1+SAVE2+SAVE3+SAVE4)

colnames(comp_SAVE)<-c("SAVE1","SAVE2","SAVE3","SAVE4")

}else{formula_SAVE=(y~ SAVE1+SAVE2+SAVE3+SAVE4+SAVE5)

colnames(comp_SAVE)<-c("SAVE1","SAVE2","SAVE3","SAVE4","SAVE5")}

formula_SAVE

Input_SAVE <-data.frame(comp_SAVE,y)

#####################################################################

###          Selection of the parameters of support vector machines is best          ###

#####################################################################

##### Support vector machines with Laplacian kernel      #####

SVM_simple_SAVE<- tune(svm, formula_SAVE,data =Input_SAVE,ranges = list(gamma
= 2^(-5:5), cost = (1:50)),tunecontrol = tune.control(sampling = "fix"))     ## Simulation
data with support vector machines

best_para_SAVE<-SVM_simple_SAVE$best.parameters            ## The best
parameters

best_gamma_SAVE<-best_para_SAVE$gamma                  ## The best gamma

best_cost_SAVE<-best_para_SAVE$cost                    ## The best C
```

```
#####################################################

#####       Support vector machines with Laplacian Kernel    #####

#####################################################

##### Parameter estimation  #####

SVM_SAVE<-ksvm(formula_SAVE, data =Input_SAVE,type="C-
svc",kernel="laplacedot",kpar=list(sigma= best_gamma_SAVE ),C=
best_cost_PLS,prob.model=TRUE)

beta_SAVE<-b(SVM_SAVE)

##### Predictive of classification      #####

ypred_SAVE<-predict(SVM_SAVE,comp_SAVE)                  ## Prediction for y

table_ypred_SAVE<-table(y,ypred_SAVE)                    ## Results of predictions

MCR_SAVE<-(1-(sum(diag(table_ypred_SAVE))/sum(table_ypred_SAVE)))    ##
Misclassification Rate

##### Graphing and finding the area under the ROC curve           ######

ypredscore_SAVE<-predict(SVM_SAVE,comp_SAVE,type="decision")

table_ypredscore_SAVE<-table(ypredscore_SAVE>0,ypred_SAVE)

pred_SAVE<-prediction(ypredscore_SAVE,y)

perf_SAVE<-performance(pred_SAVE,measure="tpr",x.measure="fpr")         ##
Graphing the ROC curve

AUC_SAVE<-auc(y,ypredscore_SAVE)

#####################################

###      Collect  the data set           ###

#####################################
```

```
Result_AUC_PCA<-c(Result_AUC_PCA,AUC_PCA)

Result_AUC_PLS<-c(Result_AUC_PLS,AUC_PLS)

Result_AUC_SAVE<-c(Result_AUC_SAVE,AUC_SAVE)

Result_MCR_PCA<-c(Result_MCR_PCA,MCR_PCA)

Result_MCR_PLS<-c(Result_MCR_SAVE,MCR_PLS)

Result_MCR_SAVE<-c(Result_MCR_SAVE,MCR_SAVE)

dim_PCA<-c(dim_PCA,q_PCA)

dim_PLS<-c(dim_PLS,q_PLS)

dim_SAVE<-c(dim_SAVE,q_SAVE)

}

###################################

###       Results of the process            ###

###################################

AUC_PCA<-matrix(Result_AUC_PCA,nrow=500,ncol=1)

AUC_PLS<-matrix(Result_AUC_PLS,nrow=500,ncol=1)

AUC_SAVE<-matrix(Result_AUC_SAVE,nrow=500,ncol=1)

Result_AUC<-cbind(AUC_PCA,AUC_PLS,AUC_SAVE)

colnames(Result_AUC)<-c("AUC PCA","AUC  PLS","AUC SAVE")

rownames(Result_AUC)<-c(1:500)

MCR_PCA<-matrix(Result_MCR_PCA,nrow=500,ncol=1)

MCR_PLS<-matrix(Result_MCR_PLS,nrow=500,ncol=1)
```

```
MCR_SAVE<-matrix(Result_MCR_SAVE,nrow=500,ncol=1)

Result_MCR<-cbind(MCR_PCA,MCR_PLS,MCR_SAVE)

colnames(Result_MCR)<-c("MCR PCA","MCR PLS","MCR SAVE")

rownames(Result_MCR)<-c(1:500)

dimention_PCA<-matrix(dim_PCA,nrow=500,ncol=1)

dimention_PLS<-matrix(dim_PLS,nrow=500,ncol=1)

dimention_SAVE<-matrix(dim_SAVE,nrow=500,ncol=1)


Result_dim<-cbind(dimention_PCA,dimention_PLS,dimention_SAVE)

colnames(Result_dim)<-c("dim PCA","dim  PLS","dim SAVE")

rownames(Result_dim)<-c(1:500)

Result_AUC

Result_MCR

Result_dim
```

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวอารยา หลงชวน เกิดวันที่ 29 เมษายน พ.ศ. 2532 สำเร็จการศึกษาปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาคณิตศาสตร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2554 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต(วท.ม.) สาขาวิชาสถิติ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2555