การวิเคราะห์องค์ประกอบบนชุดข้อมูลที่ทับซ้อน ด้วยวิธีการเลือกลักษณะสำคัญแบบพลวัต

FACTOR ANALYSIS FOR OVERLAP DATA SETS USING DYNAMIC FEATURE SELECTION

วีระยุทธ พิมพาภรณ์
นักศึกษาระดับปริญญาเอก สาขาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
Email : werayut.scisrc@gmail.com

รองศาสตราจารย์ ดร.พยุง มีสัจ
รองศาสตราจารย์ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
Email : pym@kmutnb.ac.th

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อศึกษากระบวนการวิเคราะห์องค์ประกอบ (Factor Analysis) บนชุดข้อมูลที่ทับซ้อนด้วยวิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) โดยประยุกต์ใช้กระบวนการเลือกตัวแปร (Feature Selection) และการวิเคราะห์กลุ่ม (Clustering analysis) ข้อมูลในการประมวลผลเกิดจากกิจกรรมต่างๆ ในระบบการเรียนออนไลน์ (e-Learning) โดยเน้นปัจจัยที่ส่งผลโดยตรงต่อผลสัมฤทธิ์ทางการเรียน

ผลการวิจัยพบว่าประสิทธิภาพโดยรวมของกระบวนการวิเคราะห์องค์ประกอบ โดยใช้อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต ให้ค่าความถูกต้องสูงสุดที่ 45.17% โดยใช้ 3 ตัวแปร สำหรับกระบวนการวิเคราะห์องค์ประกอบโดยวิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio ให้ค่าความถูกต้องสูงสุดที่ 44.80% โดยใช้ตัวแปร 7 ตัวแปร จากผลการวิจัยสามารถสรุปได้ว่า อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัตมีค่าความถูกต้องสูงกว่า และใช้จำนวนตัวแปรที่น้อยกว่าวิธีการคำนวณหาค่า GAIN ของข้อมูลด้วย Information Gain และ Gain ratio

คำสำคัญ : การวิเคราะห์องค์ประกอบ ชุดข้อมูลที่ทับซ้อน วิธีการเลือกลักษณะสำคัญแบบพลวัต

ABSTRACT

The purpose of this research was to study the factor analysis process on the overlapped data sets with the dynamic feature selection (DFS) method with application of the feature selection and clustering analysis processes. The data for this research were obtained from the activities in the e-Learning system, by focusing on factors that directly affect student's achievement.

The results showed that the overall efficiency of the factor analysis with the dynamic feature selection (DFS) method had the highest accuracy of 45.17% with the use of three variables. By contrast, the factor analysis process with the calculation of the information gain and the gain ratio had the highest accuracy of 44.80% with the use

of seven variables. The research results can be conclude that the dynamic feature selection (DFS) algorithm had higher accuracy and required fewer variables than the information gain and the gain ratio algorithm.

KEYWORDS : Factor analysis, Overlapped data sets, Dynamic feature selection (DFS)

บทนำ

การเกิดขึ้นอย่างรวดเร็วของเทคโนโลยีด้านคอมพิวเตอร์ท้าทายให้เกิดการแก้ปัญหาที่มีความยุ่งยากและซับซ้อน โดยเฉพาะเพราะอย่างยิ่งงานด้านการวิเคราะห์ข้อมูลเพื่อสร้างองค์ความรู้ใหม่จากข้อมูลจริงที่เกิดขึ้นด้วยเทคนิคเหมืองข้อมูล (Data Mining) โดยมีวัตถุประสงค์หลักคือการนำข้อมูลที่มีจำนวนมาก มาสืบค้นหาองค์ความรู้ที่ซ้อนอยู่ เพื่อใช้ในการอธิบายปรากฏการณ์จริงที่เกิดขึ้นผ่านชุดข้อมูล (Data Set) หรือเพื่อใช้ในการคาดคะเนสิ่งที่จะเกิดขึ้นในอนาคต

การลดขนาดข้อมูล (Data Reduction) เป็นขั้นตอนในการลดขนาดของข้อมูลด้วยวิธีต่าง ๆ เพื่อให้ข้อมูลมีความเป็นระเบียบ โดยจัดระบบและเชื่อมโยงข้อมูลตามกรอบแนวคิดในเรื่องที่จะศึกษาหรือวิเคราะห์ผลลัพธ์ ซึ่งแนวทางที่นิยมในการลดขนาดของข้อมูลสามารถทำได้ 2 แนวทางคือ การลดขนาดของข้อมูล (Data size reduction) (Sengupta, Srivastava, and Sil, 2013) เป็นการลดจำนวนแถวของข้อมูลซึ่งมีปริมาณมาก เช่น ข้อมูลของระบบตรวจจับการบุกรุกเครือข่าย (Intrusion Detection Systems) ซึ่งข้อมูลดิบที่ได้มาจะมีลักษณะซ้ำซ้อนมากส่งผลให้กระบวนการลดขนาดของข้อมูลเป็นสิ่งที่จำเป็นต้องทำก่อนการนำชุดข้อมูลไปประมวลผล และการลดขนาดมิติของข้อมูล (Dimensionality reduction) (Fodor 2002) คือ การลดขนาดของข้อมูลโดยพิจารณาเลือกเฉพาะแอตทริบิวต์หรือองค์ประกอบของข้อมูลที่ส่งผลดีต่อกระบวนการวิเคราะห์ข้อมูลและตัดแอตทริบิวต์หรือองค์ประกอบของของข้อมูลที่ส่งผลลบต่อการวิเคราะห์ข้อมูลออกไป เทคนิคในการลดขนาดมิติของข้อมูลที่นิยมในปัจจุบันมีหลายวิธี เช่น การวิเคราะห์ส่วนประกอบสำคัญ (Principal Component Analysis: PCA) การวิเคราะห์องค์ประกอบอิสระ (Independent Component Analysis: ICA) และการวิเคราะห์องค์ประกอบ (Factor Analysis) เป็นต้น

งานวิจัยชิ้นนี้ เป็นการนำเสนอผลการทดสอบประสิทธิภาพของอัลกอริทึมใหม่ที่ใช้ในการวิเคราะห์องค์ประกอบข้อมูล (Factor Analysis) ได้แก่ อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) เป็นอัลกอริทึมที่มีวัตถุประสงค์เพื่อคัดเลือกเฉพาะองค์ประกอบข้อมูล (Factor) ที่มีความเหมาะสมกับกระบวนการสกัดองค์ความรู้ที่ซ้อนอยู่ในชุดข้อมูล สำหรับงานวิจัยนี้ผู้วิจัยนำอัลกอริทึมในการจัดกลุ่มข้อมูลมาใช้ในการทดสอบเพื่อหาประสิทธิภาพในการการวิเคราะห์องค์ประกอบข้อมูล (Factor Analysis) ที่สำคัญ ได้แก่ Hierarchical Clustering (HRC) (Ward Jr, 1963; Zhao, Karypis, and Fayyad, 2005)

วัตถุประสงค์ของงานวิจัย

1. เพื่อพัฒนาวิธีใหม่ในการวิเคราะห์องค์ประกอบข้อมูล (Factor Analysis) ด้วยวิธีการเลือกลักษณะสำคัญแบบพลวัต
2. เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์องค์ประกอบข้อมูล (Factor Analysis) ด้วยวิธีการวัดค่า GAIN ของข้อมูล โดยวิธี Information Gain (IG) Gain ratio (GR) และการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS)

ทฤษฎีที่เกี่ยวข้อง

การเลือกคุณลักษณะ (Feature Selection)

การเลือกคุณลักษณะของชุดข้อมูลที่มีจำนวนมิติสูงเป็นขั้นตอนหนึ่งของกระบวนการสร้างแบบจำลองเพื่อการพยากรณ์ด้วยวิธีการเหมืองข้อมูล สำหรับการเลือกข้อมูลย่อยที่มีมิติน้อยลง (Data Reduction) (Yuan et al., 1999) กว่าข้อมูลต้นฉบับ (Original data) กระบวนการคัดเลือกคุณลักษณะถือเป็นงานสำคัญในการปรับปรุงประสิทธิภาพในการสร้างแบบจำลอง และ) อีกทั้งกระบวนการ Feature Selection ยังเป็นการช่วยในการเพิ่มความถูกต้องในการพยากรณ์ (Improving Prediction accuracy) (Koller and Sahami, 1996) เนื่องจากจุดประสงค์

สำคัญของการทำ Feature Selection เพื่อลดจำนวนมิติของข้อให้เหลือเพียงชุดข้อมูล (Feature Subset) ที่มีส่งผลต่อความถูกต้องในการพยากรณ์มากที่สุด ดังนั้นจึงกล่าวโดยสรุปได้ว่ากระบวนการเลือกคุณลักษณะถือเป็นวิธีการหนึ่งที่มีความสอดคล้องกับกระบวนการวิเคราะห์องค์ประกอบข้อมูล (Factor Analysis) (มณเฑียร รัตนศิริวงศ์วุฒิ 2553)

การเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection)

อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection) เป็นอัลกอริทึมที่ถูกออกแบบมาเพื่อใช้ในการหาตัวแปรที่ดีที่สุดสำหรับการจำแนกกลุ่ม ทั้งนี้ผู้วิจัยได้ออกแบบ วิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection) ซึ่งเป็นวิธีการในการหาลักษณะหรือองค์ประกอบข้อมูลที่ส่งผลดีที่สุดต่อ อัลกอริทึมการจัดกลุ่มข้อมูล (Clustering Algorithm) และอัลกอริทึมการจำแนกกลุ่มข้อมูล (Classification Algorithm) (Liu and Yu, 2005)

การวิเคราะห์กลุ่ม (Clustering analysis)

การวิเคราะห์กลุ่มคือเทคนิคในการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning technique) ซึ่งมีเป้าหมายเพื่อการจำแนกกลุ่มข้อมูลที่มีคุณลักษณะคล้ายกันอยู่ในกลุ่มเดียวกัน (Tsai et al., 2011) โดยข้อมูลแต่ละกลุ่มจะถูกเรียนกว่า คลัสเตอร์ (Cluster) การวิเคราะห์หรือจำแนกกลุ่มข้อมูลนั้นสามารถแบ่งออกได้เป็น 2 ประเภทได้แก่ วิธีการแบบลำดับชั้น (Hierarchical algorithms) และวิธีการแบบไม่เป็นลำดับชั้น (Non-hierarchical algorithms) (Han, 2012)

การวัดประสิทธิภาพแบบจำลอง

วิธีวัดประสิทธิภาพแบบจำลอง เป็นการเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึมที่ใช้ในการพัฒนาแบบจำลองการพยากรณ์ สามารถทำได้โดยการวัดประสิทธิภาพการจำแนกข้อมูลตามแนวคิดด้านการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งเป็นการวัดค่าต่างๆ ดังนี้ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measurement) ซึ่งเป็นการคำนวณจากตาราง Confusion Matrix (Witten, 2011)

วิธีดำเนินการวิจัย

ศึกษาปัญหาและวิเคราะห์ข้อมูล

งานวิจัยชิ้นนี้ผู้วิจัยเลือกเก็บข้อมูลจากวิชาที่จัดการเรียนการสอนขึ้นในระดับปริญญาตรี ซึ่งเป็นวิชาในกลุ่มศึกษาทั่วไป จากการจัดการศึกษาพบว่าคำถามสำคัญที่เกิดขึ้นระหว่างการเรียนการสอนในแต่ละภาคการศึกษา คือ การประมาณการผลสัมฤทธิ์ที่จะเกิดขึ้นเมื่อสิ้นสุดภาคการศึกษา ซึ่งคำถามดังกล่าวจะเกิดขึ้นตลอดระยะเวลาการศึกษา จากคำถามดังกล่าวนำมาสู่การแก้ปัญหาด้วยวิธีการพยากรณ์ผลสัมฤทธิ์ที่จะเกิดขึ้นกับผู้เรียนในอนาคต ผู้วิจัยจึงจัดทำระบบการเรียนออนไลน์ (e-Learning) และนำไปใช้ในการจัดการเรียนการสอนแบบผสมผสาน เพื่อเป็นเครื่องมือสำหรับรวบรวมข้อมูลที่เกิดขึ้นตลอด 6 ภาคการศึกษา (2/2553 ถึง 1/2556) มีจำนวนผู้เรียนทั้งสิ้น 1,605 คน และมีการเก็บรวบรวมข้อมูลผู้เรียนทุกคนสำหรับใช้ในการวิจัย ชุดข้อมูลที่เก็บรวบรวมมีรายละเอียดดังนี้

ข้อมูลที่ใช้สำหรับวิเคราะห์ผู้วิจัยแบ่งข้อมูลออกเป็น 4 ส่วนได้แก่

1. ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ความสามารถด้านพุทธิพิสัย (Cognitive Domain) ซึ่งส่งผลต่อผลสัมฤทธิ์ทางการเรียน ผู้วิจัยใช้แบบทดสอบวัดความรู้ความเข้าใจ ที่ออกแบบให้ข้อสอบมีความสอดคล้องกับวัตถุประสงค์และเป้าหมายของรายวิชา แบ่งการทดสอบออกเป็น 3 ประเภทได้แก่ 1) แบบทดสอบก่อนเรียน (Pre-test) เป็นแบบทดสอบที่ใช้ประเมินผู้เรียนก่อนดำเนินการเรียนการสอน เพื่อสำรวจความพร้อมของผู้เรียน และวัดความรู้พื้นฐานเดิมของผู้เรียน 2) แบบทดสอบประจำบทเรียน (Formative test) เป็นการทดสอบตามวัตถุประสงค์ การเรียนรู้ของแต่ละหน่วยการเรียน เพื่อสำรวจความรู้ความเข้าใจที่ผู้เรียนได้จากการเรียนผ่านระบบการเรียนแบบผสมผสาน โดยแบ่งออกเป็น 8 หน่วยการเรียนรู้ และ 3) แบบทดสอบหลังเรียน (Post-test) เป็นแบบทดสอบสำหรับประเมินผลการเรียนของผู้เรียนเมื่อสิ้นสุดรายวิชา ทั้งนี้การทดสอบทั้ง 3 ส่วน คิดค่าคะแนนเป็น 15% ของคะแนนทั้งหมดในรายวิชา

2. คะแนนการทดสอบประจำภาคการศึกษา แบ่งออกเป็น 2 ส่วนดังนี้ 1) การทดสอบกลางภาค (Midterm Examination) ครอบคลุมเนื้อหาในหน่วยการเรียนที่ 1 - 4 คิดเป็นค่าคะแนน 30% ของรายวิชา และ 2) การทดสอบปลายภาคเรียน (Final

Examination) ครอบคลุมเนื้อหาในหน่วยการเรียนที่ 5 - 8 คิดเป็นค่าคะแนน 30% ของรายวิชา

3. ค่าคะแนนการฝึกปฏิบัติ คิดเป็น 25% ของคะแนนในรายวิชา เป็นการจัดกิจกรรมโดยเน้นการฝึกปฏิบัติในห้องปฏิบัติการ เพื่อเป็นการเสริมทักษะด้านการใช้คอมพิวเตอร์และโปรแกรมประยุกต์ต่างๆ แก่ผู้เรียน

4. ข้อมูลทั่วไปของผู้เรียน ชุดข้อมูลในส่วนสุดท้ายได้แก่ ข้อมูลทั่วไปของผู้เรียน เช่น ข้อมูลภาคการศึกษา ชั้นปีที่เรียน สาขา เป็นต้น

เมื่อศึกษาในรายละเอียดของค่าคะแนนที่เกิดจากกิจกรรมตลอดภาคการศึกษาเทียบกับผลสัมฤทธิ์ทางการศึกษาที่เกิดขึ้นเมื่อผู้เรียนจบการเรียนในรายวิชานั้นๆ จะพบว่า

ตารางที่ 1 ลักษณะข้อมูลที่ปัจจัยสำหรับใช้ในงานวิจัย

ลำดับตัวแปร (Factors Order)	ปัจจัย (Factors)	ประเภทข้อมูล (Data type)	รายละเอียด	ที่มาของข้อมูล
ข้อมูลทั่วไปของผู้เรียน				
1	Major	อักษร	สาขา	ข้อมูลทั่วไปของผู้เรียน
2	Level	อักษร	ระดับชั้นปี	
3	Group.Lec	อักษร	กลุ่มบรรยาย	
4	Group.Lab	อักษร	กลุ่มปฏิบัติ	
5	Semester	อักษร	ภาคการศึกษา	
ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ความสามารถด้านพุทธิพิสัย (Cognitive Domain)				
6	Pre-test	เปอร์เซ็นต์	ทดสอบก่อนเรียน	ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ความสามารถด้านพุทธิพิสัย (Cognitive Domain)
7	FT-1	เปอร์เซ็นต์	ทดสอบบทที่ 1	
8	FT-2	เปอร์เซ็นต์	ทดสอบบทที่ 2	
9	FT-3	เปอร์เซ็นต์	ทดสอบบทที่ 3	
10	FT-4	เปอร์เซ็นต์	ทดสอบบทที่ 4	
11	FT-5	เปอร์เซ็นต์	ทดสอบบทที่ 5	
12	FT-6	เปอร์เซ็นต์	ทดสอบบทที่ 6	
13	FT-7	เปอร์เซ็นต์	ทดสอบบทที่ 7	
14	FT-8	เปอร์เซ็นต์	ทดสอบบทที่ 8	
15	Post-test	เปอร์เซ็นต์	ทดสอบหลังเรียน	
16	Avg.AllTest	เปอร์เซ็นต์	ค่าเฉลี่ยของ 6-15	
17	Avg.1-8Test	เปอร์เซ็นต์	ค่าเฉลี่ยของ 7-14	
18	NumOfPass	จำนวนเต็ม	ค่า 6-15 >= 60%	
19	Score.eLearnig	จำนวนเต็ม	คะแนนเต็ม 15	
ข้อมูลคะแนนที่ได้จากการฝึกปฏิบัติ				
20	Score.Lab	จำนวน	คะแนนเต็ม 25	ภาคปฏิบัติ
คะแนนการทดสอบประจำภาพการศึกษา				
21	Score.Mit	จำนวน	คะแนนเต็ม 30	คะแนนการทดสอบประจำภาพการศึกษา
22	Score.Final	จำนวน	คะแนนเต็ม 30	
ผลสัมฤทธิ์ทางการศึกษา				
23	Grade	ระดับ	8 ระดับ (A,B,...F)	ระดับผลสัมฤทธิ์ทางการศึกษา

ค่าคะแนนที่เกิดจากกิจกรรมตลอดภาคการศึกษา มีความสัมพันธ์โดยตรงต่อผลสัมฤทธิ์ทางการเรียนของผู้เรียน และมีค่าระดับความสัมพันธ์ที่แตกต่างกันในแต่ละประเภทของกิจกรรม อีกทั้งข้อมูลค่าคะแนนของกิจกรรมต่างๆ ยังมีลักษณะเป็น ชุดข้อมูลทับซ้อน (Overlap Data Sets) คือ ผู้เรียนที่มีระดับผลสัมฤทธิ์ที่แตกต่างกัน เช่น ผู้เรียนที่ได้เกรด A กับผู้เรียนที่ได้เกรด C มีการทำคะแนนในบางกิจกรรมใกล้เคียงกัน และบางกิจกรรมแตกต่างกัน โดยความแตกต่างอาจมีลักษณะมากกว่าหรือน้อยกว่า จากลักษณะของข้อมูลดังกล่าวจะส่งผลโดยตรงต่อค่าความผิดพลาดในการจัดกลุ่มข้อมูล ส่งผลให้ปัจจุบันการวิจัยเรื่อง การจำแนกกลุ่มข้อมูลบนชุดข้อมูลที่มีความทับซ้อน (Overlapping Data Clustering) (Lu et al. 2012; Ben N'Cir, Cleuziou, and Essoussi, 2013; Mak et al., 2011) จะเน้นในการคัดแยกข้อมูลซึ่งมีความเป็นสมาชิกของหลายกลุ่มข้อมูล เพื่อเพิ่มค่าความถูกต้องในการจำแนกข้อมูลให้สูงขึ้น

การเตรียมข้อมูล (Data Preparation)

เริ่มต้นขั้นตอนการทดสอบโดยใช้ชุดข้อมูลการเรียนการสอนของนักเรียน (Student Dataset) ไปจัดเรียงลำดับของตัวแปร (Feature Order) ใหม่ด้วยการวัดค่า GAIN ของข้อมูลโดยการคำนวณด้วยวิธี Information Gain และ Gain ratio ซึ่งคำนวณด้วยโปรแกรม Weka ซึ่งเป็นโปรแกรมโอเพนซอร์ส (Open Source) ที่ใช้ในงานเหมืองข้อมูลอย่างแพร่หลาย ค่าดังกล่าวจะถูกนำมาเป็นดัชนีในการเรียงลำดับตัวแปร โดยจะเรียงลำดับจากตัวแปรที่มีค่า GAIN มากไปหาค่าน้อย

ตารางที่ 2 : ชุดตัวแปร (Feature Set) ที่ผ่านกระบวนการจัดเรียงด้วยค่า GAIN และ DFS

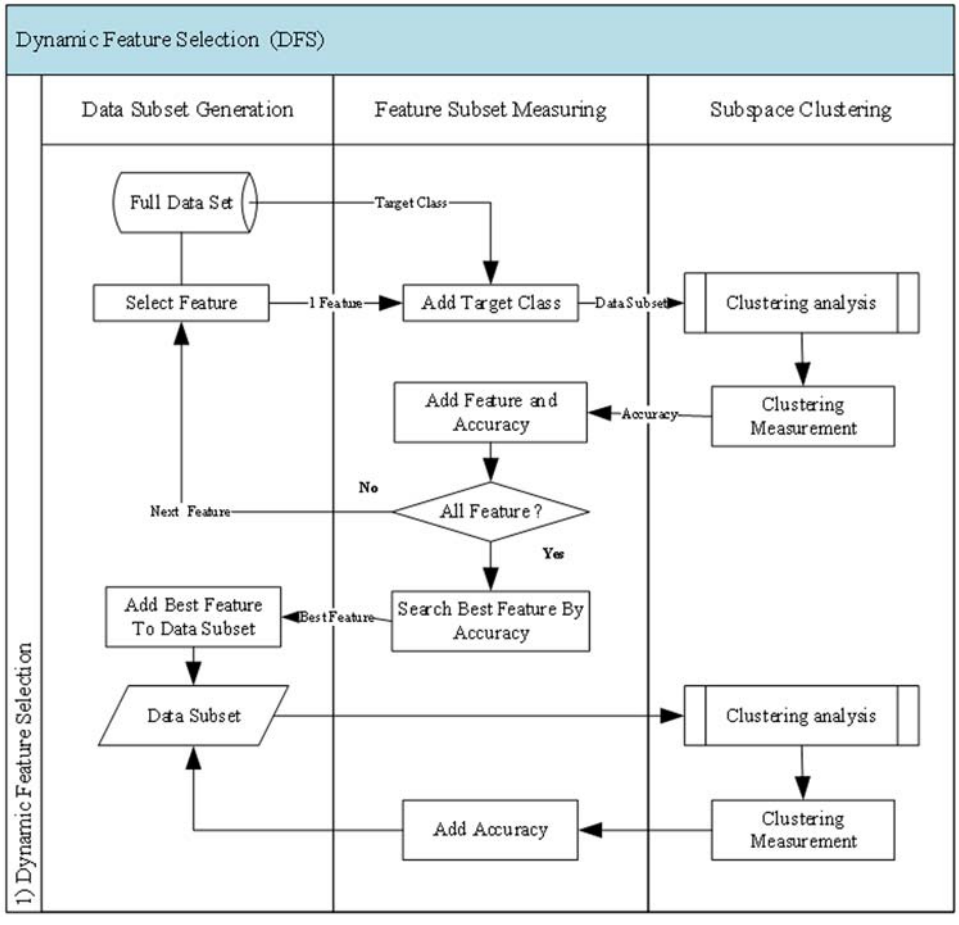
Student Original Dataset (SOD)	Student Information Gain Dataset (SIGD)	Student Gain ratio Dataset (SGRD)	Student DFS Dataset (SDFSD)
Major	Score.Mit	Post-test	Score.Final
Level	Score.Final	Score.Mit	Score.Lab
Group.Lec	Score.eLearning	Avg.1-8Test	Avg.AllTest
Group.Lab	Avg.AllTest	NumOfPass	Level
Semester	NumOfPass	Score.eLearning	FT-8
Pre-test	Post-test	Score.Final	FT-5
FT-1	Avg.1-8Test	Avg.AllTest	FT-7
FT-2	FT-5	FT-5	Pre-test
FT-3	FT-3	Score.Lab	FT-1
FT-4	Score.Lab	FT-6	Avg.1-8Test
FT-5	FT-4	FT-3	NumOfPass
FT-6	FT-7	FT-4	Score.Mit
FT-7	FT-2	FT-7	FT-4
FT-8	FT-8	FT-8	Major
Post-test	FT-6	FT-2	Post-test
Avg.AllTest	FT-1	FT-1	FT-2
Avg.1-8Test	Major	Group.Lec	FT-6
NumOfPass	Pre-test	Group.Lab	FT-3
Score.eLearning	Semester	Pre-test	Score.eLearning
Score.Lab	Group.Lab	Major	Group.Lec
Score.Mit	Group.Lec	Semester	Group.Lab
Score.Final	Level	Level	Semester

เพื่อสร้างชุดข้อมูลใหม่สำหรับเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลโดยวิธีการจัดเรียงชุดตัวแปลใหม่ด้วยอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) โดยชุดข้อมูลแต่ละชุดมีรายละเอียดลำดับของตัวแปรดังนี้

การทดสอบประสิทธิภาพของอัลกอริทึม การเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS)

จากชุดข้อมูลทั้ง 4 ที่ใช้ในการทดสอบค่าที่ใช้ในการทดสอบ คือ ค่าประสิทธิในการจัดกลุ่มข้อมูล (Accuracy) ซึ่งค่าดังกล่าวจะมีผลสอดคล้องกับลำดับของตัวแปรที่ใช้สำหรับจัดกลุ่มและจำนวนตัวแปรที่ใช้ในการจัดกลุ่ม ดังนั้นเมื่อพิจารณาถึงผลลัพธ์ที่ต้องการได้จากการทดลองนี้ คือ การทราบถึงจำนวนตัวแปรที่มีประสิทธิภาพในการจัดกลุ่ม และลำดับของตัวแปรโดยการจัดเรียงนั้นใช้วิธีการวัดค่าลำดับที่แตกต่างกัน สำหรับอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลผู้วิจัยเลือกอัลกอริทึม 2 ประเภท ได้แก่ K-Means Clustering (KMC) และ Hierarchical Clustering (HRC)

การเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) เป็นกระบวนการที่มีความสำคัญ มีวัตถุประสงค์หลักในการค้นหาชุดข้อตัวแปร (Feature Set) ที่ส่งผลดีที่สุดต่อการจัดกลุ่มข้อมูล ทั้งนี้ผู้วิจัยได้ออกแบบ วิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) โดยมุ่งเน้นให้อัลกอริทึมดังกล่าวมีความสามารถ ในการหาชุดของลักษณะหรือชุดของตัวแปรที่ส่งผลดีที่สุดต่ออัลกอริทึม การจัดกลุ่มข้อมูล (Clustering Algorithm) ซึ่งการจะตัดสินใจว่าชุดของตัวแปรใดส่งผลดีที่สุดต่อการจัดกลุ่มข้อมูลนั้น ผู้วิจัยได้นำกระบวนการวัดค่าความถูกต้องในการจัดกลุ่มข้อมูล (Clustering Evaluate) มาใช้สำหรับวิเคราะห์ผลลัพธ์ในการจัดกลุ่มเพื่อใช้ในการตัดสินใจเลือกชุดของตัวแปรในอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) รายละเอียดของตอนการทำงานของอัลกอริทึมดังกล่าวสามารถแสดงดังภาพประกอบที่ 1



ภาพประกอบที่ 1 ขั้นตอนการทำงานของอัลกอริทึม การเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS)

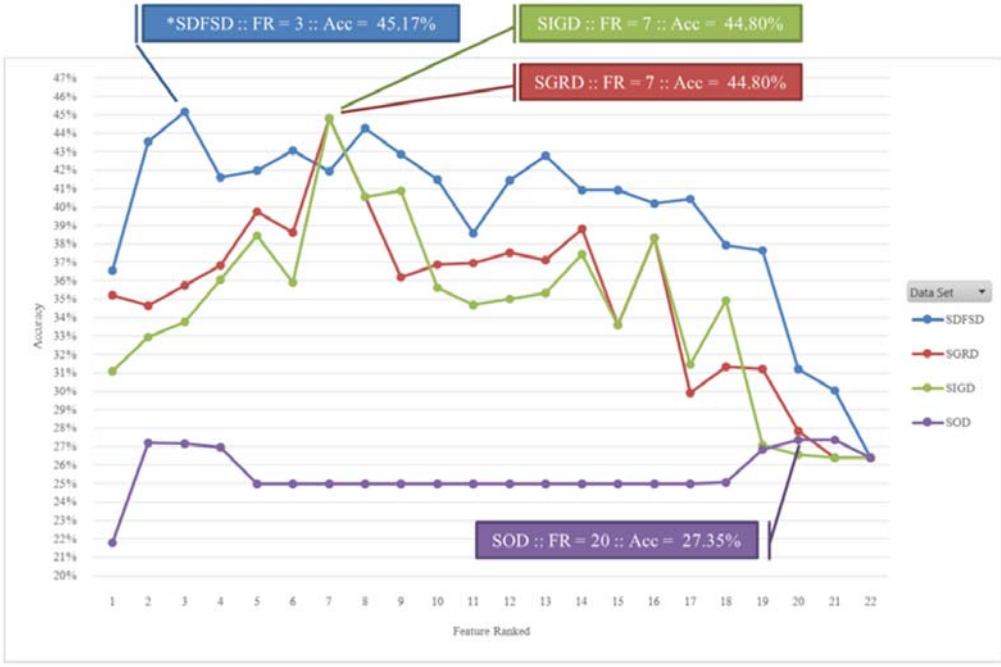
จากภาพแสดงลำดับการทำงานของอัลกอริทึม DFS ในขั้นตอนแรกจะเริ่มจากการนำเข้า ชุดข้อมูล (Data Set) ที่ต้องการนำมาวิเคราะห์ปัจจัย โดยชุดข้อมูลจะต้องประกอบไปด้วยชุดตัวแปร (Feature Set) ที่ต้องการนำมาหาปัจจัยและข้อมูลคำตอบ (Target Class) ที่เป็นผลลัพธ์ จากนั้นกระบวนการทำงานจะเลือกตัวแปรตามลำดับครั้งละ 1 ตัว (1 Feature) และนำตัวแปรดังกล่าวเพิ่มชุดคำตอบ (Target Class) เพื่อส่งต่อไปวิเคราะห์กลุ่มข้อมูล (Clustering analysis) ด้วยอัลกอริทึมจัดกลุ่มข้อมูล (Clustering Algorithm) เมื่อทำการจัดกลุ่มข้อมูลเสร็จสิ้น ตัวแปร (Feature) ดังกล่าวจะได้การประเมินค่าความถูกต้องของการจัดกลุ่ม (Accuracy) ซึ่งถือเป็นดัชนีชี้วัด ความถูกต้องสำหรับชุดตัวแปรย่อยนั้นๆ ทั้งนี้ค่าดังกล่าวสามารถคำนวณจากการนำผลลัพธ์ที่จัดกลุ่มได้จากอัลกอริทึม มาเปรียบเทียบกับค่าคำตอบหรือ Target Class ที่ได้จัดเตรียมไว้ จากนั้นทำการวนซ้ำกระบวนการข้างต้น จนครบทุกตัวแปร และเมื่อครบทุกตัวแปรจะทำการค้นหาตัวแปรที่ให้ค่าความถูกต้องมากที่สุด (Best Feature) เพื่อนำมาสร้างเป็นชุดข้อมูลย่อย (Data Subset)

ผลการวิจัย

ในขั้นตอนการทดลองผู้วิจัยทำการหาประสิทธิภาพของกระบวนการคัดเลือกตัวแปรทั้ง 3 วิธีด้วยค่าความถูกต้องในการจัดกลุ่มข้อมูล (Accuracy) โดยใช้อัลกอริทึมจัดกลุ่มข้อมูล Hierarchical Clustering Algorithm ในการจำแนกกลุ่มข้อมูล ซึ่งขั้นตอนการจำแนกกลุ่มเริ่มต้นจากการนำชุดข้อมูลที่มีการเรียงลำดับทั้ง 4 แบบ เข้าสู่อัลกอริทึมจัดกลุ่ม โดยการเลือกตัวแปรในลำดับที่ 1 นำเข้าสู่อัลกอริทึมการจัดกลุ่ม เพื่อทดสอบค่าแม่นยำในการจัดกลุ่มข้อมูล จากนั้นทำการเพิ่มตัวแปรเพิ่มขึ้นที่ละ 1 ตัวแปรตามลำดับการจัดเรียง และทำซ้ำจนครบทุกตัวแปร ซึ่งผลการวิเคราะห์ประสิทธิภาพในแต่ละชุดข้อมูลสามารถแสดงได้ดังภาพประกอบที่ 2

จากผลการทดสอบค่าประสิทธิภาพของชุดข้อมูลที่มีการจัดเรียงลำดับของชุดตัวแปรแตกต่างกัน 4 แบบ สามารถสรุปประเด็นด้านประสิทธิภาพการจัดกลุ่มได้ดังนี้

ประเด็นการจัดเรียงลำดับตัวแปรใหม่ก่อนนำเข้าสู่กระบวนการจัดกลุ่มข้อมูล โดยใช้อัลกอริทึม Dynamic Feature



ภาพประกอบที่ 2 ผลการเปรียบเทียบผลของประสิทธิภาพของการจัดเรียงตัวแปรด้วยวิธี Information Gain, Gain Ratio, Dynamic Feature Selection และ ชุดข้อมูลเดิม

Selection (DFS) สามารถให้ค่าความถูกต้องในการจัดกลุ่มข้อมูลได้สูงที่สุด เมื่อเทียบกับการจัดเรียงด้วยวิธีอื่น โดยมีค่าความถูกต้องเท่ากับ 45.17% สำหรับการจัดกลุ่มโดยใช้การคำนวณค่า GAIN ของข้อมูลแต่ละตัวแปรด้วยวิธี Information Gain และ Gain Ratio ให้ค่าผลลัพธ์ค่าประสิทธิภาพลองลงมา ที่ระดับ 44.80% ทั้ง 2 วิธี แต่สำหรับการนำชุดข้อมูลซึ่งไม่ผ่านการจัดเรียงไปวิเคราะห์ค่าประสิทธิภาพพบว่า ค่าความถูกต้องมีค่าต่ำที่สุดที่ระดับ 27.35%

ประเด็นจำนวนตัวแปรที่ใช้ในการหาค่าประสิทธิภาพที่สูงที่สุด จากภาพแสดงให้เห็นถึงค่าความถูกต้องในการจัดกลุ่มข้อมูลสูงสุดของแต่ละวิธีการจัดเรียงตัวแปร โดยวิธีการจัดเรียงตัวแปรด้วย อัลกอริทึม Dynamic Feature Selection (DFS) เรียงลำดับตัวแปรและใช้จำนวนตัวแปรน้อยที่สุด ซึ่งส่งผลต่อค่าความถูกต้องในการจัดกลุ่มข้อมูลที่สูงที่สุด โดยใช้ 3 ตัวแปร ได้แก่ Score.Final, Score.eLearning และ Avg.AllTest สำหรับวิธีการเรียงลำดับตัวแปรด้วยการใช้ค่า GAIN ของข้อมูลพบว่า วิธี Information Gain และ Gain Ratio ให้ค่าความถูกต้องเท่ากับที่ระดับ 44.80% และทั้ง 2 วิธีมีการจัดเรียงลำดับแตกต่างกัน แต่คงใช้ตัวแปรชุดเดียวกันจำนวน 7 ตัวแปร ดังนี้ Score.Mit, Score.Final, Score.eLearning, Avg.AllTest, NumOfPass, Post-test, และ Avg.1-8Test

สรุปผลการทดลอง

ผลการทดสอบประสิทธิภาพของอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) เป็นการนำชุดข้อมูลการเรียนการสอนของนักเรียน (Student Dataset) ไปจัดเรียงลำดับตัวแปร (Feature Order) ใหม่ด้วยการวัดค่า GAIN ของข้อมูล ซึ่งในงานวิจัยครั้งนี้ใช้วิธี Information Gain และ Gain ratio ในการจัดเรียงลำดับตัวแปร และนำไปทดสอบกับวิธีการจัดเรียงลำดับตัวแปรของอัลกอริทึม DFS ซึ่งผลการทดลองวัดค่าประสิทธิในการจัดกลุ่มข้อมูลด้วยค่าความถูกต้อง (Accuracy) โดยใช้อัลกอริทึมจัดกลุ่มข้อมูล Hierarchical Clustering Algorithm ในการจำแนกกลุ่มข้อมูล พบว่า การไม่เรียงลำดับของตัวแปร ส่งผลให้ค่าประสิทธิภาพสูงสุดที่ 27.35% โดยใช้ตัวแปร 20 ตัวแปร การจัดลำดับข้อมูลใหม่ โดยใช้การคำนวณหาค่า GAIN ของข้อมูลด้วยวิธี Information Gain ส่งผลให้ค่าประสิทธิภาพสูงสุดที่ 44.80% โดยใช้ตัวแปร 7 ตัวแปร การจัดเรียงลำดับตัวแปรใหม่ โดยใช้การคำนวณหาค่า GAIN ของข้อมูลด้วยวิธี Gain ratio ส่งผลให้ค่าประสิทธิภาพสูงสุดที่ 44.80% โดยใช้ตัวแปร 7 ตัวแปร และการจัดเรียงลำดับตัวแปรใหม่โดยใช้อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) ส่งผลให้ค่าประสิทธิภาพสูงสุดที่ 45.17% โดยใช้ตัวแปร 3 ตัวแปร จากตัวเลขผลลัพธ์ที่ได้จากการทดลองสามารถสรุปได้ว่า อัลกอริทึม DFS สามารถเรียงลำดับตัวแปรส่งผลต่อประสิทธิภาพที่สูงขึ้นกว่าการจัดเรียงลำดับตัวแปรด้วยการหาค่า GAIN ของข้อมูลด้วยวิธี วิธี Information Gain และ Gain ratio อยู่ 0.37% และสูงกว่าการไม่เรียงลำดับของตัวแปรอยู่ 17.82%

อภิปรายผล

เมื่อพิจารณาจากผลการทดลองในขั้นตอนการทดสอบประสิทธิภาพของอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) กับวิธีการจัดเรียงลำดับตัวแปรด้วยค่า GAIN โดยใช้วิธี Information Gain และ Gain Ratio พบว่า อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) ให้ค่าประสิทธิภาพในการจัดกลุ่มมากที่สุดด้วยวิธีการจัดเรียงลำดับตัวแปรใหม่และใช้จำนวนตัวแปรน้อยลงกว่าวิธี Information Gain และ Gain Ratio

ทั้งนี้เป็นผลมาจากรูปแบบการจัดเรียงลำดับตัวแปรของอัลกอริทึม DFS ที่มีการคำนวณหาค่าประสิทธิภาพสูงสุดที่เป็นไปได้ในแต่ละตัวแปรและแต่ละชุดตัวแปร อีกทั้งการใช้ค่าความถูกต้องในการจัดกลุ่มข้อมูลมาเป็นดัชนีในการจัดเรียงลำดับตัวแปร ซึ่งถือได้ว่าค่าความถูกต้องในการจัดกลุ่มข้อมูลเป็นส่วนส่วนประกอบที่สำคัญในการจัดเรียงลำดับตัวแปรในกระบวนการเลือกตัวแปร (Feature Selection)

อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) เป็นอัลกอริทึมที่ออกแบบมาเพื่อนำข้อดีของวิธีการเลือกตัวแปรทั้ง 2 ลักษณะ (Filter Selection และ Wrapper Selection) (Ladha and Deepa 2011) มาใช้เป็นแนวทางในการออกแบบ อีกทั้งยังใช้รูปแบบการคัดเลือกตัวแปรแบบฝังตัว (Embedded Selection) คือ การคัดเลือกตัวแปรด้วย

การสร้างดัชนีชี้วัดความถูกต้องในการจำแนกกลุ่มให้กับตัวแปรต่าง ๆ ในชุดข้อมูล สำหรับดัชนีชี้วัดประสิทธิภาพผู้วิจัยเลือกใช้ค่าความแม่นยำในการจำแนกกลุ่มข้อมูล (Accuracy) มาเป็นดัชนีในการจัดลำดับตัวแปร

ข้อเสนอแนะ

การทำงานของอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS) มีการนำอัลกอริทึมการจัดกลุ่มข้อมูล (Clustering Algorithms) มาใช้ในขั้นตอนการทำงาน ได้แก่ Hierarchical Clustering ซึ่งมีความเหมาะสมกับชุดข้อมูลที่ผู้วิจัยนำมาใช้ในการวิจัย อย่างไรก็ตามจากผลการวิจัยแสดงให้เห็นว่า ค่าประสิทธิภาพของอัลกอริทึม DFS มีผลมาจากปัจจัยในการเลือกอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลซึ่งมีความเหมาะสมกับชุดข้อมูลที่ใช้ในการทดสอบ ดังนั้นกระบวนการทดสอบประสิทธิภาพ และปรับเปลี่ยนอัลกอริทึมให้มีความเหมาะสมกับชุดข้อมูลจึงมีความจำเป็นต่อการเพิ่มประสิทธิภาพให้กับกระบวนการวิเคราะห์องค์ประกอบ (Factor Analysis) บนชุดข้อมูลที่ทับซ้อนด้วยวิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection : DFS)

เอกสารอ้างอิง

มณเฑียร รัตนศิริวงศ์วุฒิศจีมาจ ณ วิเชียร. 2553. "ปัจจัยที่มีอิทธิพลต่อสมรรถนะนักเทคโนโลยีสารสนเทศโดยการวิเคราะห์องค์ประกอบเชิงยืนยันลำดับที่สอง." **วารสารเทคโนโลยีสารสนเทศ**, 6 (12): 1-8.

Ben N'Cir, C.-E., G. Cleuziou, and N. Essoussi. 2013. "Identification of Non-Disjoint Clusters with Small and Parameterizable Overlaps." In **2013 International Conference on Computer Applications Technology (ICCAT),** 1-6. doi:10.1109/ICCAT.2013.6522010.

Fodor, Imola K. 2002. **A Survey of Dimension Reduction Techniques.** Technical Report UCRL-ID-148494, Lawrence Livermore National Laboratory. https://e-reports-ext.llnl.gov/pdf/240921.pdf.

Han, Jiawei. 2012. Data Mining: **Concepts and Techniques.** 3rd ed. Burlington, MA: Elsevier.

Koller, Daphne, and Mehran Sahami. 1996. "Toward Optimal Feature Selection." In , 284-92. Morgan Kaufmann.

Ladha, L., and T. Deepa. 2011. "Feature Selection Methods and Algorithms." **International Journal on Computer Science and Engineering,** 3 (5): 1787-97.

Liu, Huan, and Lei Yu. 2005. "Toward Integrating Feature Selection Algorithms for Classification and Clustering." **Knowledge and Data Engineering, IEEE Transactions on,** 17 (4): 491-502.

Lu, Haibing, Yuan Hong, W.N. Street, Fei Wang, and Hanghang Tong. 2012. "Overlapping Clustering with Sparseness Constraints." In **2012 IEEE 12th International Conference on Data Mining Workshops (ICDMW),** 486-94. doi:10.1109/ICDMW.2012.16.

Mak, Lee Onn, Gee-Wah Ng, G. Lim, and Kezhi Mao. 2011. "A Merging Fuzzy ART Clustering Algorithm for Overlapping Data." In **2011 IEEE Symposium on Foundations of Computational Intelligence (FOCI),** 1-6. doi:10.1109/FOCI.2011.5949461.

Sengupta, Nandita, Amit Srivastava, and Jaya Sil. 2013. "Reduction of Data Size in Intrusion Domain Using Modified Simulated Annealing Fuzzy Clustering Algorithm." In **Mobile Communication and Power Engineering,** edited by Vinu V. Das and Yogesh Chaba, 97-102. Communications in Computer and Information Science 296. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-35864-7_14.

Tsai, C.-F., C.-T. Tsai, C.-S. Hung, and P.-S. Hwang. 2011. "Data Mining Techniques for Identifying Students at Risk of Failing a Computer Proficiency Test Required for Graduation." **Australasian Journal of Educational Technology,** 27 (3): 481-98.

Ward Jr, Joe H. 1963. "Hierarchical Grouping to Optimize an Objective Function." **Journal of the American Statistical Association,** 58 (301): 236-44.

Witten, Ian H. 2011. **Data Mining: Practical Machine Learning Tools and Techniques.** 3rd ed. Morgan Kaufmann Series in Data Management Systems. Burlington, MA: Morgan Kaufmann.

Yuan, Huang, Shian-Shyong Tseng, Wu Gangshan, and Zhang Fuyan. 1999. "A Two-Phase Feature Selection Method Using Both Filter and Wrapper." In **1999 IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings,** 2:132-36 vol.2. doi:10.1109/ICSMC.1999.825221.

Zhao, Ying, George Karypis, and Usama Fayyad. 2005. "Hierarchical Clustering Algorithms for Document Datasets." **Data Mining and Knowledge Discovery,** 10 (2): 141-68.

>> วีระยุทธ พิมพาภรณ์

จบการศึกษาวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีมหานคร และวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีมหานคร

ปัจจุบันทำงานในตำแหน่ง อาจารย์ ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ คณะวิทยาศาสตร์ ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ ศรีราชา ผลงานวิชาการ เช่น A Comparative Efficiency of Clustering Using Dynamic Feature Selection Optimization of Subspace Clustering Algorithms. (2014), Selection Factors Affecting Learning Achievement Following Grade of Students Cluster by Subspace Clustering Algorithms. (2014), An Analysis of Factors Affecting Achievement by Feature Selection Optimization of Clustering Algorithm. (2013).

>> รองศาสตราจารย์ ดร.พยุง มีสัจ

จบการศึกษาหลักสูตรครุศาสตร์อุตสาหกรรมบัณฑิต สาขาวิศวกรรมไฟฟ้า จากสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ ปี พ.ศ. 2537 จบการศึกษา MS และ Ph.D. in Electrical Engineering จาก Oklahoma State University ประเทศสหรัฐอเมริกา ปี พ.ศ. 2541 และ 2545 ตามลำดับ

ปัจจุบันดำรงตำแหน่งคณบดีคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ ผลงานวิชาการ เช่น Incremental Learning Fuzzy Neural Network (ILFN) (2002), An effective neuro-fuzzy paradigm for machinery condition health monitoring (2003), Construction of Fuzzy Ontology-Based Terrorism Event Extraction (2010), A distributed data clustering based on multiple colonies swarm-like agent (2012).