การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล โดยวิธีการเลือกลักษณะสำคัญแบบพลวัตเพื่อเพิ่มประสิทธิภาพของอัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย

A Comparative Efficiency of Clustering Using Dynamic Feature Selection Optimization of Subspace Clustering Algorithms

วีระยุทธ พิมพาภรณ์ (Werayut Pimpaporn)* และ พยุง มีสัจ (Phayung Meesad)*

บทคัดย่อ

งานวิจัยชิ้นนี้เป็นการนำเสนอผลการเปรียบเทียบประสิทธิภาพระหว่างกระบวนการจัดกลุ่มข้อมูลโดยวิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) เพื่อเพิ่มประสิทธิภาพของอัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms) กับเทคนิคการจัดกลุ่มข้อมูลที่มีอยู่เดิม

ผลการวิจัยให้ผลการเปรียบเทียบประสิทธิภาพในด้านความถูกต้องของการจัดกลุ่ม อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับ Hierarchical Clustering ให้ค่าความถูกต้องในภาพรวมสูงที่สุดที่ระดับร้อยละ 94% คิดเป็นการจำแนกข้อมูลตามกลุ่มได้ถูกต้อง 1,503 ข้อมูล จาก 1,605 ข้อมูล โดยกลุ่มจำแนกได้ดีที่สุดคือ A มีค่าความถูกต้องที่ร้อยละ 100% ลองลงมาได้แก่ อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับ K-Means Clustering ให้ค่าความถูกต้องในภาพรวมที่ระดับร้อยละ 83% คิดเป็นการจำแนกข้อมูลตามกลุ่มได้ถูกต้อง 1,331 ข้อมูล จาก 1,605 ข้อมูล โดยกลุ่มจำแนกได้ดีที่สุดคือ A มีค่าความถูกต้องที่ร้อยละ 100% ทั้งนี้ทั้ง 2 อัลกอริทึมผ่านกระบวนการสร้างชุดของลักษณะสำคัญ (Feature Set) ด้วยอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection)

คำสำคัญ: การจัดกลุ่มบนปริภูมิย่อย วิธีการเลือกลักษณะสำคัญแบบพลวัต การจัดกลุ่มแบบลำดับชั้น การจัดกลุ่มแบบเคมีน

Abstract

This research presented the result of the comparative efficiency between the Clustering Using Dynamic Feature Selection (DFS) for increasing efficiency of Subspace Clustering Algorithms and the existing Clustering techniques. The result showed that the comparative efficiency in accuracy of the Subspace Clustering Algorithms worked with the Hierarchical Clustering had the overall accuracy highest rate at the level of 94%. It was classified the 1,503 accurate data from 1,605. The best group is classified as a 100% accuracy was the grade A group. Next, the Subspace Clustering Algorithms that works with K-Means Clustering had the overall accuracy level of 83%. It was classified the 1,331 accurate data from 1,605. The best group is classified as a 100% accuracy was the grade A group. However, both of algorithms through the Feature Set with the Dynamic Feature Selection (DFS).

Keyword: Dynamic Feature Selection, Subspace Clustering Algorithms, Hierarchical Clustering, K-Means Clustering.

1. บทนำ

การประยุกต์ใช้เทคโนโลยีเกี่ยวกับการประมวลผลข้อมูลในปัจจุบันมีการใช้งานอย่างแพร่หลาย เป็นผลมาจากประสิทธิภาพของเทคโนโลยีคอมพิวเตอร์ที่สูงขึ้น ส่งผลให้กระบวนการในการประมวลผลข้อมูลจำเป็นต้องปรับให้

* คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

สอดคล้องกับเทคโนโลยีที่เกิดขึ้น

เทคนิควิธีการวิเคราะห์ข้อมูลในปัจจุบันที่ได้รับความนิยมในการประมวลผลข้อมูลที่มีปริมาณมาก ได้แก่ การทำเหมืองข้อมูล (Data Mining) ทั้งนี้การทำเหมืองข้อมูลในปัจจุบันผู้วิจัยจะทำการประยุกต์ใช้เทคนิควิธีต่าง ๆ เข้ามาใช้ในการพัฒนาอัลกอริทึมใหม่ โดยเน้นไปที่ปัจจัยที่ส่งผลต่อประสิทธิภาพ และความถูกต้องของการประมวลผลข้อมูลของอัลกอริทึม

เมื่อพิจารณาถึงกระบวนการทำเหมืองข้อมูลพบว่าการทำเหมืองข้อมูลถือได้ว่าเป็นเทคนิคในการวิเคราะห์ข้อมูลขั้นสูงอย่างหนึ่ง มีลักษณะในการนำชุดฐานข้อมูลขนาดใหญ่ และมีความหลากหลายมาวิเคราะห์ เพื่อสืบค้นรูปแบบของกฎที่ซ้อนอยู่ในชุดข้อมูล เพื่อเป็นการแสดงให้เห็นถึงองค์ความรู้จากการประมวลผลข้อมูล ทั้งนี้ถือได้ว่ากระบวนการเหมืองข้อมูลเป็นกระบวนการสำคัญสำหรับการสร้างฐานความรู้ (Knowledge base) เพื่อใช้เป็นรากฐานสำคัญในการประยุกต์ใช้งานด้านต่าง ๆ

ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลที่เกี่ยวข้องกับผลสัมฤทธิ์ทางการศึกษา ทั้งนี้ลักษณะของของผลสัมฤทธิ์ทางการเรียนคือค่าที่วัดได้จากความรู้ ความเข้าใจ และความสามารถของผู้เรียนที่เกิดจากการเรียนการสอน ซึ่งวัดผลได้จากการทดสอบ โดยการนำผลไปพิจารณาเทียบกับเกณฑ์ที่ได้กำหนดไว้ เพื่อทำให้สามารถแยกผู้เรียนออกเป็นกลุ่ม ๆ ได้ตามระดับความรู้ โดยมาตรฐานระดับความรู้ของสถาบันอุดมศึกษาในประเทศไทย ใช้ระดับผลสัมฤทธิ์ทางการเรียนแยกออกเป็น 8 ระดับ คือ A (Excellent), B+ (Very Good), B (Good), C (Average) จนถึงระดับ F โดยข้อมูลดังกล่าวจะถูกนำมาเป็นชุดข้อมูลในการวิเคราะห์ประสิทธิภาพของอัลกอริทึม

งานวิจัยชิ้นนี้ เป็นการนำเสนอผลการทดสอบประสิทธิภาพของอัลกอริทึมใหม่ที่ใช้ในการสร้างแบบจำลองการพยากรณ์ ซึ่งประกอบไปด้วยอัลกอริทึม 2 อัลกอริทึม ดังนี้ อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) เป็นอัลกอริทึมที่มีวัตถุประสงค์เพื่อคัดเลือกลักษณะสำคัญ (Feature Selection) ที่ใช้ให้เหมาะสมกับการจัดกลุ่มเพื่อให้สอดคล้องกับผลลัพธ์ของการจัดกลุ่ม และอัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms) ซึ่งมุ่งเน้นในการสร้างแบบจำลอง (Model) ขององค์ความรู้ ซึ่งถือเป็นการสกัดองค์ความรู้ (Knowledge Extraction) จากชุดข้อมูล (Data Set) ซึ่งถือได้ว่าเป็นสำคัญในการพัฒนาอัลกอริทึมในกระบวนการเหมืองข้อมูล

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ประกอบไปด้วยองค์ความรู้ในด้านการทำเหมืองข้อมูล ได้แก่ การวิเคราะห์กลุ่ม (Clustering analysis) กระบวนการเลือกตัวแปร (Feature Selection) และอัลกอริทึมอื่นๆ ที่เกี่ยวข้องกับงานวิจัย ดังนั้นผู้วิจัยได้มีการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวิจัยมีรายละเอียดดังนี้

2.1 เหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล ถือเป็นกระบวนการของการกลั่นกรองสารสนเทศ (Information) ที่ซ่อนอยู่ในฐานข้อมูลขนาดใหญ่ เพื่อใช้ในการทำนายแนวโน้ม และพฤติกรรม โดยอาศัยข้อมูลในอดีตเพื่อค้นหารูปแบบความสัมพันธ์ และองค์ความรู้ใหม่จากข้อมูล [1] โดยมีขั้นตอนดังนี้ 1) ทำความเข้าใจปัญหาโดยการเลือกข้อมูลให้มีความเหมาะสมกับอัลกอริทึมที่ใช้งาน จำนวนที่ต้องการ และค่าเป้าหมายเพื่อให้ได้ผลลัพธ์ที่ต้องการ 2) ทำความเข้าใจข้อมูล โดยการรวบรวม ตรวจสอบความถูกต้อง และกำหนดคุณสมบัติที่ต้องการให้กับข้อมูล 3) เตรียมข้อมูล โดยการคัดเลือกข้อมูลเพื่อทำการแปลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์ข้อมูลด้วยเทคนิคต่าง ๆ 4) สร้างแบบจำลอง โดยแบ่งเป็น 2 ประเภทคือ การสร้างแบบจำลองเพื่อการทำนาย (Predictive Data Mining) เป็นการคาดคะแนนลักษณะหรือประมาณค่าที่ชัดเจนของข้อมูลที่จะเกิดขึ้น โดยใช้ข้อมูลในอดีต และการสร้างแบบจำลองเพื่อใช้บรรยาย (Descriptive Data Mining) เพื่อหาแบบจำลองมาอธิบายลักษณะบางอย่างของข้อมูล

2.2 การวิเคราะห์กลุ่ม (Clustering analysis)

การวิเคราะห์กลุ่มคือเทคนิคในการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning technique) ซึ่งมีเป้าหมายเพื่อการจำแนกกลุ่มข้อมูลที่มีคุณลักษณะคล้ายกันอยู่ในกลุ่มเดียวกัน [2] โดยข้อมูลแต่ละกลุ่มจะถูกเรียกว่า คลัสเตอร์ (Cluster) การวิเคราะห์หรือจำแนกกลุ่มข้อมูลนั้นสามารถแบ่งออกได้เป็น 2 ประเภทได้แก่ วิธีการแบบลำดับชั้น (Hierarchical algorithms) และ วิธีการแบบไม่เป็นลำดับชั้น (Non-hierarchical algorithms) [3]

2.3 Subspace Clustering Algorithms

การจัดกลุ่มบนปริภูมิย่อย (Subspace clustering) ถือได้ว่าเป็นส่วนขยายของการจัดกลุ่มแบบดั้งเดิม (Traditional Clustering) [4] โดยมีวัตถุประสงค์หลักเพื่อการค้นหากลุ่มข้อมูล (Clusters) ปริภูมิย่อยของชุดข้อมูล มักจะถูกนำเสนอในการค้นหากลุ่มข้อมูลที่มีลักษณะ มิติสูง (High Dimensional Data) มิติข้อมูลไม่เกี่ยวข้องกัน (Dimensions Irrelevant) รวมถึงชุดข้อมูลที่มีข้อมูลรบกวน (Noisy Data) ปะปนอยู่กับชุดข้อมูล

2.4 K-Means Clustering

K-Means Clustering [5] คือวิธีการจำแนกกลุ่มข้อมูลด้วยวิธีการแบ่งข้อมูลอัตโนมัติตามค่า k ที่กำหนด โดยกระบวนการทำงานเลือกค่า k เริ่มต้นสำหรับเป็นค่ากลางในการจัดกลุ่ม และปรับค่าตามกระบวนการดังนี้ 1) เลือกข้อมูล di สำหรับวัดระยะห่างกับค่า K เริ่มต้นทุกค่า 2) กำหนดชุดข้อมูล di ให้กับ K ที่ใกล้ที่สุด และปรับค่า K ใหม่ให้เป็นค่ากลางของกลุ่มข้อมูล และหยุดเมื่อค่า K ไม่เปลี่ยนแปลง [6] ดังนั้นการใช้ K-means clustering จึงสามารถนำมาใช้งานเมื่อทราบจำนวนกลุ่มที่ต้องการจำแนกที่แน่นอน ทั้งนี้การวัดระยะห่างระหว่างข้อมูลสามารถใช้วิธีการคำนวณระยะห่างได้หลากหลาย เช่น Euclidean, Person Correlation, Superman Rank Correlation เป็นต้น

2.5 Hierarchical Clustering

การจัดกลุ่มแบบลำดับชั้น เป็นเทคนิควิธีการที่จัดกลุ่มตามความคล้ายกันของข้อมูล ด้วยวิธีการวัดความคล้ายหรือความต่างเช่น Euclidean, Cityblock, Mahalanobis, Cosine เป็นต้น [7] รูปแบบการแสดงผลของ Hierarchical Clustering จะถูกแสดงในรูปของต้นไม้ โดยในแต่ละ class node จะประกอบไปด้วย child nodes เทคนิคนี้สามารถแบ่งวิธีการสร้างต้นไม้ได้เป็น 2 ประเภทคือ Agglomerative (Bottom-Up) และ Divisive (Top-Down) [8]

2.6 การเลือกคุณลักษณะ (Feature Selection)

การเลือกคุณลักษณะของชุดข้อมูลที่มีจำนวนมิติสูงเป็นขั้นตอนหนึ่งของกระบวนการสร้างแบบจำลองเพื่อการพยากรณ์ด้วยวิธีการเหมืองข้อมูล สำหรับการเลือกข้อมูลย่อยที่มีมิติน้อยลง (Data Reduction) [9] กว่าข้อมูลต้นฉบับ (Original data) กระบวนการคัดเลือกคุณลักษณะถือเป็นงานสำคัญในการปรับปรุงประสิทธิภาพในการสร้างแบบจำลอง อีกทั้งกระบวนการ Feature Selection ยังเป็นการช่วยในการเพิ่มความถูกต้องในการพยากรณ์ (Improving Prediction accuracy) [10] เนื่องจากจุดประสงค์สำคัญของการทำ Feature Selection เพื่อลดจำนวนมิติของข้อให้เหลือเพียงชุดข้อมูล (Feature Subset) ที่มีส่งผลต่อความถูกต้องในการพยากรณ์มากที่สุด

2.7 Dynamic Feature Selection

อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection) เป็นอัลกอริทึมที่ถูกออกแบบมาเพื่อใช้ในการหาตัวแปรที่ดีที่สุดสำหรับการจำแนกกลุ่ม ทั้งนี้ผู้วิจัยได้ออกแบบ วิธีการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection) ซึ่งเป็นวิธีการในการหาลักษณะหรือตัวแปรที่ส่งผลดีที่สุดต่อ อัลกอริทึมการจัดกลุ่มข้อมูล (Clustering Algorithm)

2.8 การวัดประสิทธิภาพแบบจำลอง

วิธีการวัดประสิทธิภาพการพยากรณ์กับชุดข้อมูลในการวิจัยครั้งนี้ใช้วัดประสิทธิภาพการพยากรณ์ข้อมูลตามแนวคิดด้านการค้นคืนสารสนเทศ (Information Retrieval) ด้วยค่าความถูกต้อง (Accuracy) ซึ่งเป็นการคำนวณจากตาราง Confusion Matrix [1] [3]

2.9 ผลสัมฤทธิ์ทางการเรียน

ผลสัมฤทธิ์ทางการเรียนเป็นสิ่งที่สำคัญสำหรับการพิจารณาผลของการจัดการเรียนการสอนในรูปแบบต่างๆ โดยเน้นที่การวัดความรู้ ความเข้าใจ และความสามารถของผู้เรียนที่เกิดจากการเรียนการสอน ด้วยวิธีการทดสอบเพื่อให้ได้ข้อมูล ซึ่งใช้ในการเปรียบเทียบกับเกณฑ์มาตรฐานที่ได้กำหนดไว้ ทำให้สามารถแยกผู้เรียนออกเป็นกลุ่มๆ ตามระดับความรู้ได้ ในการวิจัยครั้งนี้ผู้วิจัยทำการศึกษาปัจจัยเชิงพุทธิพิสัย (Cognitive Domain) ของกลุ่มผู้เรียนแต่ละกลุ่มแยกตามระดับความรู้ตามมาตรฐานเกณฑ์คะแนน 8 ระดับ คือ A (Excellent), B+ (Very Good), B (Good), C (Average) จนถึงระดับ F ซึ่งถือเป็นคลาสของชุดข้อมูลแบบหลายค่า (Multi-Class) และนำข้อมูลผลการทดสอบตลอดภาคการศึกษามาสร้างแบบจำลองสำหรับพยากรณ์ผลสัมฤทธิ์ของผู้เรียน

3. วิธีดำเนินการวิจัยและผลการทดลอง

การวิจัยครั้งนี้มุ่งเน้นการวัดประสิทธิภาพที่เกิดขึ้นจาก

อัลกอริทึมในการจำแนกข้อมูลที่พัฒนาขึ้นโดยใช้หลักการทำงานของเทคนิควิธีเหมืองข้อมูลแบบไม่มีผู้สอน (Unsupervised Learning Algorithm) ได้แก่ K-Means Clustering, Hierarchical Clustering เพื่อหาค่าประสิทธิภาพในการจำแนกกลุ่มข้อมูล และนำค่าดังกล่าวมาเปรียบเทียบเพื่อเลือกวิธีการจัดกลุ่มที่เหมาะสมที่สุดสำหรับอัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms)

3.1 การรวบรวมข้อมูลที่ใช้ในการสร้างแบบจำลอง

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

ข้อมูลที่ใช้สำหรับวิเคราะห์ผู้วิจัยแบ่งข้อมูลออกเป็น 4 ส่วนดังนี้คือ

1. ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ ความสามารถด้านพุทธิพิสัย (Cognitive Domain) ซึ่งส่งผลต่อผลสัมฤทธิ์ทางการเรียน ผู้วิจัยใช้แบบทดสอบวัดความรู้ ความเข้าใจที่ออกแบบให้ข้อสอบมีความสอดคล้องกับวัตถุประสงค์และเป้าหมายของรายวิชา แบ่งการทดสอบออกเป็น 3 ประเภท ได้แก่ 1) แบบทดสอบก่อนเรียน (Pre-test) เป็นแบบทดสอบที่ใช้ประเมินผู้เรียนก่อนดำเนินการเรียนการสอน เพื่อสำรวจความพร้อมของผู้เรียน และวัดความรู้พื้นฐานเดิมของผู้เรียน 2) แบบทดสอบประจำบทเรียน (Formative test) เป็นการทดสอบตามวัตถุประสงค์การเรียนรู้ของแต่ละหน่วยการเรียน เพื่อสำรวจความรู้ ความเข้าใจ ที่ผู้เรียนได้จากการเรียนผ่านระบบการเรียนแบบผสมผสาน 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

4. ข้อมูลทั่วไปของผู้เรียน ชุดข้อมูลในส่วนสุดท้ายได้แก่ ข้อมูลทั่วไปของผู้เรียน เช่น ข้อมูลภาคการศึกษา ชั้นปีที่เรียน สาขา เป็นต้น

จากชุดข้อมูลในตารางที่ 1 ข้อมูลทั้งหมดจะนำไปทดสอบในขั้นตอนการวิเคราะห์หาปัจจัยเพื่อการจัดกลุ่ม เพื่อทำการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม ทั้งนี้การวิจัยมีการวัดประสิทธิภาพในการจำแนกกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวม และค่าความถูกต้องในการจำแนกแยกตามกลุ่มผู้เรียนตามเกรด

3.2. การจัดเตรียมข้อมูลเพื่อการสร้างแบบจำลอง

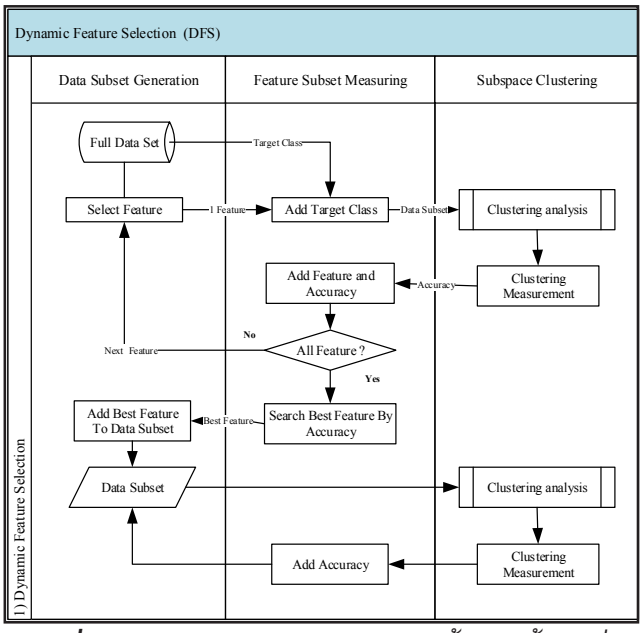
ขั้นตอนการจัดเตรียมข้อมูลเพื่อนำเข้าสู่กระบวนการสร้างแบบจำลอง ผู้วิจัยทำการออกแบบอัลกอริทึมในการสร้างชุดข้อมูล (Data Set) ที่เหมาะสมกับคลาสคำตอบ (Target Class) ทั้งนี้ผู้วิจัยได้ออกแบบ อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) มีวัตถุประสงค์ในการหาลักษณะหรือตัวแปรที่ส่งผลดีที่สุดต่ออัลกอริทึมการจัดกลุ่มข้อมูล (Clustering Algorithm) วิธีการดังกล่าวสามารถแสดงขั้นตอนการทำงานดังภาพที่ 1

จากภาพที่ 1 แสดงกระบวนการในการสร้างชุดข้อมูลย่อย (Data Subset) ซึ่งจะประกอบไปด้วย Feature ต่างๆ ตามโครงสร้างตารางที่ 1 ในขั้นตอนแรกกระบวนการทำงานจะเลือกตัวแปรตามลำดับครั้งละ 1 ตัว (1 Feature) และนำตัวแปรดังกล่าวเพิ่มชุดคำตอบ (Target Class) เพื่อส่งต่อไปยังกระบวนการเรียนรู้การจัดกลุ่มข้อมูล (Clustering Algorithm) เมื่อทำการจำแนกกลุ่มข้อมูลเสร็จสิ้น ตัวแปรดังกล่าวจะได้การประเมินค่าความถูกต้องของการจำแนกกลุ่ม (Accuracy) เป็นดัชนีชี้วัดความถูกต้อง ซึ่งคำนวณจากการนำกลุ่มที่จำแนกได้จากอัลกอริทึมมาเปรียบเทียบกับค่าคำตอบหรือ

Target Class ที่ได้จัดเตรียมไว้ จากนั้นทำการวนซ้ำกระบวนการข้างต้น จนครบทุกตัวแปร และเมื่อครบทุกตัวแปรจะทำการค้นหาตัวแปรที่ให้ค่าความถูกต้องมากที่สุด (Best Feature) เพื่อนำมาสร้างเป็นชุดข้อมูลย่อย (Data Subset)

ตารางที่ 1 *รายละเอียดปัจจัยที่ส่งผลต่อผลสัมฤทธิ์*

No.	ชื่อแอทริบิวต์	ค่าตัวแปร	รายละเอียด
1	Major	อักษร	สาขา
2	Level	อักษร	ระดับชั้นปี
3	Group.Lec	อักษร	กลุ่มบรรยาย
4	Group.Lab	อักษร	กลุ่มปฏิบัติ
5	Semester	อักษร	ภาคการศึกษา
6	Pre-test	เปอร์เซ็นต์	ทดสอบก่อนเรียน
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	ค่าเฉลี่ย 1-12
17	Avg.1-8Test	เปอร์เซ็นต์	ค่าเฉลี่ย 11-2
18	NumOfPass	จำนวน	ค่า 1-12 >=60%
19	Score.Lab	จำนวน	คะแนนเต็ม 25
20	Score.eLearnig	จำนวน	คะแนนเต็ม 15
21	Score.Mit	จำนวน	คะแนนเต็ม 30
22	Score.Final	จำนวน	คะแนนเต็ม 30
23	Grade	ระดับ	8 ระดับ (A,B,…F)



ภาพที่ 1 *แสดงกระบวนการในการสร้างชุดข้อมูลย่อย (Data Subset)*

ตารางที่ 2 *ข้อมูลจำนวนผู้เรียนและคะแนนเฉลี่ยแบ่งตามเกรด*

เกรด	จำนวนผู้เรียน	ร้อยละ	ค่าคะแนนเฉลี่ย
A	69	4%	88.7
B+	257	16%	83.4
B	374	23%	78.0
C+	357	22%	71.6
C	255	16%	64.8
D+	124	8%	58.3
D	75	5%	52.0
F	94	6%	30.2
All	1,605		

ขั้นตอนสุดท้ายของกระบวนการ ผู้วิจัยจะนำชุดข้อมูลย่อยที่ได้จากขั้นตอนข้างต้นไปเข้าสู่ อัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms) สำหรับหาจำนวนตัวแปรที่ส่งผลให้ค่าความถูกต้องของการจำแนกมากที่สุด เพื่อใช้สำหรับเลือกจำนวนตัวแปรในชุดข้อมูลย่อย ที่ส่งผลทำให้ประสิทธิภาพในการจำแนกกลุ่มข้อมูลมากที่สุด

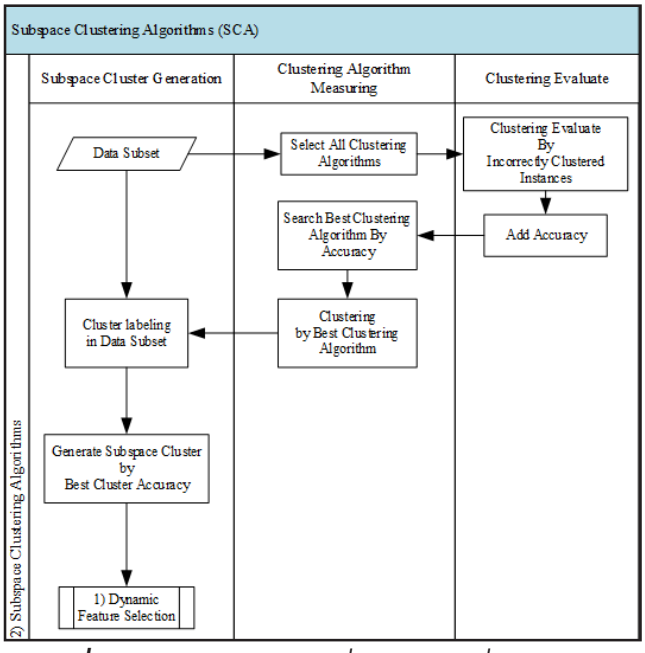
การจัดเตรียมข้อมูลเพื่อใช้ในการวิเคราะห์อัลกอริทึม ผู้วิจัยเริ่มต้นโดยการปรับค่าเกณฑ์การตัดเกรดให้อยู่ใน

มาตรฐานเดียวกัน ทั้งนี้เนื่องจากในแต่ละภาคการศึกษามีการตัดเกรดแบบอิงกลุ่มทำให้ค่าเกณฑ์การตัดเกรดมีค่าที่แตกต่างกัน ผลจากการปรับค่าเกณฑ์การตัดเกรด ทำให้จำนวนผู้เรียนในแต่ละเกรดรายละเอียดดังตารางที่ 2

จากชุดข้อมูลในตารางที่ 2 ข้อมูลทั้งหมดจะนำไปทดสอบในขั้นตอนการวิเคราะห์หาปัจจัยเพื่อการจัดกลุ่ม เพื่อทำการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่ม ทั้งนี้การวิจัยมีการวัดประสิทธิภาพในการจำแนกกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวม และค่าความถูกต้องในการจำแนกแยกตามกลุ่มผู้เรียนตามเกรด

3.3 การสร้างแบบจำลองเพื่อใช้ในการจัดกลุ่มข้อมูล

ขั้นตอนการสร้างแบบจำลองผู้วิจัยพัฒนาอัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms: SCA) ถือได้ว่าเป็นส่วนขยายของการจัดกลุ่มแบบดั้งเดิม (Traditional Clustering) โดยมีวัตถุประสงค์หลักเพื่อการค้นหากลุ่มข้อมูล (Clusters) บนปริภูมิย่อยในชุดข้อมูล อัลกอริทึมดังกล่าวสามารถแสดงได้ดังภาพที่ 2



ภาพที่ 2 *อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms)*

จากภาพที่ 2 แสดงกระบวนการในการสร้างชุดข้อมูลย่อย (Data Subset) ซึ่งจะประกอบไปด้วย Feature ต่างๆ ตามโครงสร้างตารางที่ 1 ในขั้นตอนแรกกระบวนการทำงานจะเลือกตัวแปรตามลำดับครั้งละ 1 ตัว (1 Feature) และนำตัวแปรดังกล่าวเพิ่มชุดคำตอบ (Target Class) เพื่อส่งต่อไปยังกระบวนการเรียนรู้การจัดกลุ่มข้อมูล (Clustering Algorithm) เมื่อทำการจำแนกกลุ่มข้อมูลเสร็จสิ้น ตัวแปรดังกล่าวจะได้การประเมินค่าความถูกต้องของการจำแนกกลุ่ม (Accuracy) เป็นดัชนีชี้วัดความถูกต้อง ซึ่งคำนวณจากการนำกลุ่มที่จำแนกได้จากอัลกอริทึมมาเปรียบเทียบกับค่าคำตอบหรือ Target Class ที่ได้จัดเตรียมไว้ จากนั้นทำการวนซ้ำกระบวนการข้างต้น จนครบทุกตัวแปร และเมื่อครบทุกตัวแปรจะทำการค้นหาตัวแปรที่ให้ค่าความถูกต้องมากที่สุด (Best Feature) เพื่อนำมาสร้างเป็นชุดข้อมูลย่อย (Data Subset) ในขั้นตอนสุดท้ายของกระบวนการ ผู้วิจัยจะนำชุดข้อมูลย่อยที่ได้จากขั้นตอนข้างต้นไปเข้าสู่ อัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms) สำหรับหาจำนวนตัวแปรที่ส่งผลให้ค่าความถูกต้องของการจำแนกมากที่สุด เพื่อใช้สำหรับเลือกจำนวนตัวแปรในชุดข้อมูลย่อยที่ส่งผลทำให้ประสิทธิภาพในการจำแนกกลุ่มข้อมูลมากที่สุด

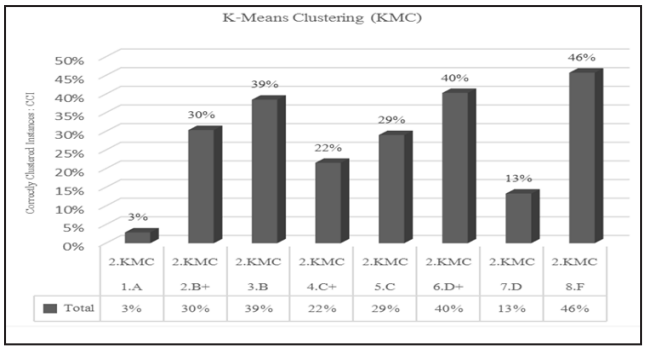
3.4. การวัดค่าประสิทธิภาพในการจัดกลุ่ม

กระบวนการวัดค่าประสิทธิภาพในงานวิจัยนี้จะแบ่งการทดสอบเพื่อหาประสิทธิภาพในการจัดกลุ่มออกเป็น 3 ขั้นตอนคือ 1) การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมแบบดั้งเดิม 2) การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) 3) การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms: SCA)

การวิจัยครั้งนี้ใช้วิธีการวัดประสิทธิภาพของการทดสอบด้วยค่าความถูกต้องในการจำแนกกลุ่ม (Correctly Clustered Instances: CCI) โดยการเปรียบเทียบกับ Target Class จำนวน 8 Class (A, B+, .., F) ของชุดข้อมูลสำหรับเรียนรู้

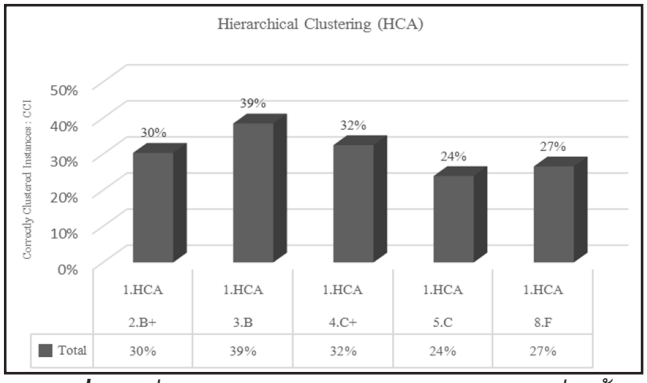
3.4.1 การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมแบบดั้งเดิม

การวิจัยนี้เลือกอัลกอริทึมในการจำแนกข้อมูลมาใช้ในการทดสอบเพื่อนำค่าประสิทธิภาพไปเปรียบเทียบ 2 อัลกอริทึม ได้แก่ K-Means Clustering (KMC), Hierarchical Clustering Algorithms (HCA) และเมื่อนำอัลกอริทึมทั้ง 2 ไปใช้ในการจัดกลุ่มข้อมูลสามารถแสดงประสิทธิภาพได้ดังภาพที่ 3



ภาพที่ 3 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย K-Means Clustering (KMC)*

จากภาพที่ 3 K-Means Clustering (KMC) สามารถจำแนกกลุ่มได้ 8 กลุ่มข้อมูล โดยเกรด F มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 46% เกรด D ร้อยละ 40 % ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่ระดับร้อยละ 30%

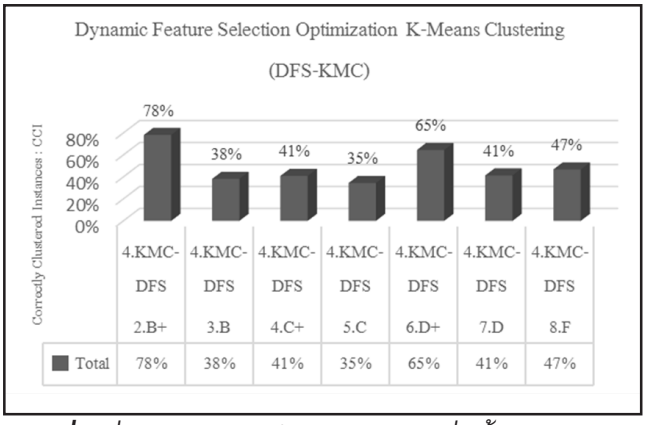


ภาพที่ 4 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย Hierarchical Clustering Algorithms*

จากภาพที่ 4 Hierarchical Clustering Algorithms สามารถจำแนกกลุ่มได้ 5 กลุ่มข้อมูล โดยเกรด B มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 39% เกรด C+ ร้อยละ 32 % ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่ระดับร้อยละ 26%

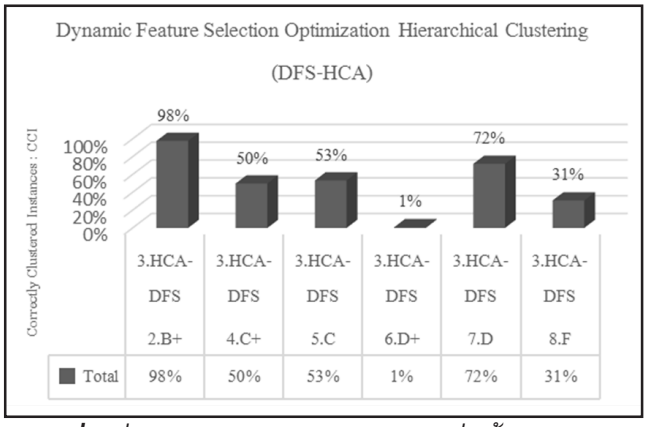
3.4.2 การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต

การทดสอบประสิทธิภาพด้วยอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) ซึ่งเป็นอัลกอริทึมที่ได้ออกแบบมาสำหรับค้นหาตัวแปร (Feature) ที่ดีที่สุดต่อการจัดกลุ่มของชุดข้อมูล โดยประยุกต์ใช้อัลกอริทึม K-Means Clustering (KMC), Hierarchical Clustering Algorithms (HCA) ในขั้นตอนการทำงาน ผลการทดสอบประสิทธิภาพแสดงดังภาพที่ 5



ภาพที่ 5 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย DFS-KMC*

จากภาพที่ 5 Dynamic Feature Selection Optimization K-Means Clustering (DFS-KMC) สามารถจำแนกกลุ่มได้ 7 กลุ่มข้อมูล โดยเกรด B+ มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 78% เกรด D+ ร้อยละ 65% ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่ระดับร้อยละ 46%



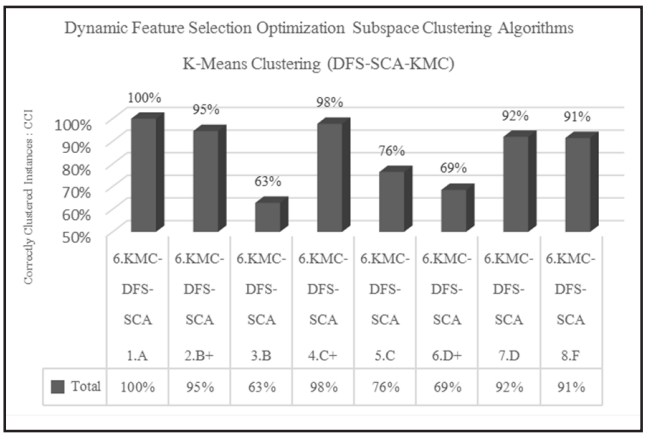
ภาพที่ 6 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย DFS-HCA*

จากภาพที่ 6 Dynamic Feature Selection Optimization Hierarchical Clustering Algorithms (DFS-HCA) สามารถจำแนกกลุ่มได้ 6 กลุ่มข้อมูล โดยเกรด B+ มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 98% เกรด D ร้อยละ 72% ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่

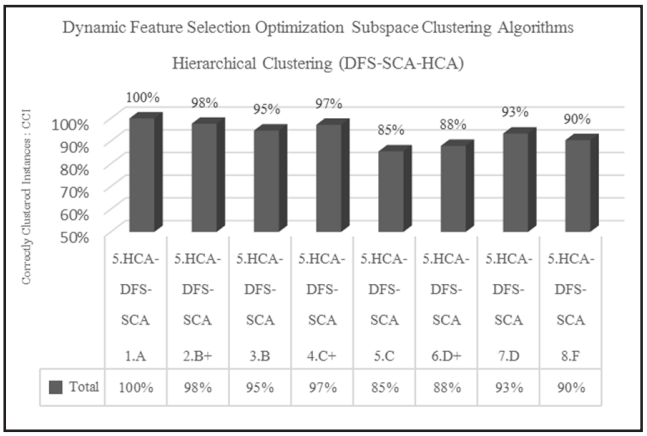
ระดับร้อยละ 40%

3.4.3 การทดสอบประสิทธิภาพในการจัดกลุ่มโดยใช้อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย

การทดสอบประสิทธิภาพด้วยอัลกอริทึมการจัดกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms: SCA) ซึ่งเป็นอัลกอริทึมที่ได้ออกแบบโดยการต่อขยายวิธีการจัดกลุ่มแบบเดิม โดยเพิ่มกระบวนการจัดลำดับในการสร้างกลุ่มข้อมูลภายใต้ชุดตัวแปร (Feature Set) ที่เหมาะสมที่สุดกับคลาสคำตอบ (Target Class) โดยอัลกอริทึมนี้จะทำงานประสานกับอัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) เพื่อเป็นการเพิ่มประสิทธิภาพในการจำแนกกลุ่มให้มากขึ้น ผลการทดสอบประสิทธิภาพแสดงดังภาพที่ 7



ภาพที่ 7 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย DFS-SCA-KMC*



ภาพที่ 8 *ค่าประสิทธิภาพในการจำแนกกลุ่มด้วย DFS-SCA-HCA*

จากภาพที่ 7 Dynamic Feature Selection Optimization Subspace Clustering Algorithms K-Means Clustering (DFS-SCA-KMC) สามารถจำแนกกลุ่มได้ 8 กลุ่มข้อมูล โดยเกรด A มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 100% เกรด C+ ร้อยละ 98% ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่ระดับร้อยละ 83%

จากภาพที่ 8 Dynamic Feature Selection Optimization Subspace Clustering Algorithms Hierarchical Clustering (DFS-SCA-HCA) สามารถจำแนกกลุ่มได้ 8 กลุ่มข้อมูล โดยเกรด A มีค่าความถูกต้องในการจำแนกกลุ่มมากที่สุดที่ระดับร้อยละ 100% เกรด B+ ร้อยละ 98% ตามลำดับ เมื่อพิจารณาประสิทธิภาพในการจัดกลุ่มด้วยค่าความถูกต้องในการจำแนกโดยรวมอยู่ที่ระดับร้อยละ 94%

4. สรุปผลการวิจัยและข้อเสนอแนะ

การวิจัยชิ้นนี้ เป็นการนำเสนอผลการทดสอบประสิทธิภาพของอัลกอริทึมใหม่ที่ใช้ในการสร้างแบบจำลองการพยากรณ์ ซึ่งประกอบไปด้วยอัลกอริทึม 2 อัลกอริทึม ดังนี้ อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS) และอัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อย (Subspace Clustering Algorithms)

งานวิจัยชิ้นนี้ผู้วิจัยเลือกเก็บข้อมูลจากวิชาที่จัดการเรียนการสอนขึ้นในระดับปริญญาตรี ซึ่งเป็นวิชาในกลุ่มศึกษาทั่วไป ทั้งนี้ผู้วิจัยจึงจัดทำระบบการเรียนออนไลน์ (e-Learning) และนำไปใช้ในการจัดการเรียนการสอนแบบผสมผสาน เพื่อเป็นเครื่องมือสำหรับรวบรวมข้อมูลที่เกิดขึ้นตลอด 6 ภาคการศึกษา (2/2553 ถึง 1/2556) มีจำนวนผู้เรียนทั้งสิ้น 1,605 คน โดยข้อมูลที่ใช้สำหรับวิเคราะห์ผู้วิจัยแบ่งข้อมูลออกเป็น 4 ส่วนดังนี้คือ 1) ข้อมูลที่ได้จากการทดสอบเพื่อวัดความรู้ความสามารถด้านพุทธิพิสัย (Cognitive Domain) 2) คะแนนการทดสอบประจำภาคการศึกษา 3) ค่าคะแนนการฝึกปฏิบัติ คิดเป็น 25% ของคะแนนในรายวิชา และ 4) ข้อมูลทั่วไปของผู้เรียน

ผลการวิจัยให้ผลการเปรียบเทียบประสิทธิภาพในด้านความถูกต้องของการจัดกลุ่ม (Correctly Clustered Instances: CCI) ที่สำคัญดังนี้ อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับ Hierarchical Clustering ให้ค่าความถูกต้อง

ในภาพรวมสูงที่สุดที่ระดับร้อยละ 94% คิดเป็นการจำแนกข้อมูลตามกลุ่มได้ถูกต้อง 1,503 ข้อมูล จาก 1,605 ข้อมูล โดยกลุ่มจำแนกได้ดีที่สุดคือ A มีค่าความถูกต้องที่ร้อยละ 100% ลองลงมาได้แก่ อัลกอริทึมการจัดกลุ่มบนปริภูมิย่อยที่ทำงานร่วมกับ K-Means Clustering ให้ค่าความถูกต้องในภาพรวมที่ระดับร้อยละ 83% คิดเป็นการจำแนกข้อมูลตามกลุ่มได้ถูกต้อง 1,331 ข้อมูล จาก 1,605 ข้อมูล โดยกลุ่มจำแนกได้ดีที่สุดคือ A มีค่าความถูกต้องที่ร้อยละ 100% ทั้งนี้ทั้ง 2 อัลกอริทึมผ่านกระบวนการสร้างชุดของลักษณะสำคัญ (Feature Set) ด้วย อัลกอริทึมการเลือกลักษณะสำคัญแบบพลวัต (Dynamic Feature Selection: DFS)

ผลที่ได้จากการวิจัยในครั้งนี้จะถูกนำไปพัฒนาต้นแบบกระบวนการวิเคราะห์ปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียนด้วยกระบวนการจัดกลุ่มข้อมูลโดยวิธีการเลือกลักษณะสำคัญแบบพลวัตเพื่อเพิ่มประสิทธิภาพของอัลกอริทึมการจำแนกกลุ่มบนปริภูมิย่อยต่อไป

5. เอกสารอ้างอิง

[1] I. H. Witten, E. Frank and M. A. Hall. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2011.

[2] C. F. Tsai, C. T. Tsai, C. S. Hung and P. S. Hwang. "Data Mining Techniques for Identifying Students at Risk of Failing a Computer Proficiency Test Required for Graduation." *Australasian Journal of Educational Technology,* Vol. 27, No. 3, pp. 481–498, 2011.

[3] J. Han, M. Kamber and J. Pei. *Data Mining: Concepts and Techniques: Concepts and Techniques.* Elsevier, 2011.

[4] L. Parsons, E. Haque and H. Liu. "Subspace clustering for high dimensional data: a review." *SIGKDD Explor Newsl,* Vol. 6, No. 1, pp. 90–105, June 2004.

[5] J. Macqueen. "Some methods for classification and analysis of multivariate observations." *In Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability,* Vol. 1, pp. 281–297, 1967.

[6] K. Wagstaff, C. Cardie, S. Rogers and S. Schrödl. "Constrained K-means Clustering with Background Knowledge." *In Proceedings of the Eighteenth International Conference on Machine Learning,* San Francisco, CA, USA, pp. 577–584, 2001.

[7] Y. Zhao, G. Karypis and U. Fayyad. "Hierarchical Clustering Algorithms for Document Datasets." *Data Mining Knowledge Discovery,* Vol. 10, No. 2, pp. 141–168, March 2005.

[8] J. H. Ward. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association,* Vol. 58, No. 301, pp. 236–244, March 1963.

[9] H. Yuan, S. S. Tseng, W. Gangshan and Z. Fuyan. "A two-phase feature selection method using both filter and wrapper." *In Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings,* Vol. 2, pp. 132–136, 1999.

[10] D. Koller and M. Sahami. "Toward Optimal Feature Selection." pp. 284–292, 1996.