อัลกอริทึมพันธุการสำหรับการจัดกลุ่มข้อมูลแบบ Partitioning

นายรติพงษ์ พูลผล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
ปีการศึกษา 2557

อัลกอริทึมพันธุการสำหรับการจัดกลุ่มข้อมูลแบบ Partitioning

นายรติพงษ์ พูลผล วศ.บ. (วิศวกรรมคอมพิวเตอร์)

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
ปีการศึกษา 2557

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการสอบวิทยานิพนธ์
(ผศ. ดร. เกรียงไกร ปอแก้ว)

..... กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์
(รศ. ดร.กิตติชัย ลวันยานนท์)

..... กรรมการ
(ผศ. ดร.ชาคริดา นุกูลกิจ)

..... กรรมการ
(ศ. ดร.บุญเสริม กิจศิริกุล)

หัวข้อวิทยานิพนธ์	อัลกอริทึมพันธุการสำหรับการจัดกลุ่มข้อมูลแบบ Partitioning
หน่วยกิต	12
ผู้เขียน	นายรติพงษ์ พูลผล
อาจารย์ที่ปรึกษา	รศ.ดร. กิตติชัย ลวันยานนท์
หลักสูตร	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
คณะ	เทคโนโลยีสารสนเทศ
ปีการศึกษา	2557

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอการพัฒนาโปรแกรมการจัดกลุ่มข้อมูล (Clustering) แบบ Partitioning ที่นำเทคนิค Divisive Analysis (Diana) และ Genetic Algorithm (GA) มาประยุกต์ใช้ โปรแกรมที่พัฒนาขึ้นมีความสามารถในการจัดกลุ่มข้อมูลเชิงจำนวน เชิงลักษณะและข้อมูลที่ประกอบด้วยเชิงจำนวนและเชิงลักษณะได้ หน่วยทำงานของ GA ในโปรแกรมที่พัฒนาขึ้นมาใหม่มีความสามารถในการปรับสภาพตัวเอง (Adaptive) ให้สามารถเปลี่ยนแนวทางการ Crossover และแนวทางการ Selection ในกรณีผลลัพธ์ไม่มีการพัฒนาสักระยะหนึ่งได้ ผลการประเมินประสิทธิภาพการจัดกลุ่มข้อมูลจำนวน 11 ชุดข้อมูลจาก Public Domain และข้อมูลนักศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พบว่าโปรแกรมที่พัฒนาขึ้นมีประสิทธิภาพการจัดกลุ่มข้อมูลดีกว่าเทคนิคที่เป็นที่รู้จักทั่วไป

คำสำคัญ : การจัดกลุ่มข้อมูล (Clustering) / ข้อมูลนักศึกษา / Divisive Analysis (Diana) / Genetic Algorithm (GA) / Partitioning Clustering

Thesis Title	A Genetic Algorithm Approach to Partitioning Clustering
Thesis Credits	12
Candidate	Mr. Ratipong Poolphol
Thesis Advisor	Assoc. Prof. Dr. Kittichai Lavangnananda
Program	Master of Science
Field of Study	Information Technology
Faculty	School of Information Technology
Academic Year	2014

Abstract

This thesis describes the implementation of a Partitioning Clustering program which utilizes Divisive Analysis (Diana) and Genetic Algorithm (GA). The program implemented is capable of clustering numerical, categorical as well as mixed data of both types. It is also adaptive in a sense that it can change the Crossover and Selection methods if no progress occurs within a given number of generations. Eleven datasets and students data from the M.Sc. (Information Technology) at School of Information Technology (SIT), King Mongut's University of Technology Thonburi (KMUTT) are used for evaluation. The program implemented is proven superior to commonly known clustering techniques.

Keywords : Clustering / Divisive Analysis (Diana) / Genetic Algorithm (GA) / Partitioning Clustering / Students Data

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลุล่วงไปได้ด้วยดี ผู้วิจัยขอขอบพระคุณ รศ.ดร. กิตติชัย ลวันยานนท์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาให้ความรู้ คำปรึกษา แนวคิด และข้อเสนอแนะอันเป็นประโยชน์ต่อการดำเนินงานโครงการวิจัยนี้ ขอขอบคุณกรรมการสอบวิทยานิพนธ์ทุกท่านที่ได้สละเวลาในการสอบวิทยานิพนธ์ ขอขอบพระคุณคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ที่สนับสนุนทุนการศึกษาและทรัพยากรคอมพิวเตอร์ที่ใช้ในการทดสอบโปรแกรมที่พัฒนา ขึ้นในโครงการวิจัยนี้

วิทยานิพนธ์สำเร็จลงได้ด้วยกำลังใจจากบุคคลรอบข้าง ผู้วิจัยขอขอบพระคุณ คุณพ่อและคุณแม่ที่คอยให้คำแนะนำและสนับสนุนด้านกำลังใจและการเงินตลอดมา ท้ายนี้ผู้วิจัยขอขอบพระคุณทุกท่านที่เป็นกำลังใจตลอดมา

สารบัญ

หน้า

รายการตาราง

ตาราง **หน้า**

รายการรูปประกอบ

บทที่ 1 บทนำ

การจัดกลุ่มข้อมูล (Clustering) [1] คือกระบวนการจัดข้อมูลโดยที่ให้ผลลัพธ์ข้อมูลภายในกลุ่มเดียวกันมีความคล้ายคลึงกันมากที่สุดและข้อมูลที่อยู่ต่างกลุ่มกันมีความแตกต่างกันมากที่สุด เทคนิคที่นิยมใช้ในการจัดกลุ่มข้อมูลมีอยู่หลายเทคนิคด้วยกัน เช่น Divisive Analysis (Diana), Partitioning Around Medoids (PAM), Self-Organizing Map (SOM) และ Evolutionary Algorithms

การเลือกใช้เทคนิคการจัดกลุ่มข้อมูลขึ้นอยู่กับธรรมชาติของข้อมูล การจัดกลุ่มข้อมูลถูกนำไปใช้ในหลาย ๆ ด้าน โดยเฉพาะอย่างยิ่งด้านการตลาดที่ความสำคัญของการจัดกลุ่มข้อมูลมีมากขึ้นในปัจจุบัน เนื่องจากการแข่งขันที่สูงขึ้นและปริมาณข้อมูลจำนวนมหาศาลที่ถูกเก็บในแต่ละวัน การค้นหารูปแบบการซื้อหรือความสนใจของลูกค้าเพื่อสร้างแคมเปญจน์การตลาดให้เหมาะสมกับลูกค้าย่อมสร้างความได้เปรียบทางการค้ากับองค์กรได้

โครงการวิจัยนี้นำเสนอเทคนิคการจัดกลุ่มข้อมูลใหม่ที่ประยุกต์เทคนิคของ Diana ร่วมกับ Genetic Algorithm (GA) โดยผลลัพธ์การจัดกลุ่มข้อมูลบนชุดข้อมูลที่ได้จาก Public Domain และข้อมูลนักศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี มีประสิทธิภาพมากกว่าเทคนิคที่นิยมใช้กันทั่วๆ ไป

1.1 วัตถุประสงค์

โครงการวิจัยนี้มีวัตถุประสงค์หลัก ดังนี้

1. เพื่อพัฒนาโปรแกรมการจัดกลุ่มข้อมูลเชิงจำนวนและเชิงลักษณะโดยใช้ GA มาประยุกต์ร่วมกันกับ Diana ในการจัดกลุ่มข้อมูล
2. เปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลของโปรแกรมที่พัฒนาขึ้นกับเทคนิคที่เป็นที่รู้จักทั่วไปในการจัดกลุ่มข้อมูล
3. เพื่อจัดกลุ่มและค้นหาคุณลักษณะของผู้สมัครและนักศึกษาของนักศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

1.2 ขอบเขตของโครงการวิจัย

โครงการวิจัยนี้มีขอบเขตหลัก 2 ประการ ดังนี้

1. โครงการวิจัยนี้อยู่บนสมมติฐานว่าเวลาที่ใช้ในการประมวลผลเพื่อจัดกลุ่มข้อมูลไม่ใช่เป็นปัจจัยหลักสำคัญ
2. ข้อมูลใช้ในโครงงานวิจัยนี้คือข้อมูลเชิงจำนวนและเชิงลักษณะจำนวน 11 ชุดที่ได้จาก Public Domain และข้อมูลนักศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรีในช่วงปี พ.ศ. 2538 - 2556

1.3 โครงร่างของวิทยานิพนธ์

โครงร่างของวิทยานิพนธ์นี้ประกอบด้วย 5 บท ดังพอกล่าวคร่าวๆ ได้ ดังนี้

บทที่ 1 บทนำ : บทนี้กล่าวถึงวัตถุประสงค์ของโครงการวิจัย ขอบเขตของโครงการวิจัย และโครงร่างของวิทยานิพนธ์

บทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง : บทนี้กล่าวถึงทฤษฎีที่เกี่ยวข้องกับโครงการวิจัยนี้ โดยประกอบด้วยการจัดกลุ่มข้อมูล (Clustering) และอัลกอริทึมพันธุการ (Genetic Algorithm: GA) นอกจากนี้ยังกล่าวถึงงานวิจัยสำคัญๆ ของงานด้านการจัดกลุ่มข้อมูลที่ใช้เทคนิคของ Evolutionary Algorithms ด้วย

บทที่ 3 Adaptive Genetic Algorithm Approach to Clustering for Numerical and Categorical data : บทนี้กล่าวถึงแนวคิดการทำงานของ Divisive Analysis (Diana) ซึ่งถูกนำมาประยุกต์ใช้ในโครงการวิจัยนี้, Similarity Measure ที่ใช้บอกความเหมือนหรือความแตกต่างระหว่างข้อมูลแต่ละตัว, องค์ประกอบของ Genetic Algorithm และการปรับตัวเพื่อพัฒนา (Adaptability) ของ GA ที่พัฒนาขึ้นในโครงการวิจัยนี้

บทที่ 4 ข้อมูลที่ใช้ในโครงงานวิจัยและการประเมินประสิทธิภาพการจัดกลุ่มข้อมูล : บทนี้กล่าวถึงข้อมูลทั้งหมดที่ใช้เพื่อประเมินประสิทธิภาพของโปรแกรมที่พัฒนาขึ้น เกณฑ์การประเมินประสิทธิภาพและผลการประเมินประสิทธิภาพการจัดกลุ่มข้อมูลของโปรแกรมที่พัฒนาขึ้นเปรียบเทียบกับผลที่ได้จากเทคนิคการจัดกลุ่มข้อมูลอื่นๆ

บทที่ 5 บทสรุป : บทนี้กล่าวถึงสิ่งที่ได้รับจากโครงการวิจัย สิ่งที่ได้เรียนรู้จากโครงการวิจัย แนวทางการพัฒนาในอนาคตและสรุปผลโครงการวิจัย

บทที่ 2 ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

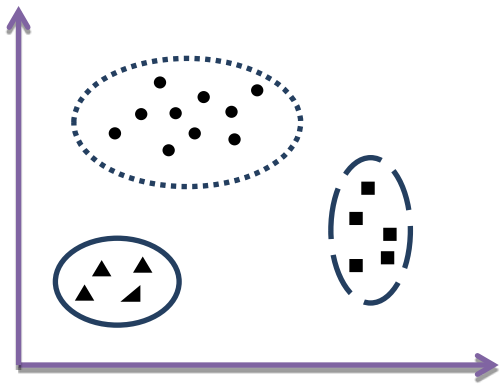
2.1 การจัดกลุ่มข้อมูล (Clustering)

การจัดกลุ่มข้อมูล (Clustering) นับได้ว่าเป็นเทคนิคหนึ่งที่นักวิทยาศาสตร์สาขาสถิติเป็นผู้คิดค้นขึ้น ในปัจจุบันการจัดกลุ่มข้อมูลกลายเป็นเทคนิคในแขนงวิชาที่ชื่อการทำเหมืองข้อมูล (Data mining หรือ Knowledge Discovery) เทคนิคนี้ได้ถูกนิยามไว้หลากหลาย โดยเฉพาะในเชิงคณิตศาสตร์ อย่างไรก็ดี นิยามที่เข้าใจง่ายที่สุดของเทคนิคนี้คือ กระบวนการจัดข้อมูลโดยที่ให้ผลลัพธ์ข้อมูลภายในกลุ่มเดียวกันมีความคล้ายคลึงกันมากที่สุดและข้อมูลที่อยู่ต่างกลุ่มกันมีความแตกต่างกันมากที่สุด [2],[3] โดยผลลัพธ์ที่ได้ในแต่ละกลุ่มน่าจะสามารถค้นหารูปแบบที่ซ่อนอยู่ในชุดข้อมูลนั้นๆ เทคนิคนี้ถูกนำไปใช้ในด้านต่างๆ ตั้งแต่วิจัยตลาด เช่น การจัดกลุ่มลูกค้าของบริษัทผู้ผลิตรถยนต์สปอร์ตเพื่อสร้างแคมเปนจ์การตลาดให้เหมาะสมกับกลุ่มลูกค้ามากขึ้น [4] จนกระทั่งถึงด้านโบราณคดี เช่น การแบ่งกลุ่มของขวานที่พบใน British-Isles ตามขนาดความยาว ความกว้างและความคมของขวาน [5] เทคนิคที่ใช้ในการจัดกลุ่มข้อมูลสามารถแบ่งคร่าวๆ ได้ 3 ลักษณะ ดังนี้

2.1.1 การจัดกลุ่มข้อมูลแบบพาร์ทิชันนอล (Partitional Clustering)

การจัดกลุ่มข้อมูลแบบพาร์ทิชันนอลจะแบ่งข้อมูลออกเป็นกลุ่มอย่างชัดเจนโดยที่ไม่มีข้อมูลใดเป็นสมาชิกมากกว่า 1 กลุ่ม สามารถอธิบายได้ดังนี้

ถ้าให้ $X = \{x_1, x_2, x_3, \ldots, x_N\}$ เป็นเซ็ตของข้อมูลจำนวน N ที่ถูกจัดกลุ่มให้เป็นสมาชิกของ $C = \{c_1, c_2, c_3, \ldots, c_k\}$ แล้ว $C_1 \cup C_2 \cup C_3 \cup \ldots C_k = X$ และทุก $C_i \cap C_j = \emptyset$ เมื่อ $i \neq j$ รูปที่ 2.1 แสดงถึงตัวอย่างของการจัดกลุ่มข้อมูลลักษณะนี้



รูปที่ 2.1 ตัวอย่างการจัดกลุ่มข้อมูลแบบพาร์ทิชันนอล

ตัวอย่างเทคนิคที่ใช้สำหรับการจัดกลุ่มข้อมูลลักษณะนี้ได้แก่ K-Means K-Medoids และ Self-Organizing Map

2.1.2 การจัดกลุ่มข้อมูลเชิงลำดับชั้น (Hierarchical Clustering)

ในการจัดกลุ่มข้อมูลแบบเชิงลำดับชั้นข้อมูลจะไม่ได้ถูกแบ่งออกเป็นพาร์ทิชันหรือเป็นกลุ่มในครั้งเดียว แต่ข้อมูลจะถูกจัดแบ่งเป็นกลุ่มย่อยที่ละขั้นเป็นลำดับชั้น โดยที่เทคนิคที่ใช้ในการจัดกลุ่มข้อมูลลักษณะนี้มีอยู่ 2 แบบด้วยกันคือ

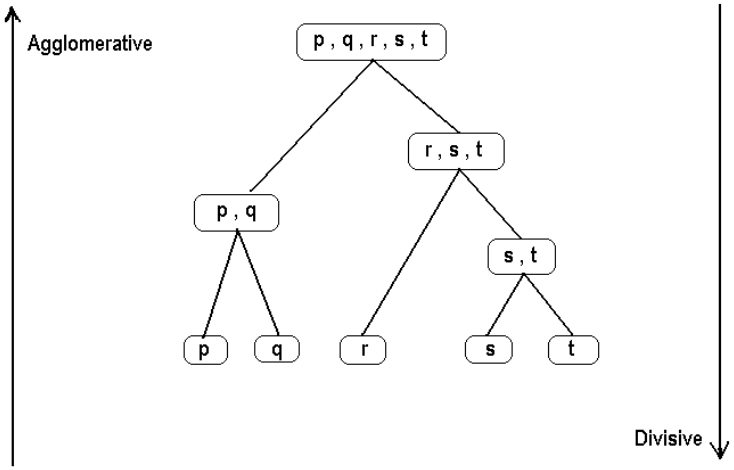
1. การจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบ Agglomerative

การจัดกลุ่มข้อมูลเชิงลำดับชั้นลักษณะนี้จะได้ชุดของกลุ่มข้อมูลย่อยๆ เป็นลำดับชั้น โดยเริ่มจากกลุ่มข้อมูลที่แต่ละกลุ่มประกอบด้วยข้อมูลเพียงตัวเดียว จากนั้นข้อมูลกลุ่มย่อยที่มีความคล้ายคลึงกันมากที่สุดจะถูกจับกลุ่มรวมกันไปทีละขั้นจนกระทั่งข้อมูลกลุ่มย่อยทั้งหมดรวมกันเหลือเพียงกลุ่มเดียว ดังแสดงได้ในรูปที่ 2.2

2. การจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบ Divisive

ขั้นตอนการจัดกลุ่มข้อมูลในลักษณะนี้กระทำตรงกันข้ามกับการจัดกลุ่มแบบ Agglomerative กล่าวคือ การจัดกลุ่มเริ่มจากชุดข้อมูลกลุ่มใหญ่ที่ประกอบด้วยข้อมูลทุกตัวเพียงกลุ่มเดียว จากนั้นข้อมูลจะถูกแยกออกเป็นกลุ่มย่อยไปทีละขั้น โดยข้อมูลที่มีความคลายคลึงกับข้อมูลอื่นในกลุ่มน้อยที่สุดจะถูกแยกออกไปก่อน ทำลักษณะนี้ไปเรื่อยๆ จนกระทั่งในแต่ละกลุ่มเหลือข้อมูลเพียงตัวเดียว ดังแสดงได้ในรูปที่ 2.2 เช่นกัน

รูปที่ 2.2 แสดงถึงการจัดกลุ่มข้อมูลเชิงลำดับชั้นแบบ Agglomerative และ Divisive

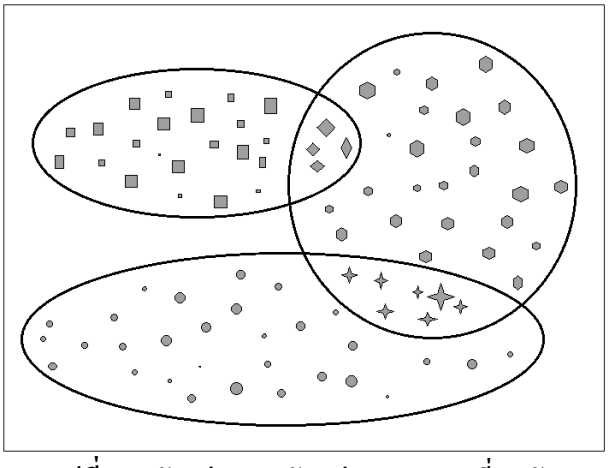


รูปที่ 2.2 ตัวอย่างการจัดกลุ่มข้อมูลเชิงลำดับชั้น

ตัวอย่างเทคนิคที่ใช้สำหรับการจัดกลุ่มข้อมูลลักษณะนี้ได้แก่ Divisive Analysis Clustering, Single Linkage Clustering, Complete Linkage Clustering และ Average Linkage Clustering

2.1.3 การจัดกลุ่มข้อมูลแบบคาบเกี่ยวกัน (Overlap Clustering)

การจัดกลุ่มข้อมูลแบบคาบเกี่ยวกันใช้ Fuzzy เซตในการจัดกลุ่มข้อมูล โดยที่ข้อมูลแต่ละตัวอาจเป็นสมาชิกได้มากกว่า 1 กลุ่มในระดับความเป็นสมาชิกที่แตกต่างกัน โดยถ้าให้ $X = \{x_1, x_2, x_3, \ldots, x_N\}$ เป็นเซ็ตของข้อมูลจำนวน N ที่ถูกจัดกลุ่มให้เป็นสมาชิกของ $C = \{c_1, c_2, c_3, \ldots, c_k\}$ แล้ว $C_1 \cup C_2 \cup C_3 \cup \ldots C_k = X$ และทุก $C_i \cap C_j = \emptyset$ เมื่อ $i \neq j$ ไม่เป็นจริงเสมอไป กลุ่มข้อมูลนั้นอาจเป็นกลุ่มข้อมูลที่คาบเกี่ยวกัน ข้อมูลที่ได้จากการจัดกลุ่มข้อมูลแบบคาบเกี่ยวกันแสดงได้ดังรูปที่ 2.3 ตัวอย่างการจัดกลุ่มข้อมูลลักษณะนี้ได้แก่ Fuzzy C-means Clustering



รูปที่ 2.3 ตัวอย่างการจัดกลุ่มแบบคาบเกี่ยวกัน

2.2 อัลกอริทึมพันธุการ (Genetic Algorithm: GA)

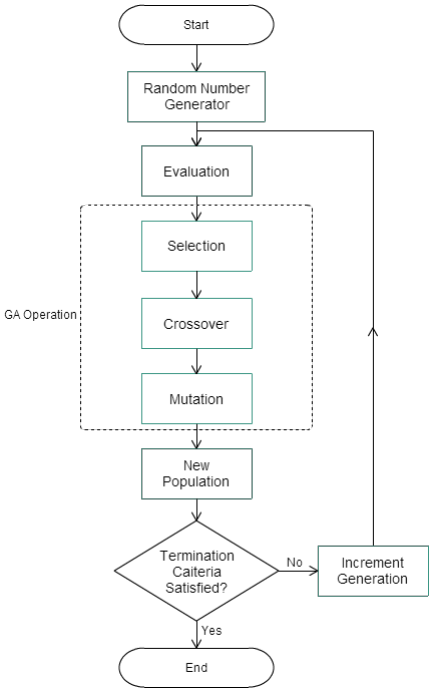
อัลกอริทึมพันธุการ หรือ Genetic Algorithm เป็นอัลกอริทึมที่จำลองการวิวัฒนาการของสิ่งมีชีวิตจากกฎของการคัดเลือกพันธุทางธรรมชาติ ถูกนำเสนอครั้งแรกในปี ค.ศ. 1975 โดยจอห์น ฮอนแลนด์ [6] GA ถูกนำไปใช้แก้ปัญหาในหลายๆ ด้าน โดยเฉพาะอย่างยิ่งปัญหาที่ยังไม่มีสมการหรืออัลกอริทึมในการแก้ปัญหาได้อย่างแน่นอน

GA อาจถูกมองเป็นเทคนิคหนึ่งในการค้นหา (Search) ของปัญญาประดิษฐ์ (Artificial Intelligence) หรือ AI และนับได้ว่าเป็นเทคนิคหนึ่งในแขนงย่อยของ AI ที่เรียกว่า Soft Computing หรือ

Computational Intelligence นอกจากนี้ GA ยังเป็นเทคนิคแรกของอัลกอริทึมประเภท Evolutionary Computation อีกด้วย

2.2.1 ขั้นตอนใน GA

ขั้นตอนการทำงานของ GA สามารถสรุปได้ดังรูปที่ 2.4



รูปที่ 2.4 องค์ประกอบและการทำงานของ GA

2.2.2 องค์ประกอบของอัลกอริทึมพันธุการ (GA Components)

จากรูปที่ 2.4 องค์ประกอบของ GA สามารถอธิบายพอสังเขปได้ ดังนี้

1. การสร้างประชากรเริ่มต้น (Initial Population)

ประชากร (Population) เริ่มต้นของระบบ GA คือเซตของ Individual โดยที่แต่ละ Individual คือ Solution ของปัญหา การแทนค่าของ Individual หรือ Chromosome ขึ้นอยู่กับลักษณะและธรรมชาติของงาน วิธีหนึ่งที่นิยมคือใช้ Bit String หรือค่า ‘0’ และ ‘1’ ตัวอย่าง เช่น ใช้ 8-bit String เพื่อแทนความฉลาด ดังนั้น Individual ที่ฉลาดที่สุด อาจสามารถแทนค่าได้เป็นโครโมโซม (Chromosome) ‘11111111’ และ Individual ที่ฉลาดน้อยที่สุดอาจสามารถแทนค่าได้เป็นโคโมโซม ‘00000000’ การใช้ค่า ‘0’ และ ‘1’ ในการแทนค่าโครโมโซมเป็นวิธีที่ธรรมดาและสะดวกที่สุด วิธีการใช้ค่าอื่นเช่น

ตัวอักษรและตัวเลขฐานสิบ ได้มีการถูกนำมาแทนค่าโครโมโซมได้เหมือนกัน ทั้งนี้ขึ้นอยู่กับลักษณะปัญหาที่ใช้ GA อยู่ในขณะนั้น โครโมโซมที่มีลักษณะอื่นนอกจาก '0' และ '1' หมายถึงขั้นตอนที่ซับซ้อนที่จะเกิดขึ้นจาก GA operation การแทนค่าโครโมโซมที่เป็นตัวเลขตัวอักษรก็มีให้เห็นกันอยู่ใน [7],[8]

2. การประเมินค่า (Evaluation)

การประเมินค่าเป็นกระบวนการสำคัญกระบวนการหนึ่งของ GA ในขั้นตอนนี้ โครโมโซม ทุกตัวจะถูกประเมินค่าคุณสมบัติโดยใช้สิ่งที่เรียกว่า Fitness function เพื่อระบุว่า โครโมโซม ตัวใดมีคุณสมบัติดีขนาดใดและสมควรแก่การเก็บไว้พัฒนาในรุ่นถัดไปมากแค่ไหน การออกแบบ Fitness function ที่เหมาะสมไม่มีหลักการที่แน่นอนขึ้นอยู่กับธรรมชาติงานที่กำลังแก้ปัญหา ไม่มี Fitness function ใดที่สามารถใช้ได้กับทุกงาน

3. การเลือกสรร (Selection)

การเลือกสรรคือการสำเนาโครโมโซมที่ดีที่สุดมาจำนวนหนึ่ง เพื่อเปิดโอกาสให้โครโมโซมที่มีคุณสมบัติที่ดี มีโอกาสส่งต่อคุณสมบัติของตัวมันเอง (Trait) ไปยัง Generation ต่อไป วิธีการหนึ่งที่นิยมใช้คือการสุ่ม (Random) แต่ทั้งนี้การเลือกสรรควรสอดคล้องกับคุณสมบัติของแต่ละโครโมโซมด้วย เช่น ถ้าโครโมโซมตัวที่ 1 มีคุณสมบัติที่ดีกว่าโครโมโซมตัวที 2 แล้ว โครโมโซมตัวที่ 1 ควรจะถูกเลือกมากกว่าโครโมโซมตัวที่ 2ดังนั้นจึงมีผู้นำเสนอวิธีการเลือกสรรแบบต่างๆ เช่น วิธีการเลือกแบบสุ่ม, วิธีการเลือกสรรแบบ Elitism, วิธีการเลือกสรรแบบ Roulette wheel [6] ฯลฯ

4. ครอสโอเวอร์ (Crossover)

ครอสโอเวอร์เป็นกระบวนการสำคัญอีกอันหนึ่งของ GA ซึ่งมีจุดประสงค์เพื่อสร้างประชากร ที่หลากหลายสำหรับ Generation ต่อไป วิธีการทั่วไปคือโครโมโซมจะถูกเลือกมาจำนวนหนึ่งจากประชากรที่มีอยู่ วิธีการเลือกโครโมโซมเพื่อการครอสโอเวอร์ไม่จำเป็นต้องมีความซับซ้อน ซึ่งการสุ่มเลือกเป็นวิธีการหนึ่งที่นิยมใช้ จากนั้นยีนต์ (Gene) ของโครโมโซมนี้จะถูกไขว้กัน เพื่อที่จะได้รุ่นลูก (Offspring) ที่เกิดมาใหม่และส่งไปเป็น ประชากร ใน Generation ต่อไป
เทคนิคการครอสโอเวอร์ที่นิยมใช้กันมี 2 วิธี ดังนี้

- 1-Point Crossover: วิธีการนี้จะแบ่งโครโมโซมออกเป็น 2 ส่วนด้วย 1 จุดครอสโอเวอร์ แล้วจึงจับแต่ละส่วนมาของแต่ละโครโมโซมมาไขว้กันเพื่อให้ได้ลูกออกมา

ตัวอย่าง เช่น ถ้า Parent1 = 100|101 และ Parent2 = 110|100 เมื่อทำการครอสโอเวอร์กันแล้ว ณ จุดที่ขั้นด้วยเครื่องหมาย | จะได้ลูกที่เกิดใหม่เป็น Offspring1 = 100|100 และ Offspring2 = 110|101

- 2-Point Crossover: ในวิธีการนี้โครโมโซมจะถูกแบ่งออกเป็น 3 ส่วนด้วย 2 จุดครอสโอเวอร์ แล้วจึงทำการไขว้เปลี่ยนยีนต์กัน

ตัวอย่างเช่น ถ้า Parent1 = 100|101|110 และ Parent2 = 110|100|111 โดย '|' แสดงถึงพื้นที่ที่การครอสโอเวอร์จะเกิดขึ้น เมื่อทำการครอสโอเวอร์แล้ว จะได้ลูกที่เกิดใหม่เป็น Offspring1 = 100|100|110 และ Offspring2 = 110|101|111

5. มิวเตชัน (Mutation)

มิวเตชันคือกระบวนการกลายพันธุ์ของโครโมโซม โดยมีจุดประสงค์เพื่อเพิ่มความหลากหลายของประชากรที่ระบบสร้างขึ้นในรุ่นต่อๆ ไป ขั้นตอนมิวเตชันเป็นเพียงการเปลี่ยนค่าในโครโมโซมของตำแหน่งที่ถูกเลือกขึ้นมา จำนวนตำแหน่งที่ถูกเลือกขึ้นมาขึ้นอยู่กับความยาวของโครโมโซมด้วย ในเบื้องต้นการมิวเตชั้นคือการเปลี่ยนค่า เช่น จาก '0' เป็น '1' และจาก '1' เป็น '0' ในกรณีที่ใช้ Binary Bit String แทนค่าโครโมโซม ตัวอย่าง ถ้า มี Parent = 1001**1**101 และตำแหน่งที่ 5 ถูกเลือกเพื่อการทำมิวเตชัน ดังนั้นโครโมโซมใหม่ที่ได้คือ Offspring = 1001**0**101

2.3 วรรณกรรมที่เกี่ยวข้อง

การจัดกลุ่มข้อมูลโดยใช้เทคนิคของ Evolutionary Algorithms ถูกนำมาใช้ในหลายๆ งานวิจัย วรรณกรรมของงานวิจัยเหล่านี้ได้ถูกรวบรวมและทำการวิเคราะห์ข้อดีและข้อด้อยของอัลกอริทึมในแต่ละวิธีไว้ใน [9] อีกทั้งยังได้สรุปและนำเสนอเทคนิคของแต่ละ Component ใน Evolutionary Algorithms ไว้ด้วย งานวิจัยที่เกิดขึ้นสามารถถูกแบ่งได้หลายๆ มิติ เช่น ลักษณะงานที่นำไปใช้ เทคนิคใน Evolutionary Algorithms ที่ใช้ ฯลฯ ในวิทยานิพนธ์นี้ขอกล่าวถึงในมิติการแทนค่าโครโมโซม (Representation) พอสังเขปดังนี้

2.3.1 Label Based Representation

การแทนค่าโครโมโซมแบบ Label Based ความยาวของโครโมโซมจะมีขนาดเท่ากับจำนวนของข้อมูลทั้งหมดใน Dataset โดยแต่ละตำแหน่งของโครโมโซมหมายถึงตำแหน่งของข้อมูลที่อยู่ใน Dataset ซึ่งค่าในแต่ละตำแหน่งนั้นๆ หมายถึง Label ของ Cluster เช่น ถ้าข้อมูลตัวที่ 1 – 4 ถูกจัดอยู่ใน Cluster1

ข้อมูลตัวที่ 5 – 8 ถูกจัดอยู่ใน Cluster2 และข้อมูลตัวที่ 9 – 10 ถูกจัดอยู่ใน Cluster3 โครโมโซมสามารถถูกแทนค่าได้เป็น [1111222233]
งานวิจัย [10] ใช้ GA เป็นเทคนิคในการจัดกลุ่มข้อมูล โดยข้อมูลที่ใช้ทดสอบประสิทธิภาพในงานวิจัยนี้เป็นข้อมูลที่ทราบกลุ่มของข้อมูลอยู่แล้ว ข้อมูลสถานะทางการเงินขององค์กรต่างๆ 46 องค์กรถูกนำมาใช้ ซึ่งในจำนวนนี้ 21 องค์กรมีสถานะล้มละลายและอีก 25 องค์กรมีสถานะทางการเงินปกติ ผลการทดลองที่ได้เปรียบเทียบกับ Traditional Clustering (Average Linkage, Single Linkage และ Complete Linkage) พบว่าให้ Error และ Hit Rate ที่ดีกว่า

GA มักถูกใช้เป็นเทคนิคในการหา Optimal Solution ของปัญหา รวมถึงค่า Optimal ของ Cluster ด้วย โดยใน [11] ใช้ Sum of Squared Euclidean เป็น Fitness function ของ GA และทำการทดลองกับข้อมูล Crude-Oil ซึ่งให้ผลลัพธ์ที่ดีกว่าวิธีการของ K-Means

นอกจากนี้ GA ยังถูกนำมารวมกับ K-Means เพื่อเพิ่มประสิทธิภาพและลดเวลาในการดำเนินการของ GA [12] และ [13] แทนที่ Crossover Operation ของ GA ด้วย K-Means ข้อมูล German Town Data (GTD) และ British Town Data (BTD) ถูกนำมาใช้เปรียบเทียบประสิทธิภาพระหว่าง GA Hybrid ที่ถูกพัฒนานี้กับ Evolutionary Strategies (ES), Evolutionary Programming (EP), GA และ Alternating Best-First (ABF) โดยเทคนิคที่พัฒนาขึ้นให้ค่า Accuracy ที่ดีกว่า

2.3.2 Centroid Based Representation

ในการแทนค่าโครโมโซมแบบ Centroid Based ข้อมูลแต่ละตำแหน่งของโครโมโซมแทนพิกัดของข้อมูลเช่น ในการจัดกลุ่มข้อมูลในระนาบ 2 มิติ โครโมโซม [1.5 2.5 10.5 1.5 5.0 5.5] ใช้แทนพิกัดของตัวแทนของ Cluster1 ที่จุด (1.5,2.5) Cluster2 ที่จุด (10.5,1.5) และ Cluster3 ที่จุด (5.0,5.5)
งานวิจัย [14] ใช้เทคนิคการเข้ารหัสโคโมโซมแบบ Centroid Based เพื่อใช้ในการจัดกลุ่มข้อมูลเชิงจำนวนที่มีจำนวน Attribute มาก ข้อมูลชนิดรูปภาพ ถูกนำมาใช้ในการประเมินประสิทธิภาพของโปรแกรมที่นำเสนอ งานวิจัยนี้ยังได้ใช้เทคนิคที่นำเสนอนี้ร่วมกับเทคนิคที่นิยมใช้กันทั่วไปอย่าง K-Means เพื่อ ปรับผลลัพธ์สุดท้ายที่ได้จาก GA เป็นการเพิ่มประสิทธิภาพโดยรวมของระบบด้วย แนวทางการปรับปรุงประสิทธิภาพของการจัดกลุ่มข้อมูลด้วยเทคนิคของ GA ยังถูกนำเสนอโดยงานวิจัย [15] โดยงานวิจัยนี้ได้นำเสนอเทคนิคที่เรียกว่า Self-Adaptive Genetic Algorithm for Clustering และนำเสนอผลการประเมินประสิทธิภาพบนข้อมูลรูปภาพเดียวกันกับ [14] ซึ่งได้ผลที่น่าพอใจมากกว่า
นอกจากข้อมูลชนิดรูปภาพแล้ว การจัดกลุ่มข้อมูลที่มีการเข้ารหัสโครโมโซมแบบ Centroid Based ยังถูกนำมาใช้ใน [16] และให้ประสิทธิภาพการจัดกลุ่มข้อมูลที่ดีกว่า จากการใช้ข้อมูล Vowel Data, Iris Data และ Crude Oil Data เมื่อเทียบกับ K-Means

2.3.3 Medoid Based Representation

ในการแทนค่าโครโมโซมแบบ Medoid Based นั้นข้อมูลแต่ละตำแหน่งของโครโมโซมแทนลำดับที่ของข้อมูลในชุดข้อมูลเพื่อใช้เป็นตัวแทน (Prototype) ของ Cluster เช่น [1 5 6 9] หมายถึงข้อมูลตัวที่ 1 5 6 และ 9 ของชุดข้อมูลจะถูกใช้เป็นตัวแทนของ Cluster

งานวิจัย [17] ใช้เทคนิคของ GA เพื่อเลือก Cluster Prototype สำหรับ Hard C-Partitioning จากข้อมูลทั้งหมด โดยมีจุดประสงค์เพื่อหา Cluster ที่มีค่า Within-Group Square Distance Criterion ที่ดีที่สุด ประสิทธิภาพของเทคนิคที่นำเสนอนี้ถูกเปรียบเทียบกับประสิทธิภาพที่ได้จาก Hard C-Means และ Random Search บนข้อมูล Goldfish พบว่าให้เทคนิคที่นำเสนอให้ค่า Within-Group Square Distance ที่ดีกว่า

การแทนค่าโครโมโซมแบบ Medoid Based ยังมีให้เห็นในงานวิจัย [18] ที่ซึ่งนำเอา GA มาประยุกต์ Dissimilarity ภายในแต่ละกลุ่มน้อยที่สุดก่อน จากนั้นจึงนำมารวม (Hybrid) กับ GA เพื่อจัดการกับปัญหา Local Optimal ที่อาจเกิดขึ้นกับ Local Search Heuristic ผลการทดลองกับข้อมูล Gene Expression จำนวน 2 ชุดเมื่อเปรียบเทียบกับอัลกอริทึมลักษณะเดียวกันอย่าง GCA [19] และ RAR_w-GA [20] พบว่าได้ผลลัพธ์การจัดกลุ่มข้อมูลที่ดีกว่า

บทที่ 3 Adaptive Genetic Algorithm Approach to Clustering for Numerical and Categorical data

เนื้อหาในบทนี้กล่าวถึงเทคนิคและวิธีการที่จะใช้ในโครงงานวิจัยนี้ เทคนิคที่ใช้นั้นได้แรงบันดาลใจมาจากอัลกอริทึม Diana (DIvisive ANAlysis) [21] ที่พัฒนาขึ้นโดย Kaufman and Rousseeuw ในปี 1990 Diana เป็นอัลกอริทึมสำหรับจัดกลุ่มข้อมูลแบบ Hierarchy – Divisive ที่สามารถจัดกลุ่มข้อมูลเชิงจำนวนและเชิงลักษณะได้ ซึ่งเหมาะสำหรับข้อมูลที่โครงสร้างโดยธรรมชาติเป็นแบบ Hierarchy แต่ด้วยข้อมูลส่วนใหญ่ที่เก็บในปัจจุบันอาจไม่เหมาะกับโครงสร้างนี้เสมอไป โครงงานวิจัยนี้จึงได้พัฒนาอัลกอริทึมขึ้นมาบนพื้นฐานแนวคิดจาก Diana โดยนำเทคนิคของ GA มาประยุกต์เพื่อให้ได้ผลลัพธ์การจัดกลุ่มในรูปแบบ Partitional เนื้อหาในบทประกอบด้วยหัวข้อหลักๆ ได้แก่ Divisive ANAlysis, Similarity Measure, Genetic Algorithm Unit, Adaptive Genetic Algorithm และแนวทางการเลือกจำนวน Cluster ที่เหมาะสม

3.1 DIvisive ANAlysis (Diana) clustering

แนวคิดการทำงานของ Diana นั้นอุปมาได้กับการแยกตัวของพรรคการเมือง กล่าวคือในพรรคการเมืองหนึ่ง เมื่อมีความขัดแย้งกันภายในเกิดขึ้น สมาชิกคนที่มีความเห็นต่างที่สุด จะแยกตัวออกมาจากพรรคการเมืองนั้นเพื่อจัดตั้งพรรคการเมืองของตนเองใหม่และสมาชิพรรคที่เหลือที่มีความคิดเห็นคล้ายกันก็จะทะยอยย้ายตามออกไปจนกระทั้งทั้งสองพรรคการเมืองมีจำนวนสมาชิกที่มีความคิดเห็นที่คล้ายกัน ดังนั้นขั้นตอนแรกของการทำงานของ Diana คือหาสมาชิกที่มีความแตกต่างมากที่สุด เพื่อใช้เป็นตัวแทนของกลุ่มใหม่ที่จะเกิดขึ้น ตัวอย่างการจัดกลุ่มด้วยวิธีของ Diana สำหรับชุดข้อมูลที่ประกอบด้วย 7 Samples สามารถแสดงได้ดังนี้

พิจารณา Distance Matrix (ตามที่ได้อธิบายไว้ในหัวข้อที่ 3.2) ของข้อมูลทั้ง 7 Sample ดังแสดงในตารางที่ 3.1

ตารางที่ 3.1 Distance Matrix

	1	2	3	4	5	6	7	Average Distance
1	-	0.45	0.57	0.49	0.48	0.31	0.77	0.51
2	0.45	-	0.38	0.61	0.79	0.76	0.36	0.56
3	0.57	0.38	-	0.67	0.69	0.63	0.46	0.57
4	0.49	0.61	0.67	-	0.43	0.55	0.46	0.54
5	**0.48**	**0.79**	**0.69**	**0.43**	**-**	**0.54**	**0.90**	**0.64**
6	0.31	0.76	0.63	0.55	0.54	-	0.81	0.60
7	0.77	0.36	0.46	0.46	0.90	0.81	-	0.63

เมื่อเริ่มต้น Samples ทุกตัวอยู่อยู่ใน Main Group จากนั้น Sample ที่มีค่าเฉลี่ยของ Distance ไปยังทุกตัวใน Main Group มากที่สุดจะถูกแยกออกไปเพื่อสร้างกลุ่มใหม่ที่เรียกว่า Splinter Group ซึ่งจากตารางที่ 3.1 พบว่า Sample ตัวที่ 5 มีค่าเฉลี่ยของ Distance มากที่สุดคือ 0.64 จึงถูกแยกออกไปสร้าง Splinter Group ที่ประกอบด้วยสมาชิกเพียงตัวเดียวคือ (5) ทำให้สมาชิกใน Main Group เหลือ 6 Samples คือ (1, 2, 3, 4, 6, 7) จากนั้นคำนวณหาค่าเฉลี่ยของ Distance ระหว่าง Samples ทุกตัวใน Main Group ไปยัง Splinter Group และระหว่าง Sample ใน Main Group ด้วยกันเอง เมื่อได้ค่าเฉลี่ย Distance ทั้ง 2 ค่าแล้วจึงคำนวณหาผลต่างของทั้งสองค่า ดังแสดงได้ดังตารางที่ 3.2

ตารางที่ 3.2 ผลต่างค่าเฉลี่ย Distance ระหว่าง Main Group และ Splinter Group ขั้นที่ 1

Sample in main group	**Average distance to splinter group (A)**	**Average distance to main group (B)**	**B - A**
1	0.48	0.52	0.04
2	0.79	0.51	-0.28
3	0.69	0.54	-0.15
4	**0.43**	**0.56**	**0.12**
6	0.54	0.61	0.07
7	0.90	0.57	-0.33

จากตารางที่ 3.2 พบว่า ผลต่างค่าเฉลี่ยของ Distance ที่มากที่สุดเป็นของ Sample ที่ 4 ดังนั้น Sample นี้จึงถูกย้ายจาก Main Group ไปยัง Splinter Group ทำให้ในขั้นนี้สมาชิกใน Splinter Group เพิ่มเป็น 2 Sample คือ (5, 4) และสมาชิกใน Main Group เหลือ 5 Sample คือ (1, 2, 3, 6, 7)

เมื่อทำตามขั้นตอนเดิมอีกครั้งหนึ่ง จะได้ผลลัพธ์ดังแสดงได้ในตารางที่ 3.3

ตารางที่ 3.3 ผลต่างค่าเฉลี่ย Distance ระหว่าง Main Group และ Splinter Group ขั้นที่ 2

Sample in main group	Average distance to splinter group (A)	Average distance to main group (B)	B - A
1	0.48	0.53	0.04
2	0.70	0.49	-0.21
3	0.68	0.51	-0.17
6	**0.55**	**0.63**	**0.08**
7	0.68	0.53	-0.15

จากขั้นตอนนี้พบว่า Sample ที่ 6 มีผลต่างค่าเฉลี่ยของ Distance ระหว่าง Main Group และ Splinter Group มากที่สุด ดังนั้น Sample นี้จึงถูกย้ายจาก Main Group ไป ยัง Splinter Group ทำให้ Main Group มีสมาชิกเหลืออยู่ 4 Sample คือ (1, 2, 3, 7) และสมาชิกใน Splinter Group มีสมาชิกเพิ่มเป็น 3 Sample คือ (5, 4, 6)

จากนั้น เมื่อเราทำขั้นตอนเดิมซ้ำอีก Sample แต่ละตัวใน Main Group ก็จะถูกทยอยย้ายไปยัง Splinter Group จนกระทั่งเมื่อผลต่างค่าเฉลี่ยระหว่าง Main Group และ Splinter Group ของทุก Sample ใน Main Group น้อยกว่าศูนย์หรือติดลบจึงหยุดการแยกกลุ่ม จากตัวอย่างข้างต้นเมื่อหยุดการทำงาน ผลลัพธ์ที่ได้คือ Main Group จะเหลือสมาชิก 2 Sample คือ (3,7) และ Splinter Group มีสมาชิกจำนวน 5 Sample คือ (5, 4, 6, 1,2) การทำงานของ Diana อาจหยุดเพียงเท่านี้หรือดำเนินการต่อตามแต่ละ Subgroup ที่แยกออกมาแล้ว จนกระทะทั่งแต่ละกลุ่มเหลือสมาชิกเพียงตัวเดียว

3.2 Similarity measure

ประเด็นสำคัญที่ต้องพิจารณาในการจัดกลุ่มข้อมูลคือ การสามารถที่จะบอกความเหมือน (Similarity) หรือความแตกต่างระหว่าง (Dissimilarity) ของข้อมูลแต่ละตัวได้อย่างไร ซึ่งการเลือกวิธีการหา Similarity ต้องให้สอดคล้องกับประเภทของข้อมูลที่กำลังจะทำการจัดกลุ่มด้วย ในงานวิจัยนี้เลือกวิธีการของ Gower สำหรับชุดข้อมูลเชิงลักษณะและชุดข้อมูลที่ผสมกันระหว่างเชิงลักษณะกับเชิงจำนวน (Mixed-Mode) เนื่องจากเป็นวิธีหนึ่งที่นิยมที่สุดในการหา Similarity ของข้อมูลแบบ Mixed-Mode และเลือกใช้ Euclidean distance สำหรับชุดข้อมูลเชิงจำนวน

3.2.1 Gower's Similarity Measure

ด้านการทำเหมืองข้อมูลนักวิทยาศาสตร์และนักสถิติได้เสนอวิธีการหา Similarity สำหรับชุดข้อมูลแบบ Mixed-Mode ไว้หลายวิธีด้วยกัน วิธีการของ Gower เป็นวิธีการหนึ่งที่นิยมนำมาใช้เพื่อหา Similarity ของชุดข้อมูลแบบ Mixed-Mode โดย Gower ได้นำเสนอสมการการหา Similarity ไว้ใน [21] ซึ่งสามารถอธิบายพอสังเขปได้ ดังนี้

สำหรับข้อมูลเชิงลักษณะ สามารถหา Similarity ได้ดังสมการที่ 3.1

$$S_{ij} = \sum_{k=1}^{p} w_{ijk} S_{ijk} / \sum_{k=1}^{p} w_{ijk} \quad (3.1)$$

เมื่อ

S_{ijk} = Similarity ระหว่างข้อมูลตัวที่ i และ j ใดๆ บนตัวแปรตำแหน่งที่ k

w_{ijk} = มีค่าเป็น 0 เมื่อค่าของตัวแปร k ในข้อมูลตัวที่ i หรือ j ไม่มี

p = จำนวนตัวแปรข้อมูลเชิงลักษณะทั้งหมด

S_{ijk}จะมีค่าเป็น 1 ก็ต่อเมื่อค่าของ i และ j บนตัวแปรตำแหน่งที่ k มีค่าเหมือนกัน มิเช่นนั้นแล้ว S_{ijk} จะมีค่าเป็น 0

สำหรับข้อมูลเชิงจำนวน สามารถหา Similarity ได้ดังสมการที่ 3.2

$$S_{ij} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k} \quad (3.2)$$

เมื่อ x_{ik} และ x_{jk} = ค่าของ i และ j ใดๆ บนตัวแปรตำแหน่งที่ k

R_k = ช่วงของข้อมูลบนตัวแปร k ทั้งหมดในชุดข้อมูล

ตัวอย่างการคำนวณหา Similarity โดยวิธีของ Gower แสดงได้ดังนี้

พิจารณาข้อมูลชนิด Mixed – Mode ของนักศึกษาที่ประกอบด้วยข้อมูลเพศ เกรดเฉลี่ย อายุ และคุณสมบัติของนักศึกษาที่ต้องการสมัครเรียนต่อหลักปริญญาโทในตารางที่ 3.4 พบว่าข้อมูลเพศและของนักศึกษาเป็นข้อมูลเชิงลักษณะ และข้อมูลเกรดเฉลี่ยและอายุเป็นข้อมูลเชิงจำนวนที่มีช่วงของเกรดเฉลี่ยคือ 1.5 และช่วงของข้อมูลอายุคือ 17

ตารางที่ 3.4 ข้อมูลนักศึกษา

#	Gender	GPA	Age	Qualification
1	Male	2.8	28	Engineer
2	Female	3.89	29	Engineer
3	Female	3.2	28	logistic
4	Male	3.7	34	IT
5	Male	3.12	40	Science
6	Male	2.5	27	Art
7	Female	4	23	IT

จากข้อมูลที่ได้สามารถคำนวณหา Similarity ระหว่างนักศึกษาคนที่ 1 และคนที่ 2 ตามสมการที่ 3.1 และ 3.2 ได้ดังนี้

$$s_{1,2} = \frac{1 \times 0 + 1 \times \left[1 - \frac{|2.8 - 3.89|}{1.5}\right] + 1 \times \left[1 - \frac{|28 - 29|}{17}\right] + 1 \times 1}{1 + 1 + 1 + 1} = 0.55$$

หลังจากคำนวณหา Similarity ระหว่างนักศึกษาทั้ง 7 คนครบแล้วก็สามารถสร้าง Similarity Matrix ของนักศึกษาทั้ง 7 คนได้ ดังตารางที่ 3.5

ตารางที่ 3.5 Similarity Matrix ของนักศึกษา

	1	2	3	4	5	6	7
1	-	0.55	0.43	0.51	0.52	0.69	0.23
2	0.55	-	0.62	0.39	0.21	0.24	0.64
3	0.43	0.62	-	0.33	0.31	0.37	0.54
4	0.51	0.39	0.33	-	0.57	0.45	0.54
5	0.52	0.21	0.31	0.57	-	0.46	0.10
6	0.69	0.24	0.37	0.45	0.46	-	0.19
7	0.23	0.64	0.54	0.54	0.10	0.19	-

เนื่องจากโปรแกรมที่นำเสนอในวิทยานิพนธ์นี้ใช้ Dissimilarity Matrix หรือ Distance Matrix ดังนั้น จากค่า Similarity ในตารางที่ 3.5 สามารถเปลี่ยนเป็น Distance ได้จากสมการที่ 3.3

$$d_{i,j} = 1 - s_{i,j} \tag{3.3}$$

ตารางที่ 3.6 แสดง Distance Matrix ของนักศึกษาที่ได้จาก Similarity Matrix ในตารางที่ 3.5

ตารางที่ 3.6 Distance Matrix ของนักศึกษา ด้วยวิธีของ Gower

	1	2	3	4	5	6	7
1	-	0.45	0.57	0.49	0.48	0.31	0.77
2	0.45	-	0.38	0.61	0.79	0.76	0.36
3	0.57	0.38	-	0.67	0.69	0.63	0.46
4	0.49	0.61	0.67	-	0.43	0.55	0.46
5	0.48	0.79	0.69	0.43	-	0.54	0.90
6	0.31	0.76	0.63	0.55	0.54	-	0.81
7	0.77	0.36	0.46	0.46	0.90	0.81	-

3.2.2 Euclidean Distance

ในงานวิจัยนี้ Euclidean Distance ถูกใช้เพื่อคำนวณหา Distance Matrix เมื่อข้อมูลเป็นเชิงจำนวนทั้งหมด โดยคำนวณได้จาก สมมการที่ 3.4

$$d_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{3.4}$$

เมื่อ

x_{ik} และ x_{jk} = ค่าของ i และ j ใดๆ บนตัวแปรตำแหน่งที่ k

ตัวอย่างการหา Distance Matrix โดยใช้ Euclidean Distance ของข้อมูลนักศึกษา 7 คนที่ประกอบด้วยข้อมูลเกรดเฉลี่ยและอายุแสดงได้ ในตารางที่ 3.7

ตารางที่ 3.7 ข้อมูลเกรดเฉลี่ยและอายุของนักศึกษา

#	GPA	Age
1	2.8	28
2	3.89	29
3	3.2	28
4	3.7	34
5	3.12	40
6	2.5	27
7	4	23

จากสมการที่ 3.4 Distance ระหว่างนักศึกษาคนที่ 1 และ 2 สามารถคำนวณได้ดังนี้

$$d_{1,2} = \sqrt{(2.8 - 3.89)^2 + (28 - 29)^2} = 1.48$$

เมื่อคำนวณหา Distance ระหว่างนักศึกษาทั้ง 7 คนครบแล้ว จะได้ Distance Matrix ดังแสดงได้ในตารางที่ 3.8

ตารางที่ 3.8 Distance Matrix ของนักศึกษาด้วยวิธี Euclidean Distance

	1	2	3	4	5	6	7
1	-	1.48	0.40	6.07	12.00	1.044	5.14
2	1.48	-	1.21	5.00	11.03	2.44	6.00
3	0.40	1.21	-	6.02	12.00	1.22	5.06
4	6.07	5.00	6.02	-	6.03	7.10	11.00
5	12.00	11.03	12.00	6.03	-	13.01	17.02
6	1.044	2.44	1.22	7.10	13.01	-	4.27
7	5.14	6.00	5.06	11.00	17.02	4.27	-

3.3 Genetic Algorithm Unit

วิธีการจัดกลุ่มแบบ Diana ถูกนำมาประยุกต์ร่วมกับ GA ในการจัดกลุ่มข้อมูล โดยใช้เทคนิคของ GA เพื่อเลือกตัวแทนของกลุ่ม (Prototype) ในจำนวน Cluster ที่เหมาะสม เพื่อสร้าง Splinter Group แทนการเลือกแบบ Diana ที่อธิบายไว้ในหัวข้อที่ 3.1 วิธีการนี้ข้อมูลจะถูกจัดกลุ่มตาม Prototype ที่ถูกเลือกมาและทำให้ผลลัพธ์ของการจัดกลุ่มที่ได้อยู่ในรูปแบบพาร์ทิชันนอลที่ไม่เป็นลำดับชั้น รายละเอียดแต่ละส่วนของ GA อธิบายได้ ดังนี้

3.3.1 โครโมโซม (Chromosome Representation)

โครโมโซมใน GA ของโครงงานวิจัยนี้เป็นแบบ Medoid based ดังที่ได้อธิบายไว้ในหัวข้อที่ 2.3.3 ความยาวของโครโมโซมในงานวิจัยนี้นับเป็นปัจจัยหนึ่งที่จำเป็นจะต้องกำหนดให้เหมาะสม

โครโมโซมที่สั้นเกินไปหมายถึงการเริ่มต้นของ Cluster ที่เล็กมาก ซึ่งจะทำให้ระบบ GA ใช้เวลาในการจัดกลุ่มนานมาก ในทางตรงกันข้าม การกำหนดโครโมโซมที่ยาวเกินไป หมายถึง Cluster เริ่มต้นที่มีขนาดใหญ่ ในกรณีนี้หมายถึงระบบ GA อาจไม่สามารถจัดกลุ่มที่เหมาะสมจากข้อมูลที่มีขึ้นได้ ดังนั้นในโครงงานวิจัยนี้จึงกำหนดให้ความยาวของโครโมโซมมีอัตราส่วนระหว่างจำนวน Cluster (k) และจำนวนของข้อมูลทั้งหมด (N) ดังนี้

$$L = \begin{cases} k \times 2, & \mathrm{N} \leq 200 \\ (k \times N)/100, & N > 200 \end{cases} \tag{3.5}$$

ตัวอย่างเช่น ขนาดความยาวของโครโมโซมที่ต้องการเพื่อใช้จัดกลุ่มของชุดข้อมูลขนาด 500 Sample ให้เป็น 3 Cluster คือ $L = \frac{(3 \times 500)}{100} = 15$ ซึ่งหมายความว่าสมาชิกตัวที่ 1 - 5, 6 - 10 และ 11 – 15 เป็น Prototype ของ Cluster ที่ 1, 2 และ 3 ตามลำดับ

3.3.2 ขนาดของประชากร (Population Size)

เช่นเดียวกันกับความยาวของโครโมโซม จำนวนของประชากรในแต่ละรุ่น (ขนาดของประชากร) เป็นอีกปัจจัยหนึ่งที่ต้องพิจารณาให้ได้ค่าที่เหมาะสม ขนาดของประชากรที่น้อยเกินไปจะไม่สามารถสร้างความหลากหลายได้ ในขณะที่ขนาดประชากรที่มากเกินไป จะทำให้ระบบ GA ใช้เวลาในการประมวลผลนานเกินไป โดยไม่มีหลักประกันว่าจะได้ผลลัพธ์ที่ดีกว่า ดังนั้นขนาดของประชากร (P) สำหรับโครงงานวิจัยถูกกำหนดให้เป็นสัดส่วนของจำนวนข้อมูลทั้งหมด (N) ดังนี้

$$P = \begin{cases} 50, & N \leq 500 \\ (10 \times N), & 500 < N \leq 2{,}000 \\ 200, & N > 2{,}000 \end{cases} \tag{3.6}$$

3.3.3 ฟิตเนสฟังก์ชัน (Fitness Function)

หลังจากที่โครโมโซมถูกสร้างขึ้นในแต่ละรุ่นแล้วโครโมโซมเหล่านั้นจะใช้เพื่อเป็น Prototype ในการจัดกลุ่มด้วยหลักการเดียวกับ Diana คือ Sample แต่ละตัวจะย้ายจาก Main Group ไปยัง Splinter Group ที่มี ใกล้ตัวมันที่สุด (ค่าเฉลี่ยของ Distance น้อยที่สุด) เมื่อได้กลุ่มของข้อมูลที่ถูกจัดเรียบร้อย สำหรับแต่ละโครโมโซมแล้วจึงสามารถวัดคุณภาพของแต่ละโครโมโซมด้วยฟิสเนสฟังก์ชัน ฟิตเนสฟังก์ชันได้มีการใช้เป็น Clustering Criteria ในหลายคุณลักษณะในการจัดกลุ่มข้อมูล ในงานวิจัยนี้เลือกใช้ Clustering Criterion หนึ่งที่เป็นที่นิยมสำหรับการวัดคุณภาพของ Cluster และถูกใช้มาก่อนใน [22] ซึ่งอธิบายได้ ดังสมการที่ 3.7

$$Fitness = \sum_{i=1}^{k}(D_{inter}(C_i) - D_{Intra}(C_i)) \quad (3.7)$$

เมื่อ

D_{Intra} = Distance ระหว่าง Sample ทุกตัวใน Cluster C_i

D_{inter} = Distance ระหว่าง Sample ทุกตัวใน Cluster C_i ไปยัง Sample อื่นทุกตัวที่ไม่ได้เป็นสมาชิกของ Cluster C_i

ดังนั้นค่า Fitness ที่มากกว่าสะท้อนถึงการจัดกลุ่มที่ดีกว่า ตัวอย่างเช่น จาก Distance Matrix ในตารางที่ 3.1 ถ้าต้องการจัดกลุ่มจำนวน 3 Cluster โดยมี Sample ตัวที่ 1, 4 และ 7 เป็น Prototype ของ Cluster ที่ k1, k2 และ k3 ตามลำดับ จะได้โครโมโซมที่มีความยาวเท่ากับ 3 และได้ Splinter Group ดังตารางที่ 3.9

ตารางที่ 3.9 Splinter Group เริ่มต้น

k1	1
k2	4
k3	7

จากนั้นคำนวณหาค่าเฉลี่ยของ Distance ระหว่าง Sample แต่ละตัวใน Main Group ไปแต่ละ Splinter Group ซึ่งจากผลการคำนวณพบว่า Sample ที่ 6 มีค่าเฉลี่ยไปยัง k1 น้อยที่สุด ดังแสดงได้ในตารางที่ 3.10 ดังนั้น Sample นี้จึงถูกย้ายจาก Main Group ไป k1 และทำให้ k1 มีสมาชิกเพิ่มเป็น 2 ตัว คือ (1,6)

ตารางที่ 3.10 ค่าเฉลี่ยของ Distance ไปยัง Splinter Group ขั้นแรก

Sample in main group	Average Distance		
	k1	k2	k3
2	0.45	0.61	0.36
3	0.57	0.67	0.46
5	0.48	0.43	0.90
6	**0.31**	0.55	0.81

เมื่อคำนวณแบบเดิมอีกครั้งพบว่า Sample ตัวที่ 2 มีค่าเฉลี่ยของ Distance ไปยัง k3 น้อยที่สุดดังแสดงได้ในตารางที่ 3.11 ดังนั้น Sample ตัวนี้จึงถูกย้ายไปยัง k3 และทำให้ k3 มีสมาชิกเพิ่มเป็น 2 ตัว คือ (7, 2)

ตารางที่ 3.11 ค่าเฉลี่ยของ Distance ไปยัง Splinter Group ขั้นที่ 2

Sample in main group	Average Distance		
	k1	k2	k3
2	0.60	0.61	**0.36**
3	0.60	0.67	0.46
5	0.51	0.43	0.90

เมื่อคำนวณแบบเดิมอีกครั้งจะพบว่า Sample ตัวที่ 3 มีค่าเฉลี่ยของ Distance ไปยัง k3 น้อยที่สุด ดังแสดงได้ในตารางที่ 3.12 ดังนั้น Sample ตัวนี้จึงถูกย้ายไปยัง k3 ทำให้ Cluster นี้มีสมาชิกเพิ่มเป็น 3 ตัว คือ (7, 2, 3) และ Main Group เหลือ สามาชิกเพียงตัวเดียวคือ Sample ตัวที่ 5

ตารางที่ 3.12 ค่าเฉลี่ยของ Distance ไปยัง Splinter Group ขั้นที่ 3

Sample in main group	Average Distance		
	k1	k2	k3
3	0.60	0.67	**0.42**
5	0.51	0.43	0.84

ขั้นสุดท้ายเมื่อคำนวณหาค่าเฉลี่ยของ Distance ระหว่าง Sample ตัวที่ 5 ใน Main Group นี้ไปยังทุก Splinter Group และพบว่า k2 เป็น Cluster ที่ใกล้ Sample ตัวที่ 5 มากที่สุด ดังแสดงได้ในตารางที่ 3.13 ดังนั้น Sample นี้จึงถูกย้ายไป k2 ทำให้ k2 มีสมาชิกเพิ่มเป็น 2 ตัวคือ (4, 5)

ตารางที่ 3.13 ค่าเฉลี่ยของ Distance ไปยัง Splinter Group ขั้นที่ 4

Sample in main group	Average Distance		
	k1	k2	k3
5	0.51	**0.43**	0.79

ผลลัพธ์สุดท้ายที่ได้จากการจัดกลุ่มคือ k1 = (1, 6), k2 = (4,5) และ k3 = (7, 2, 3) และเมื่อคำนวณหาค่าฟิตเนสจากสมการที่ 3.7 แล้วจะได้ Fitness เท่ากับ 16.46

3.3.4 การเลือกสรร (Selection)

ดังที่ได้กล่าวไว้พอสังเขปในหัวข้อที่ 2.2.2 การเลือกสรรมีหลายวิธีและยังไม่มีวิธีใดที่ถูกพิสูจน์ว่าดีที่สุด จุดประสงค์ของการเลือกสรรคือการสร้างความหลากหลายที่มากที่สุดในรุ่นต่อไป กลวิธีในการคัดเลือกที่จะใช้ในโครงงานวิจัยนี้คือ Elitism, Roulette wheel และ Random เนื่องจากมีความซับซ้อนในการคำนวณไม่มากเกินไปนัก และน่าจะเสนอความหลากหลายได้มากเพียงพอกว่าการเลือกใช้วิธีเดียว รายละเอียดของแต่ละวิธีสามารถอธิบายพอสับเขปได้ดังนี้

i. Elitism: วิธีการนี้จะทำการเก็บ โครโมโซม ที่ดีที่สุดไว้จำนวนหนึ่งเพื่อเป็น ประชากร ใหม่ใน Generation ถัดไป ทั้งนี้เพื่อป้องกันการสูญเสียโครโมโซมที่ดีที่สุดไปเนื่องจากการทำ Crossover หรือ Mutation

ii. Roulette Wheel: ในกระบวนการคัดเลือกแบบ Roulette Wheel นี้ โอกาสที่ Parent จะถูกเลือกขึ้นอยู่กับค่า fitness ของ แต่ละ โครโมโซม กล่าวคือ โครโมโซม ที่มีค่า fitness มาก จะมีโอกาสถูกเลือกมากกว่า

iii. Random: วิธีการนี้ โครโมโซม จะถูกเลือกแบบสุ่ม

3.3.5 ครอสโอเวอร์ (Crossover)

เทคนิคการครอสโอเวอร์ที่ใช้ในงานวิจัยนี้คือ 1-Point Crossover และ 2-Point Crossover วิธีการของทั้ง 2 เทคนิคนี้ได้อธิบายและยกตัวอย่างไว้ในหัวข้อที่ 2.2.2 โดยเงื่อนไขในการใช้แต่ละเทคนิคขึ้นอยู่กับการปรับ เปลี่ยนโหมดของ GA ที่ได้อธิบายไว้ในหัวข้อที่ 3.4

3.3.6 มิวเตชัน (Mutation)

เพื่อไม่ให้การทำงานของ GA ซับซ้อนเกิดไปเทคนิคการมิวเตชันที่ใช้ในงานวิจัยนี้คือการสุ่ม (Random) 1 ตำแหน่ง ดังที่ได้อธิบายและยกตัวอย่างไว้ในหัวข้อที่ 2.2.2

3.3.7 เงื่อนไขการหยุดการทำงาน (Stopping Criteria)

เงื่อนไขการหยุการทำงานของ GA ที่นิยมใช้มีหลายวิธี เช่น ใช้เวลาเป็นตัวกำหนด ใช้จำนวนรอบในการพัฒนา (Generation) และใช้ค่า Fitness แต่ในโครงงานวิจัยนี้ใช้วิธีการหนึ่งที่นิยมใช้คือ ดูการพัฒนาของ Solution หรือค่าฟิตเนสที่พัฒนาขึ้นในแต่ละ Generation กล่าวคือ การทำงานของระบบจะหยุดเมื่อมีการเปลี่ยนโหมดครบ 4 รอบ ตามเงื่อนไขของการเปลี่ยนโหมดของ GA ดังได้อธิบายในหัวข้อที่ 3.4

3.4 การปรับตัวเพื่อการพัฒนา (Adaptability)

Local maxima Plateau และ Ridge เป็นปัญหาสามัญที่สามารถเกิดขึ้นได้ในการค้นหาทั่วๆ ไป ดังนั้น ระบบ GA ที่พัฒนาขึ้นจึงจำเป็นต้องมีแนวทางหลีกเลี่ยงปัญหาเหล่านี้ โครงงานวิจัยนี้ได้พัฒนา GA ให้มีความสามารถในการปรับตัวเพื่อพัฒนาเพื่อแก้ปัญหาดั่งที่ได้กล่าวไว้ข้างต้น ความสามารถนี้คือ การเปลี่ยน Operating mode ของ GA อันได้แก่การคลอสโอเวอร์และการเลือกสรร การปรับตัวเพื่อการพัฒนาของ GA ในงานวิจัยนี้ประกอบไปด้วย 3 โหมดด้วยกัน การทำงานของระบบจะเริ่มจากโหมดที่ 1 เปลี่ยนไปเป็นโหมดที่ 2, 3 และเปลี่ยนกลับมาที่โหมดที่ 1 อีกครั้งหนึ่ง ก่อนจบการทำงานของระบบ โดยเงื่อนไขในการเปลี่ยนโหมดคือการที่ไม่สามารถสร้างโครโมโซมที่มีการพัฒนาค่าฟิตเนสให้สูงขึ้นได้ต่อเนื่องกัน 50 Generations ตารางที่ 3.14 แสดงเทคนิคการเลือกสรรและการคลอสโอเวอร์ที่ใช้ในแต่ละโหมด

ตารางที่ 3.14 โหมดการทำงานของ GA

	Mode1	Mode2	Mode3
Selection	Elitism	Roulette Wheel	Random
Crossover	1-Point	1-Point	2-Point

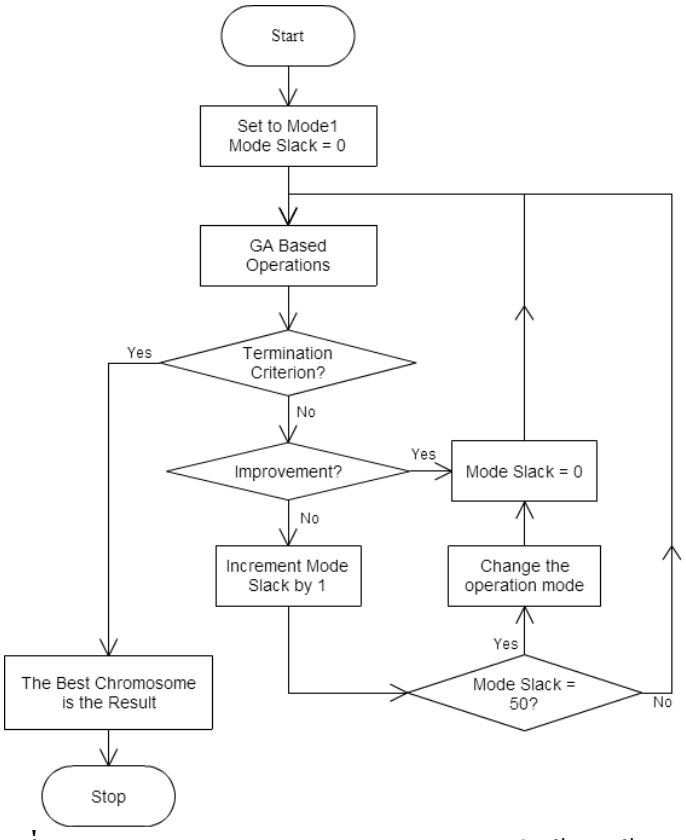
Mode 1 – เป็นการเริ่มต้นบนพื้นฐานว่าโครโมโซมที่ดีน่าจะมีโอกาสสร้างโครโมโซมที่ดีขึ้นอีกได้
Mode 2 – เปรียบเหมือนการให้โอกาสการถูกเลือกของโครโมโซมตามความเหมาะสมของแต่ละตัว ซึ่งน่าจะแก้ปัญหาเบื้องต้นได้เมื่อ GA ไม่เกิดการพัฒนาขึ้นเลยในระยะ 50 Generations
Mode 3 – Mode นี้จะถูกใช้เมื่อ Mode 1 และ Mode 2 แล้วแต่การพัฒนายังไม่เกิดขึ้นได้ การดำเนินการใน Mode นี้เหมือนกับการ Reset ส่วนใหญ่ของระบบเดิม โดยคาดว่าจะสามารถแก้ปัญหาการไม่พัฒนาเนื่องจากสภาวะต่างๆ ที่ได้อธิบายไว้ในตอนต้น

3.5 แนวทางการเลือกจำนวน Cluster ที่เหมาะสม

ในการจัดกลุ่มข้อมูลหากจำนวนของ Cluster (k) ไม่ได้ถูกระบุมาโดยผู้ใช้มาก่อนแล้ว การระบุจำนวนของ Cluster ที่เหมาะสมจึงเป็นปัญหาหนึ่งที่ต้องพิจารณา วิธีการเลือกจำนวน Cluster ที่เหมาะมีอยู่หลายวิธีด้วยกัน[23], [24] แต่หลักการโดยทั่วไปคล้ายกันคือ ทำการจัดกลุ่มโดยการระบุค่า k ที่ต่างกันและคำนวณหา Quality Index เช่น Calinski and Harabasz, C-index หรือ Silhouettes สำหรับแต่ละ k จำนวน Cluster ที่เหมาะสมคือค่า k ที่ให้ Quality Index ที่ดีที่สุด

ในงานวิจัยนี้เลือกใช้ Silhouettes Index [25] เป็น Quality Index สำหรับหาจำนวน Cluster ที่เหมาะสม เนื่องจาก Silhouettes สามารถใช้ได้กับ Distance Matrix เดียวกันกับที่ใช้ในขั้นตอนการจัดกลุ่มของ GA

การทำงานโดยรวมของระบบที่พัฒนาขึ้นแสดงได้ในรูปที่ 3.1



รูปที่ 3.1 การทำงานโดยรวมของระบบการจัดกลุ่มข้อมูลด้วย GA

บทที่ 4 ข้อมูลที่ใช้ในโครงงานวิจัยและการประเมินประสิทธิภาพการจัดกลุ่มข้อมูล

เนื้อหาบทนี้กล่าวถึงชุดข้อมูล (Datasets) ที่ใช้ทดลองในโครงงานวิจัยนี้และผลการประเมินประสิทธิภาพการจัดกลุ่มของชุดข้อมูลแต่ละประเภท หัวข้อที่ 4.1 อธิบายข้อมูลทั้งหมดที่ใช้ในโครงงานวิจัยนี้ ซึ่งข้อมูลทั้งหมดที่ใช้ในโครงงานวิจัยนี้ได้มาจาก Public Domain ที่เชื่อถือได้และนักศึกษาระดับมหาบัณฑิตของคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี หัวข้อที่ 4.3 อธิบายเกณฑ์การวัดประสิทธิภาพและเปรียบเทียบผลการวัดประสิทธภาพการจัดกลุ่มกับเทคนิคของ Partitioning Around Medoids (PAM) และ Self-Organizing Map (SOM)

4.1 ข้อมูลที่ใช้ในโครงงานวิจัย

ข้อมูลที่ใช้ในโครงงานวิจัยนี้ประกอบไปด้วยข้อมูล 2 ประเภทคือข้อมูลประเภท Labeled และ Unlabeled อย่างละ 6 ชุด โดยข้อมูลแต่ละประเภทประกอบด้วยชุดข้อมูล 3 ลักษณะๆ ละ 2 ชุด ตามประเภทของแอททริบิวท์ (Attribute) คือ

1. ชุดข้อมูลที่ทุกแอททริบิวท์เป็นข้อมูลเชิงจำนวน
2. ชุดข้อมูลที่ทุกแอททริบิวท์เป็นข้อมูลเชิงลักษณะ
3. ข้อมูลที่ประกอบด้วยแอททริบิวท์ที่เป็นเชิงจำนวนและเชิงลักษณะ

รายละเอียดของข้อมูลแต่ละประเภทอธิบายได้ ดังนี้

4.1.1 ชุดข้อมูลประเภท Labeled

ชุดข้อมูลประเภท Labeled คือชุดข้อมูลที่ทราบกลุ่มของข้อมูลอยู่แล้ว กล่าวคือสมาชิกทุกตัวในชุดข้อมูลได้ถูกจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งแล้ว ข้อมูลประเภทนี้มักถูกใช้ในงานด้าน Classification โดยมีจุดประสงค์เพื่อหารูปแบบหรือ Pattern ของข้อมูลเพื่อใช้ทำนายหรือจำแนกข้อมูลใหม่ที่จะเกิดขึ้น นอกจากนี้ข้อมูลประเภทนี้ยังถูกนำมาใช้โดยมีจุดประสงค์เพื่อทดสอบและประเมินประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มข้อมูล (Clustering) โดยเกณฑ์การประเมินประสิทธิภาพการจัดกลุ่มข้อมูลด้วยข้อมูลประเภทนี้ได้ถูกอธิบายไว้ในหัวข้อที่ 4.3.1ข้อมูลประเภท Labeled ทั้งหมดที่ใช้ในโครงงานวิจัยนี้นำมาจาก UCI Machine Learning Repository [26] ซึ่งเป็น Public domain หนึ่งที่เป็นที่นิยม รายละเอียดของแต่ละชุดข้อมูลแสดงได้ ดังตารางที่ 4.1

ตารางที่ 4.1 ชุดข้อมูล Labeled

ชุดข้อมูล	ลักษณะข้อมูล	จำนวนข้อมูล	จำนวนแอททริบิวท์	จำนวนกลุ่ม
Wisconsin Diagnostic Breast Cancer (WDBC)	เชิงจำนวน	569	32	2
Vertebral Column	เชิงจำนวน	310	6	3
Car Evaluation	เชิงลักษณะ	1,728	6	4
Congressional Voting Records	เชิงลักษณะ	435	16	2
Credit Approval	เชิงลักษณะและเชิงจำนวน	690	15	2
Cylinder bands	เชิงลักษณะและเชิงจำนวน	512	40	2

4.1.2 ข้อมูลประเภท Unlabeled

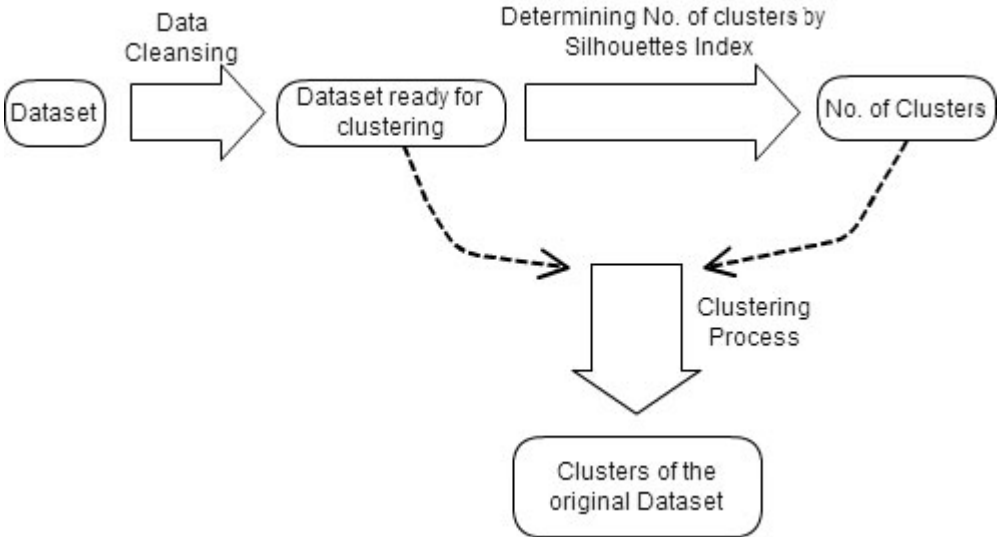
ข้อมูลประเภท Unlabeled คือข้อมูลที่แต่ละชุดยังไม่ถูกระบุกลุ่มของข้อมูลอยู่ก่อน กล่าวคือสมาชิกแต่ละตัวในชุดข้อมูลยังไม่ถูกจัดให้อยู่ในกลุ่มใดกลุ่มหนึ่งที่แน่ชัด โครงงานวิจัยนี้เลือกข้อมูลประเภทนี้เพื่อใช้ทดสอบและวัดประสิทธิภาพการจัดกลุ่มของโปรแกรมที่พัฒนาขึ้นตามเกณฑ์การประเมินประสิทธิภาพที่อธิบายไว้ ในหัวข้อที่ 4.3.2 – 4.3.4 การเลือกจำนวนกลุ่มที่เหมาะสมสำหรับแต่ชุดข้อมูลทำโดยวิธีการคำนวณและเปรียบเทียบ Silhouettes Index ที่ซึ่งได้อธิบายไว้ในบทที่ 3 แล้ว ข้อมูลประเภทนี้ทั้งหมดที่ใช้ในงานวิจัยนี้ได้มาจาก public domain ที่เป็นที่นิยมคือ UCI Machine Learning Repository, KEEL (Knowledge Extraction based on Evolutionary Learning) [27] และข้อมูลนักศึกษาระดับมหาบัณฑิตของคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี รายละเอียดของแต่ละชุดข้อมูลแสดงได้ ดังตารางที่ 4.2

ตารางที่ 4.2 ชุดข้อมูล Unlabeled

ชุดข้อมูล	ลักษณะข้อมูล	จำนวนข้อมูล	จำนวนแอททริบิวท์	จำนวนกลุ่ม
Seed	เชิงจำนวน	210	7	3
Stock Price	เชิงจำนวน	950	10	10
Weather	เชิงลักษณะ	14	5	5
Solar Flare	เชิงลักษณะ	1,066	11	4
Dresses	เชิงลักษณะและเชิงจำนวน	501	13	2
ข้อมูลนักศึกษาระดับมหาบัณฑิตของคณะเทคโนโลยีสารสนเทศมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี	เชิงลักษณะและเชิงจำนวน	2,070	8	5

4.2 ขั้นตอนการจัดกลุ่มข้อมูล

ขั้นตอนการจัดกลุ่มข้อมูลของข้อมูลทั้ง 2 ประเภท (Labeled และ Unlabeled) แสดงได้ ดังรูปที่ 4.1



รูปที่ 4.1 ขั้นตอนการจัดกลุ่มข้อมูล

จากรูปที่ 4.1 ขั้นตอนการจัดกลุ่มข้อมูลเริ่มต้นจากชุดข้อมูล (Dataset) ผ่านกระบวนการ Data Cleansing ในขั้นตอนนี้ข้อมูลที่ไม่สมบูรณ์จะถูกคัดออก เช่น ในชุดข้อมูลนักศึกษาระดับมหาบัณฑิตของคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี นักศึกษาคนใดที่มีข้อมูลไม่ครบทั้ง 8 แอททริบิวท์จะถูกตัดออกจากชุดข้อมูล ทั้งนี้สำหรับบางชุดข้อมูลที่ได้จาก Public-Domain เมื่อผ่านกระบวนการนี้แล้วทำให้จำนวนข้อมูลคงเหลือไม่เพียงพอสำหรับการจัดกลุ่ม ข้อมูลที่ไม่สมบูรณ์เหล่านั้นก็ถูกคงไว้โดยไม่เป็นอุปสรรคสำหรับการจัดกลุ่มข้อมูลสำหรับโปรแกรมที่พัฒนาขึ้น หลังจากผ่านกระบวนการ Data Cleansing แล้วข้อมูลประเภท Labeled พร้อมถูกจัดกลุ่มโดยเข้าสู่กระบวนการ Clustering Process ทันที แต่สำหรับข้อมูลประเภท Unlabeled ที่แต่ละชุดยังไม่ถูกระบุกลุ่มของข้อมูลอยู่ก่อนต้องผ่านกระบวนการระบุจำนวนกลุ่มที่เหมาะสมก่อนเริ่มกระบวนการ Clustering Process ได้

4.3 การประเมินประสิทธิภาพการจัดกลุ่มข้อมูล

การประเมินประสิทธิภาพการจัดกลุ่มของโปรแกรมที่พัฒนาขึ้นในโครงงานแยกเป็น 2 มิติ ตามประเภทของข้อมูลที่ใช้ในการประเมินประสิทธิภาพ ดังที่ได้กล่าวมาแล้วในหัวข้อที่ 4.1 การวัดประสิทธิภาพการจัดกลุ่มบนชุดข้อมูลประเภท Labeled ทำได้โดยการนำผลลัพธ์ที่ได้จากการจัดกลุ่มจากโปรแกรมที่พัฒนาขึ้นไปคำนวณหา Purity Value และเปรียบเทียบกับผลลัพธ์ที่ได้จากเทคนิคของ PAM และ SOM

ในลักษณะเดียวกันการวัดประสิทธิภาพการจัดกลุ่มข้อมูลบนข้อมูลประเภท Unlabeled ทำได้หลายวิธี โดยใช้ตัววัดประสิทธิภาพต่างๆ ที่ถูกนำเสนอในนศาสตร์นี้ ในงานวิจัยนี้นำดัชนีวัดประสิทธภาพ 3 ค่ามาพิจารณาคือค่า Inter – Intra Distance, ค่า McClain-Rao และค่า Lack of Homogeneity ดังนั้นในการเปรียบเทียบประสิทธิภาพ ดัชนีทั้ง 3 ค่านี้ถูกนำไปวัดกับผลลัพธ์ที่ได้จากโปรแกรมที่พัฒนาขึ้นและจาก PAM และ SOM ตามลำดับ รายละเอียดเบื้องตนของทั้ง 3 ดัชนีและผลการเปรียบเทียบประสิทธิภาพอธิบายไว้ในหัวข้อที่ 4.3.1 – 4.3.4

4.3.1 การประเมินชุดข้อมูลประเภท Labeled โดยใช้ค่าความบริสุทธิ์ (Purity)

การประเมินประสิทธิภาพการจัดกลุ่มด้วย Purity Value นับเป็นเกณฑ์การประเมินแบบ External Criterion ซึ่งไม่พิจารณาความเหมือนหรือความต่างกันระหว่างสมาชิกภายในกลุ่มหรือนอกกลุ่ม หากแต่พิจารณาว่าผลลัพธ์ที่ได้ถูกต้องมากน้อยเพียงใด Purity Value ที่ได้บอกถึงความสามารถของอัลกอริทึมที่สามารถจัดกลุ่มให้ข้อมูลส่วนใหญ่ในกลุ่มเป็นข้อมูลคลาสเดียวกันได้มากน้อยเพียงใด Purity Value หาได้จากสมการที่ 4.1

$$Purity(C,P) = \frac{1}{n}\sum_k \max_i |C_k \cap P_i| \quad (4.1)$$

เมื่อ

k = จำนวนกลุ่ม

i = คลาสที่ i

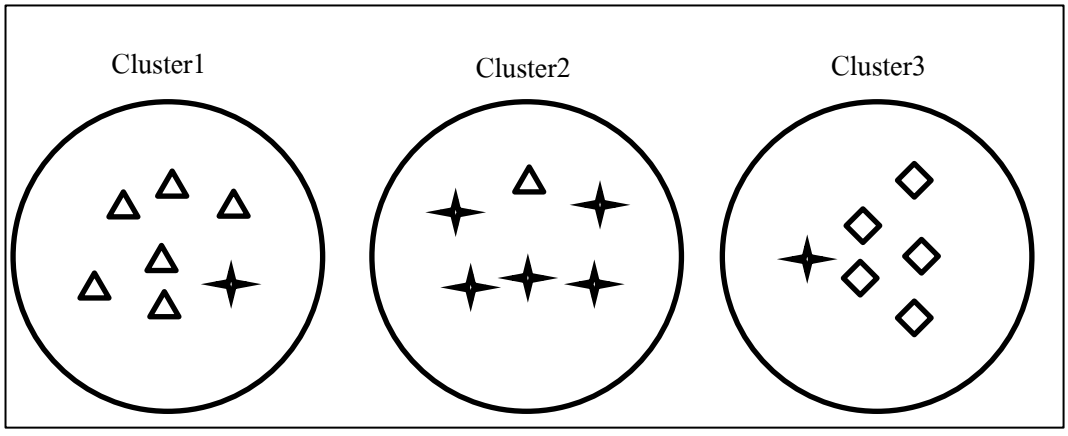
C = $\{C_1, C_2, C_3, \ldots, C_k\}$ คือผลลัพธ์การจัดกลุ่มด้วยอัลกอริทึมแบบ Clustering

P = $\{P_1, P_2, P_3, \ldots, P_i\}$คือคลาสของข้อมูล Labeled

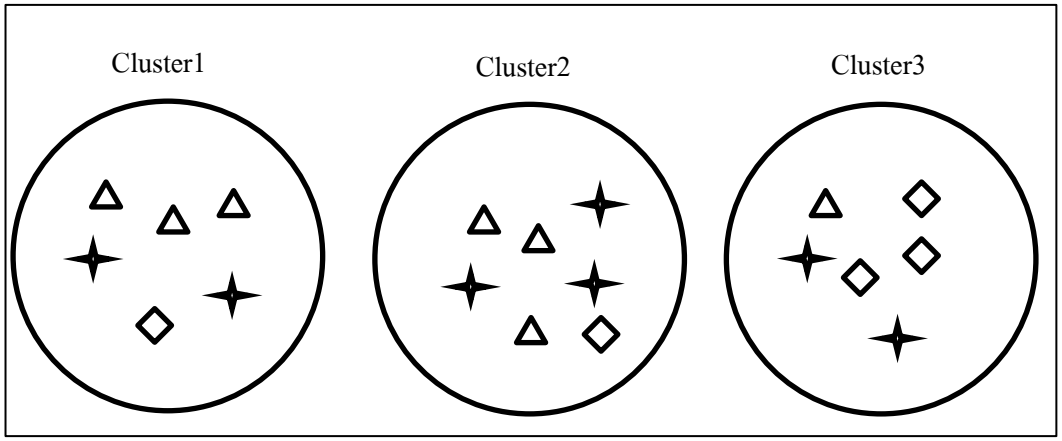
Purity Value มีค่าตั้งแต่ 0 ถึง 1 โดยที่ Purity Value ที่มากกว่าแสดงถึงประสิทธิภาพการจัดกลุ่มที่ดีกว่า ตัวอย่างการคำนวณและการเปรียบหา Purity Value แสดงได้ ดังนี้

ให้รูปที่ 4.2 และ 4.3 แสดงถึงผลลัพธ์การจัดกลุ่มข้อมูลโดย สัญลักษณ์แต่ละรูปแบบแทนข้อมูลในแต่ละคลาส (P) ดังนี้

1. สัญลักษณ์ ◇ แทนข้อมูลที่เป็นสมาชิกของคลาส 1 (P_1)
2. สัญลักษณ์ △ แทนข้อมูลที่เป็นสมาชิกของคลาส 2 (P_2)
3. สัญลักษณ์ ✦ แทนข้อมูลที่เป็นสมาชิกของคลาส 3 (P_3)



รูปที่ 4.2 การจัดกลุ่มแบบที่ 1



รูปที่ 4.3 การจัดกลุ่มแบบที่ 2

ดังนั้น ค่า Purity Value ที่ได้จากการจัดกลุ่มแบบที่ 1 (รูปที่ 6) เท่ากับ (1/19) × (6+5+5) = 0.84 และ Purity Value ที่ได้จากการจัดกลุ่มแบบที่ 2 (รูปที่ 7) เท่ากับ (1/19) × (3+3+3) = 0.47 ซึ่งหมายถึงการจัดกลุ่มแบบที่ 1 ให้ผลลัพธ์การจัดกลุ่มที่ดีกว่าการจัดกลุ่มแบบที่ 2 นั่นเอง

จากชุดข้อมูลประเภท Labeled ทั้งหมดจำนวน 6 ชุด จากตารางที่ 16 เมื่อทำการจัดกลุ่มและคำนวณหา Purity Value ที่ได้จากผลลัพธ์การจัดกลุ่มด้วยวิธีของ GA ที่พัฒนาขึ้นเปรียบเทียบกับผลลัพธ์ที่ได้จากวิธีการของ PAM และวิธีการของ SOM โดยกำหนดให้จำนวนกลุ่ม (k) เท่ากับจำนวนคลาส พบว่า Purity Value ที่ได้จากการจัดกลุ่มด้วยวิธีของ GA ที่พัฒนาขึ้นมีค่าที่สูงกว่าแสดงถึงระบบ GA ที่พัฒนาขึ้นมีประสิทธิภาพที่ดีกว่าการจัดกลุ่มโดยใช้ PAM และ SOM ตารางที่ 4.3 แสดงค่าความบริสุทธ์ของชุดข้อมูลประเภท Labeled ทั้ง 6 ชุด

ตารางที่ 4.3 ผลการเปรียบเทียบ Purity Value

Title	**Missing value**	**Purity**		
		GA	**PAM**	**SOM**
Wisconsin Diagnostic Breast Cancer (WDBC)	No	**0.91**	0.87	0.85
Vertebral Column	No	**0.72**	0.68	0.68
Car Evaluation	No	**0.71**	0.70	0.70
Congressional Voting Records	Yes	**0.87**	0.86	-
Credit Approval	Yes	**0.81**	0.79	-
Cylinder bands	Yes	**0.64**	0.58	-

4.3.2 การประเมินชุดข้อมูล Unlabeled โดยใช้ค่า Inter – Intra Distance

เนื่องจากข้อมูลประเภท Unlabeled เป็นข้อมูลที่แต่ละ Sample ไม่สามารถบอกกลุ่มได้ก่อนหน้า ดังนั้น เกณฑ์การวัดและเปรียบเทียบประสิทธิภาพการจัดกลุ่มสำหรับข้อมูลประเภทนี้จึงไม่สามารถใช้ Purity Value เช่นเดียวกับข้อมูลประเภท Labeled ได้ ค่า Inter – Intra Distance ถูกแสดงไว้ในสมการที่ 3.3 แล้ว ตารางที่ 4.4 เปรียบเทียบค่า Inter – Intra Distance ของข้อมูลประเภท Unlabeled ทั้ง 6 ชุด

ตารางที่ 4.4 ผลการเปรียบเทียบค่า Inter – Intra Distance

Title	Missing value	Inter – Intra Distance		
		GA	PAM	SOM
Seed	No	**133,140.23**	132,264.17	132,909.47
Stock Price	No	**4,154,977.415**	4,110,586.23	4,109,644.34
Weather	No	**90.4**	85.20	88.40
Solar Flare	No	**237,182.65**	232,008.18	175,870.36
Dresses	Yes	**9,119.91**	232.67	-
SIT Student	No	**1,202,798.50**	1,190,689.83	874,436.49

4.3.3 การประเมินชุดข้อมูล Unlabeled โดยใช้ค่า McClain-Rao

เกณฑ์การประเมินคุณภาพการจัดกลุ่มแบบ McClain-Rao [28] พิจารณาอัตราส่วนระหว่างผลรวมของ Intra Distance หารด้วยผลรวมของจำนวนคู่ (Pairs) ทั้งหมดระหว่าง Samples ภายในกลุ่มเดียวกันและผลรวมของ Inter Distance หารด้วยจำนวนคู่ทั้งหมดระหว่าง Samples ที่อยู่ต่างกลุ่มกัน ซึ่งสามารถคำนวณได้ด้วยสมการที่ 4.3

$$C = \frac{S_W/N_W}{S_B/N_B} \tag{4.3}$$

เมื่อ

S_W = ผลรวมของ D_{Intra}

N_W = ผลรวมคู่ภายในกลุ่ม

S_B = ผลรวมของ D_{Inter}

N_B = จำนวนคู่ทั้งหมดระหว่าง Sample ทั้งหมดที่อยู่ต่างกลุ่มกัน

ดังนั้นค่า McClain-Rao ที่น้อยกว่าแสดงถึงประสิทธิภาพของการจัดกลุ่มข้อมูลที่ดีกว่า ผลการเปรียบเทียบ McClain-Rao ของข้อมูลประเภท Unlabeled ทั้งหมด 6 ชุดแสดงได้ ดังตารางที่ 4.5

ตารางที่ 4.5 ผลการเปรียบเทียบ McClain-Rao

Title	Missing value	McClain-Rao		
		GA	PAM	SOM
Seed	No	**0.38**	0.38	0.38
Stock Price	No	**0.32**	0.33	0.32
Weather	No	**0.53**	0.56	0.56
Solar Flare	No	**0.58**	0.62	0.73
Dresses	Yes	**0.90**	0.93	-
SIT Student	No	**0.57**	0.57	0.91

4.3.4 การประเมินชุดข้อมูล Unlabeled โดยใช้ค่า Lack of Homogeneity

เกณฑ์การประเมินคุณภาพการจัดกลุ่มแบบ Lack of Homogeneity [1] พิจารณาผลรวม D_{Intra} ดังแสดงได้ในสมการที่ 4.3

$$h = \sum_{i=1}^{k} D_{Intra}(C_i) \tag{4.3}$$

เมื่อ

D_{Intra} = Distance ระหว่าง Sample ทุกตัวใน Cluster C_i

ดังนั้นค่า Lack of Homogeneity ที่น้อยกว่าแสดงถึงประสิทธิภาพของการจัดกลุ่มข้อมูลที่ดีกว่า ผลการเปรียบเทียบ Lack of Homogeneity ของข้อมูลประเภท Unlabeled ทั้งหมด 6 ชุดแสดงได้ ดังตารางที่ 4.6

ตารางที่ 4.6 ผลการเปรียบเทียบ Lack of Homogeneity

Title	Missing value	Lack of Homogeneity		
		GA	PAM	SOM
Seed	No	592.41	588.78	**587.32**
Stock Price	No	1,648.48	1,622.66	**1,602.89**
Weather	No	**0.59**	0.63	0.62
Solar Flare	No	**30.76**	33.05	46.67
Dresses	Yes	**102.09**	106.93	-
SIT Student	No	**253.51**	254.38	372.97

ข้อมูล SIT Student เป็นข้อมูลที่ไม่ได้มาจาก Public Domain ผลลัพธ์จากการจัดกลุ่มอาจมีความหมายและเป็นประโยชน์ต่อองค์กร รายละเอียดของผลลัพธ์ที่ได้จากข้อมูลชุดนี้แสดงไว้ใน [29] (ภาคผนวก ก)

4.4 ข้อสังเขป

การปรับตัวเพื่อพัฒนา (Adaptability) ของ GA ในระบบที่พัฒนาขึ้นช่วยให้ผลลัพธ์การจัดกลุ่มข้อมูลบนชุดข้อมูลจำนวน 12 ชุดที่ใช้ในโครงการวิจัยนี้มีประสิทธิภาพมากขึ้น จากการสังเกตพฤติกรรมการทำงานของแต่ละ Mode พบว่าการพัฒนาค่าฟิตเนสและเวลาในการทำงานส่วนใหญ่ของระบบอยู่ใน Mode1 เนื่องจากเป็น Mode แรกในการทำงานและเป็นพฤติกรรมทั่วไปของการเลือกแบบ Elitism ที่โครโมโซมที่ดีจะถูกเก็บไว้ ทำให้ในโครโมโซมที่ดีได้มีโอกาสพัฒนาในรุ่นถัดไป ช่วงแรกของการทำงานใน Mode1 นี้ค่าฟิตเนสพัฒนาขึ้นอย่างรวดเร็วและต่อเนื่องจากนั้นจึงค่อยๆ ลดระดับอัตราการพัฒนาลงไปทีละน้อยจนกระทั่งไม่สามารถพัฒนาค่าฟิตเนสขึ้นได้อีกแล้วจึงเปลี่ยนการทำงานไปยัง Mode2 ใน Mode นี้ค่าฟิตเนสมีการพัฒนาขึ้นบ้างเล็กน้อยเนื่องจากการพัฒนาใน Mode1 อาจใกล้เคียงค่า Optimal กล่าวคือการหาคำตอบที่ดีขึ้นอาจอยู่ในลักษณะ Local optimal หรือ Plateau หรือ Ridge การเลือกสรรแบบ Roulette Wheel ที่ใช้ใน Mode2 นี้มีพฤติกรรมที่ใกล้เคียงกับพฤติกรรมการเลือกสรรแบบ Elitism ของ Mode1 ที่ต่างก็เปิดโอกาสให้โครโมโซมที่ดีมีโอกาสถูกเลือกให้พัฒนาต่อในรุ่นถัดไป จึงอาจเป็นเหตุในการพัฒนาคำตอบเกิดขึ้นได้ไม่ดีนักโดยเฉพาะเมื่อใช้กับข้อมูลทั้ง 12 ชุดในงานวิจัยนี้ เมื่อการทำงานของระบบเปลี่ยนเป็น Mode3 ที่เปรียบเสมือนการ Reset ส่วนใหญ่พบว่ามีการพัฒนาค่าฟิตเนสขึ้นมากกว่า Mode2 การทำงานใน Mode3 นี้ใช้เวลาไม่นานเท่ากับ Mode1 ซึ่งอาจเป็นเพราะคุณสมบัติการเลือกแบบสุ่ม (Random Selection) ที่เหมาะกับการแก้ปัญหากรณี Local optimal และ Plateau อย่างไรก็ตามเมื่อปัญหาเหล่านี้ถูกแก้ไขแล้วการเลือกสรรแบบสุ่มอาจไม่เหมาะกับข้อมูลทั้ง 12 ชุดนี้อีกต่อไป หลังจากที่ระบบเปลี่ยนกลับมาทำงานใน Mode1 ซึ่งเป็น Mode การทำงานสุดท้ายของระบบ พบว่าระบบสามารถพัฒนาค่าฟิตเนสต่อได้อีกระยะหนึ่ง ทั้งนี้เนื่องจากระบบถูก Reset มาจาก Mode 3 นั่นเอง

ข้อจำกัดหนึ่งของระบบที่พัฒนาขึ้นคือเวลาที่ใช้ในการทำงานที่มากกว่าเวลาที่ใช้ใน PAM และ SOM หลายเท่าตัว เช่น การจัดกลุ่มข้อมูลนักศึกษาจำนวน 2,070 ตัวอย่างที่ใช้ในโครงการวิจัยนี้ ระบบที่พัฒนาขึ้นใช้เวลาในการทำงานประมาณ 10 ชั่วโมง ในขณะที่ PAM และ SOM ใช้เวลาเพียง 1 – 2 นาทีเท่านั้น ทั้งนี้เนื่องจากแนวคิดของ Diana ที่ต้องคำนวณค่าเฉลี่ย Distance ใหม่ทุกครั้งที่มีข้อมูลย้ายจาก Main Group ไป Splinter Group และการทำงานของ GA ต้องจัดกลุ่มข้อมูลใหม่ทุกครั้งในแต่ละ Generation สำหรับทุกๆ โครโมโซม นอกจากนี้เงื่อนไขการหยุดการทำงานของ GA ยังเป็นอีกปัจจัยหนึ่งที่ทำให้การทำงานของระบบใช้เวลานานมากกว่า PAM และ SOM กล่าวคือระบบจะหยุดการ

ทำงานก็ต่อเมื่อไม่สามารถพัฒนาค่าฟิตเนสที่ดีขึ้นได้ ดังนั้นแม้มีการพัฒนาค่าฟิตเนสขึ้นเพียง 0.001% ก็ยังถือว่าการพัฒนาเกิดขึ้น

ข้อจำกัดทางด้านเวลานี้สามารถลดลงได้โดยนิยามการพัฒนาใหม่ โดยอาจกำหนดให้ค่าฟิตเนสใหม่ที่หาได้ต้องมากกว่า 1% จึงถือว่ามีการพัฒนาขึ้น การเปลี่ยนนิยามใหม่ในแนวทางนี้จะทำให้ระบบเร็วหยุดทำงานเร็วขึ้นกว่าปัจจุบันมาก การกำหนดกลุ่มของสมาชิกในการคำนวณหาค่าเฉลี่ยระหว่าง Main Group และ Splinter Groups เพียงครั้งแรกครั้งเดียวแทนการคำนวณในแต่ละ Generation ก็จะสามารถลดเวลาการคำนวณไปได้อย่างมากเช่นกัน

อย่างไรก็ดีดังที่ได้กล่าวไว้แล้วในหัวข้อที่ 1.2 โครงการวิจัยนี้อยู่บนสมมติฐานว่าเวลาที่ใช้ในการประมวลผลเพื่อจัดกลุ่มข้อมูลไม่ใช่เป็นปัจจัยหลักสำคัญ ดังนั้นเวลาที่ใช้ในการทำงานของระบบที่มากกว่า PAM และ SOM หลายเท่าตัวจึงไม่เป็นอุปสรรคในโครงการวิจัยนี้

บทที่ 5 บทสรุป

เนื้อหาบทนี้กล่าวถึง ประโยชน์ที่ได้รับจากโครงการวิจัย สิ่งที่ได้เรียนรู้จากโครงการวิจัย แนวทางการพัฒนาในอนาคต และสรุปผลโครงการวิจัย

5.1 ประโยชน์ที่ได้รับจากโครงการวิจัย

ประโยชน์ที่ได้รับจากโครงการวิจัยนี้สามารถสรุปได้ ดังนี้

1. คุณลักษณะของผู้สมัครและนักศึกษาในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี ระบบที่พัฒนาขึ้นได้จำแนกข้อมูลนักศึกษาดังที่กล่าวไว้ออกเป็น 5 กลุ่ม ผลจากการจำแนกข้อมูลนี้ทำให้สามารถทราบถึงคุณลักษณะของแต่ละกลุ่มได้
2. โปรแกรมการจัดกลุ่มข้อมูลใหม่ที่พัฒนาขึ้นโดยนำเทคนิค Diana และ GA มาประยุกต์ใช้มีความสามารถในการจัดกลุ่มข้อมูลเชิงจำนวน, เชิงลักษณะและข้อมูลที่ประกอบด้วยเชิงจำนวนและเชิงลักษณะได้ นอกจากนั้นแล้วระบบที่พัฒนาขึ้นยังมีประสิทธิภาพในการจัดกลุ่มข้อมูลดีกว่าเทคนิคที่เป็นที่รู้จักทั่วไป โดยงานวิจัยนี้เลือกเปรียบเทียบกับ PAM และ SOM
3. GA Unit ในโปรแกรมที่พัฒนาขึ้นมาใหม่มีความสามารถในการการปรับสภาพตัวเอง (Adaptive) ให้สามารถเลือกแนวทางการ Crossover และแนวทางการ Selection ที่เหมาะสมในแต่ละรุ่น (Generation) ได้ คุณสมบัตินี้ช่วยให้การพัฒนาผลลัพธ์ของ GA สามารถเลี่ยงปัญหาที่รู้จักกันในการค้นหา กล่าวคือเมื่อผลลัพธ์ติดอยู่ใน Local maxima และ Plateau

5.2 สิ่งที่ได้เรียนรู้จากโครงการวิจัย

จากการศึกษาและวิจัยในโครงการวิจัยนี้ สามารถสรุปสิ่งที่ได้เรียนรู้พอสังเขปได้ ดังนี้

1. ความยาวของโครโมโซมและจำนวนประชากรที่เหมาะสมของ GA เป็นสิ่งที่ต้องพิจารณาระหว่างการพัฒนาโปรแกรม การกำหนดความยาวของโครโมโซมและจำนวนประชากรที่มากหรือน้อยเกินไปไม่สามารถรับประกันประสิทธิภาพการจัดกลุ่มได้ ในงานวิจัยนี้ได้กำหนดความยาวของโครโมโซมดังสมการที่ 3.5 และการกำหนดขนาดของประชากรในแต่ละ Generation ดังสมการที่ 3.6

ให้สอดคล้องกับขนาดของข้อมูลและจำนวนของกลุ่ม แนวทางนี้อาจเป็นวิธีการหนึ่งที่เหมาะสมและให้ผลการประเมินประสิทธิภาพการจัดกลุ่มข้อมูลเป็นที่น่าพอใจ

2. เนื่องจากโคโมโซมที่ใช้ในโครงการวิจัยนี้เป็นแบบ Medroid based ในระหว่างขั้นตอนการ Crossover และ Mutation แต่ละ Generation ของ GA อาจทำให้เกิดโครโมโซมที่มีสมาชิกซ้ำกัน เช่น [1, **5**, 6, 7, **5**, 2] ซึ่งโครโมโซมลักษณะนี้ทำให้เกิดผลลัพธ์การจัดกลุ่มแบบคาบเกี่ยวกัน (Overlap) การจัดการกับโครโมโซมลักษณะนี้ทำได้โดยการสร้างโครโมโซมใหม่ขึ้นมาเพื่อทดแทนโครโมโซมที่ใช้การไม่ได้นี้ อย่างไรก็ดีการสร้างโครโมโซมตัวแทนของกลุ่มซ้ำนี้อาจจะเป็นสิ่งที่ดีและสามารถนำไปประยุกต์ใช้ในการจัดกลุ่มแบบ Overlapping

5.3 แนวทางการพัฒนาในอนาคต

โครงการวิจัยนี้สามารถพัฒนาต่อในหลายๆ ด้าน ดังนี้

1. ในการพัฒนาส่วน Fitness Function ของ GA งานวิจัยนี้ใช้ค่า Inter – Intra Distance เท่านั้น Fitness Function อาจพัฒนาให้อยู่บนเกณฑ์ (Clustering Criteria) อื่นๆ ระบบ GA อาจใช้ Fitness Function มากกว่าหนึ่งและแต่ละรุ่นโครโมโซมที่ให้ค่า Fitness Function ดีที่มากกว่า 1 เกณฑ์จะถูกเก็บไว้ แนวทางนี้จะเป็นการนำเกณฑ์ต่างๆ ในการจัดกลุ่มข้อมูลมาใช้และให้โคโมโซมที่มีค่าเฉลี่ยที่ดีที่สุดเป็นผลลัพธ์ที่ดีที่สุด

2. วิธีการกำหนดกลุ่มให้กับสมาชิกจากเดิมที่ในงานวิจัยนี้ใช้แนวคิดของ Diana ที่ต้องคำนวณหาค่าเฉลี่ยระหว่าง Main Group และ Splinter Groups ใหม่ทุกครั้งเมื่อมีสมาชิกถูกย้ายออกจาก Main Group การทดลองสามารถทำได้โดยการกำหนดกลุ่มของสมาชิกในการคำนวณหาค่าเฉลี่ยระหว่าง Main Group และ Splinter Groups เพียงครั้งแรกครั้งเดียว แนวทางนี้อาจจะให้ผลดีเพียงพอกับแนวทางปัจจุบัน แต่เวลาในการคำนวณและประเมินผลจะน้อยลงมาก

3. ระบบปัจจุบันควรมีการทดลองกับชุดข้อมูลที่หลากหลายมากขึ้น โดยอาจเปลี่ยน Similarity Measure จากเดิมที่ใช้ Gower's similarity measure และ Euclidean Distance เป็นวิธีการอื่นที่มีผู้นำเสนอมาก่อนแล้ว

4. ศึกษาพฤติกรรมการทำงานแต่ละ Mode ในการปรับตัวเพื่อพัฒนาของ GA เนื่องจากผลลัพธ์ที่ได้อาจบอกถึงแนวทางเพิ่มประสิทธิภาพของระบบ อย่างไรก็ตามการศึกษาต่อในแนวทางนี้อาจต้องคำนึงถึงธรรมชาติของ Fitness Function ที่ใช้และเทคนิคของ Diana ในการจัดกลุ่มควบคู่กันไปด้วย

5. ในการวิจัยนี้จำนวนกลุ่มที่เหมาะสมได้จากการกำหนดจากการใช้ Silhouettes Index ก่อนที่จะเข้า GA Unit การพัฒนาระบบการจัดกลุ่มข้อมูลที่สามารถคำนวณจำนวนกลุ่มที่เหมาะสมควบคู่ไปด้วยกับการจัดกลุ่มข้อมูลจะเป็นการพัฒนาระบบให้มีประสิทธิภาพมากยิ่งขึ้น

5.4 สรุปผลโครงการวิจัย

การนำเทคนิคของ GA มาประยุกต์ใช้ในงานด้านการจัดกลุ่มข้อมูลนั้นมีให้เห็นอยู่หลายรูปแบบ อย่างไรก็ตามยังไม่มีงานใดที่ประยุกต์ใช้เทคนิคของ GA ร่วมกับเทคนิคการจัดกลุ่มของ Diana (การจัดกลุ่มแบบ Hierarchical) แล้วให้ผลลัพธ์การจัดกลุ่มเป็นแบบ Patitional ได้ ดังนั้นโครงการวิจัยนี้จึงมุ่นเน้นการพัฒนาโปรแกรมการจัดกลุ่มข้อมูลแบบ Patitional ใหม่ที่มีประสิทธิภาพมากกว่าเทคนิคที่เป็นที่นิยม (PAM และ SOM) ระบบที่พัฒนาขึ้นได้ถูกประเมินประสิทธิภาพโดยใช้ข้อมูลประเภท Labeled จำนวน 6 ชุดข้อมูลและข้อมูลประเภท Unlabeled จำนวน 6 ชุดข้อมูล นอกจากผลลัพธ์ที่ได้จากงานวิจัยนี้ระบบการจัดกลุ่มข้อมูลบนพื้นฐานของ GA, Gower's Similarity Measure และ Diana แล้ว งานวิจัยนี้ยังเป็นจุดเริ่มต้นของการพัฒนาต่อในหลายๆ มิติเกี่ยวกับการหา Fitness Function, การประมวลผลที่เร็วขึ้นและการกำหนดจำนวนกลุ่มที่เหมาะสม

เอกสารอ้างอิง

[1.] Everitt, B.S.,Landau, S., Lee M.and Stahl D, 2009, **Cluster Analysis**, 5th Edition, John Wiley & Sons, London, U.K.

[2.] L. J. Arabie, G. Hubert and P. DeSoete, 1999, **Clustering and Classification**, 2nd ed. ,World Scientific, Singapore.

[3.] A. K. Jain and R. C. Dubes, 1988, **Algorithms for Clustering Data**, Prentice Hall, New Jersey, USA.

[4.] Chakrapani, C., 2004, **Statistics in Market Research**, Arnold, London

[5.] Hodson, F. R., 1971, "Numerical typology and prehistoric archaeology", In **Mathematics in the Archaeological and Historical Sciences**, Edinburgh University Press, Edinburgh, pp.30–45

[6.] Eiben, A.E and Smith, J.E, 2007, **Introduction to evolutionary computing**, Springer, Berlin

[7.] กิตติชัย ลวันยานนท์ และเอกรัฐ รัฐกาญจน์, 2546, "Adaptive Genetic Programming for Inductive Learning", **National Computer Science and Engineering Conference (NCSEC 2003)**, 28-30 ตุลาคม, มหาวิยาลัยบูรพา

[8.] Lavangnananda, K., 2006, "Self-adjusting Associative Rules Generator for Classification : An Evolutionary Computation Approach", **Proc. of 2006 IEEE Mountain Workshop on Adaptive and Learning Systems (SMCals/06)**, 24-26 July, Utah, USA., PP. 237 - 242

[9.] Hruschka, E. R., Campello, R. J., Freitas, A., and De Carvalho, C. P., 2009, "A Survey of Evolutionary Algorithms for Clustering", **IEEE Transactions on Systems, Man, and Cybernetics**, Part C, Vol. 39, Issue 2, pp. 1-26.

[10.] Krovi, R., "Genetic Algorithms for Clustering: A Preliminary Investigation", **In Proc. of the 25th Hawaii Int. Conference**.

[11.] Murthy, C. A. and Chowdhury, N., 1996 , “In Search of Optimal Clusters using Genetic Algorithms”, **Pattern Recognition Letters**, Vol. 17, pp. 825-832.

[12.] Krishna, K. and Murty, N., “Genetic K-means Algorithm**”, IEEE Trans. on Systems, Man and Cybernetics – Pt. B**, Vol. 29, pp. 433-439, 1999.

[13.] Lu, Y., Lu, S., Fotouhi, F., Deng, Y. and S. J. Brown, 2004, “FGKA: A Fast Genetic K-means Clustering Algorithm”, In **Proc.ACM Symposium on Applied Computing**, 14 -17 March, 2004, NY, USA. pp. 622-623.

[14.] Fränti, P., Kivijärvi, J., Kaukoranta, T. and Nevalainen, O., 1997 , “Genetic Algorithms for Large-Scale Clustering Problems”, **The Computer Journal**, Vol. 40, pp. 547-554.

[15.] Kivijärvi, J., Fränti, P. and Nevalainen, O., 2003 , “Self-Adaptive Genetic Algorithm for Clustering”, **Journal of Heuristics**, Vol. 9, pp. 113-129.

[16.] Maulik, U. and Bandyopadhyay, S., 2000, “Genetic Algorithmbased Clustering Technique”, **Pattern Recognition**, Vol. 33, pp. 1455-1465.

[17.] Kuncheva, L. I. and J. C. Bezdek, 1997, “Selection of Cluster Prototypes from Data by a Genetic Algorithm”, In **Proc. 5th European Congress on Intelligent Techniques and Soft Computing**, pp. 1683-1688.

[18.] Sheng, W. and Liu, X., 2004, “A Hybrid Algorithm for K-Medoid Clustering of Large Data Sets”, In **Proc. IEEE Congress on Evolutionary Computation**, pp. 77-82.

[19.] Lucasius, C.B., Dane, A.D. and Kateman, G., 1993, "On k-medoid clustering of large data sets with the aid of a genetic algorithm: background, feasiblity and comparison", **Analytica Chimica Acta**, Volume 282, Issue 3, 29 October 1993, Pages 647–669

[20.] Estivill-Casuo, V., and Murray, A.T., 1998, "Spatial Clustering for Data Mining with Genetic Algorithms", **Proceedings of the International ICSC Symposium on Engineering of Intelligent Systems (1998)**

[21.] Kaufman, L. and Rousseeuw, P. J., 1990, “Finding Groups in Data: An Introduction to Cluster Analysis”, Wiley series in probability and mathematical statistics, Wiley.

[22.] Tseng,L. Y. and Yang S. B.,2001, “A Genetic Approach to the Automatic Clustering Problem”, **Pattern Recognition**, Vol. 34, pp. 415-424.

[23.] Calinski, R. B., and Harabasz, J., 1974 , “A dendrite method for cluster analysis”, **Communications in Statistics**, Vol. 3, pp. 1-27.

[24.] Duda, R. O, and Hart, P. E., 1973, Pattern classification and scene analysis, New York: Wiley.

[25.] Rousseeuw P.J. Silhouettes, 1987, “A graphical aid to the interpretation and validation of cluster analysis”, **Journal of Computational and Applied Mathematics**, Vol. 20, pp. 53–65

[26.] Bache, K. and Lichman, M., 2013, **UCI Machine Learning Repository** [Online], Available: http://archive.ics.uci.edu/ml [Oct. 2014]

[27.] Alcalá-Fdez, J., Fernandez,A., Luengo, J., Derrac, J.,García, S., Sánchez, L. and Herrera, F., 2011, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms

and Experimental Analysis Framework" **Journal of Multiple-Valued Logic and Soft Computing**, Vol.17, pp. 255-287.

[28.] McClain, J. O. and Rao, V. R., 1975 “Clustisz: A program to test for the quality of clustering of a set of objects.” **Journal of Marketing Research**, Vol. 12, pp. 56–460.

[29.] Lavangnananda, K. and Poolphol R., 2014, "A Genetic Algorithm Approach to Partitioning Clustering: A case study on M.Sc. applicants", **The 13th International Conference on Machine Learning and Applications (ICMLA'14)**, 3-6 December 2014, Detroit, MI USA.

ภาคผนวก ก

A Genetic Algorithm Approach to Partitioning Clustering:

A case study on M.Sc. applicants

A Genetic Algorithm Approach to Partitioning Clustering : A case study on M.Sc. applicants

Kittichai Lavangnananda[1] and Ratipong Poolphol[2]

Data and Knowledge Engineering Lab., School of Information Technology (SIT)
King Mongkut's University of Technology Thonburi (KMUTT)
Bangkok, Thailand
[1]Kitt@sit.kmutt.ac.th [2]ratipong.pool@gmail.com

Abstract—Acquiring a Master Degree is becoming a common practice to ensure successful life and good career path, especially in developing countries. Master Degree in Information Technology is one of the most popular programmes with prolific number of applications and students. This work has two main objectives. First is to discover the number of clusters of applicants and the characteristics of each cluster. Another is to develop a Genetic Algorithm based Partitioning Clustering Program. This is achieved by incorporating distance matrix and its application in Divisive Analysis and Gower's measure of similarity. The Genetic Algorithm based Partitioning Clustering program developed was proven superior to some common clustering techniques.

Keywords—Clustering, Divisive Analysis, Genetic Algorithms, *Gower's Measure of Similarity, Master Degree in Information* Technology, Partitioning Clustering

I. INTRODUCTION

Education is the key to the successful life and career path, especially in developing countries. In 2013, the Thai's Office of the Higher Education Commission reported that there are more than 2 million students in the country currently studying in higher education. Almost 10 percent of these are in the Master Degree programmes and this trend is likely to increase. One of the most popular Master Degree programme is the M.Sc. in Information Technology. School of Information Technology (SIT), King Mongkut's University of Technology Thonburi (KMUTT), has the highest number of students in this programme in the country. Nevertheless, apart from observation, there has never been an attempt to discover how many types of applicants/students are interested in pursuing the programme and characteristics that each type possesses.

This work has two main objectives. First is to answer the above queries and second is to develop a partitioning clustering program which incorporates Genetic Algorithm (GA) in its process. The program should also offer an advancement in clustering approach and also yields superior performance to commonly known clustering techniques.

II. CLUSTERING

Clustering is one of the well known tasks in Data Mining and Knowledge Discovery. Its definition is multifarious, from abstract to mathematical oriented descriptions. In its simple terms, clustering can be defined as dividing a dataset into an intended number of groups whereby intra-similarity is maximized while inter-similarity is minimized. [1],[2]. Clustering can be considered as NP class problem and, therefore, several techniques have been invented for this with various degrees of success and it can be domain dependent too. From machine learning perspective, clustering is an unsupervised learning method. Since clustering can also be seen from several perspectives, leaders in the machine learning had categorized clustering into various types depending on different perspectives. One perspective is from its outcome, which can be described into three categories, partitioning, overlapping and hierarchical [2]. A partition of a dataset $X=\{x_1,x_2, \ldots,x_N\}$, where N is number of samples in data set and x_j (j=1, …, N) stands for an n-attribute vector, is a collection $C=\{C_1,C_2, \ldots, C_k\}$, therefore, $C_1 \cup C_2 \cup \ldots \cup C_k = X$ is :

Partitioning Clustering, if $C_i \cap C_j = \emptyset$ for $i \neq j$. This is often referred to as hard partitioning.

Overlapping Clustering, if $C_i \cap C_j \neq \emptyset$ for $i \neq j$. This is often referred to as soft partitioning where each object may fully belong to one or more clusters.

Hierarchical Clustering, if partitioning cluster is a nested sequence of partition, each of which presents a hard partition of the data set into a set of disjoint clusters.

Applications of clustering are plentiful ranging from finance (such as determining different types of investors) to bioinformatics (such as finding number of suitable groups for DNS sequences). Clustering presents an interesting challenge to Data Mining, Knowledge Discovery and Machine Learning communities and a lot is yet to be discovered.

III. GOWER'S SIMILARITY MEASURE AND DIVISIVE ANALYSIS

As Divisive Analysis and Gower's similarity measure are utilized in the GA-based clustering program in this work, therefore, they merit a short description of each method for ease of understanding of the later Sections which describe the work in more detail.

Among several methods to construct proximities for mixed-mode data, Gower's similarity measure offers simplicity and ease of computation. In this method these proximities are known as similarity matrix. For categorical and binary attributes, a component in the matrix is given by :

$$S_{ij} = \sum_{k=1}^{p} w_{ijk} S_{ijk} / \sum_{k=1}^{p} w_{ijk} \quad (1)$$

where S_{ijk} is the similarity between the ith and jth individual as measured by the kth variable, and w_{ijk} is set to zero if the outcome of the kth variable is missing for either or both individuals i and j. S_{ijk} is one when two both individuals have the same value. For continuous variable this similarity measure is given by :

$$S_{ij} = 1 - |x_{ik} - x_{jk}|/R_k \quad (2)$$

where R_k is the range of observations for the kth variable. (i.e. the city block distance is used after scaling the kth variable to unit range.). A detail determination and application of Gower's similarity measure can be found in [3].

Divisive Analysis is a clustering method. In this method, an individual which is different from others splits out to form a new group and some others may follow. First, a distance matrix (also known as dissimilarity matrix) of a kind is constructed, the one which average distance from other individuals is highest initiates the splinter group. Next, the average distance of each individual in the main group to the individual(s) in the splinter group is determined. This may cause individual(s) to shift to a new group of create another new group altogether. This process continues until equilibrium is attained. A detail description of Divisive Analysis can be found in [4].

IV. Evolutionary Algorithms for Clustering

In the recent years, there have been a number of researches which propose clustering techniques that based on Evolutionary Algorithms (EA). There exist EA-based algorithms for all types of data. A comprehensive survey in this topic is well documented in [5]. In this section, previous EA-based clustering techniques are briefly summarized in terms of chromosome encoding scheme, as it is an aspect which can clearly differentiate different approaches and their intended applications. Chromosome encoding scheme can be categorized into three types as follows:

A. Label based Representation

In this representation, chromosome is encoded into a string length of N, where N is the number of samples in a data set. Each position in the chromosome represents a particular sample in a dataset. Each bit has a value from 1 to K, where K is cluster number. For example, if a dataset of 10 samples is to be clustered into 3 clusters and the cluster solution is the first four samples in a dataset are grouped into cluster 1, the next four samples are grouped into cluster 2 and the rest are grouped into cluster 3, so, the chromosome representation of this solution is [1111222233]. This chromosome encoding scheme is adopted in [6], [7], [8] and [9].

B. Centroid based Representation

In this representation, the value of each is position of chromosome is real number which represents a coordinate of cluster prototype and length of chromosome is n × K, where n is the number of attributes in dataset and K is the number of clusters. For example, if a 2-dimention numerical dataset is to be clustered into 2 clusters, a chromosome [5.0, 1.5, 10.5, 1.5, 5.0, 5.5] represents three coordinates of cluster prototype by [5.0, 1.5] and [10.5, 1.5] are cluster 1 and cluster 2 prototypes, respectively. The works in [10], [11] and [12] have adopted this chromosome encoding scheme and cluster results were found to be better than the results from traditional algorithms.

C. Medroid based Representation

In Medroid based representation, each position in a chromosome represents cluster prototype and the value of each position is an integer from 1 to N, where N is the number of samples in dataset. For example, if a dataset is to be clustered into 2 clusters, a chromosome [3, 5, 7, 9] represents that sample 3 and 5 are prototypes of cluster 1 while 7 and 9 are prototypes of cluster 2. This chromosome encoding scheme was adopted in [13] to select prototypes for hard c-partitioning of unlabeled data. The results showed that GA based clustering algorithm was superior than Hard C-Means and Random Search in terms of within-group square distance criterion.

To date, there has not been an EA-based approach to partitioning clustering which utilizes in Divisive Analysis and Gower's similarity measure in its chromosome encoding and manipulation of genetic operators, that is also capable of dealing with both numerical and categorical attributes.

V. Dataset

As mentioned in Section 1, one of the main objectives of this work is to find characteristics of applicants and number of types there are among applicants to M.Sc. programme. Hence, there must be past applicants available for the GA-based clustering program developed. Dataset in this work comprises 2070 applicant records from 1995 to 2013. Each record contains many attributes, however, not all attributes may be representative of any cluster which may result in. For examples, intuitively, telephone number and citizen ID number should be disregarded. Therefore, preprocessing in this work is to remove all attributes which are considered irrelevant to the outcome and to carefully consider those which are. After careful consideration of all attributes available in a record, eight are considered relevant to the clustering. Their name, type and value are shown in Table I.

TABLE I. Attributes in the Dataset

	Name	Data Type	Value
1	Gender	Categorical	Male; Female
2	Age	Integer	Range from 21 to 51
3	Marital status	Categorical	Single; Married; Divorced
4	IT related in Bachelor degree	Categorical	Yes; No
5	Bachelor degree's major	Categorical	7 different categories
6	GPA in Bachelor degree	Real	From 2.00 to 4.00
7	Type of Institutes in Bachelor degree	Categorical	6 different categories
8	Employment Status	Categorical	Employed; Unemployed

VI. Adaptive Genetic Algorithm Approach to Partitioning Clustering

The GA-based Partitioning Cluster program developed is adaptable. Its adaptability is described in the sub-Section VI.b. The main GA unit of the program developed follows the conventional GA operations, with chromosomes generation being the most crucial operation. The number of clusters (k) should be known prior to clustering. The sub-Section VI.c explains the determination of the suitable number for k. The main components of the GA-based Partitioning Clustering Program and their operations are described in the following sub-Section.

A. Components and their Operations

The program developed comprises the following :

1) Number of Population: A good number of population (i.e. chromosomes) in each generation has to be a good compromise between generating good variety and reasonable computation time. The number of population in the GA unit is proportional to number of clusters (k), and number of samples available (N). This is set to be 10% of N, but no more than 200 if N is greater than 2,000.

2) Chromosomes Generation ; Chromosome encoding is Medriod based (as described in Section IV, therefore, each value in a chromosome is a cluster prototype). Once a chromosome is generated, the distance/dissimilarity matrix is, as used in Divisive Analysis, is created. This matrix records the dissimilarity values. A component in the matrix adapts Gower's similarity measure, but it is the complement of the Gower's similarity value (i.e. 1 - S_{ij} as described in Section III, since S_{ij} is the similarity value). Hence, a distance matrix is created for every chromosome.

3) Fitness Function ; This value indicates how well the clustering represent by that chromosome is. Several methods have been suggested to evaluate the clustering, this works employs one of the most popular method and is also used in [14]. As distance matrix is now available, determination of fitness value follows the same steps as in Divisive Analysis. Hence, the fitness value is determined by the following equation.

$$Fitness = \sum_{i=1}^{k} (D_{inter}(C_i) - D_{Intra}(C_i)) \quad (3)$$

where $D_{intra}(C_i)$ is the intra-distance of cluster C_i and $D_{inter}(Ci)$ is the inter-distance between C_i and other clusters. Therefore, the fitness function in this work is actually the measure of quality of the result from clustering too.

4) Selection ; Three methods of GA selection operators are employed. These are, Elitism, Roulette Wheel and Random, their applications are discussed in sub-Section VI.b.

5) Crossover ; The program developed employed both 1-point and 2-point crossover methods, their applications are discussed in sub-Section VI.b.

6) Mutation ; Random mutation is adopted in this work.

7) Termination criterion ; The program developed incorporates an ability to change the operating mode, this is discussed in sub-Section VI.b. The process terminates when no improvement occurs after four changes of operating mode.

Several trails had been carried out to find suitable rate for each GA operators until satisfactory ones were found. Table II summarizes the parametric values and methods used for the GA based Partitioning Clustering program developed.

TABLE II. GA Parametric

GA Operator	Selection Methods		
	Elitism	**Roulette Wheel**	**Random**
Crossover Rate	90 %	95 %	95 %
Mutation Rate	10 %	10 %	10 %
Reproduction Rate	5 %	5 %	5 %
Duplication Rate	5%	-	-
Number of Populations	10% of samples (N,) but no more than 200 if N is greater than 2,000		

B. Adaptability in the GA-based Partitioning Clustering Program

As mentioned earlier, the GA-based Partitioning Clustering program developed is adaptable. The adaptability in this work is the ability to change its operating mode (i.e. the crossover and selection operations). This adaptability is the attempt to overcome situations where the GA process encounters the commonly known problems in searching, such as local optimal, plateau and ridge. Adaptability in this work comprises 3 modes. The process starts with Mode1 and, at any instance, the current operating mode is altered if no progress (i.e. fitter chromosome) is found within 50 generations. The circular switching of operating mode is : Mode1 => Mode2 => Mode3 then it goes back to the initial Mode1 and so on. Table III shows the selection and crossover operations in each operating mode.

TABLE III. The Three Operating Modes in the GA-based Clustering Program

	Mode1	**Mode2**	**Mode3**
Selection	Elitism	Roulette Wheel	Random
Crossover	1-Point	1-Point	2-Point

C. Determination of the Suitable Number of Clusters

In clustering, unless the number of clusters (k) is specified prior to clustering, there is no way of knowing what the best value of k should be. Many have suggested various methods to determine the best value for k [15], [16]. Each requires some kind of calculation which is similar to a shorter version of clustering. Nevertheless, none is proven optimal in all circumstances. This work adopts the Silhouettes Index [17] for its computation simplicity.

Prior to clustering the dataset for this work, Silhouettes Index was determined and the result was five. Therefore, the target number of clusters (k) was taken to be 5.

Figure 1 depicts the overview process of the GA-based Partitioning Clustering Program

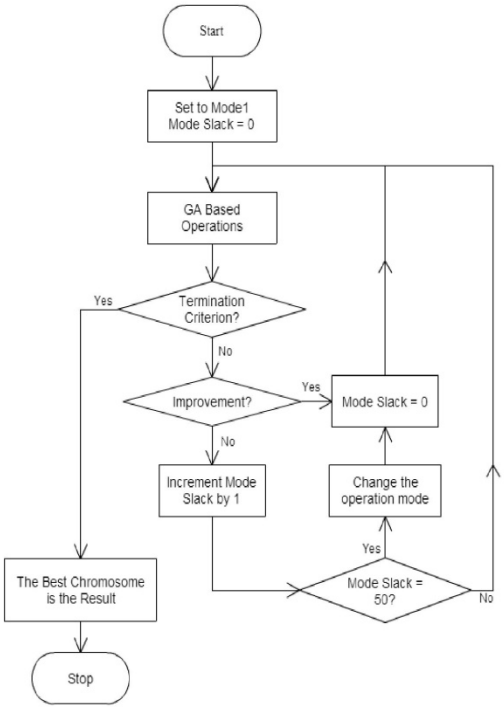


Fig. 1. Overview Process of the GA-based Partitioning Clustering Program

VII. Result and its Interpretation

The GA-based Partitioning Clustering program was applied to cluster the dataset. Several trails had been carried out to guarantee validity of the result. The best result and its comparison to other techniques were carried out and is shown in Table V. Many researchers in this field have suggested several measures of quality of the clustering results. This aspect is discussed and the comparison is shown in Section VIII.

Characteristic(s) of each cluster was determined by analyzing which attribute(s) is(are) most common among all samples in that cluster. Finally, Accuracy of each cluster is determined. This is the percentage of retrieved samples in the cluster which matches the determined characteristic(s). Table IV reveals the result of the clustering.

TABLE IV. Result from the GA-based Clustering Program

Cluster	Characteristic(s)	Number of samples	Accuracy (%)
1	(Female) AND (Employed)	472	100
2	(Bachelor Deg. in Non-IT related) AND (Employed)	383	85
3	(Bachelor Deg. from Public Uni.) AND (Unemployed)	395	78
4	(Male) AND (Bachelor Degree in IT related) AND (Employed)	485	100
5	(Bachelor Deg. in Administration) AND ((Bachelor Deg. from Private Uni.) OR (Bachelor Deg. from Rajamangala Inst.))	335	73

As M.Sc. programme under this study is operated outside working hours. Employment status is an intuitive characteristic and this works affirms this. Finding of other characteristics in each cluster is useful to management of the programme and its future sustainability. However, the implication of the finding in this work is constrained by the working culture, educational practice and the environment of the country. Hence, further analysis and any new measures that ought to be implemented from the result of the finding may not be applicable to global educational community at large but may be worthy of discussion in other forums (e.g. in the social educational field of research).

VIII. Comparison with some well known Clustering Techniques

There are plentiful of clustering techniques, some are available in a public domain software package such as Weka [18]. It was also the objective of this work to produce a GA-based Partitioning Clustering program whose performance must at least be as good as commonly known techniques. While comparison may include all kind of aspects such as computation time, memory requirement, etc. They are not of the main concern and, therefore beyond the scope of this work, the quality of each cluster the program developed produces is the prime objective.

A well known clustering k-means analysis is inappropriate for the comparison. While it can be adapted to cluster mixed data type, it is more suitable for numerical data. Two different type of clustering techniques were selected for comparison, these are Partition Around Medriods (PAM) and Neural

Network. PAM was selected as it shares similarity to this work in employing distance matrix. It is a well known clustering technique which is applicable to mixed data type and was reported to be more robust than k-means analysis [4]. Self Organizing Map (SOM) was the selected neural network architect for the comparison as its application in partitioning clustering is prolific. Both PAM and SOM are publicly available.

There are several measures of quality (i.e. how good the result is) for clustering available in the literatures [19]. Values which are used to determine these measures usually involve inter-distance and intra-distance value. The most popular one is probably the 'inter-distance – intra-distance' value (higher value indicates better performance). This work compares the results from SOM, PAM and the program developed using three quality of measures, the popular 'inter-distance – intra-distance', the McClain-Rao Index (lower value indicates better performance) and the Lack of homogeneity Index (lower value indicates better performance). Determination of these indices can be found in [20] and [21].

Several trails were performed on the dataset, with attempts to adjust for their suitable parametric values, for both PAM and SOM to ensure validity of comparison. The best result of each technique was taken. Table V shows the comparison among SOM, PAM and the GA-based Partitioning Clustering program developed.

TABLE V. COMPARISON AMONG THE THREE CLUSTERING TECHNIQUES

Clustering Technique	Inter – Intra Distance	McClain-Rao	Lack of homogeneity
Self Organizing Map (SOM)	874,436.49	0.91	372.97
Partition Around Mediods (PAM)	1,190,689.83	0.57	254.28
The GA-Based Partioning Clustering	1,202,798.50	0.57	253.51

As shown in Table V, in overall, the GA-Based Partitioning Clustering Program developed is superior in performance to the two well known clustering techniques for this particular task.

IX. CONCLUSION AND FUTURE DEVELOPMENT

The two objectives in this work have been met. The work discovers the number of types of M.Sc. applicants/students in Information Technology at SIT, KMUTT and their characteristics. The discovery is directly useful in its strategic educational planning as well as suitable syllabi development. This should enable SIT to equip itself to meet with future requirements of applicants/students better. In turn, authority may be better informed from the finding too.

From knowledge discovery and machine learning perspectives, both distance matrix and its application in Divisive Analysis and Gower's measure of similarity are separate techniques in clustering. The GA-based Partitioning Clustering Program developed incorporates these two techniques successfully. Therefore, the work offers yet another alternative GA-based approach to partitioning clustering. While the program was developed with the mentioned application in mind, in a broader context, there is no legitimate reason to prevent its application on other datasets. Hence, the work also offers a new partitioning clustering technique in general too.

Future development can be carried out on several aspects. Firstly, the program ought to be applied to other public domain datasets to find out its full potential and, perhaps, its shortcomings too. Other similarity measures and distance matrices used in clustering may be experimented in the GA process. This may lead to entirely new GA approach to clustering. Finally, GA operations in this work are a suitable candidate for parallel computation, which may lead to an even more efficient program in terms of computation time.

ACKNOWLEDGMENT

The authors gratefully acknowledge the School of Information Technology (SIT) for partial scholarship for Mr. Ratipong Poolphol and King Mongkut's University of technology Thonburi (KMUTT) for the support of this work. Preparation of this paper would not have been possible without the help and moral support from Dr. P. Lavangnanada.

REFERENCES

[1] L. J. Arabie, G. Hubert, P. DeSoete, "Clustering and Classification", World Scientific, 1999.

[2] A. K. Jain, R. C. Dubes, "Algorithms for Clustering Data", Prentice Hall, 1988.

[3] J. C. Gower, "A General Coefficient of Similarity and Some of Its Properties", Biometrics, Vol. 27, No. 4 , pp. 857-871, Dec., 1971

[4] Kaufman, L. and Rousseeuw, P. J., "Finding Groups in Data: An Introduction to Cluster Analysis", Wiley series in probability and mathematical statistics, Wiley, 1990.

[5] Hruschka, E. R., Campello, R. J., Freitas, A., and De Carvalho, C. P., "A Survey of Evolutionary Algorithms for Clustering", IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol. 39, Issue 2, 2009, pp. 1-26.

[6] R. Krovi, "Genetic Algorithms for Clustering: A Preliminary Investigation", In Proc. of the 25th Hawaii Int. Conference on System Sciences, Vol. 4, pp. 540-544, 1992.

[7] C. A. Murthy, N. Chowdhury, "In Search of Optimal Clusters using Genetic Algorithms", Pattern Recognition Letters, Vol. 17, pp. 825-832, 1996.

[8] K. Krishna, N. Murty, "Genetic K-means Algorithm", IEEE Trans. on Systems, Man and Cybernetics - Part B, Vol. 29, pp. 433-439, 1999.

[9] Y. Lu, S. Lu, F. Fotouhi, Y. Deng, S. J. Brown, "FGKA:A Fast Genetic K-means Clustering Algorithm", In Proc. ACM Symposium on Applied Computing, pp. 622-623, 2004.

[10] P. Fränti, J. Kivijärvi, T. Kaukoranta, O. Nevalainen, "Genetic Algorithms for Large-Scale Clustering Problems", The Computer Journal, Vol. 40, pp. 547-554, 1997.

[11] J. Kivijärvi, P. Fränti, O. Nevalainen, "Self-Adaptive Genetic Algorithm for Clustering", Journal of Heuristics, Vol. 9, pp. 113-129, 2003.

[12] U. Maulik, S. Bandyopadhyay, "Genetic Algorithm based Clustering Technique", Pattern Recognition, Vol. 33, pp. 1455-1465, 2000.

[13] L. I. Kuncheva, J. C. Bezdek, "Selection of Cluster Prototypes from Data by a Genetic Algorithm", In Proc. 5th European Congress on Intelligent Techniques and Soft Computing, pp. 1683-1688, 1997.

[14] L. Y. Tseng, S. B. Yang, "A Genetic Approach to the Automatic Clustering Problem", Pattern Recognition, Vol. 34, pp. 415-424, 2001.

[15] Calinski, R. B., & Harabasz, J., "A dendrite method for cluster analysis", Communications in Statistics, Vol. 3, pp. 1-27, 1974.

[16] Duda, R. O,, & Hart, P. E., "Pattern classification and scene analysis", New York: Wiley, 1973

[17] Rousseeuw P.J. Silhouettes, "A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, 1987.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update", SIGKDD Explorations, Volume 11, Issue 1, 2009.

[19] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, "Understanding of Internal Clustering Validation Measures" In. Proc. 2010 IEEE Int. Conf. on Data Mining (ICDM ',10), pp. 911-916, 2010.

[20] J. O. McClain and V. R. Rao. "Clustisz: A program to test for the quality of clustering of a set of objects." Journal of Marketing Research, Vol. 12, pp. 56–460, 1975.

[21] Brian S. Everitt, Sabine Landau, Morven Lee, Daniel Stahl, "Cluster Analysis", 5th Edition, John Wiley & Sons, London, U.K. 2011

ประวัติผู้วิจัย

ชื่อ – สกุล นายรติพงษ์ พูลผล

วัน เดือน ปีเกิด 27 มีนาคม 2530

ประวัติการศึกษา

ระดับมัธยมศึกษา มัธยมศึกษาตอนปลาย
โรงเรียนหนองสูงสามัคคีวิทยา พ.ศ. 2548

ระดับปริญญาตรี วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พ.ศ. 2552

ระดับปริญญาโท วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี พ.ศ. 2557

ประวัติการทำงาน Developer
บริษัท อีไอบิซ จำกัด พ.ศ. 2557-ปัจจุบัน

ผลงานที่ได้รับการตีพิมพ์ Lavangnananda, K. and Poolphol R., 2014, "A Genetic Algorithm Approach to Partitioning Clustering : A case study on M.Sc. applicants", **The 13th International Conference on Machine Learning and Applications (ICMLA'14)**, 3-6 December 2014, Detroit, MI USA.