อัลกอริธึมแบบปรับแต่งค่าระยะห่างระหว่างข้อมูลสำหรับคลัสเตอร์ริ่งแบบลำดับชั้น

MODIFIED DISTANCE ALGORITHM FOR HIERARCHICAL CLUSTERING

ธนิดา เรืองสวัสดิ์
บราลี จันทนาวิวัฒน์
ภูริชา หิรัญยูปกรณ์

ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2549

MODIFIED DISTANCE ALGORITHM FOR HIERARCHICAL CLUSTERING

TANIDA RUANGSAWAT

BARALI CHANTHANAVIVAT

PURICHA HIRUNYUPAKORN

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

FACULTY OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2006

หัวข้อปัญหาพิเศษ	อัลกอริธึมแบบปรับแต่งค่าระยะห่างระหว่างข้อมูลสำหรับคลัสเตอร์ริ่งแบบลำดับชั้น MODIFIED DISTANCE ALGORITHM FOR HIERARCHICAL CLUSTERING	
ชื่อนักศึกษา	นางสาวธนิดา เรืองสวัสดิ์	46050292
	นางสาวบราลี จันทนาวิวัฒน์	46050303
	นางสาวภูริชา หิรัญยูปกรณ์	46050311
ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง อนุมัติให้นับปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ประจำปีการศึกษา 2549

คณะกรรมการสอบ		ลายมือชื่อ
ประธานกรรมการ	ผศ.ดร.จีรพร ศรีสวัสดิ์	
กรรมการ	ผศ.สิริลักษณ์ อนันต์สถิตย์สิน	
กรรมการและอาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

(รองศาสตราจารย์ ดร.วีระ บุญจริง)
หัวหน้าภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

หัวข้อปัญหาพิเศษ	อัลกอริธึมแบบปรับแต่งค่าระยะห่างระหว่างข้อมูลสำหรับคลัสเตอร์ริ่งแบบลำดับชั้น	
ชื่อนักศึกษา	นางสาวธนิดา เรืองสวัสดิ์	46050292
	นางสาวบราลี จันทนาวิวัฒน์	46050303
	นางสาวภูริชา หิรัญยูปกรณ์	46050311
ปริญญา	วิทยาศาสตรบัณฑิต	
ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์	
สาขาวิชา	วิทยาการคอมพิวเตอร์	
ปีการศึกษา	2549	
อาจารย์ที่ปรึกษา	รศ.ดร.วีระ บุญจริง	

บทคัดย่อ

คลัสเตอร์ริ่งเป็นการจัดข้อมูลให้เป็นกลุ่มย่อยที่มีความคล้ายคลึงกัน โดยคำนวณจากการวัดระยะห่างระหว่างเวกเตอร์ของข้อมูลเข้า เพื่อให้ได้ความรู้นำไปใช้ประโยชน์ในด้านต่างๆ อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดนิยมนำมาใช้ในการวัดระยะห่างระหว่างเวกเตอร์ของสตริง 2 ชุด โดยจะพิจารณาจากความยาวของลำดับร่วมเท่านั้น ปัญหาพิเศษนี้จึงมีแนวคิดที่จะพัฒนาอัลกอริธึมให้สามารถจัดกลุ่มข้อมูลได้ถูกต้องมากยิ่งขึ้น โดยพิจารณาความแตกต่างของระยะห่างระหว่างตำแหน่งตัวอักษรของลำดับร่วมที่ยาวที่สุดในสตริงต้นแบบด้วย การพัฒนาอัลกอริธึมมีจุดมุ่งหมายเพื่อให้การจัดกลุ่มมีความถูกต้องมากขึ้น ซึ่งจะวัดจากค่าความบริสุทธิ์และค่าเอฟเมเซอร์ จากการทดสอบพบว่าอัลกอริธึมที่พัฒนาขึ้นให้ค่าเอฟเมเซอร์ใกล้เคียงกับอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด แต่ให้ค่าความบริสุทธิ์มากกว่าอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด ซึ่งแสดงให้เห็นว่ากลุ่มข้อมูลที่ได้จากการจัดกลุ่มของอัลกอริธึมที่พัฒนาขึ้นมีข้อมูลที่มีความคล้ายคลึงกันเป็นจำนวนมากและมีข้อมูลที่มีความแตกต่างกันเป็นจำนวนน้อย

Special Project Title	MODIFIED DISTANCE ALGORITHM FOR HIERARCHICAL CLUSTERING	
Students	Miss Tanida Ruangsawat	46050292
	Miss Barali Chanthanavivat	46050303
	Miss Puricha Hirunyupakorn	46050311
Degree	Bachelor of Science	
Department	Mathematics and Computer Science, Faculty of Science	
Programme	Computer Science	
Academic Year	2006	
Special Project Advisor	Assoc.Prof.Dr.Veera Boonjing	

ABSTRACT

Clustering is an approach to group large data into similar characteristic subgroups for gaining usefulness of data. Longest Common Subsequence is a widely-used algorithm to measure distance between vector of input strings. The algorithm only considers about length of common subsequence. Modified algorithm is developed to improve correctness of clustering results. The idea is considering both the length of common subsequence and difference of distance of common character position in the input strings. The experiment shows that F-measure of the modified algorithm is close to F-measure of Longest Common Subsequence algorithm. However, Purity of the modified algorithm is more than Purity of Longest Common Subsequence algorithm, which indicates that each cluster has much data with similar characteristic and slight data with dissimilar characteristic.

กิตติกรรมประกาศ

ปัญหาพิเศษนี้มิอาจจะสำเร็จลุล่วงไปได้ด้วยดี หากมิได้รับคำแนะนำ คำชี้แจง ความรู้ และความเอาใจใส่ จาก รศ.ดร.วีระ บุญจริง ผู้เป็นอาจารย์ที่ปรึกษา จึงใคร่ขอขอบพระคุณอย่างสูง

ขอขอบพระคุณ ผศ.ดร.จีรพร ศรีสวัสดิ์ และ ผศ.สิริลักษณ์ อนันต์สถิตย์สิน คณะกรรมการสอบปัญหาพิเศษที่กรุณาให้คำแนะนำตลอดจนข้อชี้แนะ ทำให้ปัญหาพิเศษฉบับนี้สำเร็จลงได้

ขอขอบพระคุณบิดา มารดา ที่สนับสนุนให้ศึกษาในระดับอุดมศึกษา อีกทั้งยังได้ดูแลเรื่องต่างๆเป็นอย่างดี

ขอขอบคุณเพื่อนๆที่ให้ความช่วยเหลือในการทำปัญหาพิเศษฉบับนี้

สำหรับคุณงามความดีและประโยชน์อันใดที่เกิดขึ้นจากปัญหาพิเศษฉบับนี้ คณะผู้จัดทำขอมอบให้กับบิดา มารดา อาจารย์ทุกท่านซึ่งเป็นที่เคารพรักยิ่ง ตลอดจนญาติพี่น้อง และเพื่อนๆ ทุกคน

คณะผู้จัดทำ

มีนาคม 2550

สารบัญ

หน้า

สารบัญ (ต่อ)

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญภาพ

ภาพที่ หน้า

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหา

คลัสเตอร์ริ่ง (Clustering) คือการจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกัน โดยพิจารณาจากความคล้าย (Similarity) หรือ ความใกล้ชิด (Proximity) ของข้อมูล โดยคำนวณจากการวัดระยะห่าง (Distance) ระหว่างเวกเตอร์ของข้อมูลเข้า เพื่อให้ได้ความรู้นำไปใช้ประโยชน์ในด้านต่างๆ อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด (Longest Common Subsequence – LCS) นิยมนำมาใช้ในการวัดระยะห่างระหว่างเวกเตอร์ของสตริง 2 ชุด ซึ่งทำให้สามารถพิจารณาถึงความคล้ายหรือความแตกต่างระหว่างข้อมูลและแบ่งข้อมูลออกเป็นกลุ่มได้ โดยรวมข้อมูลที่มีความคล้ายกันมากไว้เป็นกลุ่มเดียวกัน และแยกข้อมูลที่มีความคล้ายกันน้อย (มีความแตกต่างมาก) ไว้ต่างกลุ่มกัน

อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดจะวัดระยะห่างโดยพิจารณาจากความยาวของลำดับร่วม (Subsequence) เท่านั้น ปัญหาพิเศษนี้จึงมีแนวคิดที่จะพัฒนาอัลกอริธึมให้มีประสิทธิผลมากขึ้นโดยพิจารณาความแตกต่างของระยะห่างระหว่างตำแหน่งตัวอักษรของลำดับร่วมที่ยาวที่สุดในสตริงต้นแบบด้วย

1.2 วัตถุประสงค์ของการศึกษา

เพื่อพัฒนาอัลกอริธึมในการวัดระยะห่างระหว่างข้อมูล ให้คลัสเตอร์ริ่งมีประสิทธิผลมากขึ้นในด้านความถูกต้องของการจัดกลุ่มข้อมูล โดยรวมข้อมูลที่มีความคล้ายกันไว้ด้วยกัน และแยกข้อมูลที่แตกต่างกันไว้ต่างกลุ่มกัน

1.3 ขอบเขตของปัญหา

อัลกอริธึมที่พัฒนาขึ้น ใช้ในการวัดระยะห่างระหว่างข้อมูลชนิดสตริง เพื่อนำไปพิจารณาสำหรับการหาความคล้ายของข้อมูลในคลัสเตอร์ริ่ง ซึ่งจะรวมข้อมูลที่มีค่าความคล้ายมากไว้เป็นกลุ่มเดียวกัน และแยกข้อมูลที่มีค่าความคล้ายน้อยไว้ต่างกลุ่มกัน โดยใช้คลัสเตอร์ริ่งแบบลำดับชั้น การประเมินประสิทธิผลของอัลกอริธึมจะพิจารณาจากความถูกต้องของการจัดกลุ่มข้อมูล ซึ่งวัดจากค่าความบริสุทธิ์ (Purity) และค่าเอฟเมเชอร์ (F-measure)

1.4 ส่วนประกอบของปัญหาพิเศษ

เนื้อหาของปัญหาพิเศษ แบ่งออกเป็นบทดังนี้

บทที่ 2 กล่าวถึง ทฤษฎีและหลักการที่เกี่ยวข้อง ความหมายของคลัสเตอร์ริ่ง คลัสเตอร์ริ่งแบบลำดับชั้น ระยะห่างระหว่างเวกเตอร์ของข้อมูล อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

บทที่ 3 กล่าวถึง อัลกอริธึมที่พัฒนาขึ้น และข้อมูลที่นำมาใช้ในการทดสอบอัลกอริธึม

บทที่ 4 กล่าวถึง การประเมินผลการทดสอบ

บทที่ 5 สรุปผลการทดสอบ และข้อเสนอแนะในการพัฒนาต่อ

บทที่ 2
ทฤษฎีและหลักการที่เกี่ยวข้อง

2.1 คลัสเตอร์ริ่ง

2.1.1 ความหมาย

คลัสเตอร์ริ่ง คือ การจัดกลุ่มข้อมูลโดยไม่กำหนดหมวดหมู่ไว้ล่วงหน้า โดยจัดข้อมูลเป็นกลุ่ม เรียกว่า คลัสเตอร์ (Cluster) ข้อมูลที่มีความคล้ายคลึงกันจะถูกจัดอยู่ในกลุ่มเดียวกันและข้อมูลที่มีความแตกต่างกันจะถูกจัดอยู่ต่างกลุ่มกัน ซึ่งการจัดกลุ่มข้อมูลพิจารณาจากค่าความคล้ายของข้อมูล

2.1.2 คุณสมบัติของอัลกอริธึมสำหรับคลัสเตอร์ริ่ง

- ใช้กับข้อมูลจำนวนเท่าใดก็ได้
- จัดการกับแอททริบิวท์ต่างประเภทกันได้
- ค้นพบกลุ่มข้อมูลโดยไม่มีกฎเกณฑ์ของรูปร่าง
- ต้องการความรู้เกี่ยวกับโดเมนน้อยที่สุดในการกำหนดพารามิเตอร์
- สามารถจัดการกับสิ่งรบกวน (Noise) และข้อมูลที่ผิดปกติ (Outlier) ได้
- ไม่ขึ้นกับลำดับของข้อมูลเข้า
- ทำความเข้าใจได้ง่ายและใช้ได้ง่าย

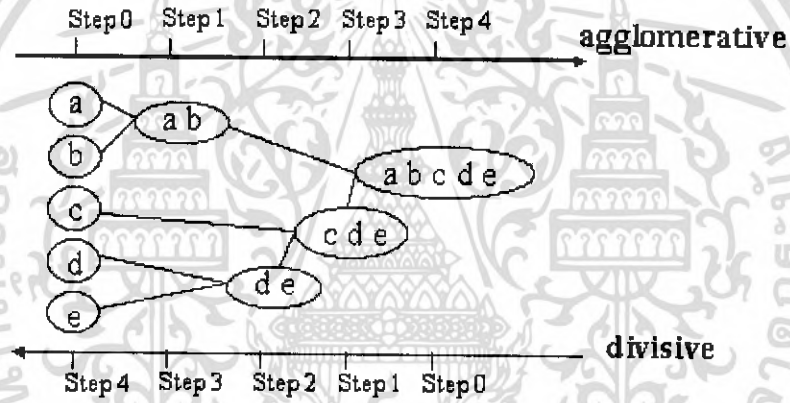
2.1.3 ประเภทของคลัสเตอร์ริ่ง

1. คลัสเตอร์ริ่งแบบลำดับชั้น (Hierarchical clustering) จะจัดคลัสเตอร์ใหม่จากคลัสเตอร์ที่มีอยู่แล้ว

2. คลัสเตอร์ริ่งแบบแบ่งกลุ่ม (Partitional clustering) จะแบ่งข้อมูลทั้งหมดออกเป็นคลัสเตอร์ในครั้งเดียว

2.2 คลัสเตอร์ริ่งแบบลำดับชั้น

คลัสเตอร์ริ่งแบบลำดับชั้น เป็นการจัดกลุ่มของข้อมูลต่างๆให้เป็นลำดับชั้น โดยทำการจัดกลุ่มแบบเดิมซ้ำไปเรื่อยๆ อัลกอริธึมของคลัสเตอร์ริ่งแบบลำดับชั้น แบ่งเป็น 2 ลักษณะ คือ แบบรวมกลุ่ม (Agglomerative หรือ Bottom-up) และแบบแยกกลุ่ม (Divisive หรือ Top-down) โดยแบบรวมกลุ่ม จะเริ่มจากการมองแต่ละข้อมูลเป็นคลัสเตอร์แล้วจัดกลุ่มโดยการรวม 2 คลัสเตอร์ที่มีความคล้ายคลึงกันมากที่สุดเป็นคลัสเตอร์เดียวกัน จากนั้นทำการจัดกลุ่มแบบเดิมไปเรื่อยๆ จนกระทั่งได้จำนวนคลัสเตอร์ที่ต้องการหรือตรงกับเงื่อนไขที่กำหนดไว้ ส่วนแบบแยกกลุ่ม จะทำตรงข้ามกัน คือ จะเริ่มจากคลัสเตอร์เดียว แล้วใช้เกณฑ์บางอย่างในการแบ่งกลุ่มข้อมูลที่มีความคล้ายคลึงกันให้เป็นคลัสเตอร์ย่อยๆ จากนั้นทำการจัดกลุ่มแบบเดิมไปเรื่อยๆจนกระทั่งได้จำนวนคลัสเตอร์ที่ต้องการหรือตรงกับเงื่อนไขที่กำหนดไว้



ภาพที่ 2.1 คลัสเตอร์ริ่งแบบรวมกลุ่มและแบบแยกกลุ่ม

โดยทั่วไปแล้ว อัลกอริธึมแบบแยกกลุ่มมีแนวโน้มว่าจะทำงานรวดเร็วกว่าแบบรวมกลุ่ม แต่ผลที่ได้จะมีความถูกต้องน้อยกว่า การจัดกลุ่มทั้งสองแบบนี้จะใช้เมทริกซ์ความใกล้ชิด (Proximity matrix) เป็นตัวช่วย ซึ่งเมทริกซ์ความใกล้ชิด ก็คือเมทริกซ์ที่แสดงค่าความคล้ายระหว่างแต่ละคลัสเตอร์ ซึ่งตัวเลขในเมทริกซ์ตำแหน่งที่ a_{xy} จะแสดงค่าความคล้ายระหว่างคลัสเตอร์แถวที่ x และคอลัมน์ที่ y หากตัวเลขในเมทริกซ์มีค่ามากแสดงว่ามีความคล้ายกันมาก คลัสเตอร์ริ่งแบบลำดับชั้นจะแสดงผลลัพธ์ออกมาในรูปของแผนภาพเดนโดแกรม (Dendogram) ซึ่งแสดงถึงการจัดกลุ่มของคลัสเตอร์ตามลำดับ แผนภาพนี้จะช่วยให้ผู้ใช้เข้าใจการจัดกลุ่มของข้อมูลได้ง่ายขึ้นและสามารถช่วยในการตัดสินใจได้ว่าควรจะแบ่งคลัสเตอร์ในลักษณะใด แต่อาจเกิดข้อบกพร่องจากการเลือกจุดที่จะแบ่งกลุ่มจากแผนภาพเดนโดแกรม ซึ่งตรวจสอบโดยใช้สายตา

มากกว่าการใช้เกณฑ์หรืออัลกอริธึมที่เป็นมาตรฐาน ทำให้การแบ่งคลัสเตอร์จากแผนภาพเดนโดแกรมเดียวกันทำได้หลายแบบ จึงมีความไม่แน่นอน

ปัญหาสำคัญของคลัสเตอร์ริ่งแบบลำดับชั้นคือ เมื่อมีการรวมกลุ่มหรือแยกกลุ่มเกิดขึ้นแล้วจะไม่สามารถแก้ไขได้ หรือกล่าวอีกอย่างหนึ่งก็คือ ข้อผิดพลาดของการตัดสินใจในการจัดกลุ่มจะไม่สามารถแก้ไขได้ ดังนั้น ผลลัพธ์สุดท้ายของคลัสเตอร์ริ่งแบบลำดับชั้นอาจทำให้ได้คลัสเตอร์ที่ไม่สามารถใช้ประโยชน์ได้

2.3 วิธีการของคลัสเตอร์ริ่งแบบลำดับชั้น

มี 3 แบบ คือ

1. ซิงเกิลลิงค์ (Single-link) เมื่อจัดกลุ่มแล้วจะเลือกค่าความคล้ายระหว่างคู่คลัสเตอร์ที่มีค่ามากที่สุดเป็นตัวแทนค่าความคล้ายของกลุ่ม

2. คอมพลีทลิงค์ (Complete-link) เมื่อจัดกลุ่มแล้วจะเลือกค่าความคล้ายระหว่างคู่คลัสเตอร์ที่มีค่าน้อยที่สุดเป็นตัวแทนค่าความคล้ายของกลุ่ม

3. เอเวอเรจลิงค์ (Average-link) เมื่อจัดกลุ่มแล้วจะใช้ค่าเฉลี่ยของค่าความคล้ายระหว่างคู่คลัสเตอร์เป็นตัวแทนค่าความคล้ายของกลุ่ม

2.4 ระยะห่างระหว่างเวกเตอร์ของข้อมูล

ระยะห่างระหว่างเวกเตอร์ของข้อมูล คือ ข้อมูลประเภทตัวเลขซึ่งบอกถึงค่าความแตกต่างของบางสิ่งบางอย่าง ในทางคณิตศาสตร์ค่าความแตกต่างของข้อมูลต้องได้มาจากเกณฑ์หรือสูตรที่แน่นอน ส่วนในทางกายภาพหรือสิ่งที่เราพบเห็นทั่วไปรอบๆตัว ค่าความแตกต่างของข้อมูลอาจจะหมายถึง ความยาว ช่วงเวลา เป็นต้น

การคำนวณค่าระยะห่างระหว่างสตริงสามารถทำได้โดยมองสตริงแต่ละชุดเป็นเวกเตอร์แล้วคำนวณค่าระยะห่างระหว่างเวกเตอร์ของสตริงนั้น โดยใช้การวัดระยะแบบต่างๆ เช่น ในทางพีชคณิตสามารถหาค่าความแตกต่างของข้อมูล ระหว่างจุด 2 จุด บนระนาบ XY โดยการใช้สูตรในการหาค่าความแตกต่างของข้อมูล ซึ่งค่าความแตกต่าง (d) ของข้อมูลระหว่างจุด (x_1, y_1) และ (x_2, y_2) ถูกกำหนดโดย

$$d = \sqrt{(\Delta x)^2 + (\Delta y)^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

ในกรณีเดียวกัน ถ้ากำหนดจุด (x_1, y_1, z_1) และ (x_2, y_2, z_2) ค่าความแตกต่างของข้อมูลคำนวณได้จาก

$$d = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

ซึ่งสามารถพิสูจน์โดยการสร้างสามเหลี่ยมมุมฉาก โดยการนำทฤษฎีพิธากอรัสมาประยุกต์ใช้

ในทางคณิตศาสตร์โดยเฉพาะทางเรขาคณิต ฟังก์ชันในการหาค่าความแตกต่างของข้อมูลบนเซต M ที่กำหนดให้ จะเป็นฟังก์ชัน $d: M \times M \rightarrow R$ ซึ่ง R แทนเซตของจำนวนจริง โดยมีเงื่อนไขดังนี้

- $d(x, y) \geq 0$ โดย $d(x, y) = 0$ เมื่อ $x = y$ เท่านั้น (ค่าความแตกต่างของข้อมูล ซึ่งเป็นค่าระยะห่างระหว่างจุด 2 จุด จะเป็นค่าบวกและจะเป็นศูนย์เมื่อจุด x และ y เป็นจุดเดียวกัน)

- $d(x, y) = d(y, x)$ คือ จะสมมาตรกัน (ค่าความแตกต่างของข้อมูลระหว่าง x และ y จะเหมือนกันในแต่ละทิศทาง)

- $d(x, z) \leq d(x, y) + d(y, z)$ คือ เมื่อพิจารณาค่าความแตกต่างของข้อมูลของกรณีหนึ่ง ค่าความแตกต่างของข้อมูลระหว่างจุด 2 จุดจะสั้นที่สุด คือจะสั้นกว่าหรือเท่ากับค่าความแตกต่างของข้อมูลของเส้นทางที่ประกอบกันด้วยจำนวนจุดที่มากกว่า โดยจุดต้นและจุดปลายเป็นจุดเดียวกัน

นิยามของค่าความแตกต่างของข้อมูลระหว่างจำนวนจริง x และ y คือ $d(x, y) = |x - y|$ ซึ่งนิยามนี้เป็นไปตามเงื่อนไขทั้ง 3 ข้อด้านบน และตรงกับโครงสร้างมาตรฐานของเส้นจำนวนจริง (Real line) แต่ค่าความแตกต่างของข้อมูลของเซตที่กำหนดให้จะเป็นทางเลือกที่แน่นอน ทางเลือกอื่นที่เป็นไปได้คือนิยามว่า $d(x, y) = 0$ ถ้า $x = y$ และจะเท่ากับ 1 ในกรณีอื่น ซึ่งกรณีเหล่านี้นิยามได้ด้วยเมทริกซ์ แต่ยอมรับได้ด้วยโครงสร้างไม่ต่อเนื่อง

2.5 อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

ส่วนของลำดับ คือ ส่วนที่เหมือนกันของสตริง 2 ชุด ซึ่งสามารถมีบางส่วนหายไป หรืออาจไม่มีส่วนใดหายไปเลย

<u>ตัวอย่าง</u>

$X = \langle A, B, C, B, D, A, B \rangle$

$Z = \langle B, C, D, B \rangle$

$\therefore$ Z เป็นส่วนของลำดับของ X

การหาส่วนของลำดับร่วมที่ยาวที่สุด หรือเรียกว่า แอลซีเอส (LCS) คือ การหาส่วนของลำดับที่เหมือนกัน ซึ่งมีความยาวมากที่สุด จากสตริง 2 ชุด

<u>ตัวอย่าง</u>

- misspell กับ mispell

ได้ส่วนของลำดับร่วมที่ยาวที่สุด เป็น mispell (ความยาวเท่ากับ 7)

- misspelled กับ misinterpreted

ได้ส่วนของลำดับร่วมที่ยาวที่สุด เป็น mispeed (ความยาวเท่ากับ 7)

<u>ตัวอย่าง</u>

ภาพที่ 2.2 ตัวอย่างของลำดับร่วมที่ยาวที่สุด

อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด จะเป็นแบบเวียนบังเกิด (Recursive) และมักจะใช้เมทริกซ์มาช่วย

การกำหนดค่าในเมทริกซ์ของอัลกอริธึม เป็นดังนี้

$$c[i,j] = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0 \\ c[i-1, j-1]+1 & \text{if } i, j > 0 \text{ and } x_i = y_j \\ \max(c[i, j-1], c[i-1, j]) & \text{if } i, j > 0 \text{ and } x_i \neq y_j \end{cases}$$

โดย c คือ เมทริกซ์ที่ใช้ในการหาค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด

อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

ให้ X และ Y เป็นสตริงที่ต้องการหาส่วนของลำดับร่วมที่ยาวที่สุด

c เป็นเมทริกซ์ที่เก็บค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด

r เป็นเมทริกซ์ที่เก็บค่าทิศทาง (แทนด้วยลูกศร)

LCS-LENGTH (X, Y)

```
1.   m = length[X]
2.   n = length[Y]
3.   for i = 1 to m
4.       c[i, 0] = 0
5.   for j = 0 to n
6.       c[0, j] = 0
7.   for i = 1 to m
8.       for j = 1 to n
9.           if      x_i = y_j
10.              then  c[i, j] = c[i-1, j-1] + 1
11.                   r[i, j] = " ↖ "
12.          else if  c[i-1, j] ≥ c[i, j-1]
13.              then  c[i, j] = c[i-1, j]
14.                   r[i, j] = " ↑ "
15.          else     c[i, j] = c[i, j-1]
16.                   r[i, j] = " ← "
17.  return c and r
```

ตัวอย่าง

	j	0	1	2	3	4	5	6
i		y_i	B	D	C	A	B	A
0	x_i	0	0	0	0	0	0	0
1	A	0	0↑	0↑	0↑	1↖	←1	1↖
2	B	0	1↖	←1	←1	1↑	2↖	←2
3	C	0	1↑	1↑	2↖	←2	2↑	2↑
4	B	0	1↖	1↑	2↑	2↑	3↖	←3
5	D	0	1↑	2↖	2↑	2↑	3↑	3↑
6	A	0	1↑	2↑	2↑	3↖	3↑	4↖
7	B	0	1↖	2↑	2↑	3↑	4↖	4↑

ภาพที่ 2.3 ตัวอย่างเมทริกซ์ที่ใช้ในการหาส่วนของลำดับร่วมที่ยาวที่สุด

ลูกศรจะชี้ไปด้านบนซ้ายในช่องที่สตริงทั้งสองมีตัวอักษรที่เหมือนกัน ลูกศรที่ชี้ไปด้านบนหรือทางซ้ายมือคือนำค่าตัวเลขจากช่องนั้นๆมา จากนั้น หาเส้นทางที่ยาวที่สุดจากช่องสุดท้ายไปยังช่องแรกสุด แล้วเรียงลำดับตัวอักษรที่เหมือนกัน ในที่นี้ LCS(x, y) = "BCBA"

ตัวอย่างงานที่ใช้การหาส่วนของลำดับร่วมที่ยาวที่สุด มีดังนี้

- ชีวโมเลกุล (Molecular biology) ลำดับของดีเอ็นเอ (ยีน) สามารถแทนได้ด้วยลำดับของ 4 ตัวอักษรคือ ACGT ซึ่งเป็นส่วนประกอบในการสร้างดีเอ็นเอ เมื่อนักชีววิทยาพบลำดับของดีเอ็นเอแบบใหม่ และต้องการที่จะรู้ว่าคล้ายกับลำดับดีเอ็นเอแบบใดมากที่สุด วิธีหนึ่งในการคำนวณความคล้ายของ 2 ลำดับคือหาความยาวของส่วนของลำดับร่วมที่ยาวที่สุด

- การเปรียบเทียบไฟล์ (File comparison) ในโปรแกรมยูนิกซ์ คำสั่ง "diff" ใช้ในการเปรียบเทียบ 2 เวอร์ชันของไฟล์ชนิดเดียวกัน เพื่อหาว่ามีการเปลี่ยนแปลงเกิดขึ้นกับไฟล์หรือไม่ คำสั่งนี้ทำงานโดยหาส่วนของลำดับร่วมที่ยาวที่สุดของไฟล์ทั้งสอง ซึ่งจะมองแต่ละแถวในไฟล์เป็นสตริง

บทที่ 3
การพัฒนาอัลกอริธึม

ปัญหาพิเศษนี้ จะพัฒนาอัลกอริธึมในการหาค่าระยะห่างระหว่างข้อมูลชนิดสตริง เพื่อให้สามารถรวมกลุ่มข้อมูลที่มีความคล้ายคลึงกันได้ถูกต้องมากขึ้น โดยค่าที่คำนวณได้จากอัลกอริธึมแสดงถึงความใกล้ชิดของข้อมูล ซึ่งข้อมูลที่มีค่าความใกล้ชิดมาก จะเป็นข้อมูลที่มีความคล้ายกันมาก

การพัฒนาปัญหาพิเศษ มีขั้นตอนดังนี้

3.1 การศึกษารวบรวมข้อมูล

ทำการศึกษาข้อมูลจากแหล่งความรู้ต่างๆ เช่น เอกสารงานวิจัย หนังสือ ข้อมูลทางอินเทอร์เน็ต โดยจะศึกษาข้อมูลเกี่ยวกับวิธีคลัสเตอร์ริ่งข้อมูล อัลกอริธึมในการหาค่าระยะห่างระหว่างข้อมูล ข้อมูลตัวอย่างที่สามารถนำมาใช้ในการทดสอบอัลกอริธึม และวิธีการวัดประสิทธิผลของอัลกอริธึม

3.2 การทำงานของอัลกอริธึม

ขั้นตอนที่ 1 รับข้อมูลชนิดสตริง 2 ข้อมูลมาตัดแบ่งเป็นเซตของตัวอักษร จะได้ $X = \{x_1, x_2, ..., x_m\}$ เป็นเซตของตัวอักษรของข้อมูลที่ 1 และ $Y = \{ y_1, y_2, ..., y_n\}$ เป็นเซตของตัวอักษรของข้อมูลที่ 2

ขั้นตอนที่ 2 สร้างเมทริกซ์ c และ r ซึ่งเป็นเมทริกซ์ 2 มิติ ขนาด m x n โดย m คือ จำนวนสมาชิกของเซต X และ n คือ จำนวนสมาชิกของเซต Y ซึ่งเมทริกซ์ c จะใช้สำหรับเก็บค่าที่คำนวณได้เพื่อหาค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด ส่วนเมทริกซ์ r จะใช้สำหรับเก็บทิศทาง

ขั้นตอนที่ 3 คำนวณหาค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุดและทิศทางทีละแถวจากซ้ายไปขวาและจากบนลงล่าง โดยใช้เกณฑ์ดังนี้

- ถ้า i หรือ j เป็น 0 ให้ $c[i, j] = 0$
- ถ้า i และ j มากกว่า 0 และ $x_i = y_j$ ให้ $c[i, j] = c[i-1, j-1] + 1$ และ $r[i, j]$ ทิศทางเป็น ↖
- ถ้า i และ j มากกว่า 0 และ $x_i \neq y_j$ ให้ $c[i, j] = \max(c[i, j-1], c[i-1, j])$

- ถ้า c[i-1, j] ≥ c[i, j-1] แล้ว r[i, j] มีทิศทางเป็น ↑

- ถ้า c[i-1, j] < c[i, j-1] แล้ว r[i, j] มีทิศทางเป็น ←

ขั้นตอนที่ 4 หาค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุดระหว่าง X และ Y ซึ่งจะเท่ากับค่าในช่องสุดท้ายของเมทริกซ์ c

ขั้นตอนที่ 5 เลือกเส้นทางในเมทริกซ์ r จากช่องสุดท้ายไปยังช่องแรกสุด โดยดูตามทิศทางของลูกศร

- เรียงลำดับตัวอักษรที่ตรงกันคือช่องที่มีทิศทางเป็น ↖ จะได้ส่วนของลำดับร่วมที่ยาวที่สุด

- ถ้า r[i, j] มีทิศทางเป็น ↖ (มีตัวอักษรตรงกัน) และ r[i-k, j-l] เป็นช่องถัดไปที่มีทิศทางเป็น ↖ จะหาค่าระยะห่างระหว่างตำแหน่งตัวอักษรที่เป็นลำดับร่วมได้จาก i – (i -k) และ j – (j - l) ซึ่งเป็นค่าระยะห่างของข้อมูลที่ 1 และ 2 ตามลำดับ

- หาค่าความแตกต่างของทั้ง 2 ค่า โดยนำมาลบกันแล้วหาค่าสัมบูรณ์

- ถ้าค่าความแตกต่างที่ได้มากกว่าหรือเท่ากับค่าเธรชโฮลด์ (Threshold) จะนำค่าความแตกต่างไปคูณกับค่าน้ำหนัก (Weight) แล้วนำค่าที่ได้ไปลบออกจากค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด

ขั้นตอนที่ 6 ส่งค่าที่ได้กลับไปเก็บยังเมทริกซ์ที่เก็บค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูล

ขั้นตอนที่ 7 ทำขั้นตอนที่ 1 – 6 ซ้ำจนกระทั่งได้ค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูลทั้งหมด

ขั้นตอนที่ 8 จัดกลุ่มข้อมูลที่มีความคล้ายกันเข้าด้วยกัน โดยใช้ค่าจากเมทริกซ์ที่เก็บค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูลทั้งหมด ข้อมูลที่มีค่าความใกล้ชิดระหว่างข้อมูลมาก ซึ่งเป็นข้อมูลที่มีความคล้ายกันมาก จะถูกรวมเข้าเป็นกลุ่มเดียวกัน จากนั้นทำการปรับค่าในเมทริกซ์โดยใช้วิธีคอมพลีทลิงค์ และทำการจัดกลุ่มไปจนกว่าข้อมูลทั้งหมดจะถูกรวมเข้าเป็นกลุ่มเดียว

3.3 ตัวอย่างการทำงานของอัลกอริธึม

สมมติให้ชุดข้อมูลทั้งหมดมีดังนี้

t,a,c,t,a,g,c,a,a,t,+

g,t,a,c,t,a,g,a,g,a,+

t,g,c,t,a,t,c,c,t,g,-

c,t,c,g,t,c,c,t,c,a,-

t,a,a,c,a,t,t,a,a,t,-

กำหนดให้เป็นข้อมูลที่ [0] ถึง [4] ดังนี้

[0] tactagcaat

[1] gtactagaga

[2] tgctatcctg

[3] ctcgtcctca

[4] taacattaat

แสดงตัวอย่างการคำนวณ ของข้อมูล 2 ข้อมูล คือ [0] และ [1]

ขั้นตอนที่ 1 รับข้อมูลชนิดสตริงทั้ง 2 ข้อมูลมาตัดแบ่งเป็นเซตของตัวอักษร

X = {t, a, c, t, a, g, c, a, a, t} เป็นเซตของตัวอักษรของข้อมูล [0]

Y = {g, t, a, c, t, a, g, a, g, a} เป็นเซตของตัวอักษรของข้อมูล [1]

ขั้นตอนที่ 2 สร้างเมทริกซ์ c และ r ซึ่งเป็นเมทริกซ์ 2 มิติ ขนาด 10 x 10

ขั้นตอนที่ 3 คำนวณหาค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุดและทิศทางทีละแถวจากซ้ายไปขวาและจากบนลงล่าง จะได้ข้อมูลที่อยู่ในเมทริกซ์ c และ r ดังนี้

เมทริกซ์ c

	j	0	1	2	3	4	5	6	7	8	9	10
i			g	t	a	c	t	a	g	a	g	a
0		0	0	0	0	0	0	0	0	0	0	0
1	t	0	0	1	1	1	1	1	1	1	1	1
2	a	0	0	1	2	2	2	2	2	2	2	2
3	c	0	0	1	2	3	3	3	3	3	3	3
4	t	0	0	1	2	3	4	4	4	4	4	4
5	a	0	0	1	2	3	4	5	5	5	5	5
6	g	0	1	1	2	3	4	5	6	6	6	6
7	c	0	1	1	2	3	4	5	6	6	6	6
8	a	0	1	1	2	3	4	5	6	7	7	7
9	a	0	1	1	2	3	4	5	6	7	7	8
10	t	0	1	2	2	3	4	5	6	7	7	8

ภาพที่ 3.1 เมทริกซ์ c

เมทริกซ์ r

	j	0	1	2	3	4	5	6	7	8	9	10
i			g	t	a	c	t	a	g	a	g	a
0												
1	t		↑	↖	←	←	↖	←	←	←	←	←
2	a		↑	↑	↖	←	←	↖	←	↖	←	↖
3	c		↑	↑	↑	↖	←	←	←	←	←	←
4	t		↑	↖	↑	↑	↖	←	←	←	←	←
5	a		↑	↑	↖	↑	↑	↖	←	↖	←	↖
6	g		↖	↑	↑	↑	↑	↑	↖	←	↖	←
7	c		↑	↑	↑	↖	↑	↑	↑	↑	↑	↑
8	a		↑	↑	↖	↑	↑	↖	↑	↖	←	↖
9	a		↑	↑	↖	↑	↑	↖	↑	↖	↑	↖
10	t		↑	↖	↑	↑	↖	↑	↑	↑	↑	↑

ภาพที่ 3.2 เมทริกซ์ r

	j	0	1	2	3	4	5	6	7	8	9	10
i			g	t	a	c	t	a	g	a	g	a
0		0	0	0	0	0	0	0	0	0	0	0
1	t	0	0 ↑	↖1	←1	←1	↖1	←1	←1	←1	←1	←1
2	a	0	0 ↑	1 ↑	↖2	←2	←2	↖2	←2	↖2	←2	↖2
3	c	0	0 ↑	1 ↑	2 ↑	↖3	←3	←3	←3	←3	←3	←3
4	t	0	0 ↑	↖1	2 ↑	3 ↑	↖4	←4	←4	←4	←4	←4
5	a	0	0 ↑	1 ↑	↖2	3 ↑	4 ↑	↖5	←5	↖5	←5	↖5
6	g	0	↖1	1 ↑	2 ↑	3 ↑	4 ↑	5 ↑	↖6	←6	↖6	←6
7	c	0	1 ↑	1 ↑	2 ↑	↖3	4 ↑	5 ↑	6 ↑	6 ↑	6 ↑	6 ↑
8	a	0	1 ↑	1 ↑	↖2	3 ↑	4 ↑	↖5	6 ↑	↖7	←7	↖7
9	a	0	1 ↑	1 ↑	↖2	3 ↑	4 ↑	↖5	6 ↑	↖7	7 ↑	↖8
10	t	0	1 ↑	↖2	2 ↑	3 ↑	↖4	5 ↑	6 ↑	7 ↑	7 ↑	8 ↑

ภาพที่ 3.3 ค่าที่คำนวณได้ในเมทริกซ์ c และ r

ขั้นตอนที่ 4 ได้ค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด = 8

ขั้นตอนที่ 5

- ได้ส่วนของลำดับร่วมที่ยาวที่สุด คือ tactagaa

- หาค่าระยะห่างระหว่างตำแหน่งตัวอักษรที่เป็นลำดับร่วมที่ยาวที่สุดเรียงไปตามลูกศร แล้วนำมาคำนวณหาค่าความแตกต่างของทั้ง 2 ค่า

a – a	=	| (9 – 8) – (10 – 8) |	=	1
a – g	=	| (8 – 6) – (8 – 7) |	=	1
g – a	=	| (6 – 5) – (7 – 6) |	=	0
a – t	=	| (5 – 4) – (6 – 5) |	=	0
t – c	=	| (4 – 3) – (5 – 4) |	=	0
c – a	=	| (3 – 2) – (4 – 3) |	=	0
a – t	=	| (2 – 1) – (3 – 2) |	=	0

- ค่าความแตกต่างระหว่าง a – a และ a – g เท่ากับค่าเธรชโฮลด์ (= 1) จะนำค่าความแตกต่างไปคูณกับค่าน้ำหนัก (= 0.05) แล้วนำค่าที่ได้ไปลบออกจากค่าความยาวของส่วนของลำดับร่วมที่ยาวที่สุด

ค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูล = 8 – (1 * 0.05) – (1*0.05) = 7.9

ขั้นตอนที่ 6 ส่งค่าที่ได้กลับไปเก็บยังเมทริกซ์ที่เก็บค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูล

ขั้นตอนที่ 7 ทำขั้นตอนที่ 1 – 6 ซ้ำจนกระทั่งได้ค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูลทั้งหมด

ขั้นตอนที่ 8 ได้เมทริกซ์ที่ใช้เก็บค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูลทั้งหมด จากนั้นจัดกลุ่มข้อมูลที่มีความคล้ายคลึงกันเข้าด้วยกัน คือ เลือกคู่ข้อมูลที่มีค่าความใกล้ชิดระหว่างเวกเตอร์ของข้อมูลมากที่สุด แสดงดังภาพ

Step 0	[0]	[1]	[2]	[3]	[4]
[0]	0.0	**7.9**	5.85	4.85	6.7
[1]	0.0	0.0	5.7	4.85	5.8
[2]	0.0	0.0	0.0	5.8	4.65
[3]	0.0	0.0	0.0	0.0	4.8
[4]	0.0	0.0	0.0	0.0	0.0

ภาพที่ 3.4 เมทริกซ์ความใกล้ชิดก่อนการรวมกลุ่มข้อมูล

ค่ามากที่สุดคือ 7.9 ซึ่งเป็นค่าความใกล้ชิดระหว่างข้อมูล [0] และ [1] ทำการรวมกลุ่มข้อมูล [0] และ [1] เข้าด้วยกัน แล้วใช้วิธีคอมพลีทลิงค์ในการปรับค่าความใกล้ชิดระหว่างข้อมูล [0][1] และข้อมูลอื่นๆ คือ เลือกค่าความใกล้ชิดที่น้อยที่สุด

จากตัวอย่าง ค่าความใกล้ชิดระหว่างข้อมูล [0] และ [2] คือ 5.85 ค่าความใกล้ชิดระหว่างข้อมูล [1] และ [2] คือ 5.7 จะเลือกค่าที่น้อยกว่า คือ 5.7 เป็นค่าความใกล้ชิดระหว่างข้อมูล [0][1] และ [2]

Step 1	[2]	[3]	[4]	[0][1]
[2]	0.0	5.8	4.65	**5.7**
[3]	0.0	0.0	4.8	4.85
[4]	0.0	0.0	0.0	5.8
[0][1]	0.0	0.0	0.0	0.0

ภาพที่ 3.5 การปรับค่าในเมทริกซ์ความใกล้ชิดในการรวมกลุ่มข้อมูลครั้งที่ 1

ทำการรวมกลุ่มไปจนกว่าข้อมูลทั้งหมดจะถูกรวมเข้าเป็นกลุ่มเดียว

Step 1	[2]	[3]	[4]	[0][1]
[2]	0.0	**5.8**	**4.65**	5.7
[3]	0.0	0.0	4.8	4.85
[4]	0.0	0.0	0.0	5.8
[0][1]	0.0	0.0	0.0	0.0

ภาพที่ 3.6 เมทริกซ์ความใกล้ชิดหลังจากการรวมกลุ่มข้อมูลครั้งที่ 1

Step 2	[4]	[0][1]	[2][3]
[4]	0.0	**5.8**	4.65
[0][1]	0.0	0.0	4.85
[2][3]	0.0	0.0	0.0

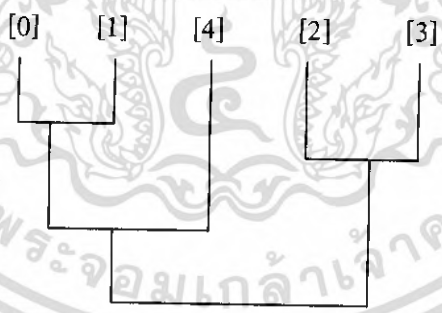
ภาพที่ 3.7 เมทริกซ์ความใกล้ชิดหลังจากการรวมกลุ่มข้อมูลครั้งที่ 2

Step 3	[2][3]	[4][0][1]
[2][3]	0.0	**4.65**
[4][0][1]	0.0	0.0

ภาพที่ 3.8 เมทริกซ์ความใกล้ชิดหลังจากการรวมกลุ่มข้อมูลครั้งที่ 3

Step 4	[2][3][4][0][1]
[2][3][4][0][1]	0.0

ภาพที่ 3.9 เมทริกซ์ความใกล้ชิดหลังจากการรวมกลุ่มข้อมูลครั้งที่ 4



ภาพที่ 3.10 แผนภาพเดนโดแกรมแสดงการจัดกลุ่มของข้อมูล

3.4 การทดสอบความถูกต้อง

ทดสอบความถูกต้องโดยเปรียบเทียบผลลัพธ์ที่ได้จากการจัดกลุ่มซึ่งคำนวณระยะห่างระหว่างข้อมูลด้วยอัลกอริธึมที่พัฒนาขึ้นและอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดกับข้อมูลตัวอย่างซึ่งได้มีการแบ่งประเภทไว้แล้ว

3.5 ข้อมูลที่ใช้ในการทดสอบ

ใช้ชุดข้อมูลมาตรฐาน 10 ชุดข้อมูล จาก UCI (http://www.ics.uci.edu/~mlearn/MLRepository.html) และ GRAPPA (http://www.grappa.univ-lille3.fr/~torre/guide.php?id=datasets) ข้อมูลทุกชุดมีการแบ่งประเภทไว้แล้ว แต่ละประเภทเรียกว่า คลาส ซึ่งส่วนของข้อมูลที่เป็นการระบุคลาสจะถูกลบออกในการทดสอบการจัดกลุ่มข้อมูล เนื่องจากคลัสเตอร์ริ่งจะทำการจัดกลุ่มข้อมูลโดยไม่กำหนดหมวดหมู่ไว้ล่วงหน้า

ข้อมูลแต่ละชุดไม่จำเป็นต้องมีจำนวนข้อมูล, จำนวนแอททริบิวท์ และจำนวนคลาสเท่ากัน และชุดข้อมูลซึ่งมีแอททริบิวท์ชนิดตัวเลขจะถูกพิจารณาโดยมองเป็นสตริง

1. **Car Evaluation** ข้อมูลชนิดสตริง 172 ข้อมูล แต่ละข้อมูลมี 6 แอททริบิวท์ แบ่งข้อมูลเป็น 4 คลาส

2. **Credit Approval** ข้อมูลชนิดสตริงและตัวเลข 68 ข้อมูล แต่ละข้อมูลมี 15 แอททริบิวท์ (สตริง 9 แอททริบิวท์ และตัวเลข 6 แอททริบิวท์) แบ่งข้อมูลเป็น 2 คลาส

3. **Hepatitis** ข้อมูลชนิดตัวเลข 155 ข้อมูล แต่ละข้อมูลมี 19 แอททริบิวท์ แบ่งข้อมูลเป็น 2 คลาส

4. **Iris Plant** ข้อมูลชนิดตัวเลข 150 ข้อมูล แต่ละข้อมูลมี 4 แอททริบิวท์ แบ่งข้อมูลเป็น 3 คลาส

5. **Liver Disorders** ข้อมูลชนิดตัวเลข 68 ข้อมูล แต่ละข้อมูลมี 6 แอททริบิวท์ แบ่งข้อมูลเป็น 2 คลาส

6. **Pendigits** ข้อมูลชนิดตัวเลข 90 ข้อมูล แต่ละข้อมูลมี 16 แอททริบิวท์ แบ่งข้อมูลเป็น 10 คลาส

7. **Pima Indians Diabetes** ข้อมูลชนิดตัวเลข 77 ข้อมูล แต่ละข้อมูลมี 8 แอททริบิวท์ แบ่งข้อมูลเป็น 2 คลาส

8. **Promoter Gene Sequence** ข้อมูลชนิดสตริง 106 ข้อมูล แต่ละข้อมูลมี 57 แอททริบิวท์ แบ่งข้อมูลเป็น 2 คลาส

9. Sonar Mines vs. Rocks ข้อมูลชนิดตัวเลข 21 ข้อมูล แต่ละข้อมูลมี 60 แอททริบิวท์ แบ่งข้อมูลเป็น 2 คลาส

10. Wine Recognition ข้อมูลชนิดตัวเลข 178 ข้อมูล แต่ละข้อมูลมี 13 แอททริบิวท์ แบ่งข้อมูลเป็น 3 คลาส

บทที่ 4
การประเมินผล

การประเมินประสิทธิผลของอัลกอริธึม ทำได้โดยนำข้อมูลซึ่งมีการแบ่งประเภทของข้อมูลไว้แล้วมาลบข้อมูลส่วนที่เป็นการบอกประเภทออก แล้วคลัสเตอร์ริ่งโดยใช้อัลกอริธึมที่พัฒนาขึ้น พิจารณาความถูกต้องของการจัดกลุ่มในลำดับชั้นที่ให้ผลลัพธ์ของจำนวนกลุ่มเท่ากับจำนวนของประเภทข้อมูล โดยวัดจากค่าความบริสุทธิ์และค่าเอฟเมเชอร์ และไม่ประเมินผลประสิทธิภาพในด้านความเร็วและปริมาณการใช้ทรัพยากร

4.1 หลักการวัดค่าความแม่นยำและค่าความครบถ้วน

4.1.1 ค่าความแม่นยำ

ค่าความแม่นยำ (Precision) เป็นค่าที่แสดงว่าอัลกอริธึมสามารถเลือกคำตอบที่ถูกต้องได้มากเท่าใดจากคำตอบที่เลือกมาทั้งหมด

$$p_k = \frac{\max_c\{CF_k(c)\}}{N_k}$$

โดย p_k คือ ค่าความแม่นยำ

c คือ อินเด็กซ์ของคลาส (คลาส คือ กลุ่มของข้อมูล ซึ่งได้มีการกำหนดประเภทไว้แล้ว)

k คือ อินเด็กซ์ของคลัสเตอร์ (คลัสเตอร์ คือ กลุ่มของข้อมูลที่ได้จากอัลกอริธึม)

$CF_k(c)$ คือ จำนวนข้อมูลคลาส c ในคลัสเตอร์ k

N_k คือ จำนวนข้อมูลในคลัสเตอร์ k

4.1.2 ค่าความครบถ้วน

ค่าความครบถ้วน (Recall) เป็นค่าที่แสดงว่าอัลกอริธึมสามารถเลือกคำตอบที่ถูกต้องได้มากเท่าใดจากคำตอบที่ถูกต้องทั้งหมด

$$r_c = \frac{\max_k \{CF_k(c)\}}{N_c}$$

โดย r_c คือ ค่าความครบถ้วน

c คือ อินเด็กซ์ของคลาส (คลาส คือ กลุ่มของข้อมูล ซึ่งได้มีการกำหนดประเภทไว้แล้ว)

k คือ อินเด็กซ์ของคลัสเตอร์ (คลัสเตอร์ คือ กลุ่มของข้อมูลที่ได้จากอัลกอริธึม)

$CF_k(c)$ คือ จำนวนข้อมูลคลาส c ในคลัสเตอร์ k

N_c คือ จำนวนข้อมูลในคลาส c

4.2 หลักการวัดค่าความบริสุทธิ์และค่าเอฟเมเชอร์

4.2.1 ค่าความบริสุทธิ์

คำนวณจากค่าความแม่นยำของทุกคลัสเตอร์

$$\text{Purity} = \frac{1}{N} \sum p_k$$

โดย p_k คือ ค่าความแม่นยำ

k คือ อินเด็กซ์ของคลัสเตอร์ (คลัสเตอร์ คือ กลุ่มของข้อมูลที่ได้จากอัลกอริธึม)

N คือ จำนวนคลัสเตอร์

4.2.2 ค่าเอฟเมเชอร์

คำนวณได้จากค่าความแม่นยำและค่าความครบถ้วน

$$\text{F-measure} = \frac{2PR}{P+R}$$

โดย P คือ ค่าความแม่นยำ

R คือ ค่าความครบถ้วน

4.3 ตัวอย่างการคำนวณ

ข้อมูลมาตรฐานจำนวน 106 ข้อมูล แบ่งเป็น 2 คลาส คือ + และ - คลาสละ 53 ข้อมูล จะพิจารณาในลำดับชั้นที่มีจำนวนคลัสเตอร์เท่ากับจำนวนคลาส คือ ลำดับชั้นที่มีจำนวนคลัสเตอร์เท่ากับ 2

ผลการจัดกลุ่มข้อมูลที่ได้จากอัลกอริธึมเป็นดังนี้

คลัสเตอร์ 1 : มีข้อมูลทั้งหมด 48 ข้อมูล แบ่งเป็นคลาส + จำนวน 8 ข้อมูล และคลาส – จำนวน 40 ข้อมูล

คลัสเตอร์ 2 : มีข้อมูลทั้งหมด 58 ข้อมูล แบ่งเป็นคลาส + จำนวน 45 ข้อมูล และคลาส – จำนวน 13 ข้อมูล

คำนวณค่าความแม่นยำและค่าความครบถ้วน โดยคำนวณจากจำนวนข้อมูลคลาสเดียวกันที่มีมากที่สุดในแต่ละคลัสเตอร์

ค่าความแม่นยำ

คลัสเตอร์ 1 : ค่าความแม่นยำ = 40 / 48 = 0.83

คลัสเตอร์ 2 : ค่าความแม่นยำ = 45 / 58 = 0.77

ค่าความครบถ้วน

คลัสเตอร์ 1 : ค่าความครบถ้วน = 40 / 53 = 0.75

คลัสเตอร์ 2 : ค่าความครบถ้วน = 45 / 53 = 0.84

คำนวณค่าความบริสุทธิ์และค่าเอฟเมเชอร์ ได้ดังนี้

ค่าความบริสุทธิ์

$$\text{ค่าความบริสุทธิ์} = \frac{0.83 + 0.77}{2} = 0.80459$$

ค่าเอฟเมเชอร์

$$\text{คลัสเตอร์ 1 : ค่าเอฟเมเชอร์} = \frac{2\,(0.83)\,(0.75)}{0.83 + 0.75} = 0.792$$

$$\text{คลัสเตอร์ 2 : ค่าเอฟเมเชอร์} = \frac{2\,(0.77)\,(0.84)}{0.77 + 0.84} = 0.810$$

$$\text{ค่าเอฟเมเชอร์เฉลี่ย} = \frac{2\,(0.792)\,(0.810)}{2} = 0.80144$$

4.4 ค่าความบริสุทธิ์และค่าเอฟเมเชอร์ที่ได้จากการทดสอบ

คลัสเตอร์ริ่งโดยใช้อัลกอริธึมที่พัฒนาขึ้น เปรียบเทียบกับคลัสเตอร์ริ่งโดยใช้อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด ซึ่งการเปรียบเทียบจะทำในลำดับชั้นที่มีจำนวนกลุ่มเท่ากับจำนวนคลาสที่มีการกำหนดไว้

แต่ละชุดข้อมูลจะทดลอง 5 ครั้ง แล้วนำค่าความบริสุทธิ์และค่าเอฟเมเชอร์ที่ได้ในแต่ละครั้งมาหาค่าเฉลี่ย ได้ผลดังนี้

ตารางที่ 4.1 ค่าความบริสุทธิ์และค่าเอฟเมเชอร์ของข้อมูลแต่ละชุด

ชุดข้อมูล	อัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด		อัลกอริธึมที่พัฒนาขึ้น	
	ค่าความบริสุทธิ์	ค่าเอฟเมเชอร์	ค่าความบริสุทธิ์	ค่าเอฟเมเชอร์
Car Evaluation	0.662	0.342	0.696	0.354
Credit Approval	0.67	0.482	0.706	0.435
Hepatitis	0.809	0.553	0.802	0.533
Iris Plant	0.764	0.725	0.775	0.656
Liver Disorders	0.601	0.559	0.633	0.555
Pendigits	0.365	0.325	0.434	0.364
Pima Indians Diabetes	0.648	0.555	0.647	0.52
Promoter Gene Sequence	0.773	0.716	0.788	0.715
Sonar Mines vs. Rocks	0.548	0.52	0.576	0.538
Wine Recognition	0.632	0.578	0.655	0.592

4.5 การประเมินผลการทดสอบ

4.5.1 การหาช่วงความเชื่อมั่น

เพื่อให้การสรุปผลมีความหมายมากขึ้นนั้น มีความจำเป็นที่จะต้องนำช่วงความเชื่อมั่น (Confidence interval) ซึ่งเป็นเรื่องของสถิติอ้างอิง (Inferential statistics) เข้ามาใช้ในการวิเคราะห์ค่าเฉลี่ยของค่าความบริสุทธิ์และค่าเอฟเมเซอร์จากการทดสอบอัลกอริธึม ซึ่งการหาช่วงความเชื่อมั่น สามารถหาได้จากสมการ ดังนี้

$$\text{ช่วงความเชื่อมั่น} = \overline{x} \pm t_{n-1,\,\alpha/2}\,\frac{s}{\sqrt{n}}$$

โดย $\overline{x}$ คือ ค่าเฉลี่ยของค่าความบริสุทธิ์หรือค่าเอฟเมเซอร์ที่ได้จากการทดสอบอัลกอริธึมกับข้อมูลทุกชุด

t คือ ค่าเฉลี่ยของการแจกแจงที่ได้จากตารางการแจกแจงแบบที

n คือ จำนวนข้อมูลทั้งหมดที่นำมาใช้ในการทดสอบ

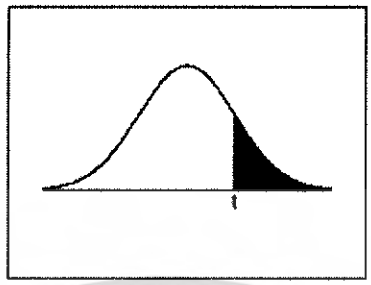
α คือ ระดับความเชื่อมั่น

s คือ ส่วนเบี่ยงเบนมาตรฐาน ซึ่งสามารถหาได้จากสมการ

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

ตารางที่ 4.2 การแจกแจงแบบที

t-Distribution Table



The shaded area is equal to α for $t = t_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
32	1.309	1.694	2.037	2.449	2.738
34	1.307	1.691	2.032	2.441	2.728
36	1.306	1.688	2.028	2.434	2.719
38	1.304	1.686	2.024	2.429	2.712
∞	1.282	1.645	1.960	2.326	2.576

4.5.2 การหาช่วงความเชื่อมั่นของค่าความบริสุทธิ์ที่ได้จากการทดสอบอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

เมื่อแทนค่าที่ได้จากการทดสอบลงในสมการโดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x} = 0.647$

$s = 0.128$

ช่วงความเชื่อมั่น $= 0.647 \pm t_{10-1,\,1-0.05/2} \dfrac{0.128}{\sqrt{10}}$

$= 0.647 \pm 2.262 \left(\dfrac{0.128}{3.16} \right)$

$= 0.647 \pm 0.092$

$= [0.647 - 0.092,\ 0.647 + 0.092]$

$= [0.555,\ 0.739]$

แสดงว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดนั้น สามารถมั่นใจได้ถึงร้อยละ 95 ว่าจะมีค่าความบริสุทธิ์อยู่ในช่วงตั้งแต่ 0.555 ถึง 0.739

4.5.3 การหาช่วงความเชื่อมั่นของค่าความบริสุทธิ์ที่ได้จากการทดสอบอัลกอริธึมที่พัฒนาขึ้น

เมื่อแทนค่าที่ได้จากการทดสอบลงในสมการโดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x} = 0.671$

$s = 0.111$

ช่วงความเชื่อมั่น $= 0.671 \pm t_{10-1,\,1-0.05/2} \dfrac{0.111}{\sqrt{10}}$

$= 0.671 \pm 2.262\left(\dfrac{0.111}{3.16}\right)$

$= 0.671 \pm 0.079$

$= [0.671 - 0.079,\ 0.671 + 0.079]$

$= [0.592,\ 0.75]$

แสดงว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมที่พัฒนาขึ้นนั้น สามารถมั่นใจได้ถึงร้อยละ 95 ว่าจะมีค่าความบริสุทธิ์อยู่ในช่วงตั้งแต่ 0.592 ถึง 0.75

4.5.4 การหาช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ที่ได้จากการทดสอบอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

เมื่อแทนค่าที่ได้จากการทดสอบลงในสมการโดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x} = 0.536$

$s = 0.132$

ช่วงความเชื่อมั่น $= 0.536 \pm t_{10-1,\,1-0.05/2} \frac{0.132}{\sqrt{10}}$

$= 0.536 \pm 2.262 \left(\frac{0.132}{3.16} \right)$

$= 0.536 \pm 0.095$

$= [0.536 - 0.095, 0.536 + 0.095]$

$= [0.441, 0.631]$

แสดงว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดนั้น สามารถมั่นใจได้ถึงร้อยละ 95 ว่าจะมีค่าเอฟเมเซอร์อยู่ในช่วงตั้งแต่ 0.441 ถึง 0.631

4.5.5 การหาช่วงความเชื่อมั่นของค่าเอฟเมเชอร์ที่ได้จากการทดสอบอัลกอริธึมที่พัฒนาขึ้น

เมื่อแทนค่าที่ได้จากการทดสอบลงในสมการ โดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x} = 0.526$

$s = 0.117$

ช่วงความเชื่อมั่น $= 0.526 \pm t_{10-1,\,1-0.05/2} \frac{0.117}{\sqrt{10}}$

$= 0.526 \pm 2.262\left(\frac{0.117}{3.16}\right)$

$= 0.526 \pm 0.083$

$= [0.526 - 0.083,\ 0.526 + 0.083]$

$= [0.443,\ 0.609]$

แสดงว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมที่พัฒนาขึ้นนั้น สามารถมั่นใจได้ถึงร้อยละ 95 ว่าจะมีค่าเอฟเมเชอร์อยู่ที่ในช่วงตั้งแต่ 0.443 ถึง 0.609

4.5.6 การหาช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของค่าความบริสุทธิ์ระหว่างอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดและอัลกอริธึมที่พัฒนาขึ้น

การหาช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่าง สามารถคำนวณได้โดยนำผลต่างของทุกข้อมูลมาหาค่าเฉลี่ย แล้วนำไปแทนค่าในสมการเดียวกับการหาช่วงความเชื่อมั่น

เมื่อแทนค่าที่ได้ลงในสมการ โดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x_1 - x_2} = -0.024$

$s = 0.0216$

โดย x_1 คือ ค่าความบริสุทธิ์ของอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

x_2 คือ ค่าความบริสุทธิ์ของอัลกอริธึมที่พัฒนาขึ้น

$$\text{ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่าง} = -0.024 \pm t_{10-1,\,1-0.05/2}\,\frac{0.0216}{\sqrt{10}}$$

$$= -0.024 \pm 2.262\left(\frac{0.0216}{3.16}\right)$$

$$= -0.024 \pm 0.0154$$

$$= [-0.024 - 0.0154, -0.024 + 0.0154]$$

$$= [-0.0394, -0.0086]$$

จากช่วงความเชื่อมั่นที่ได้ พบว่าไม่มีค่า 0 อยู่ในช่วง ซึ่งทำให้สรุปได้ว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมทั้งสอง สามารถมั่นใจได้ถึงร้อยละ 95 ว่ามีค่าความบริสุทธิ์แตกต่างกัน โดยอัลกอริธึมที่พัฒนาขึ้นให้ค่าความบริสุทธิ์มากกว่า

4.5.7 การหาช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดและอัลกอริธึมที่พัฒนาขึ้น

การหาช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่าง สามารถคำนวณได้โดยนำผลต่างของทุกข้อมูลมาหาค่าเฉลี่ย แล้วนำไปแทนค่าในสมการเดียวกับการหาช่วงความเชื่อมั่น

เมื่อแทนค่าที่ได้ลงในสมการ โดยทำการคำนวณค่าที่ระดับความเชื่อมั่นร้อยละ 95 จะได้ผลลัพธ์ ดังนี้

$\overline{x_1 - x_2} = 0.0093$

$s = 0.0332$

โดย x_1 คือ ค่าเอฟเมเซอร์ของอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

x_2 คือ ค่าเอฟเมเซอร์ของอัลกอริธึมที่พัฒนาขึ้น

ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่าง $= 0.0093 \pm t_{10-1,\,1-0.05/2} \frac{0.0332}{\sqrt{10}}$

$= 0.0093 \pm 2.262 \left(\frac{0.0332}{3.16}\right)$

$= 0.0093 \pm 0.0238$

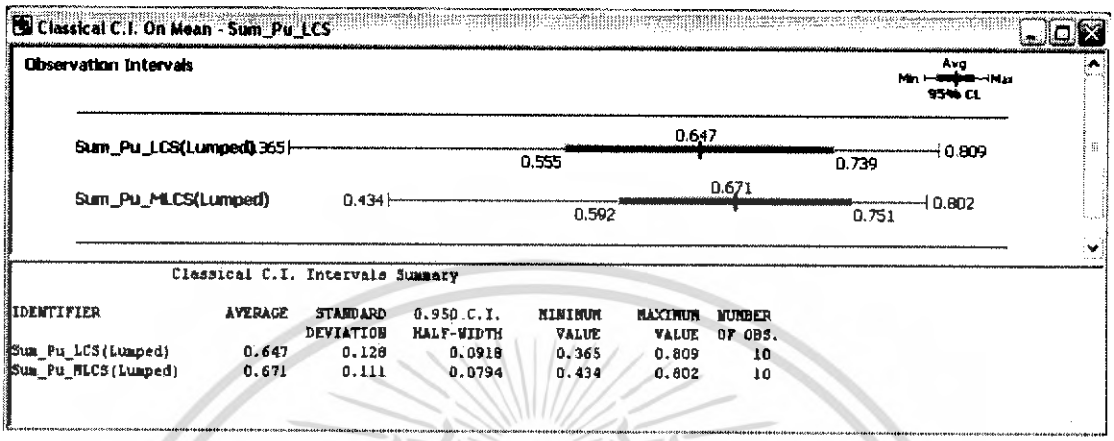
$= [0.0093 - 0.0238, 0.0093 + 0.0238]$

$= [-0.0145, 0.0331]$

จากช่วงความเชื่อมั่นที่ได้ พบว่ามีค่า 0 อยู่ในช่วง ซึ่งทำให้สรุปได้ว่าผลลัพธ์ที่ได้จากการทดสอบอัลกอริธึมทั้งสอง สามารถมั่นใจได้ถึงร้อยละ 95 ว่ามีค่าเอฟเมเซอร์ไม่แตกต่างกัน

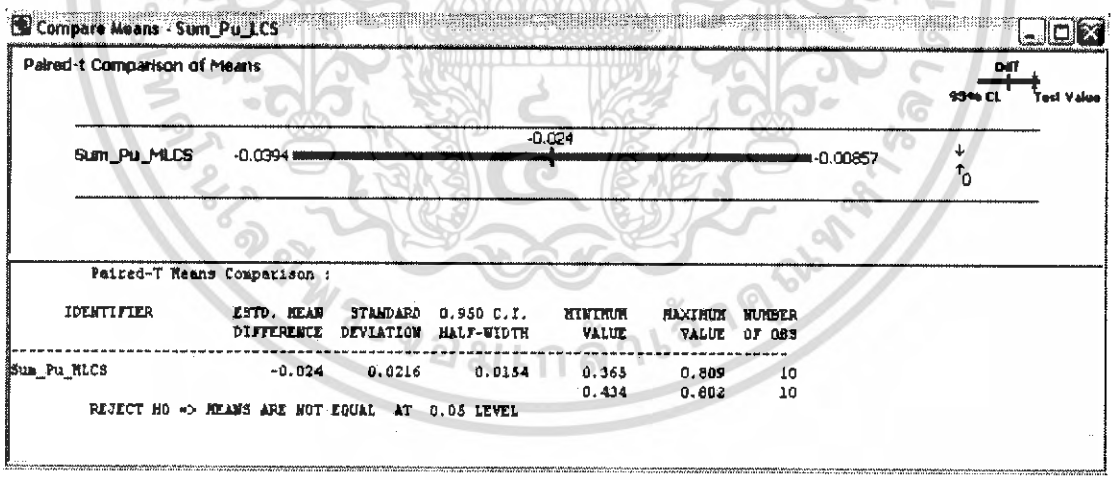
4.6 ภาพแสดงการเปรียบเทียบช่วงความเชื่อมั่น

4.6.1 ค่าความบริสุทธิ์



ภาพที่ 4.1 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของทุกชุดข้อมูล

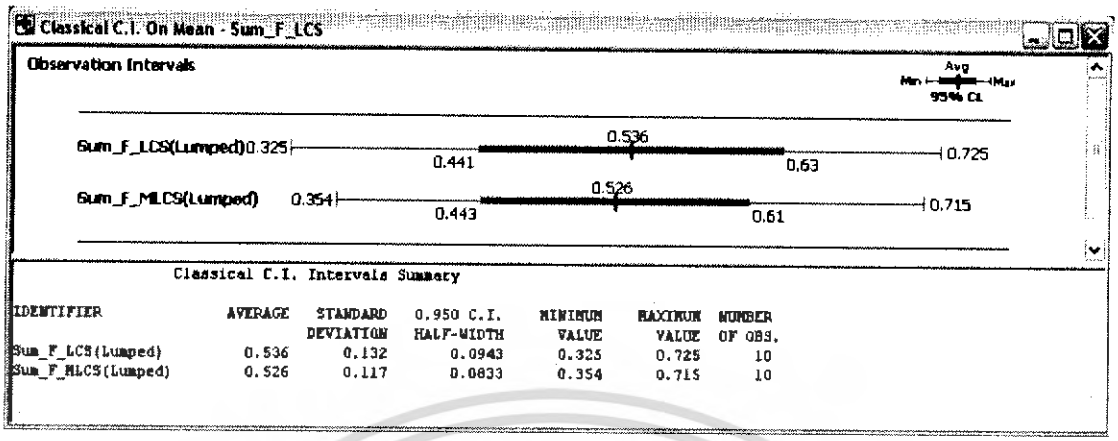
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ 4.2 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของทุกชุดข้อมูล

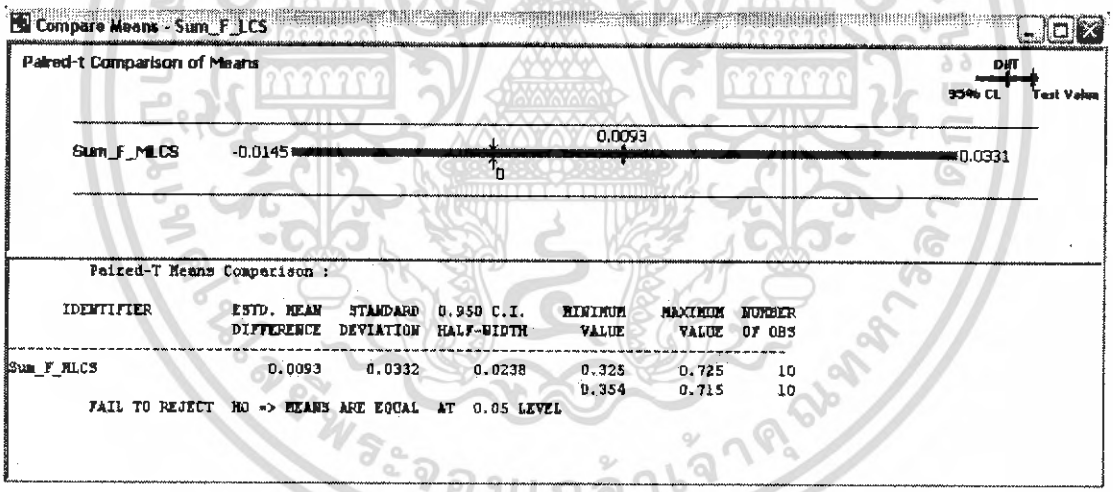
เนื่องจากไม่มีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์แตกต่างกัน โดยอัลกอริธึมที่พัฒนาขึ้นให้ค่าความบริสุทธิ์มากกว่าอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด

4.6.2 ค่าเอฟเมเชอร์



ภาพที่ 4.3 ช่วงความเชื่อมั่นของค่าเอฟเมเชอร์ของทุกชุดข้อมูล

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเชอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ 4.4 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของทุกชุดข้อมูล

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเชอร์ไม่แตกต่างกัน

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุป

ปัญหาพิเศษนี้มีแนวคิดที่จะพัฒนาอัลกอริธึมในการหาระยะห่างระหว่างเวกเตอร์ของข้อมูลประเภทสตริงให้มีประสิทธิผลมากขึ้น โดยพิจารณาความแตกต่างของระยะห่างระหว่างตำแหน่งตัวอักษรของลำดับร่วมที่ยาวที่สุดในสตริงต้นแบบด้วย แต่จะไม่พิจารณาถึงประสิทธิภาพในด้านอื่น เช่น ความเร็วในการทำงานของอัลกอริธึม เป็นต้น

การพัฒนาอัลกอริธึมในการหาระยะห่างระหว่างเวกเตอร์ของข้อมูลประเภทสตริงจำเป็นต้องอาศัยความรู้เรื่องคลัสเตอร์ริ่ง ซึ่งในปัญหาพิเศษนี้ ประเภทของคลัสเตอร์ริ่งที่นำมาใช้ในการจัดกลุ่มเป็นคลัสเตอร์ริ่งแบบลำดับชั้น นอกจากนี้ยังรวมถึงความรู้เรื่องอัลกอริธึมที่ใช้ในการหาระยะห่างระหว่างเวกเตอร์ของข้อมูลประเภทสตริง ซึ่งในที่นี้เลือกอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด เพื่อนำมาใช้ในการเปรียบเทียบกับอัลกอริธึมที่พัฒนาขึ้น

วิธีการวัดประสิทธิผลของอัลกอริธึมที่พัฒนาขึ้นในปัญหาพิเศษนี้ พิจารณาความถูกต้องของการจัดกลุ่มในลำดับชั้นที่ให้ผลลัพธ์ของจำนวนกลุ่มเท่ากับจำนวนของประเภทข้อมูล โดยวัดจากค่าความบริสุทธิ์และค่าเอฟเมเชอร์ จากผลการทดสอบพบว่า อัลกอริธึมที่พัฒนาขึ้นให้ค่าเอฟเมเชอร์ใกล้เคียงกับอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด แต่ให้ค่าความบริสุทธิ์มากกว่าอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุด ซึ่งแสดงให้เห็นว่ากลุ่มข้อมูลที่ได้จากการจัดกลุ่มของอัลกอริธึมที่พัฒนาขึ้นมีข้อมูลที่มีความคล้ายคลึงกันเป็นจำนวนมากและมีข้อมูลที่มีความแตกต่างกันเป็นจำนวนน้อย

5.2 ข้อเสนอแนะ

ปัญหาพิเศษนี้ไม่ได้พัฒนาประสิทธิภาพของอัลกอริธึมในด้านอื่น นอกจากด้านความถูกต้องในการจัดกลุ่ม และไม่ได้ทำการทดสอบอัลกอริธึมกับคลัสเตอร์ริ่งแบบอื่น ผู้ที่ทำการพัฒนาต่ออาจปรับปรุงอัลกอริธึมให้ทำงานได้รวดเร็วมากยิ่งขึ้น เนื่องจากคลัสเตอร์ริ่งมักใช้เพื่อแบ่งข้อมูลจำนวนมากออกเป็นกลุ่มย่อย ซึ่งต้องใช้การคำนวณสูง หรือนำอัลกอริธึมไปประยุกต์ใช้กับคลัสเตอร์ริ่งแบบอื่นต่อไป

บรรณานุกรม

พิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย. 2547. "การปรับปรุงการแบ่งกลุ่มเอกสารโดยใช้ฐานความรู้เวิร์ดเนต." หน้า 380-388. ใน **NCSEC 2004.**

Bergroth, L. et. al. 2000. **"A Survey of Longest Common Subsequence Algorithms."** IEEE Transactions.

Consens M. and Navarro G. 2005. "Evaluating Hicrarchical Clustering of Search Results." 49-54 in **SPIRE 2005.** LNCS 3772. Springer-Verlag Berlin Heidelberg 2005.

Crabtree, Daniel. et. al. 2005. "Standardized Evaluation Method for Web Clustering Results." in Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05).

Jain, Anil K. and Dubes, Richard C. 1988. **Algorithms for Clustering Data.** New Jersey : Prentice Hall.

Jiang, D. et. al. 2004. "Cluster Analysis for Gene Expression Data : A Survey." **IEEE Transactions on Knowledge and Data Engineering.** 16(11).

Hierarchical Clustering. [Online]. Available : https://www.resample.com/xlminer/help/HClst/HClst_ex.htm. 2006.

Hirschberg, Daniel S. 1977. "Algorithms for the Longest Common Subsequence Problem." **Journal of the Association for Computing Machinery**. 24(4) : 664-675.

Meunier, Bruno. et. al. 2007. "Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining." **Journal of Proteome Research.** 6(1) : 358-366.

Needleman, Sual B. and Wunsch, Christian D. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." **J. Mol. Biol.** 1970(48) : 443-453.

Sousa, Fernanda and Tendeiro, Jorge. 2005. "A Validation Methodology in Hierarchical Clustering." 396-403 in Proceedings ASMDA 2005 Part V. Clustering.

ภาคผนวก ก.
ตัวอย่างข้อมูลที่ใช้ในการทดสอบอัลกอริธึม

1. Car Evaluation ข้อมูลชนิดสตริง 172 ข้อมูล แต่ละข้อมูลมี 6 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 4 คลาส คือ unacc, acc, good และ vgood ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

vhigh,vhigh,2,2,small,low,unacc
vhigh,vhigh,2,2,small,med,unacc
vhigh,vhigh,2,2,small,high,unacc
vhigh,vhigh,2,2,med,low,unacc
vhigh,vhigh,2,2,big,low,unacc
high,low,4,more,small,low,unacc
high,low,4,more,small,med,unacc
high,low,4,more,small,high,acc
med,vhigh,3,4,big,high,acc
low,vhigh,3,more,med,med,acc
low,low,5more,more,small,high,good
low,low,5more,more,med,med,good
low,low,5more,more,med,high,vgood
low,low,5more,more,big,med,good
low,low,5more,more,big,high,vgood
low,med,5more,4,big,med,good
low,med,5more,4,big,high,vgood
low,med,5more,more,med,low,unacc
low,med,5more,more,med,med,good
low,low,2,4,med,med,acc

2. Credit Approval ข้อมูลชนิดสตริงและตัวเลข 68 ข้อมูล แต่ละข้อมูลมี 15 แอททริบิวท์ (สตริง 9 แอททริบิวท์ และตัวเลข 6 แอททริบิวท์) แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ + และ - ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

b,30.83,0,u,g,w,v,1.25,t,t,01,f,g,00202,0,+

a,58.67,4.46,u,g,q,h,3.04,t,t,06,f,g,00043,560,+

a,24.50,0.5,u,g,q,h,1.5,t,f,0,f,g,00280,824,+

b,27.83,1.54,u,g,w,v,3.75,t,t,05,t,g,00100,3,+

b,20.17,5.625,u,g,w,v,1.71,t,f,0,f,s,00120,0,+

b,32.08,4,u,g,m,v,2.5,t,f,0,t,g,00360,0,+

b,33.17,1.04,u,g,r,h,6.5,t,f,0,t,g,00164,31285,+

a,22.92,11.585,u,g,cc,v,0.04,t,f,0,f,g,00080,1349,+

b,54.42,0.5,y,p,k,h,3.96,t,f,0,f,g,00180,314,+

b,42.50,4.915,y,p,w,v,3.165,t,f,0,t,g,00052,1442,+

a,20.08,1.25,u,g,c,v,0,f,f,0,f,g,00000,0,-

b,19.50,0.29,u,g,k,v,0.29,f,f,0,f,g,00280,364,-

b,27.83,1,y,p,d,h,3,f,f,0,f,g,00176,537,-

b,17.08,3.29,u,g,i,v,0.335,f,f,0,t,g,00140,2,-

b,36.42,0.75,y,p,d,v,0.585,f,f,0,f,g,00240,3,-

b,40.58,3.29,u,g,m,v,3.5,f,f,0,t,s,00400,0,-

b,21.08,10.085,y,p,e,h,1.25,f,f,0,f,g,00260,0,-

a,22.67,0.75,u,g,c,v,2,f,t,02,t,g,00200,394,-

a,25.25,13.5,y,p,ff,ff,2,f,t,01,t,g,00200,1,-

b,17.92,0.205,u,g,aa,v,0.04,f,f,0,f,g,00280,750,-

3. Hepatitis ข้อมูลชนิดตัวเลข 155 ข้อมูล แต่ละข้อมูลมี 19 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ 1 และ 2 ซึ่งระบุไว้ที่ตำแหน่งแรกของแต่ละข้อมูล และ ? คือแอททริบิวท์ที่ไม่ทราบค่า

2,30,2,1,2,2,2,2,1,2,2,2,2,2,2,1.00,85,18,4.0,?,1

2,50,1,1,2,1,2,2,1,2,2,2,2,2,2,0.90,135,42,3.5,?,1

2,78,1,2,2,1,2,2,2,2,2,2,2,2,2,0.70,96,32,4.0,?,1

2,31,1,?,1,2,2,2,2,2,2,2,2,2,2,0.70,46,52,4.0,80,1

2,34,1,2,2,2,2,2,2,2,2,2,2,2,2,1.00,?,200,4.0,?,1

2,34,1,2,2,2,2,2,2,2,2,2,2,2,2,0.90,95,28,4.0,75,1

1,51,1,1,2,1,2,1,2,2,1,1,2,2,?,?,?,?,?,1

2,23,1,2,2,2,2,2,2,2,2,2,2,2,2,1.00,?,?,?,?,1

2,39,1,2,2,1,2,2,2,1,2,2,2,2,2,0.70,?,48,4.4,?,1

2,30,1,2,2,2,2,2,2,2,2,2,2,2,2,1.00,?,120,3.9,?,1

2,39,1,1,1,2,2,2,1,1,2,2,2,2,2,1.30,78,30,4.4,85,1

2,32,1,2,1,1,2,2,2,1,2,1,2,2,2,1.00,59,249,3.7,54,1

2,41,1,2,1,1,2,2,2,1,2,2,2,2,2,0.90,81,60,3.9,52,1

2,30,1,2,2,1,2,2,2,1,2,2,2,2,2,2.20,57,144,4.9,78,1

2,47,1,1,1,2,2,2,2,2,2,2,2,2,?,?,60,?,?,1

2,38,1,1,2,1,1,1,2,2,2,2,1,2,2,2.00,72,89,2.9,46,1

2,66,1,2,2,1,2,2,2,2,2,2,2,2,2,1.20,102,53,4.3,?,1

2,40,1,1,2,1,2,2,2,1,2,2,2,2,2,0.60,62,166,4.0,63,1

2,38,1,2,2,2,2,2,2,2,2,2,2,2,2,0.70,53,42,4.1,85,2

2,38,1,1,1,2,2,2,1,1,2,2,2,2,2,0.70,70,28,4.2,62,1

4. **Iris Plant** ข้อมูลชนิดตัวเลข 150 ข้อมูล แต่ละข้อมูลมี 4 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 3 คลาส คือ Iris-setosa, Iris-versicolor และ Iris-virginica ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

5.1,3.5,I.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.I,I.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica

5. Liver Disorders ข้อมูลชนิดตัวเลข 68 ข้อมูล แต่ละข้อมูลมี 6 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ 1 และ 2 ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

85,92,45,27,31,0.0,1
85,64,59,32,23,0.0,2
86,54,33,16,54,0.0,2
91,78,34,24,36,0.0,2
87,70,12,28,10,0.0,2
98,55,13,17,17,0.0,2
88,62,20,17,9,0.5,1
88,67,21,11,11,0.5,1
92,54,22,20,7,0.5,1
90,60,25,19,5,0.5,1
89,52,13,24,15,0.5,1
82,62,17,17,15,0.5,1
90,64,61,32,13,0.5,1
86,77,25,19,18,0.5,1
96,67,29,20,11,0.5,1
91,78,20,31,18,0.5,1
89,67,23,16,10,0.5,1
89,79,17,17,16,0.5,1
91,107,20,20,56,0.5,1
94,116,11,33,11,0.5,1

6. Pendigits ข้อมูลชนิดตัวเลข 90 ข้อมูล แต่ละข้อมูลมี 16 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 10 คลาส คือ 0-9 ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

88,92,2,99,16,66,94,37,70,0,0,24,42,65,100,100,8
80,100,18,98,60,66,100,29,42,0,0,23,42,61,56,98,8
0,94,9,57,20,19,7,0,20,36,70,68,100,100,18,92,8
95,82,71,100,27,77,77,73,100,80,93,42,56,13,0,0,9
68,100,6,88,47,75,87,82,85,56,100,29,75,6,0,0,9
70,100,100,97,70,81,45,65,30,49,20,33,0,16,0,0,1
40,100,0,81,15,58,100,57,47,87,50,88,40,42,36,0,4
3,71,0,95,45,100,100,99,79,78,48,53,31,24,54,0,7
79,87,98,81,71,100,72,73,100,66,91,21,48,0,0,13,9
92,95,30,100,34,68,87,89,84,78,100,35,64,0,0,19,9
58,64,100,96,27,100,0,63,79,65,91,72,48,36,10,0,9
34,89,3,70,1,25,49,0,100,23,100,67,56,99,0,100,0
0,90,46,100,88,92,79,69,60,48,39,27,47,6,100,0,2
20,71,0,29,31,0,78,12,100,51,84,93,37,100,8,66,0
100,100,67,98,41,80,44,50,78,42,68,16,35,2,0,0,5
91,69,48,57,9,79,60,100,100,75,95,40,64,8,0,0,9
30,74,55,100,89,87,66,56,100,38,92,8,41,0,0,20,3
5,65,0,89,37,100,88,97,100,79,71,53,48,26,59,0,7
42,93,19,88,0,42,24,0,83,11,100,56,75,100,17,97,0
4,100,0,72,15,44,79,50,100,76,90,51,83,22,85,0,4

7. Pima Indians Diabetes ข้อมูลชนิดตัวเลข 77 ข้อมูล แต่ละข้อมูลมี 8 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ 0 และ 1 ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

4,144,82,32,0,38.5,0.554,37,1

14,175,62,30,0,33.6,0.212,38,1

8,100,74,40,215,39.4,0.661,43,1

7,114,64,0,0,27.4,0.732,34,1

1,199,76,43,0,42.9,1.394,22,1

1,144,82,46,180,46.1,0.335,46,1

7,129,68,49,125,38.5,0.439,43,1

1,102,74,0,0,39.5,0.293,42,1

3,187,70,22,200,36.4,0.408,36,1

6,190,92,0,0,35.5,0.278,66,1

1,85,66,29,0,26.6,0.351,31,0

5,116,74,0,0,25.6,0.201,30,0

1,103,80,11,82,19.4,0.491,22,0

0,100,88,60,110,46.8,0.962,31,0

4,146,85,27,100,28.9,0.189,27,0

7,62,78,0,0,32.6,0.391,41,0

3,113,44,13,0,22.4,0.140,22,0

4,123,80,15,176,32.0,0.443,34,0

1,96,122,0,0,22.4,0.207,27,0

4,97,60,23,0,28.2,0.443,22,0

8. Promoter Gene Sequence ข้อมูลชนิดสตริง 106 ข้อมูล แต่ละข้อมูลมี 57 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ + และ - ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

t,a,c,t,a,g,c,a,a,t,a,c,g,c,t,t,g,c,g,t,t,c,g,g,t,g,g,t,t,a,a,g,t,a,t,g,t,a,t,a,a,t,g,c,g,c,g,g,g,c,t,t,g,t,c,g,t,+

t,g,c,t,a,t,c,c,t,g,a,c,a,g,t,t,g,t,c,a,c,g,c,t,g,a,t,t,g,g,t,g,t,c,g,t,t,a,c,a,a,t,c,t,a,a,c,g,c,a,t,c,g,c,c,a,a,+

g,t,a,c,t,a,g,a,g,a,a,c,t,a,g,t,g,c,a,t,t,a,g,c,t,t,a,t,t,t,t,t,t,t,g,t,t,a,t,c,a,t,g,c,t,a,a,c,c,a,c,c,c,g,g,c,g,+

a,a,t,t,g,t,g,a,t,g,t,g,t,a,t,c,g,a,a,g,t,g,t,g,t,t,g,c,g,g,a,g,t,a,g,a,t,g,t,t,a,g,a,a,t,a,c,t,a,a,c,a,a,a,a,c,t,c,+

t,c,g,a,t,a,a,t,t,a,a,c,t,a,t,t,g,a,c,g,a,a,a,a,a,g,c,t,g,a,a,a,a,a,c,c,a,c,t,a,g,a,a,t,g,c,g,c,c,t,c,c,g,t,g,g,t,a,g,+

a,g,g,g,g,c,a,a,g,g,a,g,g,a,t,g,g,a,a,a,g,a,g,g,t,t,g,c,c,g,t,a,t,a,a,a,g,a,a,a,c,t,a,g,a,g,t,c,c,g,t,t,t,a,g,g,t,+

c,a,g,g,g,g,g,t,g,g,a,g,g,a,t,t,t,a,a,g,c,c,a,t,c,t,c,c,t,g,a,t,g,a,c,g,c,a,t,a,g,t,c,a,g,c,c,c,a,t,c,a,t,g,a,a,t,+

t,t,t,c,t,a,c,a,a,a,a,c,a,c,t,t,g,a,t,a,c,t,g,t,a,t,g,a,g,c,a,t,a,c,a,g,t,a,t,a,a,t,t,g,c,t,t,c,a,a,c,a,g,a,a,c,a,+

c,g,a,c,t,t,a,a,t,a,t,a,c,t,g,c,g,a,c,a,g,g,a,c,g,t,c,c,g,t,t,c,t,g,t,g,t,a,a,a,t,c,g,c,a,a,t,g,a,a,a,t,g,g,t,t,t,+

t,t,t,t,a,a,a,t,t,t,c,c,t,c,t,t,g,t,c,a,g,g,c,c,g,g,a,a,t,a,a,c,t,c,c,c,t,a,t,a,a,t,g,c,g,c,c,a,c,c,a,c,t,g,a,c,a,+

c,c,g,a,g,t,a,g,a,c,c,c,t,t,a,g,a,g,a,g,c,a,t,g,t,c,a,g,c,c,t,c,g,a,c,a,a,c,t,t,g,c,a,t,a,a,a,t,g,c,t,t,t,c,t,t,g,-

c,g,c,t,a,g,g,a,c,t,t,t,c,t,t,g,t,t,g,a,t,t,t,t,c,c,a,t,g,c,g,g,t,g,t,t,t,t,g,c,g,c,a,a,t,g,t,t,a,a,t,c,g,c,t,t,t,-

t,a,t,g,a,c,c,g,a,a,c,g,a,g,t,c,a,a,t,c,a,g,a,c,c,g,c,t,t,t,g,a,c,t,c,t,g,g,t,a,t,t,a,c,t,g,t,g,a,a,c,a,t,t,a,t,t,-

a,g,a,g,g,g,t,g,t,a,c,t,c,c,a,a,g,a,a,g,a,g,g,a,a,g,a,t,g,a,g,g,c,t,a,g,a,c,g,t,c,t,c,t,g,c,a,t,g,g,a,g,t,a,t,g,a,-

g,a,g,a,g,c,a,t,g,t,c,a,g,c,c,t,c,g,a,c,a,a,c,t,t,g,c,a,t,a,a,a,t,g,c,t,t,t,c,t,t,g,t,a,g,a,c,g,t,g,c,c,c,t,a,c,g,-

c,c,t,c,a,a,t,g,g,c,c,t,c,t,a,a,a,c,g,g,g,t,c,t,t,g,a,g,g,g,g,t,t,t,t,t,g,c,t,g,a,a,a,g,g,a,g,g,g,a,a,c,t,a,t,a,t,-

g,t,a,t,t,c,t,c,a,a,c,a,a,g,a,t,t,a,a,c,c,g,a,c,a,g,a,t,t,c,a,a,t,c,t,c,g,t,g,g,a,t,g,g,a,c,g,t,t,c,a,a,c,a,t,t,g,-

c,g,c,g,a,c,t,a,c,g,a,t,g,a,g,a,t,g,c,c,t,g,a,g,t,g,c,t,t,c,c,g,t,t,a,c,t,g,g,a,t,t,g,t,c,a,c,c,a,a,g,g,c,t,t,c,c,-

c,t,c,g,t,c,c,t,c,a,a,t,g,g,c,c,t,c,t,a,a,a,c,g,g,g,t,c,t,t,g,a,g,g,g,g,t,t,t,t,t,t,g,c,t,g,a,a,a,g,g,a,g,g,a,a,c,-

t,a,a,c,a,t,t,a,a,t,a,a,a,t,a,a,g,g,a,g,g,c,t,c,t,a,a,t,g,g,c,a,c,t,c,a,t,t,a,g,c,c,a,a,t,c,a,a,t,c,a,a,g,a,a,c,t,-

9. Sonar Mines vs. Rocks ข้อมูลชนิดตัวเลข 21 ข้อมูล แต่ละข้อมูลมี 60 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 2 คลาส คือ M และ R ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

0.0587,0.1210,0.1268,0.1498,0.1436,0.0561,0.0832,0.0672,0.1372,0.2352,0.3208,0.4257,0.52
01,0.4914,0.5950,0.7221,0.9039,0.9111,0.8723,0.7686,0.7326,0.5222,0.3097,0.3172,0.2270,0
.1640,0.1746,0.1835,0.2048,0.1674,0.2767,0.3104,0.3399,0.4441,0.5046,0.2814,0.1681,0.263
3,0.3198,0.1933,0.0934,0.0443,0.0780,0.0722,0.0405,0.0553,0.1081,0.1139,0.0767,0.0265,0.
0215,0.0331,0.0111,0.0088,0.0158,0.0122,0.0038,0.0101,0.0228,0.0124,M
0.0116,0.0179,0.0449,0.1096,0.1913,0.0924,0.0761,0.1092,0.0757,0.1006,0.2500,0.3988,0.38
09,0.4753,0.6165,0.6464,0.8024,0.9208,0.9832,0.9634,0.8646,0.8325,0.8276,0.8007,0.6102,0
.4853,0.4355,0.4307,0.4399,0.3833,0.3032,0.3035,0.3197,0.2292,0.2131,0.2347,0.3201,0.445
5,0.3655,0.2715,0.1747,0.1781,0.2199,0.1056,0.0573,0.0307,0.0237,0.0470,0.0102,0.0057,0.
0031,0.0163,0.0099,0.0084,0.0270,0.0277,0.0097,0.0054,0.0148,0.0092,M
0.0336,0.0294,0.0476,0.0539,0.0794,0.0804,0.1136,0.1228,0.1235,0.0842,0.0357,0.0689,0.17
05,0.3257,0.4602,0.6225,0.7327,0.7843,0.7988,0.8261,1.0000,0.9814,0.9620,0.9601,0.9118,0
.9086,0.7931,0.5877,0.3474,0.4235,0.4633,0.3410,0.2849,0.2847,0.1742,0.0549,0.1192,0.115
4,0.0855,0.1811,0.1264,0.0799,0.0378,0.1268,0.1125,0.0505,0.0949,0.0677,0.0259,0.0170,0.
0033,0.0150,0.0111,0.0032,0.0035,0.0169,0.0137,0.0015,0.0069,0.0051,R
0.0253,0.0808,0.0507,0.0244,0.1724,0.3823,0.3729,0.3583,0.3429,0.2197,0.2653,0.3223,0.55
82,0.6916,0.7943,0.7152,0.3512,0.2008,0.2676,0.4299,0.5280,0.3489,0.1430,0.5453,0.6338,0
.7712,0.6838,0.8015,0.8073,0.8310,0.7792,0.5049,0.1413,0.2767,0.5084,0.4787,0.1356,0.229
9,0.2789,0.3833,0.2933,0.1155,0.1705,0.1294,0.0909,0.0800,0.0567,0.0198,0.0114,0.0151,0.
0085,0.0178,0.0073,0.0079,0.0038,0.0116,0.0033,0.0039,0.0081,0.0053,R

10. Wine Recognition ข้อมูลชนิดตัวเลข 178 ข้อมูล แต่ละข้อมูลมี 13 แอททริบิวท์ แต่ละแอททริบิวท์คั่นด้วยเครื่องหมายจุลภาค (,) แบ่งข้อมูลเป็น 3 คลาส คือ 1, 2 และ 3 ซึ่งระบุไว้ที่ตำแหน่งสุดท้ายของแต่ละข้อมูล

14.75,1.73,2.39,11.4,91,3.1,3.69,0.43,2.81,5.4,1.25,2.73,1150,1

12.85,1.6,2.52,17.8,95,2.48,2.37,0.26,1.46,3.93,1.09,3.63,1015,1

13.05,1.77,2.1,17,107,3,3,0.28,2.03,5.04,0.88,3.35,885,1

14.21,4.04,2.44,18.9,111,2.85,2.65,0.3,1.25,5.24,0.87,3.33,1080,1

13.83,1.65,2.6,17.2,94,2.45,2.99,0.22,2.29,5.6,1.24,3.37,1265,1

13.32,3.24,2.38,21.5,92,1.93,0.76,0.45,1.25,8.42,0.55,1.62,650,3

13.08,3.9,2.36,21.5,113,1.41,1.39,0.34,1.14,9.40,0.57,1.33,550,3

13.23,3.3,2.28,18.5,98,1.8,0.83,0.61,1.87,10.52,0.56,1.51,675,3

13.84,4.12,2.38,19.5,89,1.8,0.83,0.48,1.56,9.01,0.57,1.64,480,3

13.34,0.94,2.36,17,110,2.53,1.3,0.55,0.42,3.17,1.02,1.93,750,2

13.49,1.66,2.24,24,87,1.88,1.84,0.27,1.03,3.74,0.98,2.78,472,2

12,0.92,2,19,86,2.42,2.26,0.3,1.43,2.5,1.38,3.12,278,2

12.08,1.83,2.32,18.5,81,1.6,1.5,0.52,1.64,2.4,1.08,2.27,480,2

12.47,1.52,2.2,19,162,2.5,2.27,0.32,3.28,2.6,1.16,2.63,937,2

12.29,3.17,2.21,18,88,2.85,2.99,0.45,2.81,2.3,1.42,2.83,406,2

11.03,1.51,2.2,21.5,85,2.46,2.17,0.52,2.01,1.9,1.71,2.87,407,2

13.56,1.73,2.46,20.5,116,2.96,2.78,0.2,2.45,6.25,0.98,3.03,1120,1

14.22,1.7,2.3,16.3,118,3.2,3,0.26,2.03,6.38,0.94,3.31,970,1

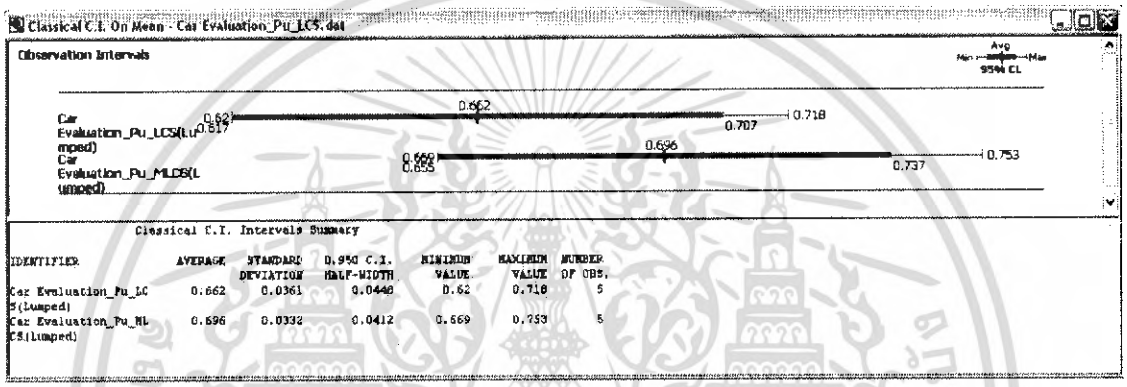
12.84,2.96,2.61,24,101,2.32,0.6,0.53,0.81,4.92,0.89,2.15,590,3

12.96,3.45,2.35,18.5,106,1.39,0.7,0.4,0.94,5.28,0.68,1.75,675,3

ภาคผนวก ข.
การหาช่วงความเชื่อมั่นของแต่ละชุดข้อมูล

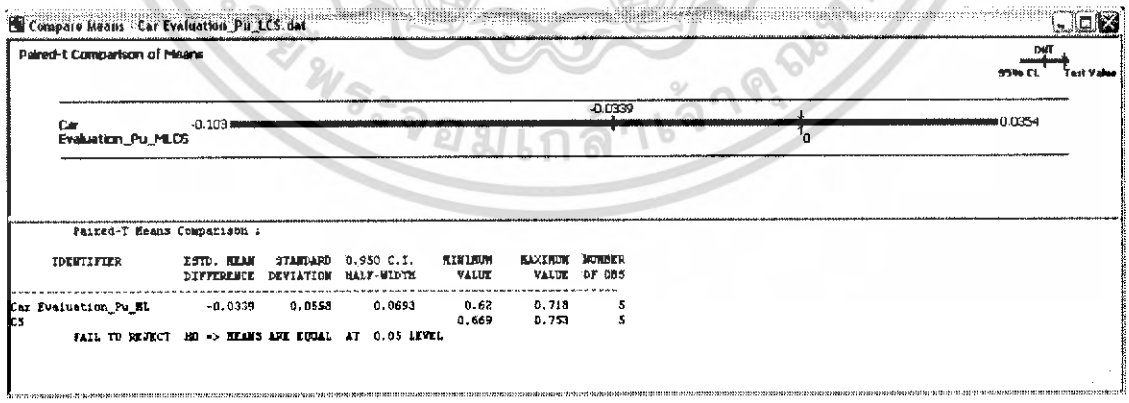
นำค่าความบริสุทธิ์และค่าเอฟเมเซอร์ที่ได้จากการจัดกลุ่มข้อมูลแต่ละชุดมาหาช่วงความเชื่อมั่นโดยใช้โปรแกรมเอาท์พุทอนาไลเซอร์ (Output Analyzer) ซึ่งอยู่ในชุดโปรแกรมอารีนา 7.0 (Arena 7.0) ได้ผลดังนี้

1. Car Evaluation



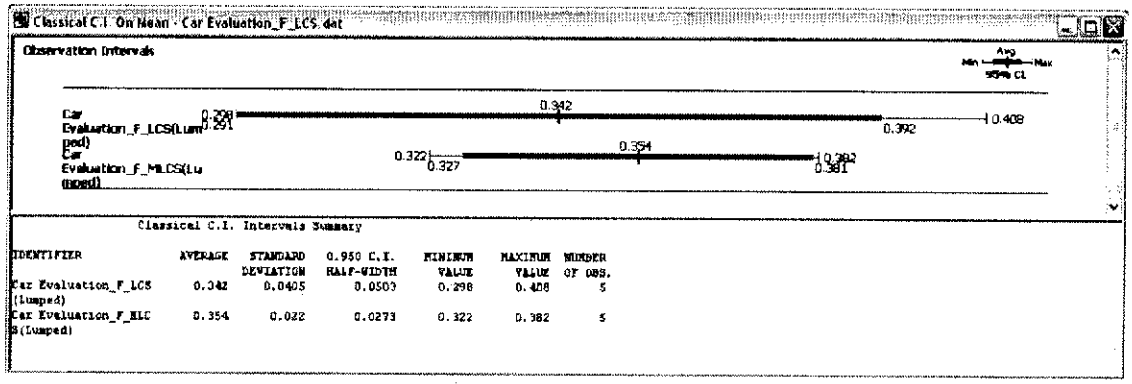
ภาพที่ ข.1 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Car Evaluation

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



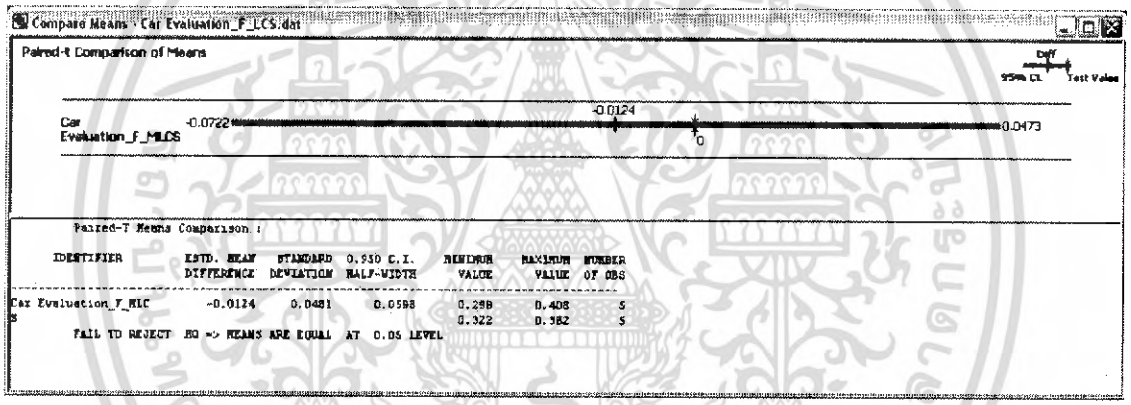
ภาพที่ ข.2 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Car Evaluation

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.3 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Car Evaluation

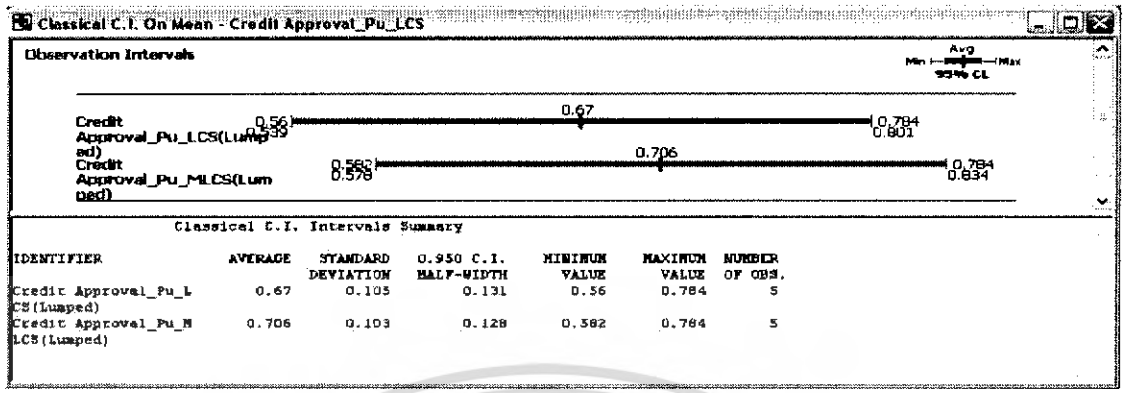
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.4 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Car Evaluation

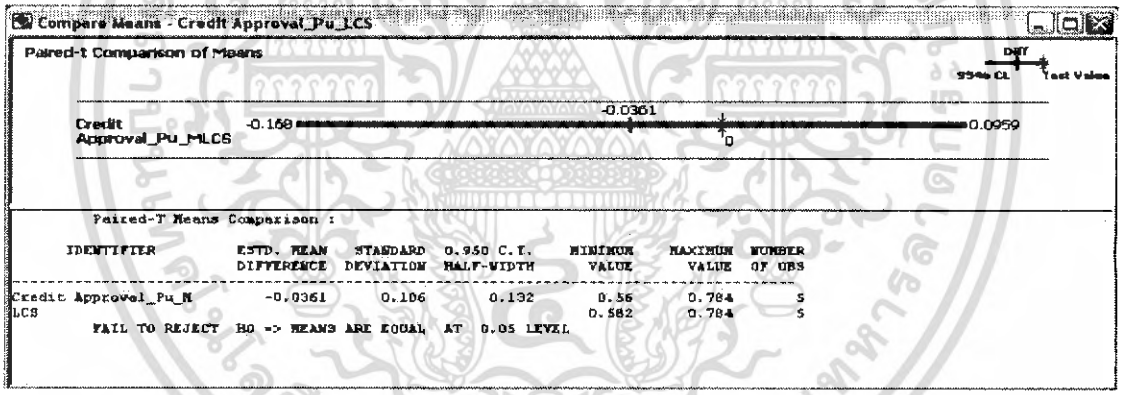
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

2. Credit Approval



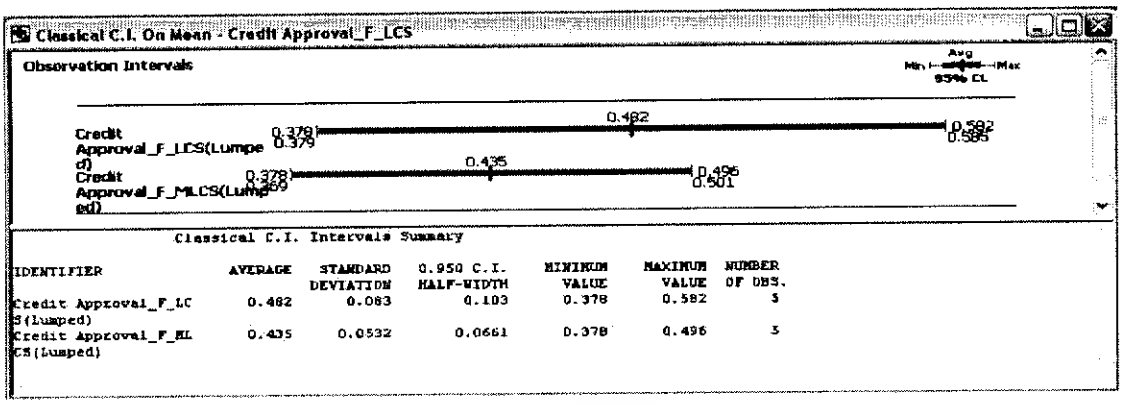
ภาพที่ ข.5 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Credit Approval

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



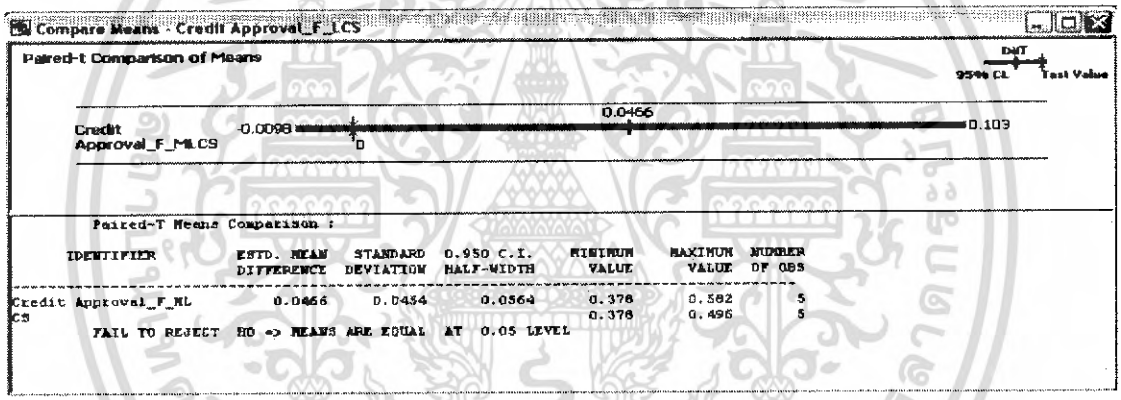
ภาพที่ ข.6 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Credit Approval

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.7 ช่วงความเชื่อมั่นของค่าเอฟเมเชอร์ของชุดข้อมูล Credit Approval

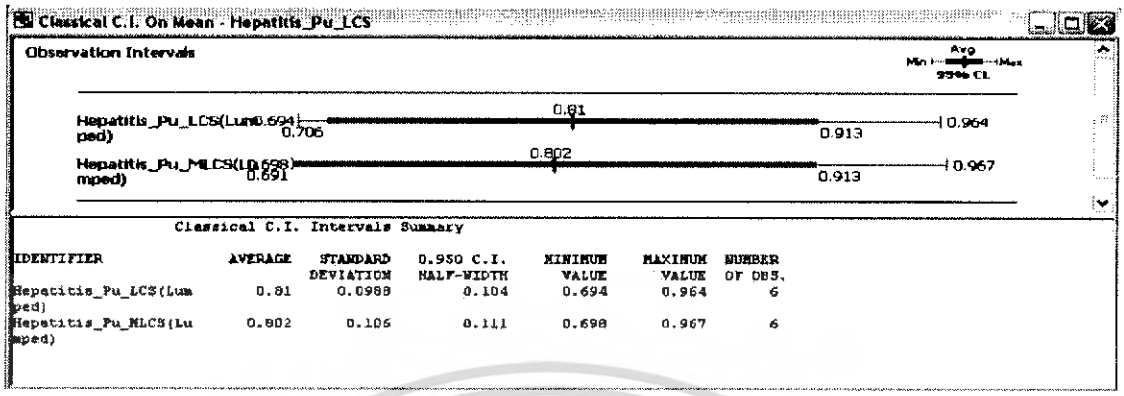
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเชอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.8 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Credit Approval

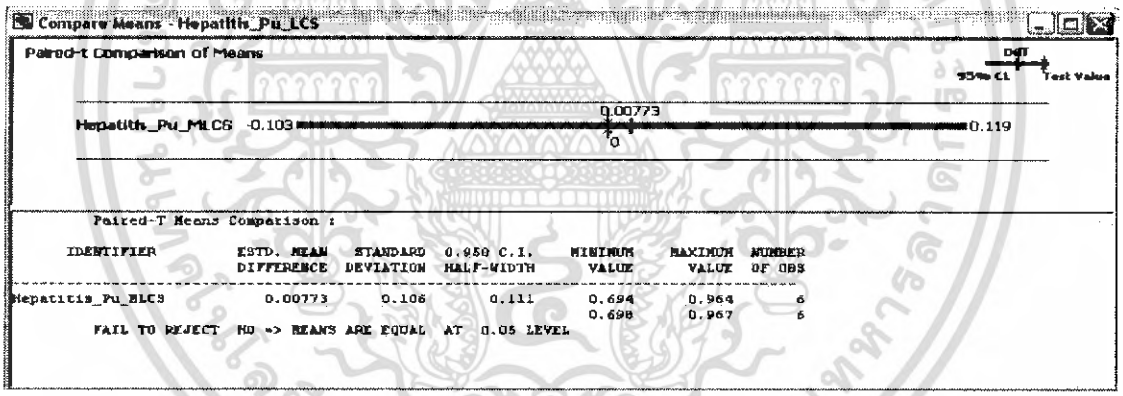
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเชอร์ไม่แตกต่างกัน

3. Hepatitis



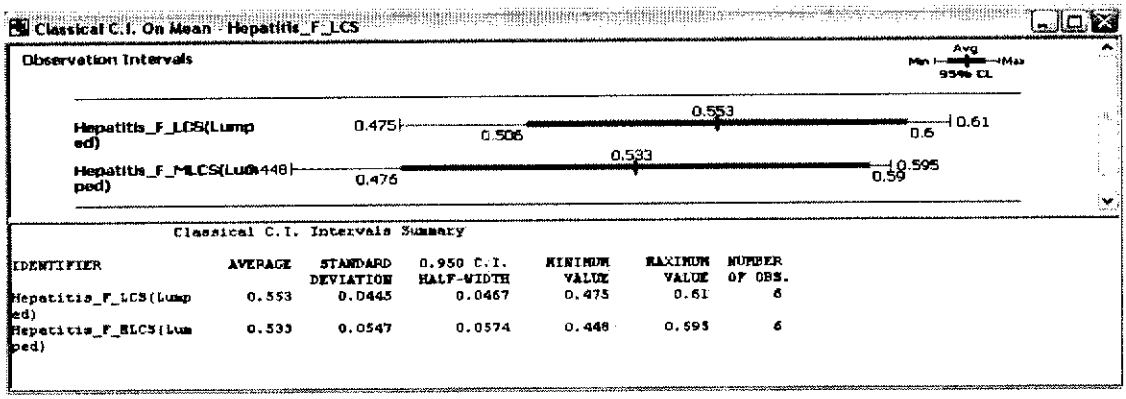
ภาพที่ ข.9 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Hepatitis

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



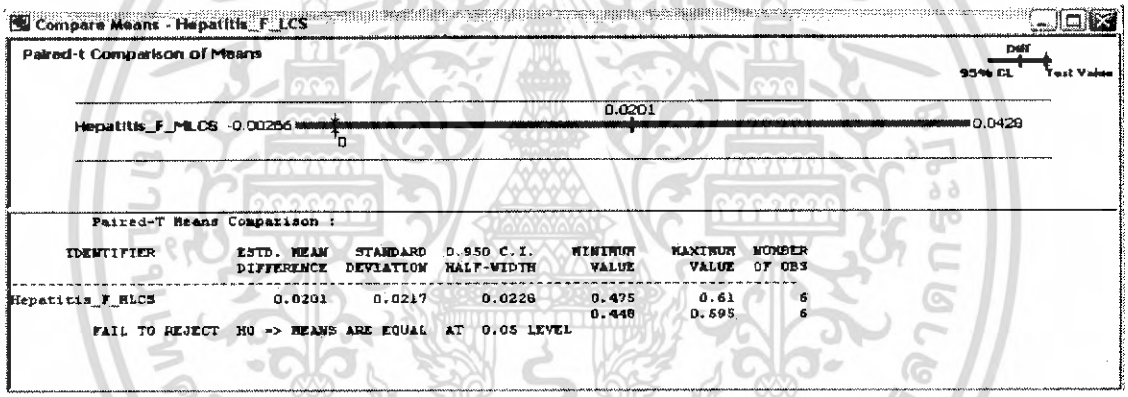
ภาพที่ ข.10 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Hepatitis

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.11 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Hepatitis

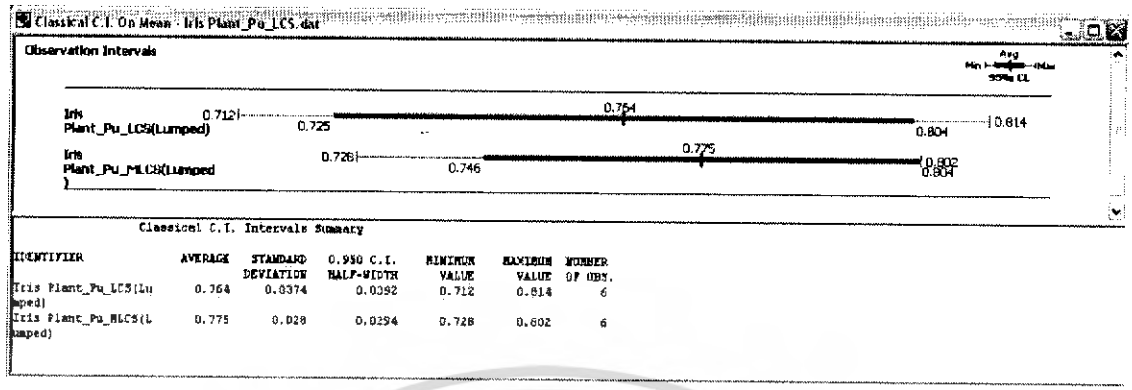
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.12 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Hepatitis

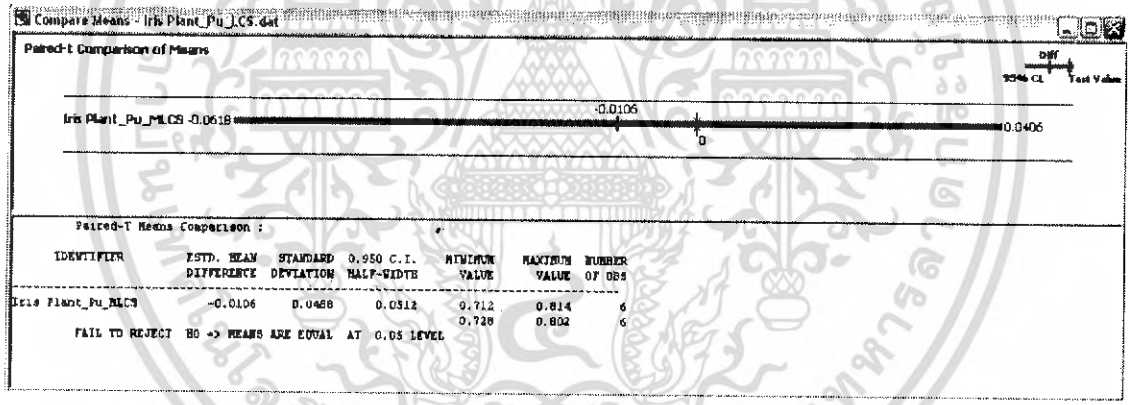
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

4. Iris Plant



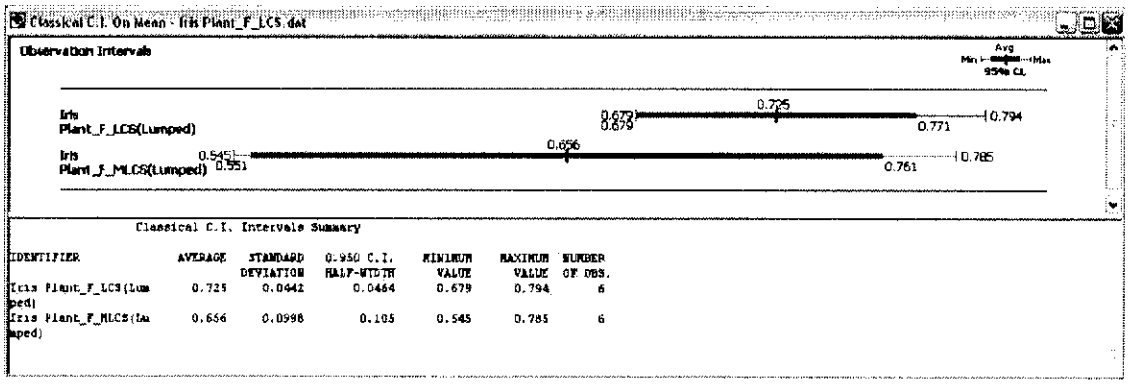
ภาพที่ ข.13 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Iris Plant

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



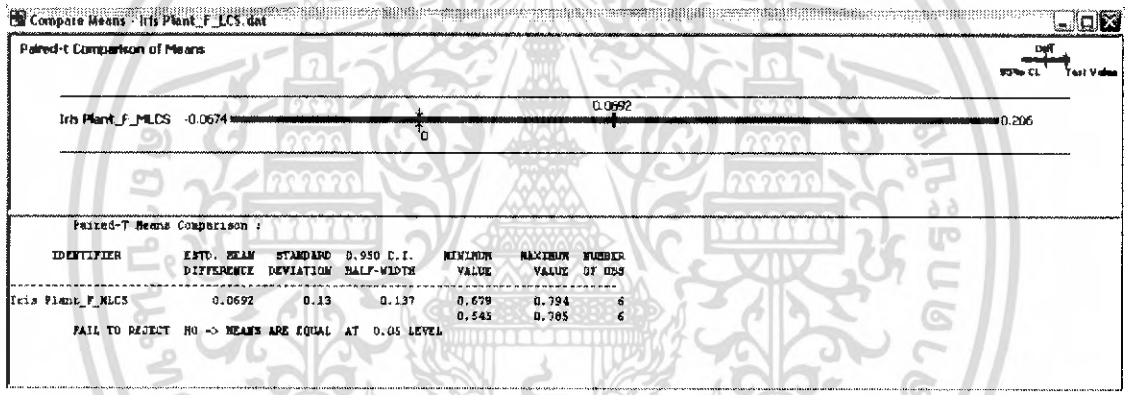
ภาพที่ ข.14 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Iris Plant

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.15 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Iris Plant

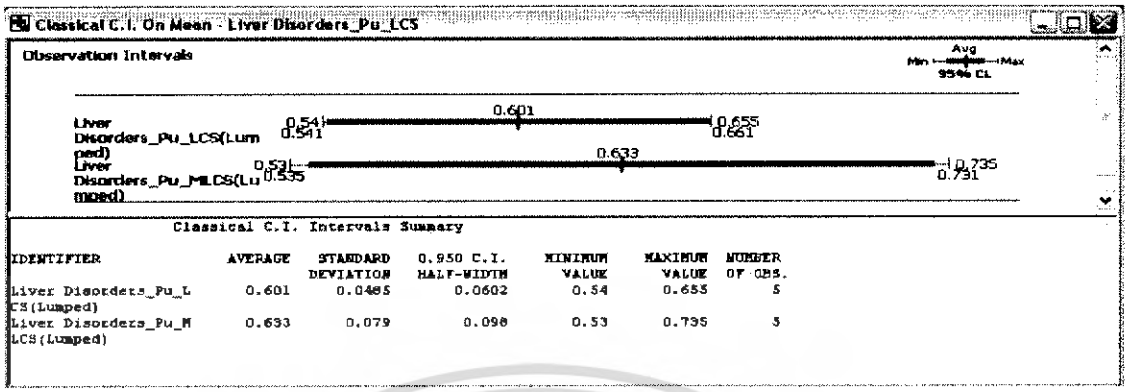
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.16 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Iris Plant

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

5. Liver Disorders

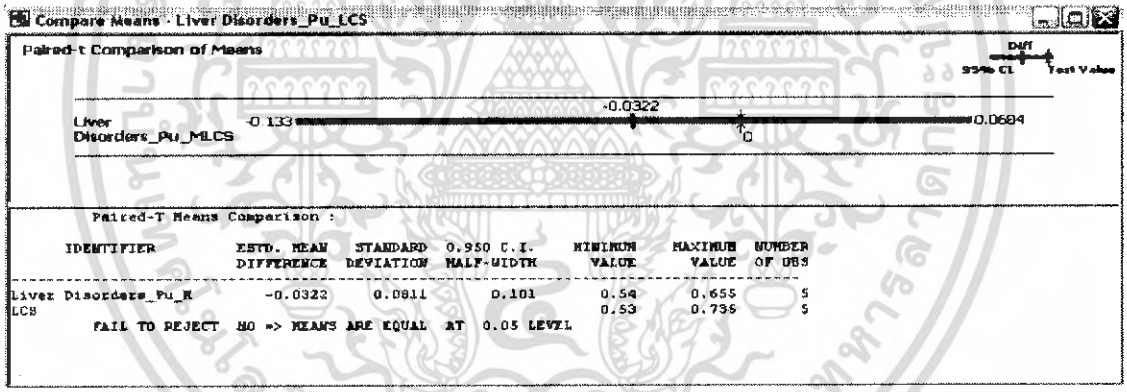


Classical C.I. Intervals Summary

IDENTIFIER	AVERAGE	STANDARD DEVIATION	0.950 C.I. HALF-WIDTH	MINIMUM VALUE	MAXIMUM VALUE	NUMBER OF OBS.
Liver Disorders_Pu_LCS(Lumped)	0.601	0.0485	0.0602	0.54	0.655	5
Liver Disorders_Pu_MLCS(Lumped)	0.633	0.079	0.098	0.53	0.735	5

ภาพที่ ข.17 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Liver Disorders

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



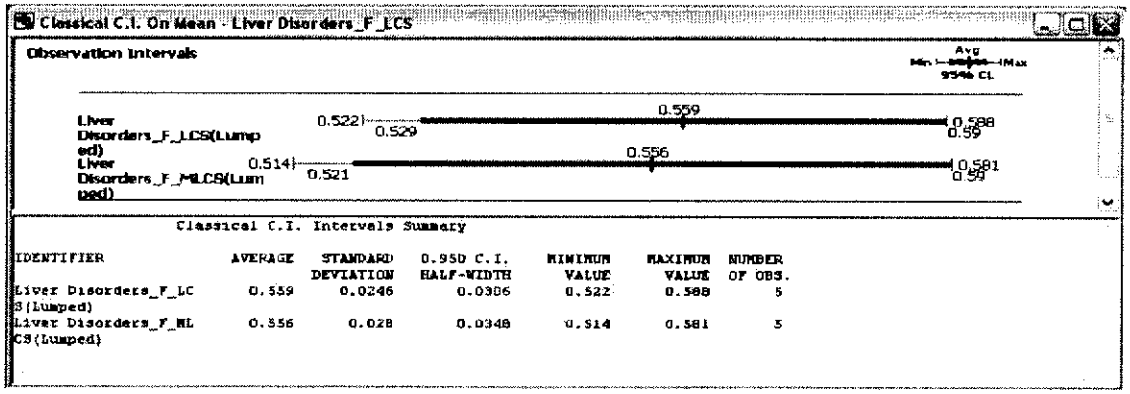
Paired-T Means Comparison :

IDENTIFIER	ESTD. MEAN DIFFERENCE	STANDARD DEVIATION	0.950 C.I. HALF-WIDTH	MINIMUM VALUE	MAXIMUM VALUE	NUMBER OF OBS
Liver Disorders_Pu_MLCS	-0.0322	0.0811	0.101	0.54 0.53	0.655 0.735	5 5

FAIL TO REJECT H0 => MEANS ARE EQUAL AT 0.05 LEVEL

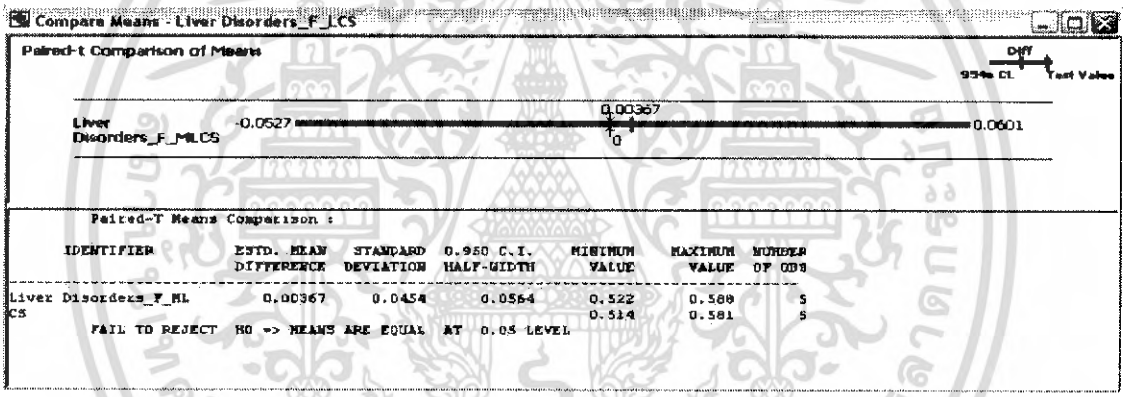
ภาพที่ ข.18 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Liver Disorders

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.19 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Liver Disorders

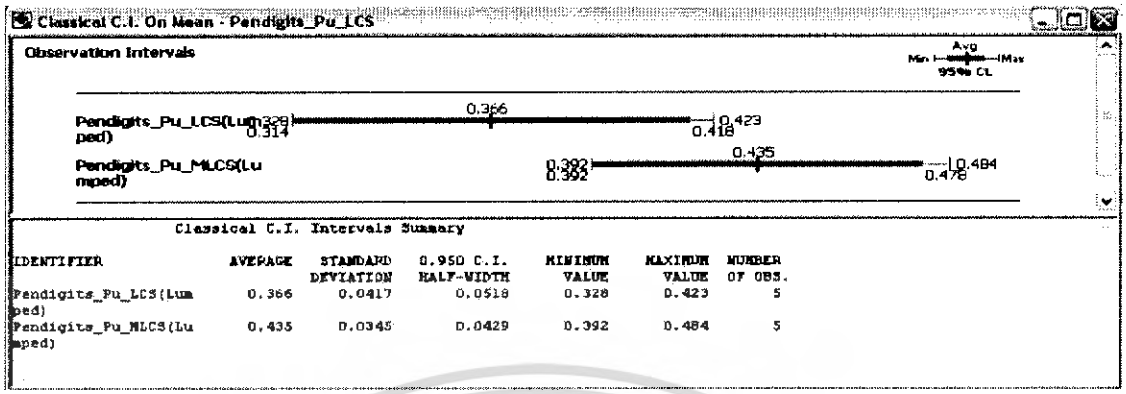
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.20 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Liver Disorders

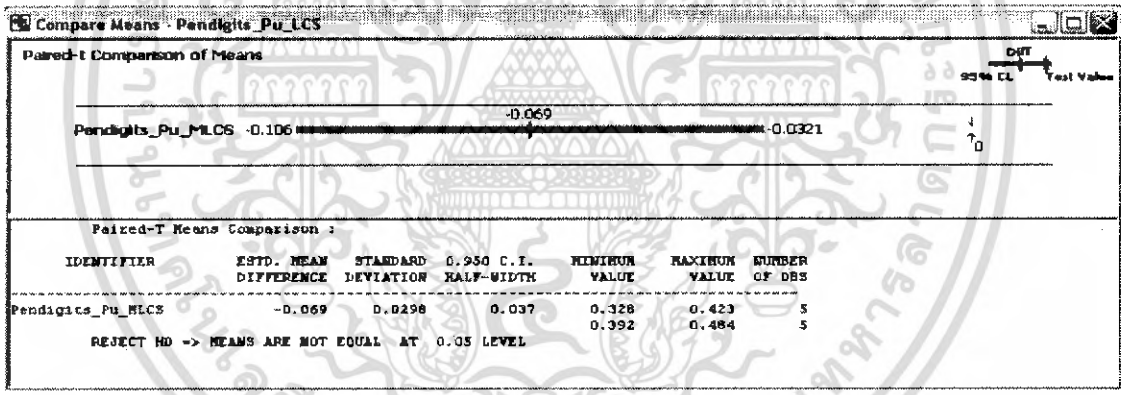
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

6. Pendigits



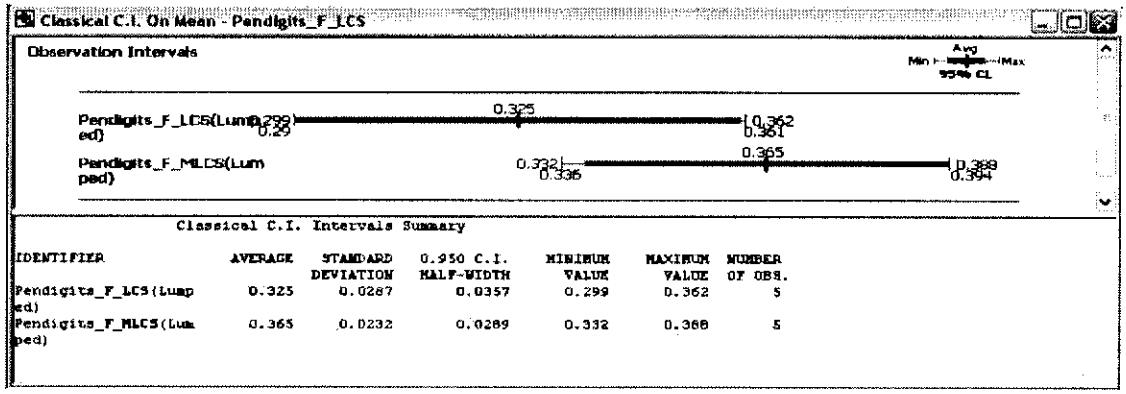
ภาพที่ ข.21 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Pendigits

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



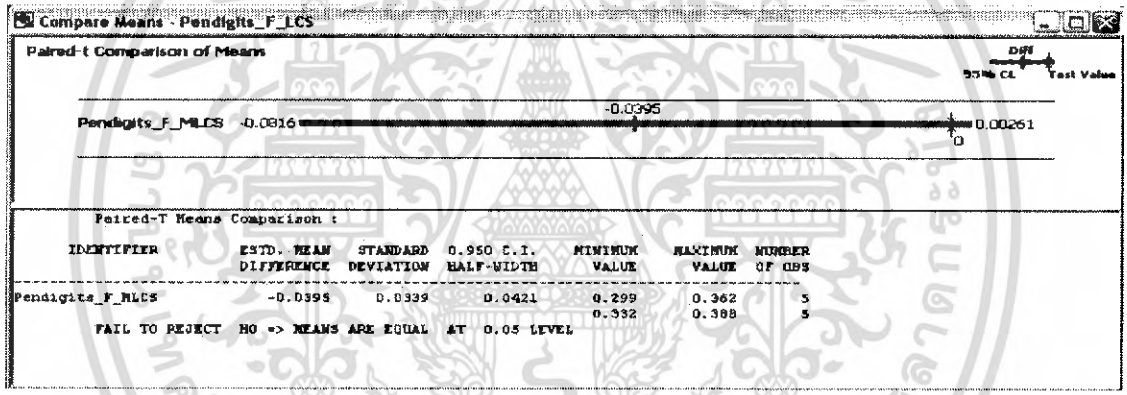
ภาพที่ ข.22 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Pendigits

เนื่องจากไม่มีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ที่แตกต่างกัน โดยอัลกอริธึมที่พัฒนาขึ้นให้ค่าความบริสุทธิ์มากกว่า



ภาพที่ ข.23 ช่วงความเชื่อมั่นของค่าเอฟเมเชอร์ของชุดข้อมูล Pendigits

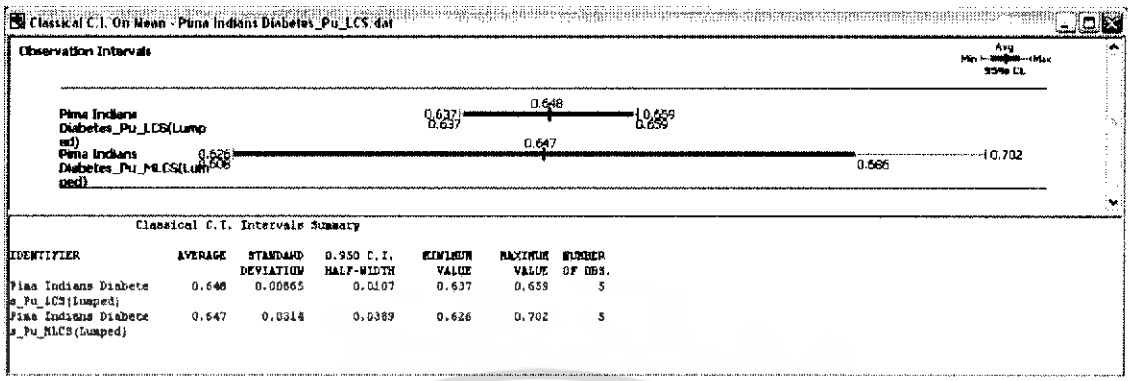
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเชอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.24 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Pendigits

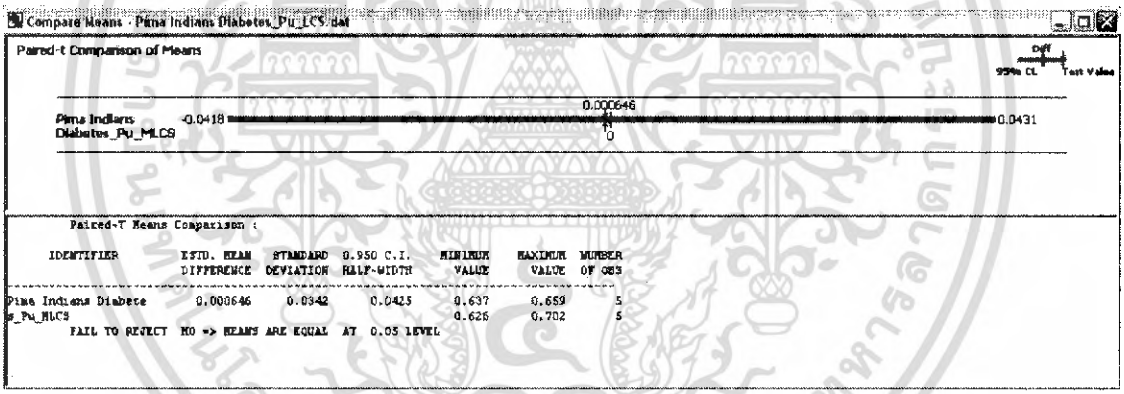
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเชอร์ไม่แตกต่างกัน

7. Pima Indians Diabetes



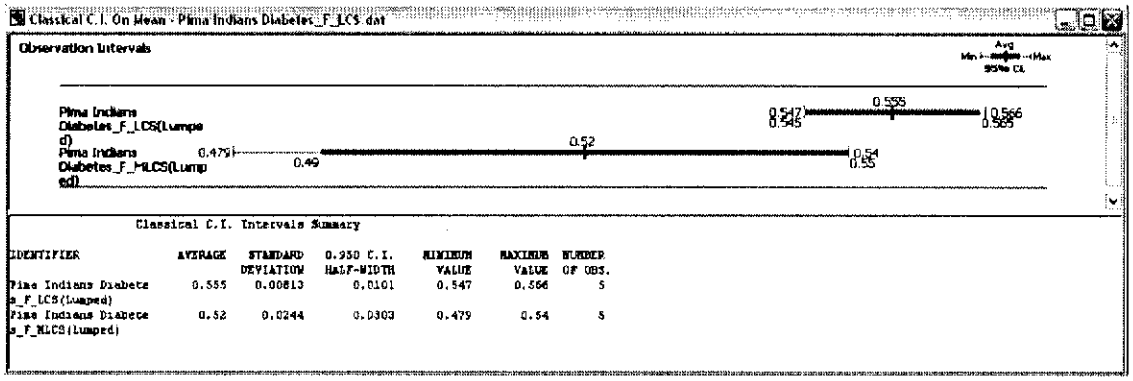
ภาพที่ ข.25 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Pima Indians Diabetes

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



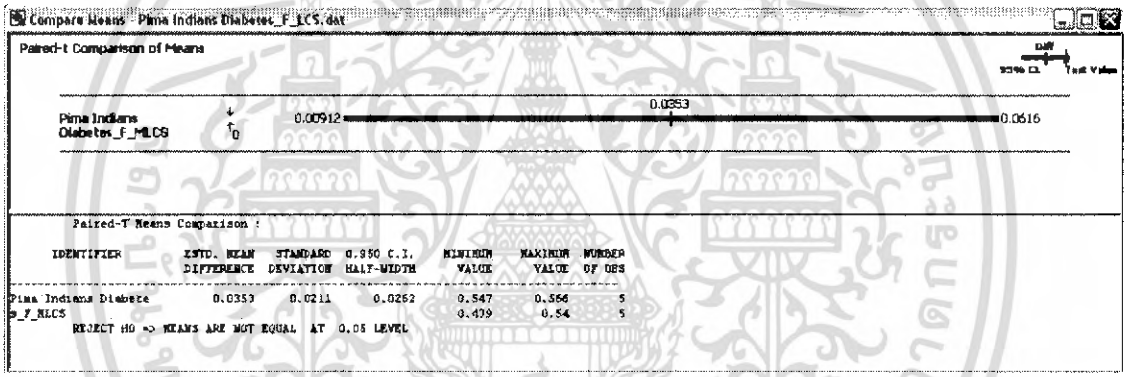
ภาพที่ ข.26 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Pima Indians Diabetes

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.27 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Pima Indians Diabetes

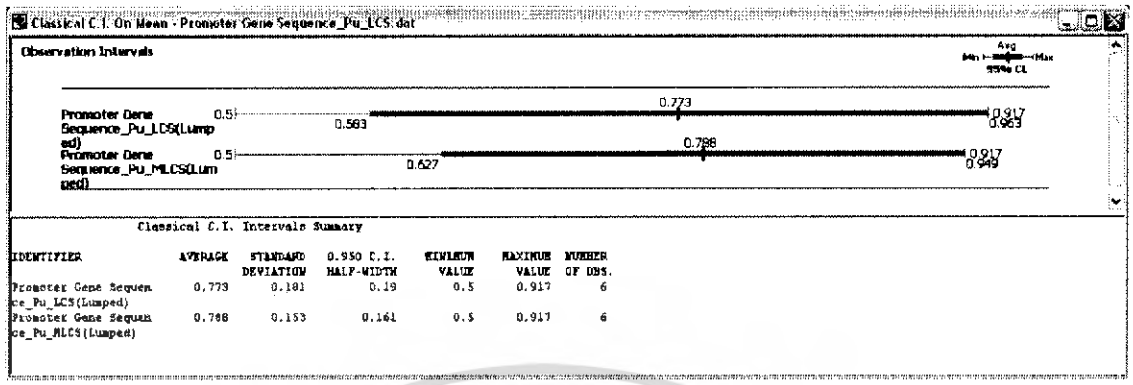
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.28 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Pima Indians Diabetes

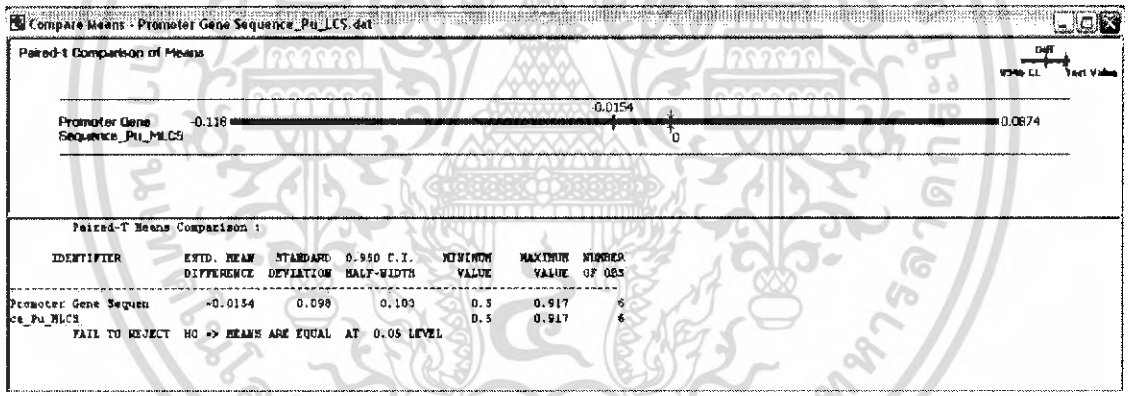
เนื่องจากไม่มีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ที่แตกต่างกัน โดยอัลกอริธึมในการหาส่วนของลำดับร่วมที่ยาวที่สุดให้ค่าเอฟเมเซอร์มากกว่า

8. Promoter Gene Sequence



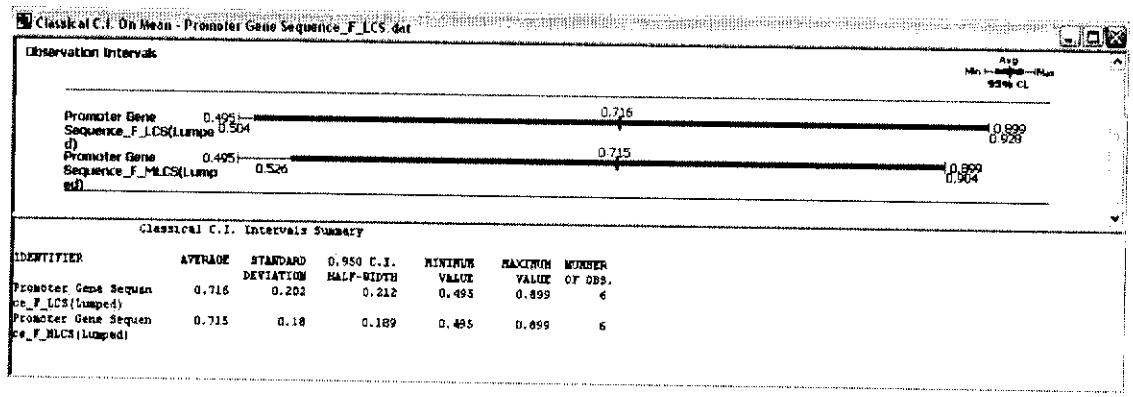
ภาพที่ ข.29 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Promoter Gene Sequence

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



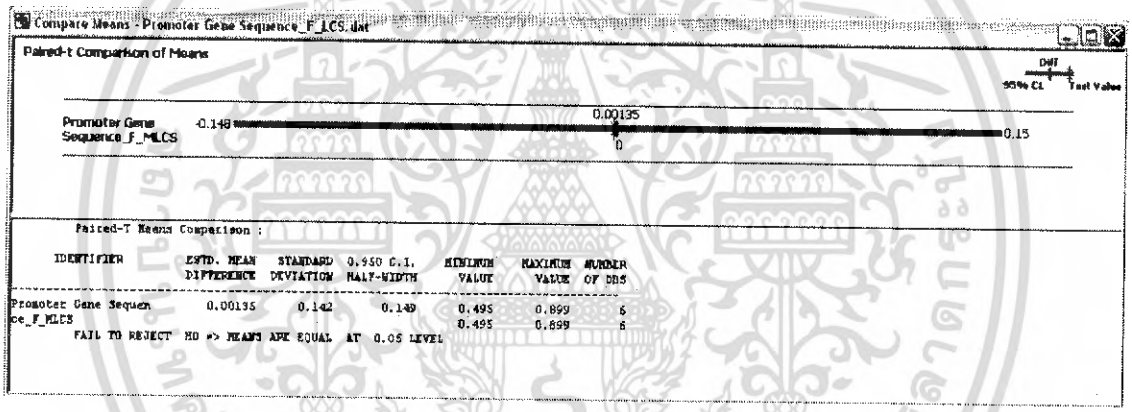
ภาพที่ ข.30 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Promoter Gene Sequence

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.31 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Promoter Gene Sequence

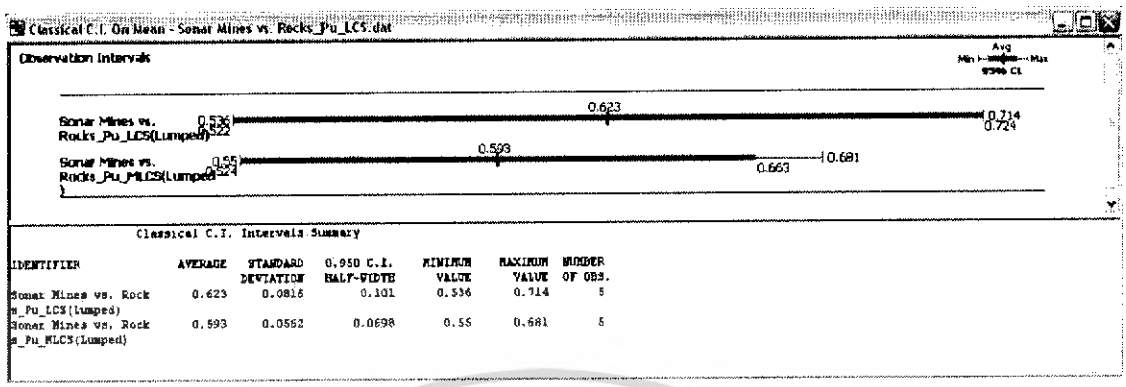
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.32 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Promoter Gene Sequence

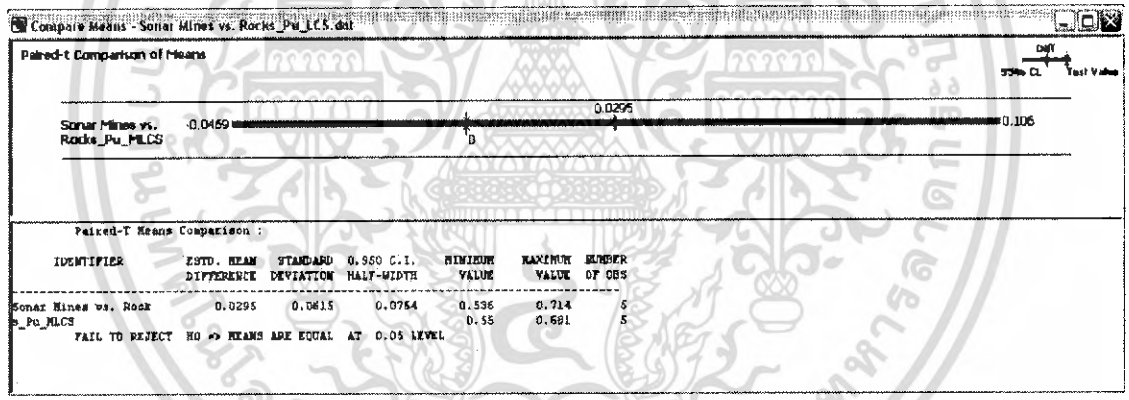
เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

9. Sonar Mines vs. Rocks



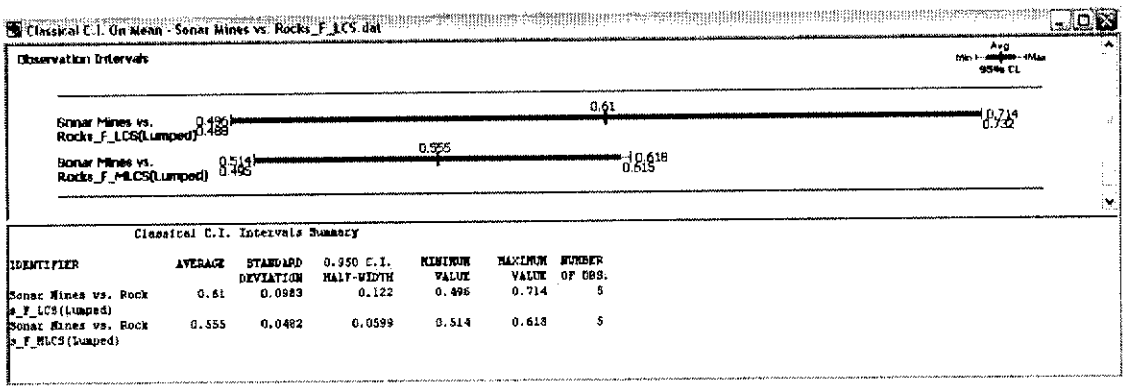
ภาพที่ ข.33 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Sonar Mines vs. Rocks

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



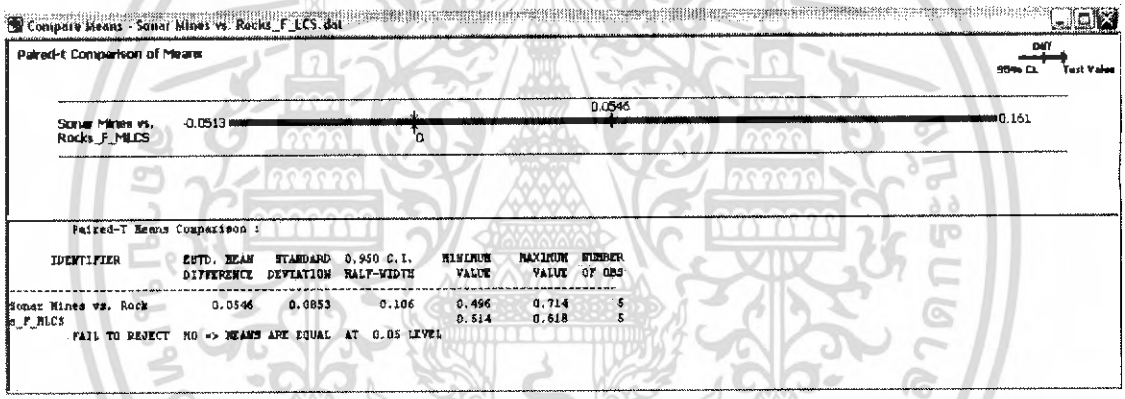
ภาพที่ ข.34 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Sonar Mines vs. Rocks

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.35 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Sonar Mines vs. Rocks

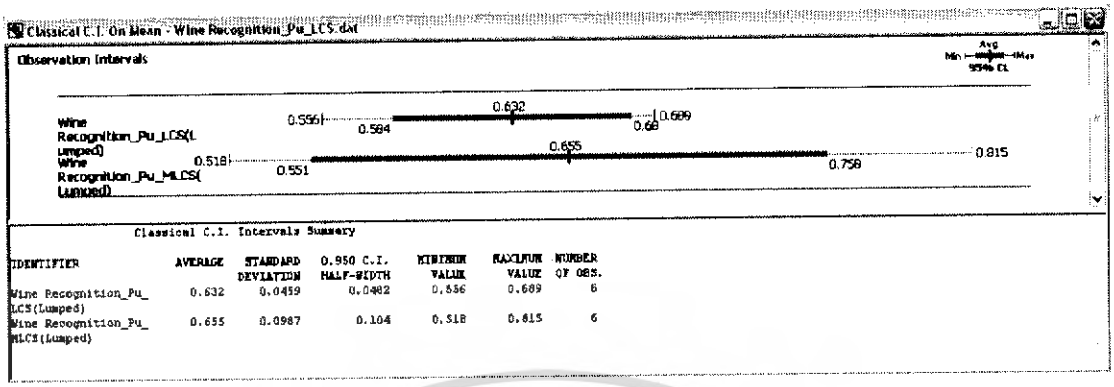
เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.36 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Sonar Mines vs. Rocks

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน

10. Wine Recognition

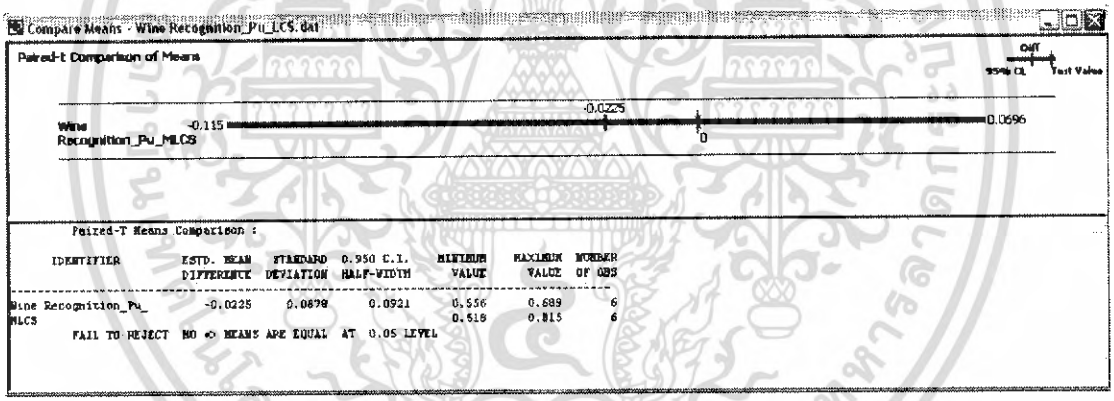


Classical C.I. Intervals Summary

IDENTIFIER	AVERAGE	STANDARD DEVIATION	0.950 C.I. HALF-WIDTH	MINIMUM VALUE	MAXIMUM VALUE	NUMBER OF OBS.
Wine Recognition_Pu_LCS(Lumped)	0.632	0.0459	0.0482	0.536	0.689	6
Wine Recognition_Pu_MLCS(Lumped)	0.655	0.0987	0.104	0.518	0.815	6

ภาพที่ ข.37 ช่วงความเชื่อมั่นของค่าความบริสุทธิ์ของชุดข้อมูล Wine Recognition

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าความบริสุทธิ์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



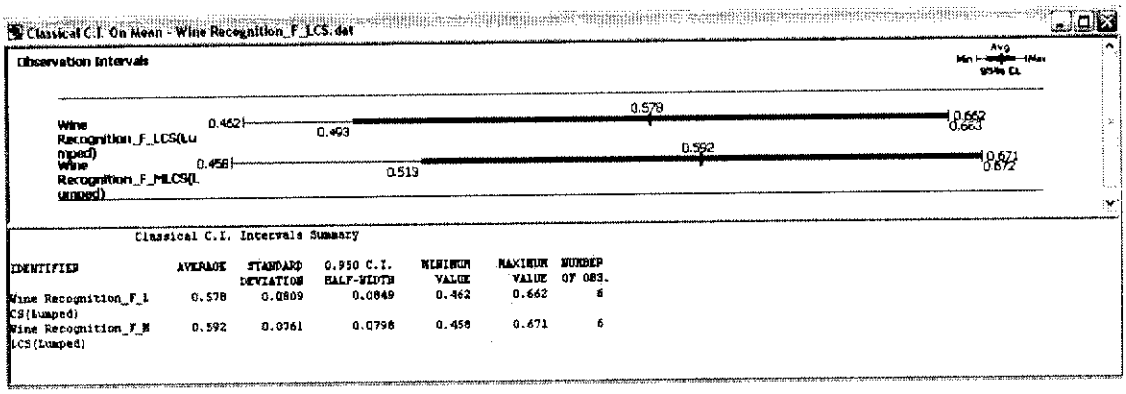
Paired-T Means Comparison :

IDENTIFIER	ESTD. MEAN DIFFERENCE	STANDARD DEVIATION	0.950 C.I. HALF-WIDTH	MINIMUM VALUE	MAXIMUM VALUE	NUMBER OF OBS
Wine Recognition_Pu_MLCS	-0.0225	0.0879	0.0921	0.556 0.518	0.689 0.815	6 6

FAIL TO REJECT HO <> MEANS ARE EQUAL AT 0.05 LEVEL

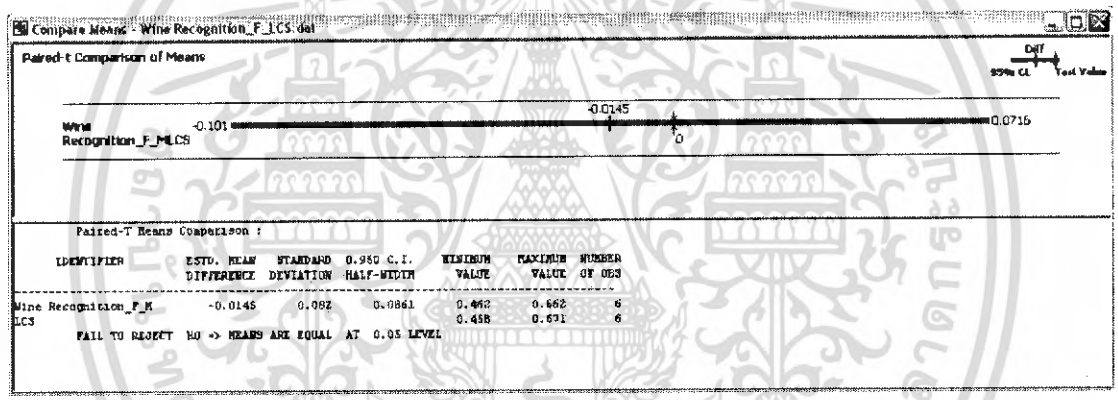
ภาพที่ ข.38 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Wine Recognition

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าความบริสุทธิ์ไม่แตกต่างกัน



ภาพที่ ข.39 ช่วงความเชื่อมั่นของค่าเอฟเมเซอร์ของชุดข้อมูล Wine Recognition

เนื่องจากมีบางส่วนของช่วงความเชื่อมั่นซ้ำกัน จึงไม่สามารถสรุปความสัมพันธ์ของค่าเอฟเมเซอร์ระหว่างอัลกอริธึมทั้งสองได้ จึงต้องทำการเปรียบเทียบค่าเฉลี่ยของผลต่าง



ภาพที่ ข.40 ช่วงความเชื่อมั่นของค่าเฉลี่ยของผลต่างของชุดข้อมูล Wine Recognition

เนื่องจากมีค่า 0 อยู่ในช่วงจึงสามารถสรุปได้ว่าอัลกอริธึมทั้งสองให้ค่าเอฟเมเซอร์ไม่แตกต่างกัน