การหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน

A NEW INITIALIZATION METHOD FOR K-MEANS CLASSIFICATION

พรรณี เกษีสังข์

PANNEE KESISUNG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2557

KMITL-2014-IT-M-001-001

การหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน

A NEW INITIALIZATION METHOD FOR K-MEANS CLASSIFICATION

พรรณี เกษีสังข์

PANNEE KESISUNG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2557

KMITL-2014-IT-M-001-001

A NEW INITIALIZATION METHOD FOR K-MEANS CLASSIFICATION

PANNEE KESISUNG

A THESIS SUBMITTED IN PATIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2014

KMITL-2014-IT-M-001-001

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน

A new Initialization method for K-means Classification

นักศึกษา นางสาวพรรณี เกษีสังข์

รหัสประจำตัว ๕๒๖๖๐๔๐๔

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ รองศาสตราจารย์ ดร.อาริต ธรรมโน

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
รองศาสตราจารย์ ดร.วรพจน์ กรีสุระเดช	
รองศาสตราจารย์ ดร.พีระพนธ์ โสพัศสถิตย์	
รองศาสตราจารย์ ดร.อาริต ธรรมโน	
รองศาสตราจารย์ ดร.พรฤดี เนติโสภากุล	

วัน / เดือน / ปี ที่สอบ วันอังคารที่ ๔ มีนาคม ๒๕๕๗ เวลา ๐๙.๓๐ น.

สถานที่สอบ ณ ห้อง ๓๓๓ ชั้น ๓ คณะเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

(รองศาสตราจารย์ ดร.จันทรัปพรณ์ สถิตวิริยวงศ์)

คณบดีคณะเทคโนโลยีสารสนเทศ

วันที่ 11 **เดือน** มีนาคม **พ.ศ.** 2557

หัวข้อวิทยานิพนธ์	การหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน
นักศึกษา	นางสาวพรรณี เกษีสังข์
รหัสนักศึกษา	52660404
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2557
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการจำแนกประเภทข้อมูล โดยการสร้างเป็นแบบจำลอง ซึ่งแบบจำลองดังกล่าว ประกอบด้วยอัลกอริทึมเคมีนและค่าเอนโทรปีเป็นโครงสร้างหลัก ทั้งนี้เนื่องจากประสิทธิภาพการจัดกลุ่มของอัลกอริทึมเคมีนมีการแปรผันไปตามจุดเริ่มต้น ดังนั้นงานวิจัยนี้ จึงได้นำเสนอวิธีการเลือกจุดเริ่มต้นและนำแนวคิดการใช้ค่าเอนโทรปีเพื่อปรับอัลกอริทึมเคมีนแบบดั้งเดิมให้เป็นเทคนิคการจำแนกประเภท ในส่วนของผลการทดลองได้นำไปเปรียบเทียบกับอัลกอริทึม C4.5 และอัลกอริทึม LVQ (Learning Vector Quantization) โดยใช้วิธีการวัดประสิทธิภาพความถูกต้องของการจำแนกประเภท

Thesis	A new Initialization method for K-means Classification
Student	Pannee Kesisung
Student ID	52660404
Degree	Master of Science
Program	Information Technology
Year	2014
Thesis Advisor	Assoc.Prof.Dr.Arit Thammano

ABSTACT

This thesis proposes a method for solving classification problems. The proposed algorithm constructs a learning model which is based on K-means algorithm and the concept of entropy. As the performance of K-means algorithm depends heavily on the selection of initial centroids, therefore this thesis proposes a new scheme to select the initial cluster centers. In the proposed model, the entropy concept is employed to adapt the traditional K-means algorithm to be used as a classification technique. The experimental results have been compared with C4.5 and LVQ (Learning Vector Quantization) algorithms to measure the performance of the classification accuracy of the data.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความกรุณาจาก รศ.ดร.อาริต ธรรมโน อาจารย์ที่ปรึกษา ที่คอยช่วยเหลือ เสอนแนะวิธีการแก้ปัญหาและให้คำปรึกษาแก่ข้าพเจ้า

ขอขอบคุณบิดา มารดา และญาติพี่น้องทุกคนที่ให้การสนับสนุนด้านทุนการศึกษาตลอดจนให้กำลังใจในการทำวิทยานิพนธ์ฉบับนี้ให้สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบคุณพี่ๆและน้องๆในแลปที่ช่วยสอนในบางเนื้อหาที่ไม่เข้าใจ รวมทั้งช่วยตรวจทานเล่มเพื่อปรับปรุงแก้ไขให้สมบรูณ์

ขอขอบคุณ เจ้าหน้าที่ คณะเทคโนโลยีสารสนเทศ สถาบัณพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกท่านที่คอยให้ความช่วยเหลือ ด้านบริการงานการศึกษาและงานด้านอื่นๆ

วิทยานิพนธ์ฉบับนี้ ขอมอบให้ให้บิดา มารดา และญาติพี่น้องและเพื่อนๆ พี่ๆ น้องๆ รวมถึงอาจารย์ที่เคารพทุกท่านที่เคยอบรมสั่งสอนข้าพเจ้า

พรรณี เกษีสังข์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง(ต่อ)

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อัลกอริทึมเคมีนรู้จักกันดีในปัญหาการจัดกลุ่ม (Clustering) ซึ่งประสิทธิภาพของการจัดกลุ่มมีการแปรผันไปตามจุดเริ่มต้น งานวิจัยจำนวนมากมุ่งเน้นไปที่การศึกษาเพื่อเพิ่มประสิทธิภาพความถูกต้องและเวลาของการจัดกลุ่ม โดยเสนอวิธีการเลือกจุดเริ่มต้น ส่วนใหญ่วิธีที่เสนอจะให้ประสิทธิภาพของการจัดกลุ่มดีกว่าอัลกอริทึมเคมีนแบบดั้งเดิม

แต่เมื่อไม่นานมานี้พบว่า มีการนำอัลกอริทึมเคมีนและอัลกอริทึมของการจัดกลุ่มมาใช้กับปัญหาการจำแนกประเภท (Classification) เช่น จัดกลุ่มข้อมูลก่อน นำข้อมูลไปฝึกสอน โดยใช้เทคนิคของการจำแนกประเภท

ด้วยเหตุนี้ผู้วิจัยจึงสนใจ นำอัลกอริทึมเคมีนมาใช้กับปัญหาการจำแนกประเภท ทั้งนี้ได้เสนอวิธีการเลือกจุดเริ่มต้น และนำแนวคิดค่าเอนโทรปีของข้อมูลมาใช้เพื่อปรับอัลกอริทึมเคมีนแบบดั้งเดิมให้เป็นเทคนิคสำหรับปัญหาการจำแนกประเภทข้อมูล

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้มีความมุ่งหมายและวัตถุประสงค์ของการศึกษาดังนี้

1. เพื่อศึกษาทฤษฎีเกี่ยวกับการจำแนกประเภทข้อมูล
2. เพื่อศึกษางานวิจัยที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล
3. ประยุกต์อัลกอริทึมเคมีนเพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูล

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้มีขอบเขตของการวิจัยดังนี้

1. นำเสนอวิธีในการหาจุดเริ่มต้นสำหรับอัลกอริทึมเคมีนแทนการสุ่มเพื่อสร้างแบบจำลองการจำแนกประเภทข้อมูล
2. ใช้ค่าเอนโทรปีเพื่อปรับอัลกอริทึมเคมีนแบบดั้งเดิมให้เป็นเทคนิคการจำแนกประเภทข้อมูล

1.4 ขั้นตอนของการศึกษา

งานวิจัยนี้มีขั้นตอนของการศึกษาดังนี้

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. กำหนดปัญหา วัตถุประสงค์และขอบเขตการวิจัย
3. พัฒนาโปรแกรมโดยใช้ซอฟต์แวร์ MATLAB
4. เตรียมข้อมูลสำหรับการทดลอง
5. ทำการทดลอง ปรับปรุง เปรียบเทียบและสรุปผล
6. จัดทำเอกสารประกอบวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

งานวิจัยนี้มีประโยชน์ที่คาดว่าจะได้รับดังนี้

1. ได้แบบจำลองการจำแนกประเภทข้อมูล ที่เกิดจากการพัฒนาอัลกอริทึมเคมีนร่วมกับแนวคิดค่าเอนโทรปี (Entropy)
2. ผู้วิจัยมีความรู้เรื่องการจำแนกประเภทข้อมูลเพิ่มขึ้น
3. เพื่อใช้เป็นแหล่งค้นหาและอ้างอิงสำหรับบุคคลที่สนใจ

บทที่ 2
ทฤษฎีพื้นฐานและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีพื้นฐาน

2.1.1 การจำแนกประเภทข้อมูล (Classification)

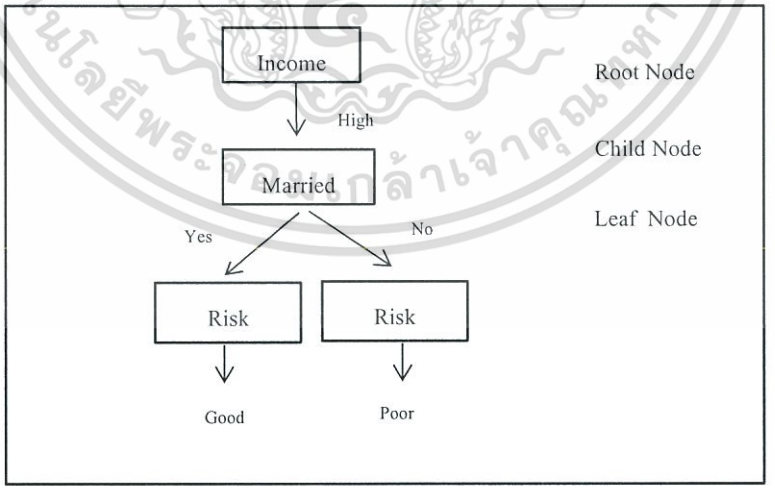
การจำแนกประเภทข้อมูล คือเทคนิคหนึ่งที่ใช้ในการทำเหมืองข้อมูล (Data mining) เพื่อหารูปแบบในการอธิบายและทำนายกลุ่มข้อมูล ซึ่งปัญหาการจำแนกประเภทมีเทคนิคหลายเทคนิคให้เลือกใช้ แต่ละเทคนิคก็จะมีหลายอัลกอริทึม ซึ่งแต่ละอัลกอริทึมจะให้ผลลัพธ์ที่ต่างกัน ตัวอย่างเทคนิคที่ใช้แก้ปัญหาการจำแนกประเภทข้อมูล ได้แก่

1. **ต้นไม้ตัดสินใจ (Decision Tree)** คือการนำเอาข้อมูลมาสร้างแบบจำพยากรณ์ในโครงสร้างต้นไม้ซึ่งต้นไม้ตัดสินใจ ปกติมักประกอบด้วยกฎในรูปแบบ
"ถ้า เงื่อนไข แล้ว ผลลัพธ์" เช่น

"If Income = High and Married = No THEN Risk = Poor"

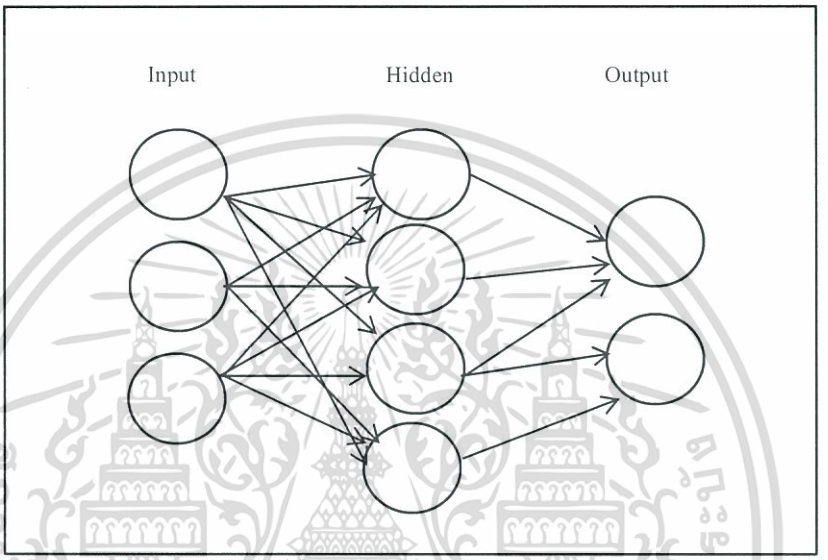
"If Income = High and Married = Yes THEN Risk = Good"

ซึ่งโครงสร้างแบบต้นไม้ประกอบด้วย Root Node, Child Node และ Leaf Node ดังรูป 2.1



รูปที่ 2.1 แสดงโครงสร้างต้นไม้ตัดสินใจ

2. โครงข่ายประสาทเทียม (Neural Networks) คือการจำลองการทำงานโครงข่ายประสาทเทียมในสมองของมนุษย์ให้กับคอมพิวเตอร์ทำงานมีโครงสร้างของโหนดที่เชื่อมโยงถึงกันในแต่ละชั้น (Layer) คือ Input layer, Hidden layer, Output layer ดังรูปที่ 2.2



รูปที่ 2.2 แสดงโครงสร้างโครงข่ายเส้นประสาทเทียม

3. นาอีฟเบย์ (Naive-Bayes) คือการวิเคราะห์ความน่าจะเป็นจากสิ่งที่ยังไม่เกิดโดยการคาดเดาจากสิ่งที่เกิดขึ้นมาก่อน

4. K-nearest neighbor (K-NN) คือการจำแนกประเภทข้อมูลโดยแยกจำนวนเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุดเพื่อหาผลรวมของเงื่อนไขและกรณีต่างๆสำหรับแต่ละคลาสและกำหนดเงื่อนไขใหม่ๆให้คลาสที่เหมือนและใกล้เคียงกับมันมากที่สุด k-nn ค่อนข้างใช้ปริมาณงานในการคำนวณสูงเพราะเวลาสำหรับการคำนวณจะเพิ่มขึ้นเป็นแฟคทอเรียลตามจำนวนจุดทั้งหมดของข้อมูล

2.1.2 อัลกอริทึมเคมี (K-means Algorithm)

อัลกอริทึมเคมีน(K-means Algorithm) คือวิธีการจัดกลุ่มข้อมูลทั้งหมดในชุดข้อมูล จำนวน K กลุ่ม และค่า K ต้องน้อยกว่าจำนวนข้อมูลทั้งหมด (N) ซึ่งจำนวนกลุ่มต้องเป็นจำนวนเต็มบวก และการจัดกลุ่มใช้คุณสมบัติความเหมือนของข้อมูลโดยวิธีการวัดระยะทางที่ใกล้ที่สุดระหว่างข้อมูลทั้งหมดกับจุดศูนย์กลางของแต่ละกลุ่ม โดยใช้การวัดระยะทางแบบยูคลิด (Euclidean distance) ดังสมาการที่ 2.1

$$d(x_i, z_j) = \sqrt{\sum_{q=1}^{Q}(x_{iq} - z_{jq})^2} \quad (2.1)$$

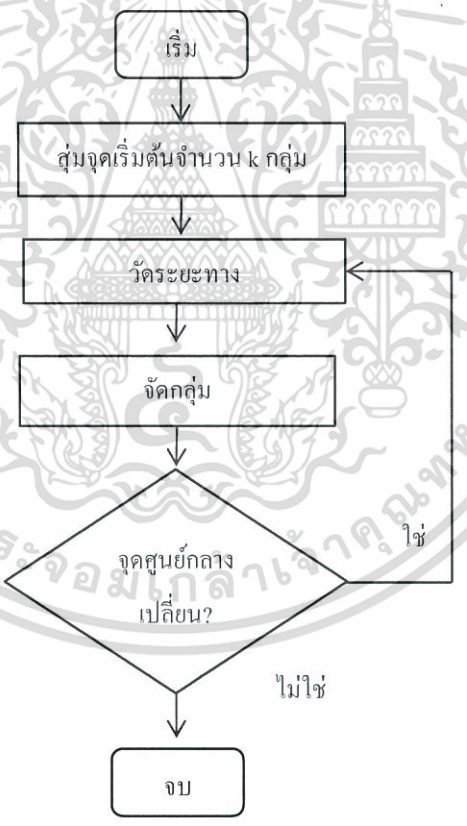
เมื่อ $d(x_i, z_j)$ คือระยะทางระหว่าง จุด x_i กับ z_j

x_i คือข้อมูลที่หนึ่ง

z_j คือจุดข้อมูลที่สอง

Q คือมิติของข้อมูล

จุดศูนย์กลางของแต่ละกลุ่มที่ใช้เป็นตัวแทนของกลุ่มเกิดจากค่าเฉลี่ยของสมาชิกที่อยู่ในกลุ่ม ซึ่งโครงสร้างกระบวนการทำงานอัลกอริทึมเคมีมีขั้นตอนดังรูป 2.3



รูปที่ 2.3 แสดงโครงสร้างกระบวนการทำงานอัลกอริทึมเคมีน

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 งานวิจัยเรื่อง Improving the Accuracy and Efficiency of the K-means Clustering Algorithm [1]

งานวิจัยนี้ได้นำเสนอการปรับปรุงประสิทธิภาพและความถูกต้องสำหรับการจัดกลุ่มของอัลกอริทึมเคมีน โดยนำไปประยุกต์ใช้กับแอพพลิเคชั่นสำหรับการค้นคืนข้อมูลในฐานข้อมูลขนาดใหญ่ ซึ่งแบ่งออกเป็น 2 ขั้นตอนได้แก่ การหาจุดเริ่มต้นของอัลกอริทึมเคมีนและการกำหนดกลุ่มให้กับข้อมูล โดยมีรายละเอียดดังต่อไปนี้

1. การหาจุดเริ่มต้น

1.1 กำหนดให้ m=1

1.2 คำนวณระยะทางระหว่างแต่ละจุดข้อมูลกับข้อมูลทั้งหมดในชุดข้อมูล และกำหนดให้เป็นเซตข้อมูล D

1.3 หาคู่ของข้อมูลที่มีระยะทางใกล้กันที่สุดจากเซตข้อมูล D และกำหนดเซตข้อมูล Am โดยที่เซตข้อมูล Am ต้องมีจำนวนสมาชิกเท่ากับ (1<=m<=k) แล้วใส่คู่ของข้อมูลดังกล่าว จากนั้นลบคู่ข้อมูลดังกล่าวออกจากเซตข้อมูล D

1.4 หาจุดข้อมูลในเซตข้อมูล D ที่มีระยะทางใกล้กับเซตข้อมูล Am และใส่ให้เซตข้อมูล Am พร้อมทั้ง ลบออกจากเซตข้อมูล D

1.5 ทำซ้ำข้อ 1.4 จนกระทั่งจุดของข้อมูลในเซตข้อมูล Am มีจำนวนสมาชิกเท่ากับ 0.75*(n/k)

1.6 ถ้า m<k แล้ว m=m+1 จากนั้นหาคู่ของข้อมูลอื่นในเซตข้อมูล D ที่มีระยะทางใกล้สุดระหว่างข้อมูลอื่นในเซตข้อมูล Am พร้อมทั้งลบคู่ข้อมูลนั้นออกจากเซตข้อมูล D แล้วกลับไปทำข้อ 1.4

1.7 แต่ละจุดข้อมูลในเซตข้อมูล Am (1<=m<=k) คำนวณหาค่าเฉลี่ยของเวกเตอร์ของจุดข้อมูล ค่าเฉลี่ยที่ได้เป็นจุดศูนย์กลางเริ่มต้น

2. การกำหนดกลุ่มให้กับข้อมูล

2.1 คำนวณระยะทางระหว่างแต่ละจุดของข้อมูล di (1<=i<=n) กับจุดศูนย์กลางเริ่มต้นทั้งหมด cj (1<=j<=k) คือ d(di ,cj)

2.2 แต่ละจุดของข้อมูล di ที่มีระยะทางใกล้กับจุดศูนย์กลาง cj จัดให้อยู่กลุ่มคลัสเตอร์ j

2.3 แต่ละคลัสเตอร์ j (1<=j<=k) คำนวณจุดศูนย์กลางใหม่

2.4 แต่ละจุดข้อมูล di นำมาคำนวณหาระยะทางระหว่างจุดศูนย์กลางของคลัสเตอร์ปัจจุบัน

2.4.1 ถ้าระยะทางน้อยกว่าหรือเท่ากับคลัสเตอร์ปัจจุบัน ข้อมูลจุดดังกล่าวยังคงอยู่คลัสเตอร์เดิม

2.4.2 ถ้าไม่ใช่ นำจุดศูนย์กลางทุกจุด cj (1<=j<=k) มาคำนวณหาระยะทางใหม่ d(di,dj)

2.4.3 แต่ละคลัสเตอร์คำนวณจุดศูนย์กลางใหม่

2.5 ทำซ้ำข้อ 2.4 จนกระทั่ง จุดศูนย์กลางไม่เปลี่ยน

งานวิจัยนี้ทำการทดลองโดยใช้ชุดข้อมูลจาก UCI ซึ่งผลการทดลองแสดงดังตารางที่ 2.1

ตารางที่ 2.1 แสดงผลการทดลองของงานวิจัยเรื่อง Improving the Accuracy and Efficiency of the K-means Clustering Algorithm

Algorithm	Initial Centroids	Accuracy (%)	Time taken (ms)
k-means algorithm (executed 7 times with randomly selected initial centroid	5.1, 3.5, 1.4, 0.2, 4.3, 3, 1.1, 0.1, 6.6, 2.9, 4.6, 1.3	52.6	71
	7, 3.2, 4.7, 1.4,6.7, 3.1, 4.4, 1.4,5.1, 3.5, 1.4, 0.2	88.7	69
	7, 3.2, 4.7, 1.4,6.7, 3.1, 4.4, 1.4,7.4, 2.8, 6.1, 1.9	89.3	70
	7.4, 2.8, 6.1, 1.9,6, 3, 4.8, 1.8,6.7, 3.1, 4.4, 1.4	89.3	72
	5.1, 3.5, 1.4,0.2,4.3,3.0,1.1, 0.1,6.0, 3, 4.8, 1.8	52.7	70
	6, 3, 4.8, 1.8,5.8, 2.7, 5.1, 1.9,5.1, 3.5, 1.4, 0.2	89.3	72
	5.1, 3.5, 1.4, 0.2,7, 3.2, 4.7, 1.4,6.3, 3.3, 6, 2.5	89.3	71

ตารางที่ 2.1(ต่อ) แสดงผลการทดลองของงานวิจัยเรื่อง Improving the Accuracy and Efficiency of the K-means Clustering Algorithm

Algorithm	Initial Centroids	Accuracy (%)	Time taken (ms)
Mean value	-	78.7	70.7
Enhanced algorithm	computed by the program	88.6	67

ข้อสังเกต อัลกอริทึมที่งานวิจัยนี้นำเสนอ ต้องกำหนดจำนวนกลุ่ม k

2.2.2 งานวิจัยเรื่อง enhancing the K-means clustering algorithm by using a O (n logn) heuristic method for finding better initial centroids [2]

งานวิจัยนี้ได้นำเสนอการหาจุดเริ่มต้นของ k-means algorithm โดยมีรายละเอียดดังต่อไป

1. พิจารณาในแต่ละคอลัมน์ของข้อมูลโดยแบ่งเป็นช่วงของข้อมูล ระหว่างช่วงข้อมูลสูงสุดกับต่ำสุด
2. แยกคอลัมน์ที่มีช่วงข้อมูลสูงสุด
3. เรียงข้อมูลทั้งหมดโดยใช้วิธีเรียงแบบฮีฟ (Heap sort) บนพื้นฐานของช่วงข้อมูลสูงสุด
4. แบ่งช่วงข้อมูลที่เรียงแล้ว จำนวน K ส่วน
5. นำส่วนที่แบ่งจากข้อ 4 มาหาค่าเฉลี่ย (C1,.., Ck) เป็นจุดศูนย์กลางเริ่มต้นสำหรับอัลกอริทึมเคมีน
6. จัดกลุ่มข้อมูลโดยพิจารณาข้อมูลที่อยู่ใกล้กับจุดศูนย์กลาง
7. ทำซ้ำข้อ 6 จนกว่าจุดศูนย์กลางจะไม่เปลี่ยน

งานวิจัยนี้ทำการทดลองกับ 3 ชุดข้อมูล โดยวัดประสิทธิภาพของความถูกต้องและเวลาที่ใช้ในการทำงานของอัลกอริทึมที่งานวิจัยนี้นำเสนอเปรียบเทียบกับ K-means และ Enhanced K-means แสดงดังตารางที่ 2.2

ตารางที่ 2.2 แสดงผลการทดลองของงานวิจัยเรื่อง Enhancing the k-means clustering algorithm by using a O (n logn) heuristic method for finding better initial centroid เปรียบเทียบกับอัลกอริทึม เปรียบเทียบกับ K-means และ Enhanced K-means

Data Sets	Algorithms					
	K-means		Enhanced K-means		อัลกอริทึมที่งานวิจัยนี้นำเสนอ	
	Accuracy (%)	Time Taken(ms)	Accuracy (%)	Time Taken(ms)	Accuracy (%)	Time Taken(ms)
E-Coli	79.7	64	81.5	48	81.5	40
Breast Cancer	96	68	96.2	56	96.2	42
Thyroid	75	60	82.3	56	86	52

ข้อสังเกต อัลกอริทึมที่งานวิจัยนี้นำเสนอ ต้องกำหนดจำนวนกลุ่ม k

2.2.3 งานวิจัยเรื่อง clustering for Classification [3]

งานวิจัยนี้นำอัลกอริทึมของการจัดกลุ่มมาใช้เพื่อรวบรวมข้อมูลที่เป็นตัวแทนของกลุ่มข้อมูลในฐานข้อมูลขนาดใหญ่ เปรียบเทียบกับการสุ่มเลือก ซึ่งมีรายละเอียดดังนี้

เตรียมจำนวนคลัสเตอร์เพื่อใช้สำหรับกระบวนการทดสอบระบบ โดยใช้ 5 อัลกอริทึมได้แก่

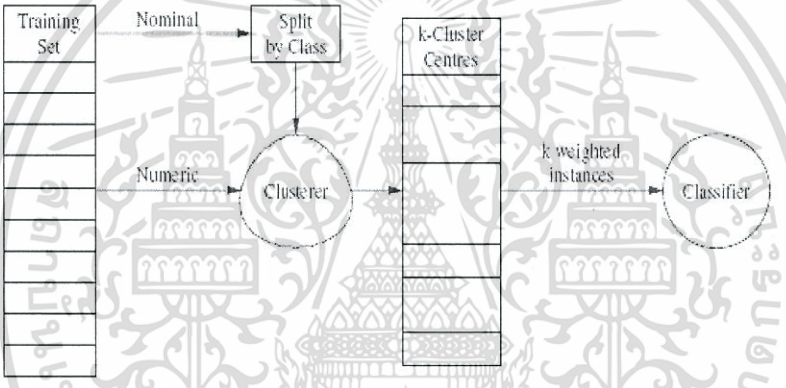
1. Fist K
2. K-means
3. Farthest first
4. Bisecting K-means
5. Expectation Maximization

ออกแบบการทดลองโดยเลือกใช้อัลกอริทึมตามลักษณะข้อมูลแสดงดังตารางที่ 2.3

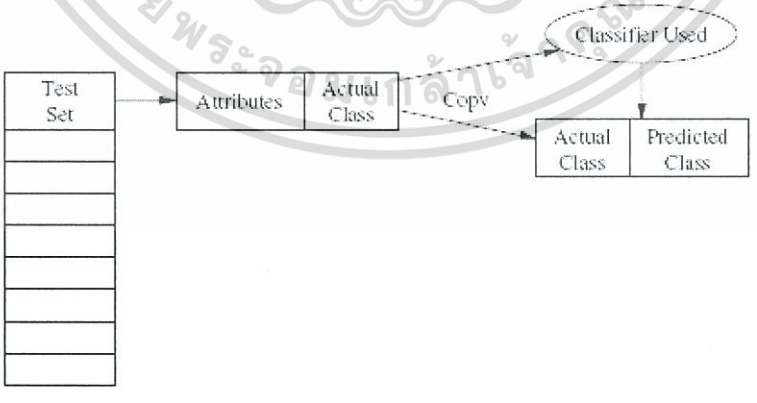
ตารางที่ 2.3 แสดงอัลกอริทึมสำหรับการทดลอง

	Nominal	Numeric
Simple	Naïve Bayes	Linear Regression
Complex	Logistic Regression	M5

จากตารางที่ 2.3 แสดงอัลกอริทึมสำหรับการฝึกสอนระบบซึ่งแยกตามลักษณะข้อมูลแสดงดังรูปที่ 2.4



รูปที่ 2.4 แสดงกระบวนการฝึกสอนระบบสำหรับคลาสเป้าหมายทุกประเภทในชุดข้อมูล



รูปที่ 2.5 แสดงกระบวนการทดสอบระบบของคลาสเป้าหมายที่เป็นข้อมูลเชิงคุณภาพ

สรุปผลการทดลอง อัลกอริทึมที่งานวิจัยนี้นำเสนอเปรียบเทียบกับการสุ่มเลือก ผลปรากฏว่าอัลกอริทึมที่งานวิจัยนี้นำเสนอให้ผลดีกว่า

ข้อสังเกต การนำอัลกอริทึมของการจัดกลุ่มมาใช้จะขึ้นอยู่กับโครงสร้างของข้อมูลซึ่งทำให้เสียเวลาในการจัดเตรียมข้อมูล

2.2.4 งานวิจัยเรื่อง Fuzzy Clustering and Fuzzy Entropy based Classification Model [4]

งานวิจัยนี้เสนอแบบจำลองการจำแนกประเภทโดยใช้ Fuzzy Clustering and Fuzzy Entropy เพื่อเพิ่มประสิทธิภาพของความถูกต้องให้วิธีการร่วมกันตัดสินใจ (ensemble methods) เพื่อการจำแนกประเภทข้อมูล สำหรับงานด้าน Pattern Recognition โดยมีรายละเอียดดังต่อไปนี้

1. คำนวณความเป็นสมาชิกของคลาส (Class Membership) โดยใช้ Fuzzy C Mean แสดงดังสมการที่ 2.2

$$J_0 = \sum_{i=1}^{d} \sum_{j=1}^{c} \mu_{ij}^{m} D_{ij}^{2} \tag{2.2}$$

เมื่อ J_0 คือฟังก์ชั่นเป้าหมาย

μ_{ij} คือค่าความเป็นสมาชิก

D คือระยะทางระหว่างข้อมูลกับจุดศูนย์กลาง

ซึ่งค่า Membership คำนวณได้จากสมการที่ 2.3 ดังนี้

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \frac{D_{ij}}{D_{kj}}^{\frac{2}{(m-1)}}} \tag{2.3}$$

เมื่อ μ_{ij} คือฟังก์ชั่นของความเป็นสมาชิก

D คือระยะทางของข้อมูลกับจุดศูนย์กลาง

k คือจำนวนจุดศูนย์กลาง

2. คำนวณ Fuzzy Entropy ของข้อมูลทั้งหมด คำนวณได้จากสามการที่ 2.4 ดังนี้

$$H(x) = -\sum_{j=1}^{c} \mu_{ij} \log_2 \mu_{ij} \tag{2.4}$$

เมื่อ $H(x)$ คือค่าเอนโทรปี

μ_{ij} คืออัตราส่วนระหว่างคลาสของข้อมูลกับจำนวนข้อมูลทั้งหมดในกลุ่ม

3. คำนวณค่า Mean Fuzzy Entropy

4. กำหนดให้ข้อมูลที่มีค่าเอนโทรปีน้อยกว่าค่าเอนโทรปีเฉลี่ย (Mean Entropy) ให้เป็น Core หรือ Easy Ensemble

5. กำหนดให้ข้อมูลที่มีค่าเอนโทรปีมากกว่าค่าเอนโทรปีเฉลี่ย ให้เป็น Hard หรือ Boundary Ensemble

6. ฝึกสอนข้อมูลที่ผ่านการแบ่งจากข้อ 5 โดยใช้ SVM (Support Vector Machine Classifier)

7. รวม Classification decision โดยใช้ Mean rule

งานวิจัยนี้ ทดลองกับ 4 ชุดข้อมูล แสดงดังตารางที่ 2.4

ตารางที่ 2.4 แสดงชุดข้อมูลสำหรับการทดลองของงานวิจัยเรื่อง Fuzzy Clustering and Fuzzy Entropy based Classification Model

Dataset	Featers	Instances
WDBC	30	569
WPBC	32	198
Parkinsons	21	197
Inoospheres	34	351

สรุปผลการทดลอง วัดความถูกต้องโดยใช้ 10- Fold cross validation เปรียบเทียบกับ วิธีการร่วมกันตัดสินใจ (ensemble methods) ที่เป็นมาตราฐานคือ bagging และ boosting ซึ่งงานวิจัยนี้ให้ประสิทธิภาพความถูกต้องดีกว่า

ข้อสังเกต ค่าความเป็นสมาชิกของ Hard Ensemble มีการกระจายตัวมาก อาจจะมีผลต่อการประสิทธิภาพของการจำแนกประเภท

2.2.5 งานวิจัยเรื่อง Model using K-Means Clustering Algorithm [5]

งานวิจัยนี้เสนอแบบจำลองการจำแนกประเภทข้อมูล เพื่อทำนายสภาวะอากาศสำหรับการเล่นเทนนิส โดยใช้อัลกอลิทึมเคมีนร่วมกับฟังก์ชั่น Probability Density Function (PDF) เปรียบเทียบกับการทำนายตามลักษณะของคลาส ซึ่งฟังก์ชั่น Probability Density Function แสดงดังสมการที่ 2.5

$$F(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}} \quad (2.5)$$

โดย F(x) คือฟังก์ชั่นความหนาแน่นของความน่าจะเป็น

X คือข้อมูล

μ คือค่าเฉลี่ย (Means)

σ คือค่าเบี่ยงเบนมาตรฐาน (standard deviation)

ข้อมูล 720 เรคคอต 5 แอตทริบิวต์

1. Outlook
2. Temperature
3. Humidity
4. Windy
5. Outcome (play): yes or no

รายละเอียดการสร้างแบบจำลองดังนี้

1. กำหนดเป้าหมายเพื่อการทำนายกลุ่มโดยเลือกที่ แอตทริบิวต์ outcome

2. ข้อมูลที่เป็นจำนวนเต็ม ไปหาผลรวม, ค่าเฉลี่ย, ค่าส่วนเบี่ยงเบนมาตราฐาน โดยแบ่งตามแอตทริบิวต์ outcome

3. คำนวณหา PDF ของข้อมูลที่เป็นตัวเลข โดยแบ่งตามแอตทริบิวต์ outcome: yes or no

4. นำข้อมูลที่เป็น Categories ไปหาความน่าจะเป็นของการเกิดเหตุการณ์ที่เป็นอิสระต่อกัน (Probability of Independent) โดยแบ่งตามแอตทริบิวต์ outcome: yes or no

5. หาความน่าจะเป็นของ แอตทริบิวต์ outcome ทั้งหมด โดยแบ่งตามแอตทริบิวต์ outcome: yes or no

6. ผลการคำนวณ จากข้อ 5 คือ ข้อมูลที่จะนำมาใช้กับอัลกอลิทึมเคมีน ซึ่งจะแบ่ง k

ตามแอตทริบิวต์ outlook

7. ประเมินความน่าจะเป็นของผลการทำนาย

8. ประเมินประสิทธิภาพของการทำนาย

ผลการทดลอง

1. ผลการทดลองโดยการทำนายตามลักษณะคลาส

Overall accuracy	77.36 %
Precision	85.63%
Recall	91.63%

2. ผลการทดลองโดยใช้อัลกอริทึมเคมีนจำแนกประเภทร่วมกับฟังก์ชั่น Probability Density Function

Mean square error	29.69%
Root mean square error	5.45%
Mean absolute error	39.58%

สรุปผลการทดลอง อัลกอริทึมเคมีนให้ค่าความคาดเคลื่อนน้อย เป็นไปตามสมติฐานที่ตั้งไว้

ข้อสังเกต การทำนายอาจจะคาดเคลื่อนมาก ถ้าหากข้อมูลที่ใช้มีความหลากหลายประเภท

บทที่ 3
การหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน

3.1 แนวคิดการหาจุดศูนย์กลางเริ่มต้นสำหรับการจำแนกประเภทแบบเคมีน

3.1.1 การหาจุดเริ่มต้นของอัลกอริทึมเคมีนสำหรับการจำแนกประเภทข้อมูล

เนื่องจากประสิทธิภาพการจัดกลุ่มของอัลกอริทึมเคมีนมีการแปรผันไปตามจุดเริ่มต้นดังนั้น งานวิจัยนี้จึงเสนอวิธีการเลือกจุดเริ่มต้น โดยการแปลงข้อมูล (Normalize) ให้อยู่ในช่วง 0 ถึง1 จากนั้นแบ่งข้อมูลออกเป็นจำนวน N ส่วน แล้วนับความถี่และจำนวนสมาชิกของแต่ละคลาสในชุดข้อมูล จากนั้นเรียงความถี่ของช่วงข้อมูลจากมากไปหาน้อย แล้วเลือกตัวแทนของคลาส โดยนำความถี่สามอันดับแรกมาหาสัดส่วนระหว่าง สมาชิกของคลาสที่พิจารณากับสมาชิกของคลาสอื่นที่เหลือทั้งหมดในชุดข้อมูล ความถี่สามอันดับแรกจะต้องเป็นสมาชิกของคลาสที่พิจารณา จากนั้นเลือกช่วงข้อมูลที่มีค่าสัดส่วนที่มากที่สุดเป็นตัวแทนคลาสเพื่อเป็นจุดเริ่มต้นของแอตทริบิวต์ที่พิจารณา ทำลักษณะเดียวกันนี้จนครบทุกคลาสและทุกแอตทริบิวต์ของชุดข้อมูล รายละเอียดการเลือกจุดเริ่มต้นของอัลกอริทึมเคมีนสำหรับการจำแนกประเภทข้อมูลแสดงในหัวข้อที่ 3.3.2.2

3.1.2 แนวคิดการใช้ค่าเอนโทรปีเพื่อปรับอัลกอริทึมเคมีนแบบดั้งเดิมให้เป็นเทคนิคการจำแนกประเภทข้อมูล

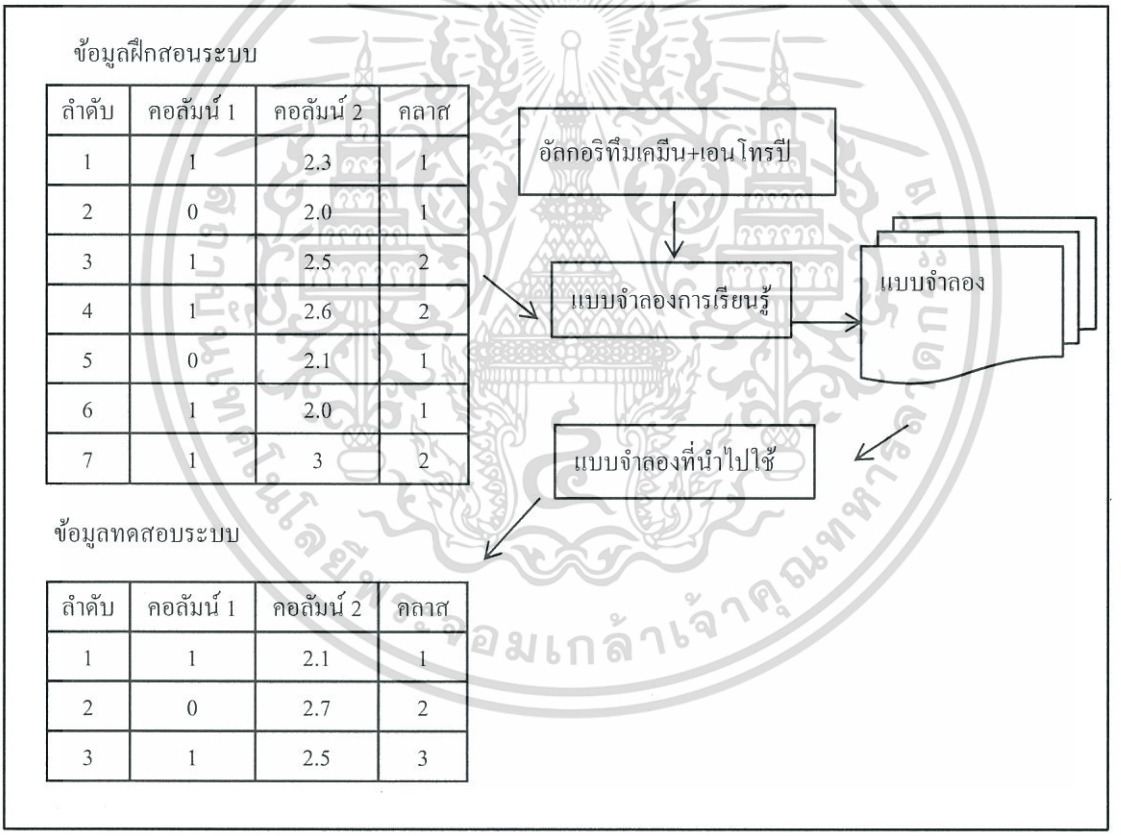
การพัฒนาอัลกอริทึมเคมีนสำหรับปัญหาการจำแนกประเภทข้อมูลในงานวิจัยนี้ ได้นำแนวคิดการใช้ค่าเอนโทรปีเพื่อปรับอัลกอริทึมเคมีนแบบดั้งเดิมให้เป็นเทคนิคการจำแนกประเภทข้อมูล กล่าวคืออัลกอริทึมเคมีนแบบดั้งเดิมใช้กับปัญหาการจัดกลุ่ม (Clustering) ซึ่งการจัดกลุ่มจัดเป็นการเรียนรู้แบบไม่มีผู้สอน(unsupervised Learning) โดยการเรียนรู้ของเครื่อง (Machine learning) จะเรียนรู้จากการนำข้อมูลที่ไม่มีการกำหนดคลาส (Class label) ให้กับข้อมูล ไปผ่านกระบวนการหาความคล้ายคลึง จนกระทั่งได้กลุ่มที่เหมาะสม ต่างจากการเรียนรู้แบบมีผู้สอน (supervised Learning) ที่มีการกำหนดคลาสให้กับข้อมูลเพื่อให้เครื่องเรียนรู้รูปแบบและนำรูปแบบดังกล่าว ไปใช้งานกับข้อมูลในอนาคต เช่นเดียวกับการคำนวณค่าเอนโทรปี ต้องพิจารณาจากคลาสข้อมูลเพื่อวัดความบริสุทธิ์ของข้อมูล ดังนั้นจึงนำเทคนิคดังกล่าวมาปรับอัลกอริทึมเคมีนเพื่อใช้สำหรับปัญหาการจำแนกประเภทข้อมูล

3.2 โครงสร้างการทำงานของอัลกอริทึมที่นำเสนอ

โครงสร้างการทำงานของอัลกอริทึมที่เสนอ ในหัวข้อนี้จะกล่าวถึงโครงสร้างแบบจำลองการจำแนกประเภทข้อมูลและโครงสร้างการทำงานของอัลกอริทึมที่นำเสนอ ดังนี้

3.2.1 โครงสร้างแบบจำลองการจำแนกประเภทข้อมูล

แบบจำลองการจำแนกประเภทข้อมูล โดยทั่วไปสร้างโดยนำชุดข้อมูลฝึกสอนระบบ (Training Data) ผ่านกระบวนการเรียนรู้ของแบบจำลองการเรียนรู้ จากนั้นนำแบบจำลองที่ผ่านการเรียนรู้ ดังกล่าวไปทำนายชุดข้อมูลทดสอบระบบ (Testing Data) โครงสร้างแบบจำลองการจำแนกประเภทข้อมูลแสดงดังรูปที่ 3.1



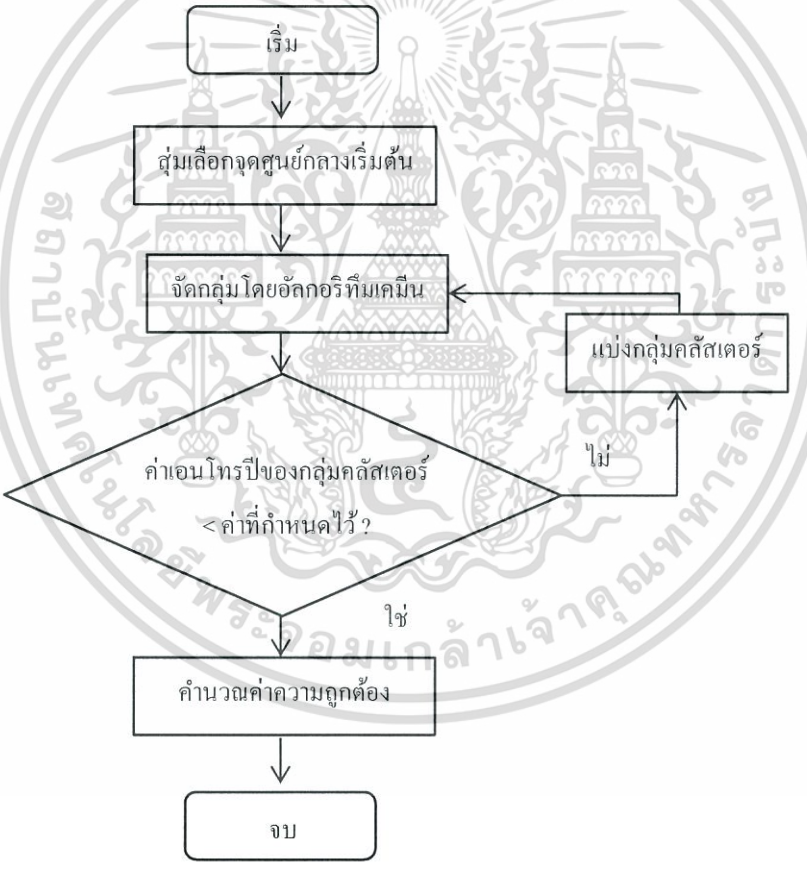
รูปที่ 3.1 แสดงโครงสร้างแบบจำลองการจำแนกประเภทข้อมูล

3.2.2 โครงสร้างการทำงานของอัลกอริทึมที่หนึ่ง

โครงสร้างการทำงานอัลกอริทึมที่หนึ่ง มีขั้นตอนการทำงานดังนี้

1. สุ่มเลือกจุดศูนย์กลางเริ่มต้น
2. จัดกลุ่มโดยใช้อัลกอริทึมเคมีน
3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์
4. แบ่งกลุ่มคลัสเตอร์ เมื่อค่าเอนโทรปีมากกว่าค่าที่กำหนดไว้
5. คำนวณค่าความถูกต้อง

โครงสร้างการทำงานของอัลกอริทึมที่หนึ่ง มีขั้นตอนการทำงานแสดงดังรูปที่ 3.2



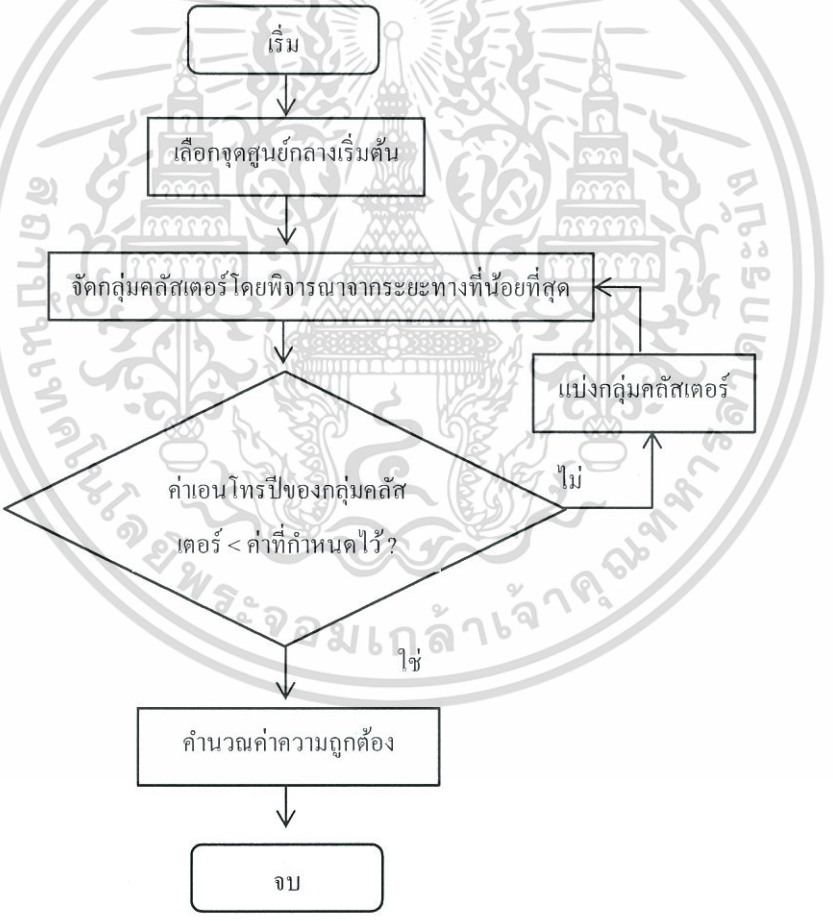
รูปที่ 3.2 แสดงโครงสร้างการทำงานของอัลกอริทึมที่หนึ่ง

3.2.3 โครงสร้างการทำงานของอัลกอริทึมที่สอง

โครงสร้างการทำงานอัลกอริทึมที่สอง มีขั้นตอนการทำงานดังนี้

1. เลือกจุดศูนย์กลางเริ่มต้นโดยใช้วิธีที่นำเสนอ รายละเอียดแสดงในหัวข้อที่ 3.3.2.2
2. จัดกลุ่มคลัสเตอร์ โดยพิจารณาจากระยะทางที่น้อยที่สุด
3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์
4. แบ่งกลุ่มคลัสเตอร์ เมื่อค่าเอนโทรปีมากกว่าค่าที่กำหนดไว้
5. คำนวณค่าความถูกต้อง

โครงสร้างการทำงานของอัลกอริทึมที่สอง มีขั้นตอนการทำงานแสดงดังรูปที่ 3.3



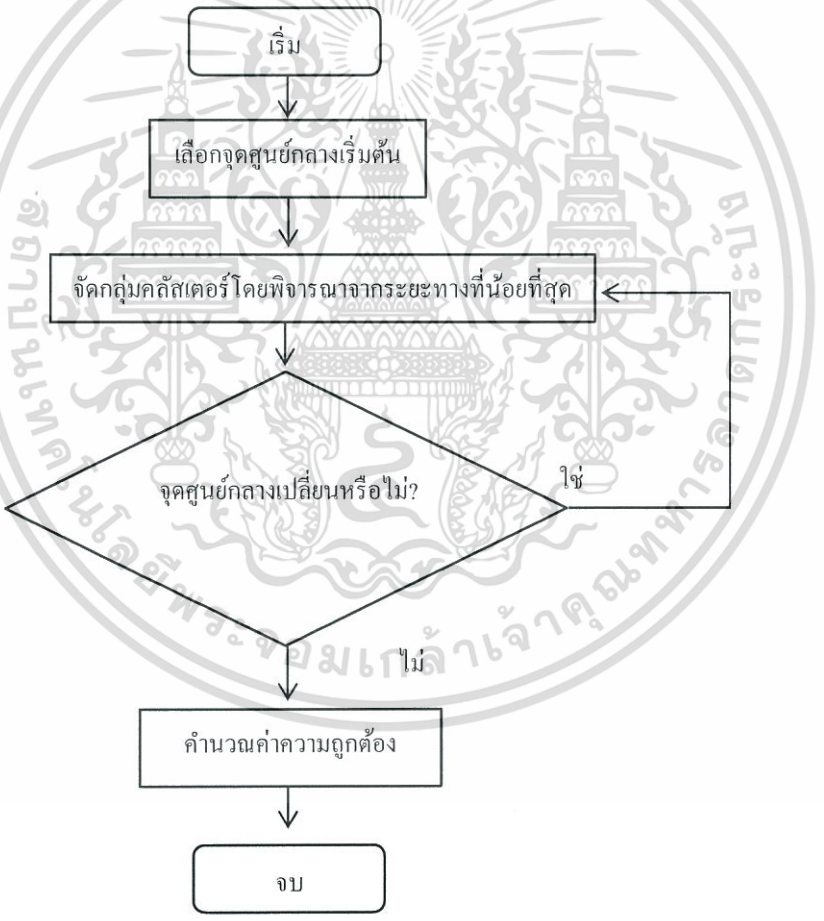
รูปที่ 3.3 แสดงโครงสร้างการทำงานของอัลกอริทึมที่สอง

3.2.4 โครงสร้างการทำงานของอัลกอริทึมที่สาม

โครงสร้างการทำงานอัลกอริทึมที่สอง มีขั้นตอนการทำงานดังนี้

1. เลือกจุดศูนย์กลางเริ่มต้นโดยใช้วิธีที่นำเสนอ รายละเอียดแสดงในหัวข้อที่ 3.3.2.2
2. จัดกลุ่มคลัสเตอร์ โดยพิจารณาจากระยะทางที่น้อยที่สุด
3. พิจารณาจุดศูนย์กลางเปลี่ยนหรือไม่ ถ้าเปลี่ยน ทำซ้ำข้อ 2 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยน
4. คำนวณค่าความถูกต้อง

โครงสร้างการทำงานของอัลกอริทึมที่สาม มีขั้นตอนการทำงานแสดงดังรูปที่ 3.4



รูปที่ 3.4 แสดงโครงสร้างการทำงานของอัลกอริทึมที่สาม

3.3 วิธีการของอัลกอริทึมที่นำเสนอ

วิธีการของอัลกอริทึมที่นำเสนอ ในหัวข้อนี้จะกล่าวถึง ขั้นตอนการเตรียมข้อมูล ขั้นตอนการเรียนรู้และขั้นตอนการทดสอบของอัลกอริทึมที่นำเสนอ โดยมีรายละเอียดดังนี้

3.3.1 ขั้นตอนการเตรียมข้อมูล

การเตรียมข้อมูลโดยแปลงข้อมูล (Normalize)ให้อยู่ในช่วง 0 ถึง 1 โดยใช้สมการที่ 3.1

$$v' = \frac{V - Min_A}{Max_A - Min_A} \left(New_{max_A} - New_{min_A} \right) + New_{min_A} \quad (3.1)$$

เมื่อ v' คือข้อมูลผ่านการแปลงที่มีค่าอยู่ระหว่าง 0 ถึง 1

v คือข้อมูลก่อนการแปลง

Min_A คือข้อมูลที่มีค่าต่ำสุดในชุดข้อมูล

Max_A คือข้อมูลที่มีค่าสูงสุดในชุดข้อมูล

New_{max_A} คือค่ามากสุดของข้อมูลที่ต้องการ

New_{min_A} คือค่าน้อยสุดของข้อมูลที่ต้องการ

จากนั้นแบ่งข้อมูลเป็น 2 ส่วน โดยร้อยละ 80 สำหรับใช้ฝึกสอนระบบ และร้อยละ 20 สำหรับใช้ทดสอบระบบ

3.3.2 ขั้นตอนการเรียนรู้

3.3.2.1 อัลกอริทึมที่หนึ่ง

ขั้นตอนการเรียนรู้ของอัลกอริทึมที่หนึ่ง มีขั้นตอนดังนี้

1. เลือกจุดศูนย์กลางเริ่มต้น

การเลือกจุดเริ่มต้นของอัลกอริทึมที่หนึ่ง ใช้วิธีการสุ่มเลือก เท่ากับจำนวนคลาสของชุดข้อมูล

2. จัดกลุ่มโดยอัลกอริทึมเคมีน

การจัดกลุ่มโดยอัลกอริทึมเคมีน คือพิจารณาระยะทางที่ใกล้สุดระหว่างจุดศูนย์กลางกับข้อมูล โดยใช้สมการที่ 3.2

$$J = \arg^{K} \min_{j=1} \left[d(x_i, z_j) \right] \quad (3.2)$$

เมื่อ J คือคลัสเตอร์

K คือจำนวนข้อมูล

$d(x_i, z_j)$ คือระยะทางระหว่างจุดข้อมูล x_i กับจุดศูนย์กลาง z_j

โดยระยะทางระหว่างจุดศูนย์กลางกับข้อมูล คำนวณจากสมการที่ 3.3

$$d(x_{i,}, z_j) = \sqrt{\sum_{q=1}^{Q}(x_{iq,} - z_{jq})^2} \quad (3.3)$$

เมื่อ $d(x_{i,}, z_j)$ คือระยะทางระหว่าง จุด x_i กับ z_j

x_i คือข้อมูลที่หนึ่ง

z_j คือจุดข้อมูลที่สอง

Q คือมิติของข้อมูล

3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์

การคำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์เพื่อพิจารณาการแบ่งกลุ่ม การแบ่งกลุ่มเกิดขึ้นก็ต่อเมื่อค่าเอนโทรปีของคลัสเตอร์มากกว่า E_γ

เมื่อ E_γ คือค่าที่กำหนดไว้

ค่าเอนโทรปีของกลุ่มคลัสเตอร์ คำนวณได้จากสมการที่ 3.4

$$H(C_j) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (3.4)$$

เมื่อ $H(C_j)$ คือค่าเอนโทรปีของกลุ่ม C_j

p_i คือค่าสัดส่วนของสมาชิกของคลาสกับจำนวนสมาชิกทั้งหมดในกลุ่มคลัสเตอร์

m คือจำนวนคลาสที่อยู่ในกลุ่ม

4. แบ่งกลุ่มคลัสเตอร์ เมื่อค่าเอนโทรปีมากกว่า E_γ

วิธีการแบ่งกลุ่มคลัสเตอร์ คือตรวจสอบคลาสของข้อมูล จากนั้นนำข้อมูลที่มีคลาสเหมือนกัน จัดอยู่ให้อยู่กลุ่มเดียวกัน

หลังจากการแบ่งกลุ่มจุดศูนย์กลางใหม่เกิดจากค่าเฉลี่ยของสามาชิกในกลุ่ม จากนั้นทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเอนโทรปีของกลุ่มคลัสเตอร์น้อยกว่า E_γ

5. คำนวณค่าความถูกต้อง

การคำนวณค่าความถูกต้องของการจำแนกประเภทข้อมูลคือค่าสัดส่วนระหว่างข้อมูลที่จำแนกประเภทได้ถูกต้องกับข้อมูลทั้งหมด โดยใช้สมการที่ 3.5

$$\text{Accuracy} = \frac{\text{TC}}{\text{TN}} \tag{3.5}$$

เมื่อ Accuracy คือค่าความถูกต้องของข้อมูล

TC คือจำนวนข้อมูลที่จำแนกประเภทได้ถูกต้อง

TN คือจำนวนข้อมูลทั้งหมด

3.3.2.2 อัลกอริทึมที่สอง

ขั้นตอนการเรียนรู้ของอัลกอริทึมที่สอง มีขั้นตอนดังนี้

1. เลือกจุดศูนย์กลางเริ่มต้นโดยใช้วิธีที่นำเสนอ มีขั้นตอนดังนี้
   1.1 กำหนดจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนคลาสของชุดข้อมูล
   1.2 แบ่งข้อมูลออกเป็น N ช่วง เมื่อ N คือค่าที่กำหนดไว้
   1.3 เริ่มทำทีละแอตทริบิวต์ (Attribute) โดยเริ่มจากแอตทริบิวต์ที่หนึ่ง จากนั้นนับจำนวนความถี่ในช่วงข้อมูลและเรียงข้อมูลจากความถี่สูงไปต่ำ
   1.4 เลือกสามอันดับแรกที่มีความถี่สูงสุดและต้องมีความเป็นสมาชิกของคลาสที่พิจารณา โดยพิจารณาจากคลาสที่หนึ่งของชุดข้อมูลก่อน
   1.5 นำข้อมูลจากข้อ 1.4 มาหาสัดส่วนกับคลาสอื่นๆที่เหลือทั้งหมดในชุดข้อมูล
   1.6 เลือกช่วงข้อมูลที่มีค่าสัดส่วนมากที่สุดเป็นตัวแทนของคลาส ซึ่งจุดศูนย์กลางที่ได้เกิดจากค่าเฉลี่ยของข้อมูลทั้งหมดในช่วงข้อมูลที่ถูกเลือก
   1.7 ลบช่วงข้อมูลที่ถูกเลือกในขั้นตอนที่ 1.4
   1.8 ทำขั้นตอนที่ 1.4 -1.7 เพื่อหาจุดศูนย์กลางของคลาสอื่นๆที่เหลือทั้งหมดในชุดข้อมูล (2,3,4…K) จนครบทุกคลาสในชุดข้อมูล

ทำขั้นตอนที่ 1.3 - 1.8 กับทุกแอตทริบิวต์ของชุดข้อมูล

หลังจากทำครบทั้ง 8 ขั้นตอนข้างต้น จะได้จุดศูนย์กลางเริ่มต้นเป็นตัวแทนของแต่ละคลาส

2. จัดกลุ่มคลัสเตอร์

การจัดกลุ่มคลัสเตอร์พิจารณาจากระยะทางระหว่างจุดศูนย์กลางที่ได้จากขั้นตอนก่อนหน้านี้กับข้อมูลทั้งหมดในชุดข้อมูล ซึ่งข้อมูลที่มีระยะทางใกล้กับจุดศูนย์กลางจะถูกจัดให้อยู่กลุ่มเดียวกัน ระยะทางระหว่างจุดศูนย์กลางกับข้อมูล คำนวณจากสมการที่ 3.3

3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์

การคำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์เพื่อพิจารณาการแบ่งกลุ่ม การแบ่งกลุ่มเกิดขึ้นก็ต่อเมื่อค่าเอนโทรปีของคลัสเตอร์มากกว่า E_γ

เมื่อ E_γ คือค่าที่กำหนดไว้

ค่าเอนโทรปีของกลุ่มคลัสเตอร์ คำนวณได้จากสมการที่ 3.4

4. แบ่งกลุ่มคลัสเตอร์ เมื่อค่าเอนโทรปีมากกว่า E_γ

วิธีการแบ่งกลุ่มคลัสเตอร์ คือตรวจสอบคลาสของข้อมูล จากนั้นนำข้อมูลที่มีคลาสเหมือนกัน จัดอยู่ให้อยู่กลุ่มเดียวกัน

หลังจากการแบ่งกลุ่ม จุดศูนย์กลางใหม่เกิดจากค่าเฉลี่ยของสามาชิกในกลุ่ม จากนั้นทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเอนโทรปีของกลุ่มคลัสเตอร์น้อยกว่า E_γ

5. คำนวณค่าความถูกต้อง

การคำนวณค่าความถูกต้องของการจำแนกประเภทข้อมูล คือค่าสัดส่วนระหว่างข้อมูลที่จำแนกประเภทได้ถูกต้องกับข้อมูลทั้งหมดในชุดข้อมูล โดยใช้สมการที่ 3.5

3.3.2.3 อัลกอริทึมที่สาม

ขั้นตอนการเรียนรู้ของอัลกอริทึมที่สาม มีขั้นตอนดังนี้

1. เลือกจุดศูนย์กลางเริ่มต้นโดยใช้วิธีที่นำเสนอ มีขั้นตอนเช่นเดียวกับอัลกอริทึมที่สอง รายละเอียดแสดงในหัวข้อ 3.3.2.2
2. จัดกลุ่มคลัสเตอร์ โดยพิจารณาจากระยะทางที่น้อยที่สุดดังสมการที่ 3.2
3. พิจารณาจุดศูนย์กลางเปลี่ยนหรือไม่ ถ้าเปลี่ยน ทำซ้ำข้อ 2 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยน
4. คำนวณค่าความถูกต้อง ดังสมการที่ 3.5

3.3.3 ขั้นตอนการทดสอบ

การทดสอบ คือการนำตัวแทนของกลุ่มข้อมูลที่ผ่านการเรียนรู้ มาทำนายกลุ่มซึ่งการทำนายกลุ่มทำได้โดย การวัดระยะทางระหว่างข้อมูลทดสอบระบบกับตัวแทนของกลุ่มข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมการเรียนรู้เพื่อจำแนกประเภทโดยพิจารณาระยะทางที่ใกล้สุด จากนั้นคำนวณค่าความถูกต้องโดยใช้สมการที่ 3.5

3.4 ตัวอย่างการทำงาน

1. ขั้นตอนการเตรียมข้อมูล

การเตรียมข้อมูล คือการแปลงข้อมูลเพื่อให้อยู่ในช่วง 0 ถึง 1 โดยใช้สมการที่ 3.1 จากนั้นแบ่งข้อมูลโดยร้อยละ 80 สำหรับใช้ฝึกสอนระบบ แสดงดังตารางที่ 3.1 และร้อยละ 20 สำหรับใช้ทดสอบระบบ แสดงดังตารางที่ 3.2

ตารางที่ 3.1 แสดงข้อมูลฝึกสอนระบบ

ลำดับ	แอตทริบิวต์ 1	แอตทริบิวต์ 2	แอตทริบิวต์ 3	แอตทริบิวต์ 4	คลาส
1	0.3361	0.5882	0.0556	0.0415	1
2	0.2173	0.5882	0.0556	0.081	1
3	0.3124	0.6618	0.0874	0.0415	1
4	0.1936	0.4412	0.0874	0.002	1
5	0.4549	0.5515	0.5639	0.5949	2
6	0.4786	0.3309	0.4845	0.4763	2
7	0.5499	0.3676	0.5322	0.4763	2
8	0.1936	0.1838	0.3733	0.3577	2
9	0.5499	0.3309	0.7387	0.7925	3
10	0.5261	0.3676	0.7387	0.6739	3
11	0.4311	0.4044	0.6593	0.6739	3
12	0.6449	0.4779	0.7863	0.8715	3

ตารางที่ 3.2 ข้อมูลทดสอบระบบ

ลำดับ	แอตทริบิวต์ 1	แอตทริบิวต์ 2	แอตทริบิวต์ 3	แอตทริบิวต์ 4	คลาส
1	0.2411	0.5515	0.0874	0.0415	1
2	0.4074	0.2574	0.4845	0.4368	2
3	0.6211	0.5147	0.7546	0.9506	3

2. ขั้นตอนการเรียนรู้

2.1 ตัวอย่างการเรียนรู้ของอัลกอริทึมที่หนึ่ง

1. เลือกจุดศูนย์กลางเริ่มต้น

การเลือกจุดศูนย์กลางเริ่มต้นของอัลกอริทึมที่หนึ่ง เลือกโดยการสุ่มเท่ากับจำนวนคลาสของชุดข้อมูล โดยใช้ข้อมูลฝึกสอนระบบในตารางที่ 3.1 จุดศูนย์กลางเริ่มต้นของอัลกอริทึมที่หนึ่งแสดงดังตารางที่ 3.3

ตารางที่ 3.3 แสดงจุดศูนย์กลางเริ่มต้นของอัลกอริทึมที่หนึ่งโดยการสุ่ม

ลำดับกลุ่ม	แอตทริบิวต์ 1	แอตทริบิวต์ 2	แอตทริบิวต์ 3	แอตทริบิวต์ 4	คลาส
1	0.3124	0.6618	0.0874	0.0415	1
2	0.4311	0.4044	0.6593	0.6739	3
3	0.4549	0.5515	0.5639	0.5949	2

2. จัดกลุ่มโดยอัลกอริทึมเคมีน

เมื่อได้จุดศูนย์กลางจากการสุ่ม แล้วนำไปจัดกลุ่มโดยใช้อัลกอริทึมเคมีนซึ่งพิจารณาระยะทางที่ใกล้สุดระหว่างจุดศูนย์กลางกับข้อมูล ดังสมการที่ 3.2 และระยะทางระหว่างจุดศูนย์กลางกับข้อมูล คำนวณจากสมการที่ 3.3 ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมดของการจัดกลุ่มโดยใช้อัลกอริทึมเคมีและกลุ่มของข้อมูลแสดงดังตารางที่ 3.4

ตารางที่ 3.4 แสดงระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมดและกลุ่มของข้อมูลที่ได้จากการจัดกลุ่มโดยอัลกอริทึมเคมีนในขั้นตอนการทำงานของอัลกอริทึมที่หนึ่ง

ลำดับ	ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมด			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	0.0753	1.0198	0.6604	1
2	0.0664	1.0234	0.6612	1
3	0.1047	1.0208	0.6733	1
4	0.1531	1.0481	0.6660	1
5	0.7650	0.2901	0.2414	3
6	0.6800	0.3806	0.0656	3
7	0.7235	0.3418	0.1381	3
8	0.5875	0.6688	0.3298	3
9	1.0712	0.0768	0.4248	2
10	0.9769	0.0849	0.3364	2
11	0.8947	0.1512	0.2655	2
12	1.1631	0.1882	0.5568	2

จากตารางที่ 3.4 แสดงกลุ่มคลัสเตอร์โดยใช้อัลกอริทึมเคมีนในรอบแรก จากนั้นรันอัลกอริทึมเคมีนจนกระทั่งจุดศูนย์กลางไม่เปลี่ยน

3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์

จากตารางที่ 3.4 คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์โดยใช้สมการ 3.4 ค่าเอนโทรปีแสดงดังตารางที่ 3.5

ตารางที่ 3.5 แสดงจำนวนสมาชิกของกลุ่มคลัสเตอร์และค่าเอนโทรปีของอัลกอริทึมที่หนึ่ง

คลัสเตอร์	จำนวนสมาชิก	คลาส 1	คลาส 2	คลาส 3	เอนโทรปี
คลัสเตอร์ 1	4	4	0	0	0
คลัสเตอร์ 2	4	0	0	4	0
คลัสเตอร์ 3	4	0	4	0	0

จากตารางที่ 3.5 พิจารณาค่าเอนโทรปีของแต่ละคลัสเตอร์เพื่อแบ่งกลุ่ม

4. แบ่งกลุ่มคลัสเตอร์

การแบ่งกลุ่มคลัสเตอร์จะแบ่งก็ต่อเมื่อค่าเอนโทรปีของคลัสเตอร์มากกว่า E_γ สำหรับการทดลองนี้ให้ E_γ เท่ากับ 0.2 วิธีการแบ่งกลุ่มคลัสเตอร์ คือตรวจสอบคลาสของข้อมูล จากนั้นนำข้อมูลที่มีคลาสเหมือนกัน จัดให้อยู่กลุ่มเดียวกัน

หลังจากการแบ่งกลุ่มคลัสเตอร์ จุดศูนย์กลางใหม่เกิดจากค่าเฉลี่ยของสามาชิกในกลุ่ม จากนั้นทำซ้ำขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเอนโทรปีของกลุ่มคลัสเตอร์น้อยกว่า E_γ

จากตารางที่ 3.5 ค่าเอนโทรปีของคลัสเตอร์ เท่ากับศูนย์ซึ่งถูกต้องตามเงื่อนไขของการหยุดการทำงาน และข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่ง แสดงดังตารางที่ 3.6

ตารางที่ 3.6 แสดงข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่ง

ลำดับกลุ่ม	แอตทริบิวต์1	แอตทริบิวต์2	แอตทริบิวต์3	แอตทริบิวต์4	คลาส
1	0.2649	0.5699	0.0715	0.0415	1
2	0.5380	0.3952	0.7308	0.7530	3
3	0.4193	0.3585	0.4885	0.4763	2

5. คำนวณค่าความถูกต้อง

เมื่อค่าเอนโทรปีของแต่ละกลุ่มคลัสเตอร์น้อยกว่า E_γ นำจุดศูนย์กลางของแต่ละกลุ่มคลัสเตอร์ไปทำนายกลุ่มกับข้อมูลทดสอบระบบและหาค่าความถูกต้อง ซึ่งแสดงในขั้นตอนการทดสอบ

2.2 ตัวอย่างการเรียนรู้ของอัลกอริทึมที่สอง

1. เลือกจุดศูนย์กลางเริ่มต้นโดยวิธีที่นำเสนอ มีขั้นตอนดังนี้

1.1. กำหนดจุดศูนย์กลางเริ่มต้นเท่ากับจำนวนคลาสของชุดข้อมูล K=3

1.2. แบ่งช่วงข้อมูลออกเป็น N=10 ช่วง

1.3. เริ่มทำทีละแอตทริบิวต์เริ่มจากแอตทรบิวต์ที่หนึ่ง นับจำนวนความถี่และเรียงความถี่จากสูงไปต่ำ ตัวอย่างแสดงดังตารางที่ 3.7

ตารางที่ 3.7 แสดงความถี่ที่เรียงจากค่าสูงไปต่ำของแอตทริบิวต์ที่สอง

ลำดับ	ช่วงข้อมูล	ความถี่	จำนวนสมาชิกของคลาส		
			คลาส 1	คลาส 2	คลาส 3
1	0.1900-0.1999	2	1	1	0
2	0.5400-0.5499	2	0	1	1
3	0.2100-0.2199	1	1	0	0
4	0.3100-0.3199	1	1	0	0
5	0.3300-0.3399	1	1	0	0
6	0.4300-0.4399	1	0	0	1
7	0.4500-0.4599	1	0	1	0
8	0.4700-0.4799	1	0	1	0
9	0.5200-0.5299	1	0	0	1
10	0.6400-0.6499	1	0	0	1

1.4 เลือกสามอันดับแรกที่มีความถี่สูงสุดและต้องมีความเป็นสมาชิกของคลาสที่พิจารณา โดยพิจารณาจากคลาสที่หนึ่งของชุดข้อมูลก่อน ตัวอย่างแสดงดังตารางที่ 3.8

ตารางที่ 3.8 แสดงช่วงข้อมูลที่มีความถี่สูงสุดและเป็นสมาชิกของคลาส 1

ลำดับ	ช่วงข้อมูล	ความถี่	จำนวนสมาชิกของคลาส		
			คลาส 1	คลาส 2	คลาส 3
1	0.1900-0.1999	2	1	1	0
2	0.2100-0.2199	1	1	0	0
3	0.3100-0.3199	1	1	0	0

1.5 นำข้อมูลที่ได้จากข้อ 1.4 มาหาสัดส่วนกับคลาสอื่นๆที่เหลือทั้งหมดในชุดข้อมูล โดยรายละเอียดการหาสัดส่วนดังนี้

จากตารางที่ 3.8 นำความถี่ของสมาชิกคลาส 1 มาหาสัดส่วนกับคลาสอื่น ดังนี้

ลำดับที่ 1 ช่วงข้อมูล 0.1900-0.1999 สัดส่วนระหว่างคลาส 1กับคลาส 2 เท่ากับ 1/1=1

สัดส่วนระหว่างคลาส 1กับคลาส 3 เท่ากับ 1/0=0

ลำดับที่ 2 ช่วงข้อมูล 0.2100-0.2199 สัดส่วนระหว่างคลาส 1กับคลาส 2 เท่ากับ 1/0=0

สัดส่วนระหว่างคลาส 1กับคลาส 3 เท่ากับ 1/0=0

ลำดับที่ 3 ช่วงข้อมูล 0.3100-0.3199 สัดส่วนระหว่างคลาส 1กับคลาส 2 เท่ากับ 1/0=0

สัดส่วนระหว่างคลาส 1กับคลาส 3 เท่ากับ 1/0=0

1.6 เลือกช่วงข้อมูลที่มีสัดส่วนมากที่สุดเป็นตัวแทนของคลาส ซึ่งจุดศูนย์กลางที่ได้เกิดจากค่าเฉลี่ยของข้อมูลทั้งหมดในช่วงข้อมูล

จากข้อมูลข้างต้นจะเห็นว่าช่วงข้อมูลลำดับที่ 1 มีค่าสูงสุด นำข้อมูลที่อยู่ในช่วง 0.1900-0.1999 มาหาค่าเฉลี่ย ได้ดังนี้

ช่วง 0.1900-0.1999 มีความถี่เท่ากับ 2 นำมาหาค่าเฉลี่ย (0.1936+0.1936)/2=0.1936

ดังนั้น 0.1936 เป็นตัวแทนของคลาส1 สำหรับแอตทริบิวต์ที่ 1

1.7 ลบช่วงข้อมูลที่ถูกเลือกในขั้นตอน 1.4

1.8 ทำขั้นตอนที่ 1.4 -1.7 เพื่อหาจุดศูนย์กลางของคลาสอื่นๆ (2, 3, 4...K) จนครบทุกคลาสในชุดข้อมูล

ทำขั้นตอนที่ 1.3 -1.8 กับทุกแอตทริบิวต์ของชุดข้อมูล

หลังจากทำครบทั้ง 8 ขั้นตอนข้างต้นจะได้จุดศูนย์กลางเริ่มต้นในแต่กลุ่มคลาส ตัวอย่างแสดงดังตารางที่ 3.9

ตารางที่ 3.9 แสดงจุดเริ่มต้นสำหรับอัลกอริทึมที่สอง

ลำดับกลุ่ม	แอตทริบิวต์1	แอตทริบิวต์2	แอตทริบิวต์3	แอตทริบิวต์4	คลาส
1	0.1936	0.5882	0.0556	0.0415	1
2	0.5499	0.3309	0.3733	0.4763	2
3	0.4311	0.3676	0.7387	0.6739	3

2. จัดกลุ่มคลัสเตอร์

การจัดกลุ่มคลัสเตอร์พิจารณาจากระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมดในชุดข้อมูลซึ่งข้อมูลที่มีระยะทางใกล้กับจุดศูนย์กลางจะถูกจัดให้อยู่กลุ่มเดียวกัน ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลโดยใช้สมการที่ 3.3 ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมดและกลุ่มของข้อมูลในขั้นตอนการทำงานของอัลกอริทึมที่สอง แสดงดังตารางที่ 3.10

ตารางที่ 3.10 แสดงระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมดและกลุ่มของข้อมูลในขั้นตอนการทำงานของอัลกอริทึมที่สอง

ลำดับ	ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมด			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	0.1425	0.6340	0.9614	1
2	0.0461	0.6588	0.9553	1
3	0.1433	0.6608	0.9616	1
4	0.1555	0.6677	0.9682	1
5	0.7964	0.3288	0.2668	3
6	0.7214	0.1321	0.3275	2
7	0.7693	0.1631	0.3095	2
8	0.6037	0.4033	0.5689	2
9	1.1062	0.4832	0.1718	3
10	1.0128	0.4177	0.0950	3
11	0.9244	0.3746	0.0875	3
12	1.1994	0.5978	0.3149	3

3. คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์

จากตารางที่ 3.10 คำนวณค่าเอนโทรปีของกลุ่มคลัสเตอร์โดยใช้สมการ 3.4 ค่าเอนโทรปีแสดงดังตารางที่ 3.11

ตารางที่ 3.11 แสดงสมาชิกของกลุ่มคลัสเตอร์และค่าเอนโทรปีของอัลกอริทึมที่สอง

กลุ่ม	จำนวนสมาชิก	คลาส1	คลาส2	คลาส3	เอนโทรปี
คลัสเตอร์1	4	4	0	0	0
คลัสเตอร์2	3	0	3	0	0
คลัสเตอร์3	5	0	1	4	0.7219

จากตารางที่ 3.11 พิจารณาค่าเอนโทรปีของแต่ละคลัสเตอร์เพื่อแบ่งกลุ่ม

4. แบ่งกลุ่มคลัสเตอร์

การแบ่งกลุ่มก็ต่อเมื่อค่าเอนโทรปีของคลัสเตอร์มากกว่า E_γ สำหรับการทดลองนี้ให้ E_γ เท่ากับ 0.2 วิธีการแบ่งกลุ่มคลัสเตอร์ คือตรวจสอบคลาสของข้อมูลจากนั้นนำข้อมูลที่มีคลาสเหมือนกัน จัดให้อยู่กลุ่มเดียวกัน หลังจากการแบ่งกลุ่มจุดศูนย์กลางใหม่เกิดจากค่าเฉลี่ยของสามาชิกในกลุ่ม

จากนั้นทำซ้ำ ขั้นตอนที่ 2 และ 3 จนกระทั่งค่าเอนโทรปีของกลุ่มคลัสเตอร์น้อยกว่า E_γ และข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง แสดงดังตารางที่ 3.12

ตารางที่ 3.12 แสดงข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง

ลำดับ	แอตทริบิวต์ 1	แอตทริบิวต์ 2	แอตทริบิวต์ 3	แอตทริบิวต์ 4	คลาส
1	0.1936	0.5882	0.0556	0.0415	1
2	0.5499	0.3309	0.3733	0.4763	2
3	0.4311	0.3676	0.7387	0.6739	3
4	0.4549	0.5515	0.5639	0.5949	2
5	0.5380	0.3952	0.7308	0.7530	3

5. คำนวณค่าความถูกต้อง

เมื่อค่าเอนโทรปีของแต่ละกลุ่มคลัสเตอร์น้อยกว่า E_γ นำจุดศูนย์กลางของแต่ละกลุ่มคลัสเตอร์ไปทำนายกลุ่มกับข้อมูลทดสอบระบบและหาค่าความถูกต้อง ซึ่งแสดงในขั้นตอนการทดสอบ

2.3 ตัวอย่างการเรียนรู้ของอัลกอริทึมที่สาม

1. เลือกจุดศูนย์กลางเริ่มต้นโดยใช้วิธีที่นำเสนอ มีขั้นตอนเช่นเดียวกับอัลกอริทึมที่สอง

ดังนั้นจุดศูนย์กลางเริ่มต้นของอัลกอริทึมที่สาม ในตัวอย่างนี้แสดงดังตารางที่ 3.9

2. จัดกลุ่มคลัสเตอร์ โดยพิจารณาจากระยะทางที่น้อยที่สุดระหว่างจุดศูนย์กลางจากตารางที่ 3.9 กับข้อมูลฝึกสอนระบบทั้งหมดจากตารางที่ 3.1 โดยใช้สมการที่ 3.2 ระยะทางและการจัดกลุ่มในรอบแรก แสดงดังตารางที่ 3.13

ตารางที่ 3.13 แสดงระยะทางจุดศูนย์กลางกับข้อมูลทั้งหมดและกลุ่มของข้อมูลที่ได้จากการจัดกลุ่มในขั้นตอนการทำงานของอัลกอริทึมที่สาม

ลำดับ	ระยะทางระหว่างจุดศูนย์กลางกับข้อมูลทั้งหมด			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	0.0753	0.6244	0.9226	1
2	0.0664	0.6130	0.9282	1
3	0.1047	0.6494	0.9238	1
4	0.1531	0.5856	0.9544	1
5	0.7650	0.3880	0.2045	3
6	0.6800	0.1798	0.2927	2
7	0.7235	0.2684	0.2522	3
8	0.5875	0.1798	0.5974	2
9	1.0712	0.5367	0.1588	3
10	0.9769	0.4585	0.0848	3
11	0.8947	0.3870	0.0966	3
12	1.1631	0.6916	0.2606	3

3. พิจารณาจุดศูนย์กลางเปลี่ยนหรือไม่ ถ้าเปลี่ยน ทำซ้ำข้อ 2 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยน จากตารางที่ 3.13 ทำซ้ำจนกระทั่งจุดศูนย์กลางไม่เปลี่ยน จุดศูนย์กลางสุดท้ายที่ได้จากการเรียนรู้ของอัลกอริทึมที่สาม แสดงดังตารางที่ 3.14

ตารางที่ 3.14 แสดงข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สาม

ลำดับ	แอตทริบิวต์1	แอตทริบิวต์2	แอตทริบิวต์3	แอตทริบิวต์4	คลาส
1	0.2649	0.5699	0.0715	0.0415	1
2	0.3361	0.2574	0.4289	0.4170	2
3	0.5261	0.4167	0.6699	0.6805	3

4. คำนวณค่าความถูกต้อง โดยใช้สมการที่ 3.5

ขั้นตอนการคำนวณค่าความถูกต้อง คือการนำข้อมูลที่ได้จากการรียนรู้จากอัลกอริทึมที่สามไปทำนายกลุ่มกับข้อมูลทดสอบระบบและคำนวณค่าความถูกต้อง ซึ่งจะแสดงในขั้นตอนการทดสอบ

3. ขั้นตอนการทดสอบ

การทดสอบ คือการนำข้อมูลที่ผ่านการเรียนรู้ ไปทำนายกลุ่มซึ่งการทำนายกลุ่มทำได้โดยการวัดระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมการเรียนรู้เพื่อจำแนกประเภทโดยพิจารณาระยะทางที่ใกล้สุด จากนั้นคำนวณค่าความถูกต้อง

3.1 ตัวอย่างการทดสอบของอัลกอริทึมที่หนึ่ง

นำข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่งจากตารางที่ 3.6 และข้อมูลทดสอบระบบ จากตารางที่ 3.2 มาคำนวณหาระยะทาง แสดงดังตารางที่ 3.15 ดังนี้

ตารางที่ 3.15 แสดงระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่งและกลุ่มของข้อมูล

ลำดับ	ระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่ง			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	<u>0.0340</u>	1.0162	0.6472	1
2	0.6669	0.4434	<u>0.1092</u>	3
3	1.1929	<u>0.2466</u>	0.6008	2

จากตารางที่ 3.15 การทดสอบอัลกอริทึมที่หนึ่ง แสดงได้ดังตารางที่ 3.16 ดังนี้

ตารางที่ 3.16 แสดงผลการทดสอบอัลกอริทึมที่หนึ่ง

ข้อมูลทดสอบระบบ		ข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่หนึ่ง		ผลการทำนายกลุ่ม		
ลำดับข้อมูล	คลาส	ลำดับกลุ่ม	คลาส	ลำดับกลุ่ม	คสาส	ถูกต้อง
1	1	1	1	1	1	1
2	2	2	3	3	2	1
3	3	3	2	2	3	1

จากตารางที่ 3.16 คำนวณค่าความถูกต้องโดยใช้สมการที่ 3.5 ดังนี้

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่จำแนกกลุ่มได้ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$\text{แทนค่า} = \frac{3}{3}$$

ค่าความถูกต้อง คือ 1

3.2 ตัวอย่างการทดสอบของอัลกอริทึมที่สอง

กรณีตัวอย่างการทดสอบของอัลกอริทึมที่สอง จากตารางที่ 3.12 ข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง มีจำนวนมากกว่าข้อมูลทดสอบระบบ จากตารางที่ 3.2 ดังนั้นจึงเลือกข้อมูลที่ผ่านขั้นตอนการเรียนรู้จากอัลกอริทึมที่สองสำหรับเป็นตัวแทนของแต่ละคลาส โดยการสุ่มเลือกแสดงดังตารางที่ 3.17

ตารางที่ 3.17 แสดงข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สองโดยการสุ่มเลือกเพื่อนำไปทดสอบระบบ

ลำดับ	แอตทริบิวต์ 1	แอตทริบิวต์ 2	แอตทริบิวต์ 3	แอตทริบิวต์ 4	คลาส
1	0.1936	0.5882	0.0556	0.1936	คลาส 1
2	0.5499	0.3309	0.3733	0.5499	คลาส 2
3	0.4311	0.3676	0.7387	0.4311	คลาส 3

จากตารางที่ 3.17 ระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง แสดงดังตารางที่ 3.18

ตารางที่ 3.18 แสดงระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สองและกลุ่มของข้อมูล

ลำดับ	ระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	0.0679	0.6441	0.9455	1
2	0.7038	0.1991	0.3654	2
3	1.2261	0.6397	0.3668	3

จากตารางที่ 3.18 การทดสอบอัลกอริทึมที่สอง แสดงได้ดังตารางที่ 3.19 ดังนี้

ตารางที่ 3.19 แสดงผลการทดสอบอัลกอริทึมที่สอง

ข้อมูลทดสอบระบบ		ข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สอง		ผลการทำนายกลุ่ม		
ลำดับข้อมูล	คลาส	ลำดับกลุ่ม	คลาส	ลำดับกลุ่ม	คลาส	ถูกต้อง
1	1	1	1	1	1	1
2	2	2	2	2	2	1
3	3	3	3	3	3	1

จากตารางที่ 3.19 คำนวณค่าความถูกต้องโดยใช้สมการที่ 3.5 ดังนี้

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่จำแนกกลุ่มได้ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$\text{แทนค่า} = \frac{3}{3}$$

ค่าความถูกต้อง คือ 1

3.3 ตัวอย่างการทดสอบของอัลกอริทึมที่สาม

นำข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สามจากตารางที่ 3.14 และข้อมูลทดสอบระบบ จากตารางที่ 3.2 มาคำนวณหาระยะทาง แสดงดังตารางที่ 3.20 ดังนี้

ตารางที่ 3.20 แสดงระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สามและกลุ่มของข้อมูล

ลำดับ	ระยะทางระหว่างข้อมูลทดสอบระบบกับข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สาม			จัดอยู่ในกลุ่ม
	กลุ่ม 1	กลุ่ม 2	กลุ่ม 3	
1	0.0340	0.5943	0.9203	1
2	0.6669	0.0926	0.3650	2
3	1.1929	0.7336	0.3143	3

จากตารางที่ 3.20 การทดสอบอัลกอริทึมที่สาม แสดงได้ดังตารางที่ 3.21 ดังนี้

ตารางที่ 3.21 แสดงผลการทดสอบอัลกอริทึมที่สาม

ข้อมูลทดสอบระบบ		ข้อมูลที่ผ่านการเรียนรู้จากอัลกอริทึมที่สาม		ผลการทำนายกลุ่ม		
ลำดับข้อมูล	คลาส	ลำดับกลุ่ม	คลาส	ลำดับกลุ่ม	คลาส	ถูกต้อง
1	1	1	1	1	1	1
2	2	2	2	2	2	1
3	3	3	3	3	3	1

จากตารางที่ 3.21 คำนวณค่าความถูกต้องโดยใช้สมการที่ 3.5 ดังนี้

$$\text{ค่าความถูกต้อง} = \frac{\text{จำนวนข้อมูลที่จำแนกกลุ่มได้ถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมด}}$$

$$\text{แทนค่า} = \frac{3}{3}$$

ค่าความถูกต้อง คือ 1

บทที่ 4

การทดลองและผลการทดลอง

4.1 การทดลอง

การทดลองจะกล่าวถึงรายละเอียดชุดข้อมูลที่ใช้สำหรับการทดลอง วิธีการวัดประสิทธิภาพของการทดลองและการออกแบบผลการทดลอง ดังนี้

4.1.1 ชุดข้อมูลที่ใช้สำหรับการทดลอง

ชุดข้อมูลที่ใช้สำหรับการทดลองมีทั้งหมด 11 ชุดข้อมูล ซึ่งเป็นชุดข้อมูลมาตรฐานจาก UCI Machine Learning [6] จำนวน 7 ชุดข้อมูลและอีก 4 ชุดข้อมูลที่สร้างขึ้น [7] โดยมีรายละเอียดของข้อมูลดังนี้

4.1.1.1 ชุดข้อมูลมาตรฐาน

1. Iris data

ชุดข้อมูลดอกไอริส คือข้อมูลของดอกไอริส 3 สายพันธุ์ได้แก่ Iris-setosa Iris-versicolor และ Iris-virginica จำนวน 150 เรคคอต ประกอบด้วย 4 แอตทริบิวต์ ได้แก่ ความยาวของดอก ความกว้างของดอก ความกว้างของกลีบเลี้ยงและความยาวของกลีบเลี้ยง โดยการทดลองแบ่งข้อมูลแป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 120 เรคคอต และข้อมูลทดสอบระบบ จำนวน 30 เรคคอต

2. Wine recognition data

ชุดข้อมูลไวน์ คือข้อมูลการวิเคราะห์ทางเคมีของไวน์ที่เติบโตในพื้นที่เดียวกันจากไวน์ 3 สายพันธุ์ จำนวน 178 เรคคอต โดยการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 90 เรคคอต และข้อมูลทดสอบระบบ จำนวน 88 เรคคอต

3. Haber man's survival data

ชุดข้อมูล Haber man's survival คือชุดข้อมูลที่เก็บรวบรวมจากกรณีศึกษาจำนวน 306 กรณีระหว่างปี ค.ศ. 1958 ถึง 1970 ที่โรงพยาบาลบิลลิลส์ของมหาวิทยาลัยชิคาโก ซึ่งศึกษาการอยู่รอดของผู้ป่วยหลังจากได้รับการผ่าตัดมะเร็งเต้านม จำนวน 333 เรคคอต โดยการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 214 เรคคอต และข้อมูลทดสอบระบบ จำนวน 92 เรคคอต

4. Heart disease problem

ชุดข้อมูลปัญหาโรคหัวใจ คือข้อมูลที่รวบรวมปัจจัยที่เกี่ยวข้องกับการเกิดโรคหัวใจ โดยแพทย์ได้นำมาทดสอบกับผู้ป่วยที่มีแนวโน้มเป็นโรคหัวใจ ซึ่งชุดข้อมูลมีจำนวน 270 เรคคอต โดยการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 135 เรคคอต และข้อมูลทดสอบระบบ จำนวน 135 เรคคอต

5. Balance Scale Weight and Distance data

ชุดข้อมูลการวัดความสมดุลของน้ำหนักและระยะทาง เป็นข้อมูลที่สร้างมาเพื่อการทดลองทางจิตวิทยาโดยแต่ละตัวอย่างจะเลื่อนสมดุลไปทางซ้ายหรือทางขวา หรือสมดุล ซึ่งข้อมูลเป็นตัวเลขจำนวน 4 แอตทริบิวต์ ได้แก่ น้ำหนักทางซ้าย ระยะทางซ้าย น้ำหนักทางขวา และระยะทางขวา จำนวน 625 เรคคอต โดยการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 313 เรคคอต และข้อมูลทดสอบระบบ จำนวน 312 เรคคอต

6. Vehicle Silhouettes data

ชุดข้อมูลภาพเงาของพาหนะ คือชุดข้อมูลที่เก็บรวบรวมภาพเงาของพาหนะเพื่อใช้สำหรับการทำนายประเภทของพาหนะ 4 ประเภท ได้แก่ รถบัสdouble-decker รถตู้เชฟโรเลต Saab 9000 และ Opel Manta 400 โดยข้อมูลมีทั้งหมด 18 แอตทริบิวต์ จำนวน 846 เรคคอต ซึ่งการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 424 เรคคอต และข้อมูลทดสอบระบบ จำนวน 422 เรคคอต

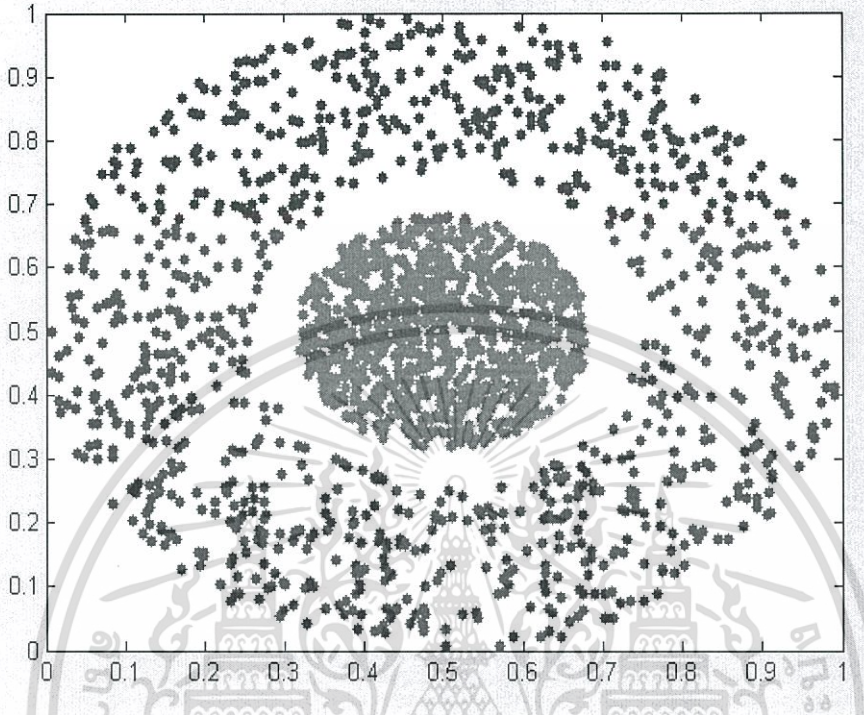
7. Pima Indians diabetes data

ชุดข้อมูลโรคเบาหวาน คือชุดข้อมูลที่ใช้สำหรับการทำนายโอกาสการเป็นโรคเบาหวานตามเกณฑ์มาตรฐานขององค์การอนามัยโลก โดยข้อมูลมีทั้งหมด 8 แอตทริบิวต์ จำนวน 768 เรคคอต โดยการทดลองแบ่งข้อมูลเป็น 2 ส่วน ดังนี้ ข้อมูลฝึกสอนระบบ จำนวน 384 เรคคอต และข้อมูลทดสอบระบบ จำนวน 384 เรคคอต

4.1.1.2 ชุดข้อมูลที่สร้างขึ้น

1. Donut

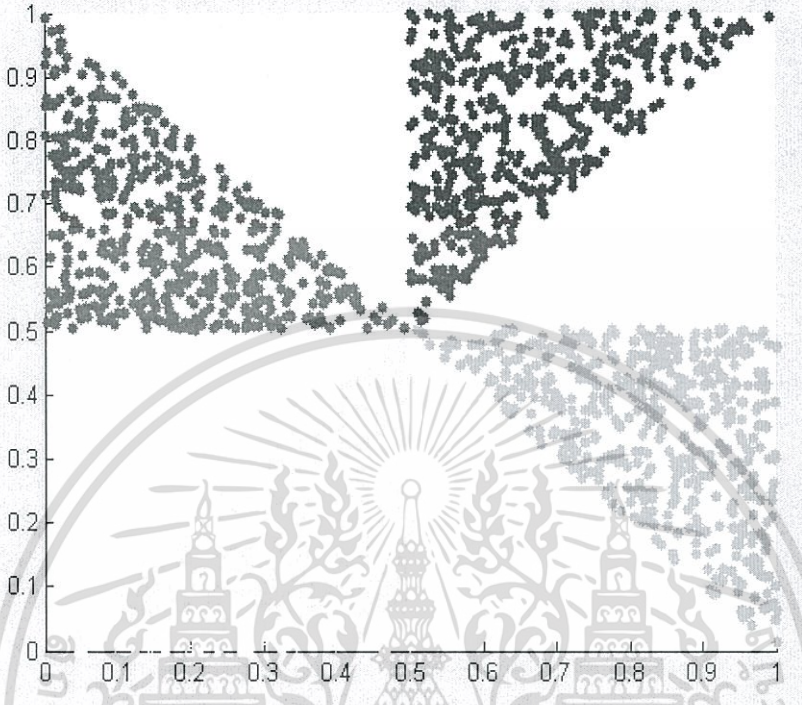
ชุดข้อมูลโดนัท คือข้อมูลสองมิติมีลักษณะคล้ายรูปโดนัท ดังรูปที่ 4.1 โดยแบ่งเป็นข้อมูลฝึกสอนระบบ จำนวน 2000 เรคคอต และข้อมูลทดสอบระบบ จำนวน 2000 เรคคอต มีจำนวนกลุ่ม 2 กลุ่ม



รูปที่ 4.1 แสดงชุดข้อมูลรูปโดนัท

2. Fan

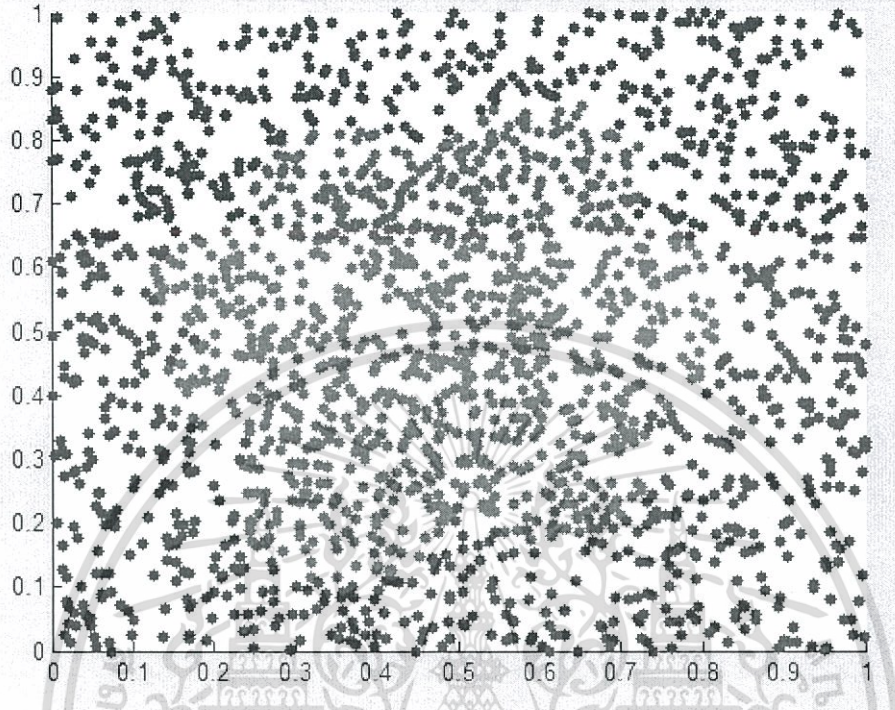
ชุดข้อมูลรูปใบพัดลม คือข้อมูลสองมิติมีลักษณะคล้ายรูปใบพัดลม ดังรูปที่ 4.2 โดยแบ่งเป็นข้อมูลฝึกสอนระบบ จำนวน 2000 เรคคอต และข้อมูลทดสอบระบบ จำนวน 2000 เรคคอต มีจำนวนกลุ่ม 4 กลุ่ม



รูปที่ 4.2 แสดงชุดข้อมูลรูปใบพัดลม

3. Flower 1

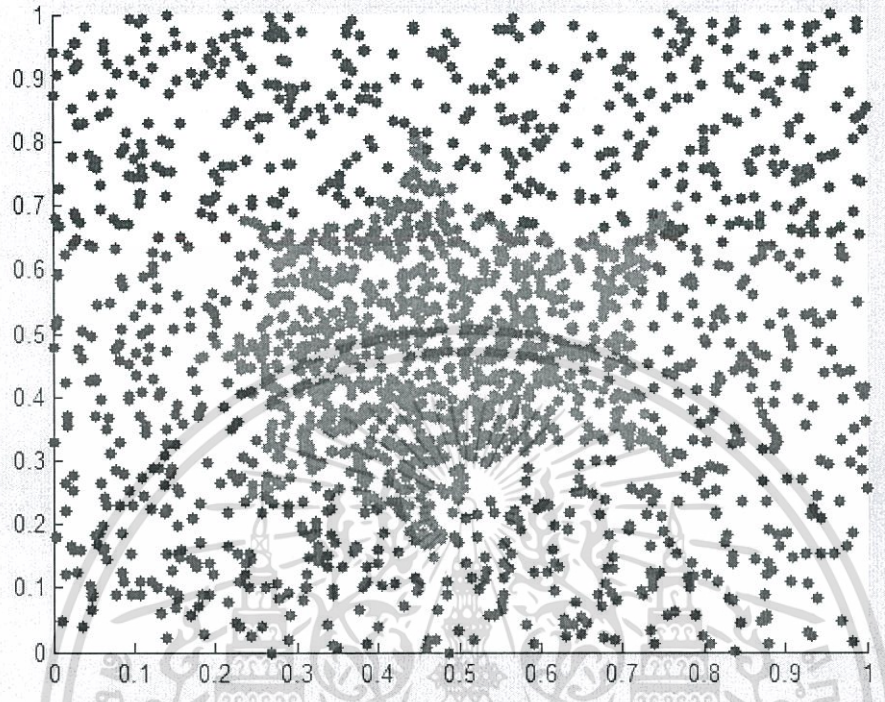
ชุดข้อมูลรูปดอกไม้แบบที่หนึ่ง คือข้อมูลสองมิติมีลักษณะคล้ายรูปดอกไม้แบบที่หนึ่ง ดังรูปที่ 4.3 โดยแบ่งเป็นข้อมูลฝึกสอนระบบ จำนวน 2000 เรคคอต และข้อมูลทดสอบระบบ จำนวน 2000 เรคคอต มีจำนวนคลุ่ม 2 กลุ่ม



รูปที่ 4.3 แสดงชุดข้อมูลรูปดอกไม้แบบที่หนึ่ง

4. Flower2

ชุดข้อมูลรูปดอกไม้แบบที่สอง คือข้อมูลสองมิติมีลักษณะคล้ายรูปดอกไม้แบบที่สอง ดังรูปที่ 4.4 โดยแบ่งเป็นข้อมูลฝึกสอนระบบ จำนวน 2000 เรคคอต และข้อมูลทดสอบระบบ จำนวน 2000 เรคคอต มีจำนวนกลุ่ม 2 กลุ่ม



รูปที่ 4.4 แสดงชุดข้อมูลรูปดอกไม้แบบที่สอง

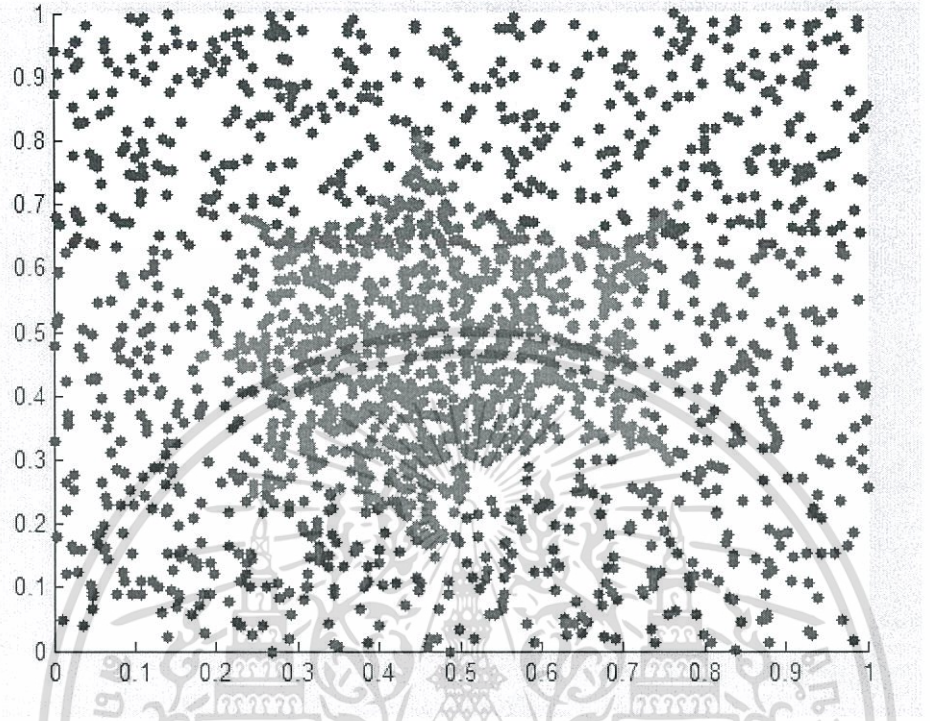
4.1.2 วิธีการวัดประสิทธิภาพของการทดลอง

วิธีการวัดประสิทธิภาพของการทดลองของการจำแนกประเภท ใช้วิธีการวัดค่าความถูกต้อง เวลาและ10 Fold cross validation

1. ความถูกต้องของการจำแนกประเภทข้อมูล คือค่าอัตราส่วนของจำนวนข้อมูลที่จำแนกประเภทได้ถูกต้องกับจำนวนข้อมูลทั้งหมดที่นำมาทดสอบ ค่าความถูกต้องของการจำแนกประเภท แสดงดังสมการที่ 4.1

$$\text{Accuracy} = \frac{TC}{TN} \tag{4.1}$$

เมื่อ	Accuracy	คือค่าความถูกต้องของการจำแนกประเภทข้อมูล
	TC	คือจำนวนข้อมูลที่จำแนกประเภทได้ถูกต้อง
	TN	คือจำนวนข้อมูลทั้งหมด



รูปที่ 4.4 แสดงชุดข้อมูลรูปดอกไม้แบบที่สอง

4.1.2 วิธีการวัดประสิทธิภาพของการทดลอง

วิธีการวัดประสิทธิภาพของการทดลองของการจำแนกประเภท ใช้วิธีการวัดค่าความถูกต้อง เวลาและ10 Fold cross validation

1. ความถูกต้องของการจำแนกประเภทข้อมูล คือค่าอัตราส่วนของจำนวนข้อมูลที่จำแนกประเภทได้ถูกต้องกับจำนวนข้อมูลทั้งหมดที่นำมาทดสอบ ค่าความถูกต้องของการจำแนกประเภท แสดงดังสมการที่ 4.1

$$\text{Accuracy} = \frac{\text{TC}}{\text{TN}} \tag{4.1}$$

เมื่อ Accuracy คือค่าความถูกต้องของการจำแนกประเภทข้อมูล

TC คือจำนวนข้อมูลที่จำแนกประเภทได้ถูกต้อง

TN คือจำนวนข้อมูลทั้งหมด

2. 10 Fold Cross Validation คือวิธีการวัดประสิทธิภาพของความถูกต้อง โดยแบ่งข้อมูลออกเป็น 10 ส่วน จากนั้นทำการทดลองโดย นำข้อมูล 9 ส่วนแรกไปฝึกสอนระบบและที่เหลือ 1 ส่วนนำมาทดสอบระบบ ทำซ้ำโดยเปลี่ยนจากข้อมูลส่วนที่ 2, 3...k ไปทดสอบระบบและที่เหลือ k-1ไปฝึกสอนระบบและคำนวณหาค่าเฉลี่ยของค่าความถูกต้อง

4.1.3 การออกแบบการทดลอง

การทดลองทำบนเครื่องคอมพิวเตอร์โปรเซสเซอร์ความเร็ว 2.53 กิกะเฮิร์ต หน่วยความจำ 4 กิกะไบต์ ระบบปฏิบัติการ Windows 7 Professional และใช้โปรแกรม MATLAB เวอร์ชั่น R2010a

รายละเอียดของข้อมูลที่ใช้สำหรับการทดลองได้แก่ จำนวนแอตทริบิวต์ จำนวนคลาส จำนวนข้อมูลสำหรับการฝึกสอนระบบและจำนวนข้อมูลทดสอบระบบ แสดงดังตารางที่ 4.1 ดังนี้

ตารางที่ 4.1 แสดงรายละเอียดของชุดข้อมูลสำหรับการทดลอง

ลำดับ	ชุดข้อมูล	แอตทริบิวต์	จำนวนคลาส	ข้อมูลฝึกสอนระบบ	ข้อมูลทดสอบระบบ
	ชุดข้อมูลมาตรฐาน				
1	Balance scale	4	3	313	312
2	Haber man's survival	3	2	214	92
3	Heart disease	13	2	135	135
4	Iris	4	3	120	30
5	Pima Indians diabetes	8	2	384	384
6	Vehicle silhouettes	18	4	424	422
7	Wine recognition	13	3	90	88
	ข้อมูลที่สร้างขึ้น				
1	Donut	2	2	2000	2000
2	Fan	2	4	2000	2000
3	Flower1	2	2	2000	2000
4	Flower2	2	2	2000	2000

จากตารางที่ 4.1 แสดงรายละเอียดของชุดข้อมูลสำหรับการทดลอง โดยข้อมูลที่ใช้สำหรับการทดลองจะทำการแปลงข้อมูลให้มีค่าอยู่ระหว่างช่วง 0 -1 ซึ่งวิธีการแปลงข้อมูลแสดงรายละเอียดในหัวข้อ 3.3.1 จากนั้นนำชุดข้อมูลฝึกสอนระบบไปเรียนรู้ ซึ่งวิธีการเรียนรู้ในแต่ละอัลกอริทึมแสดงรายละเอียดในหัวข้อ 3.3 จากนั้นข้อมูลที่ผ่านการเรียนรู้ จะเรียกว่าแบบจำลอง ซึ่งจะนำแบบจำลองดังกล่าวไปทดสอบกับข้อมูลทดสอบระบบในแต่ละชุดข้อมูล โดยรายละเอียดการทดสอบระบบ แสดงในหัวข้อที่ 3.3.3 จากนั้นจึงวัดค่าความถูกต้อง

4.2 ผลการทดลอง

4.2.1 ผลการทดลองของอัลกอริทึมที่นำเสนอ

4.2.1.1 ผลการทดลองของอัลกอริทึมที่หนึ่ง

ทำการทดลองจำนวน 10 ครั้ง วัดประสิทธิภาพความถูกต้อง เวลาและ 10 Fold Cross Validation มีผลการทดลองดังนี้

1. ชุดข้อมูลมาตรฐาน

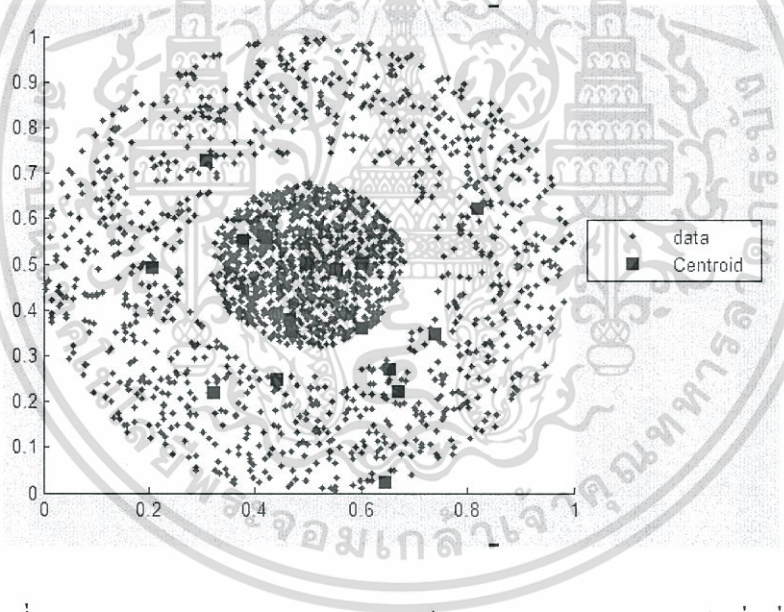
ตารางที่ 4.2 แสดงผลการทดลองของอัลกอริทึมที่หนึ่งโดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง(เฉลี่ย)	เวลา (เฉลี่ย) วินาที	10 Fold Cross Validation
1	Balance scale weight and distance	0.8500	0.547	0.8654
2	Haber man's survival	0.8345	0.6564	0.8525
3	Heart disease	0.7800	0.4368	0.7926
4	Iris	0.9800	0.2742	0.9800
5	Pima Indians diabetes	0.7586	1.7212	0.8229
6	Vehicle silhouettes	0.7806	6.2504	0.7891
7	Wine recognition	0.9167	2.165	0.9167

2. ชุดข้อมูลที่สร้างขึ้น

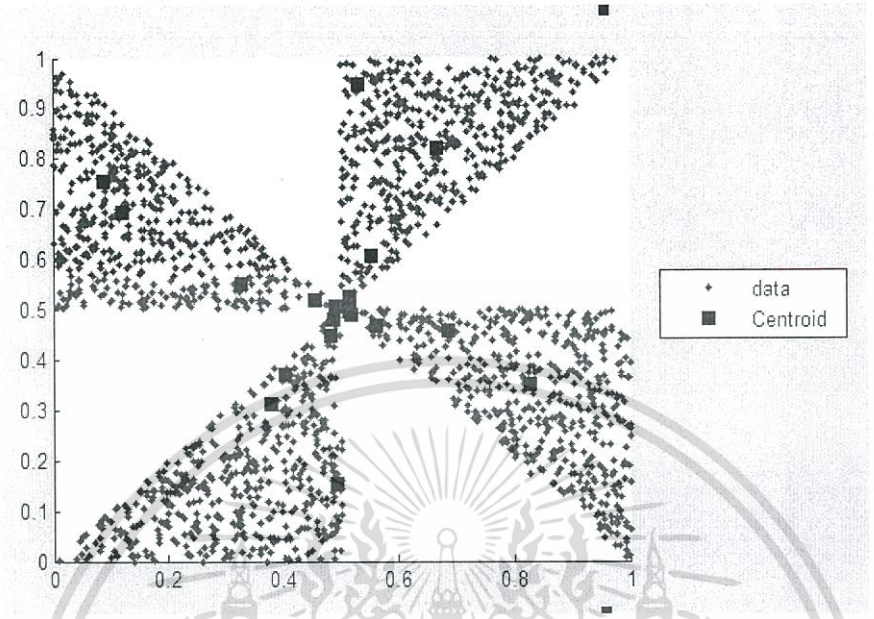
ตารางที่ 4.3 แสดงผลการทดลองของอัลกอริทึมที่หนึ่งโดยใช้ชุดข้อมูลที่สร้างขึ้น

ลำดับ	ข้อมูล	ค่าความถูกต้อง (เฉลี่ย)	เวลา (เฉลี่ย) วินาที	10 Fold Cross Validation
1	Donut	0.9975	1.0668	1
2	Fan	0.9995	3.5474	1
3	Flower1	0.9995	3.8364	0.9880
4	Flower2	0.9890	3.5268	1



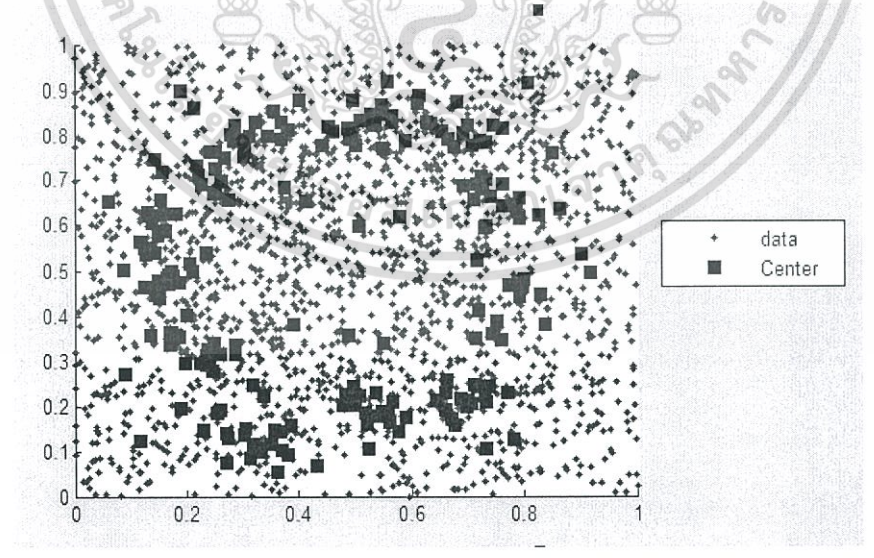
รูปที่ 4.5 แสดงผลการจำแนกประเภทของข้อมูลรูปโดนัทจากอัลกอริทึมที่หนึ่ง

จากรูปที่ 4.5 แสดงผลการจำแนกประเภทข้อมูลรูปโดนัทจากอัลกอริทึมที่หนึ่ง ให้ค่าความถูกต้อง 0.9975 และจำนวน 18 กลุ่ม



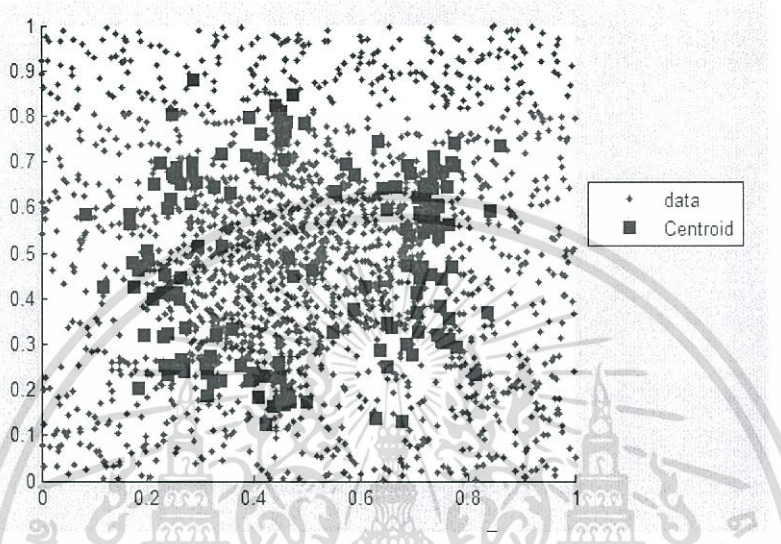
รูปที่ 4.6 แสดงผลการจำแนกประเภทของข้อมูลรูปใบพัดจากอัลกอริทึมที่หนึ่ง

จากรูปที่ 4.6 แสดงผลการจำแนกประเภทข้อมูลรูปใบพัดจากอัลกอริทึมที่หนึ่ง ให้ค่าความถูกต้อง 0.9995 และจำนวน 19 กลุ่ม



รูปที่ 4.7 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่หนึ่ง

จากรูปที่ 4.7 แสดงผลการจำแนกประเภทข้อมูลรูปดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่หนึ่ง ให้ค่าความถูกต้อง 0.9865 และจำนวน 234 กลุ่ม



รูปที่ 4.8 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่สองจากอัลกอริทึมที่หนึ่ง

จากรูปที่ 4.8 แสดงผลการจำแนกประเภทข้อมูลรูปดอกไม้แบบที่สองจากอัลกอริทึมที่หนึ่ง ให้ค่าความถูกต้อง 0.9890 และจำนวน 184 กลุ่ม

4.2.1.2 ผลการทดลองของอัลกอริทึมที่สอง

1. ชุดข้อมูลมาตรฐาน

ตารางที่ 4.4 แสดงผลการทดลองของอัลกอริทึมที่สองโดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Balance scale weight and distance	0.8558	0.2735	0.8600
2	Haber man's survival	0.9016	0.3282	0.9010
3	Heart disease	0.8222	0.2184	0.8222
4	Iris	1	0.1371	0.9670

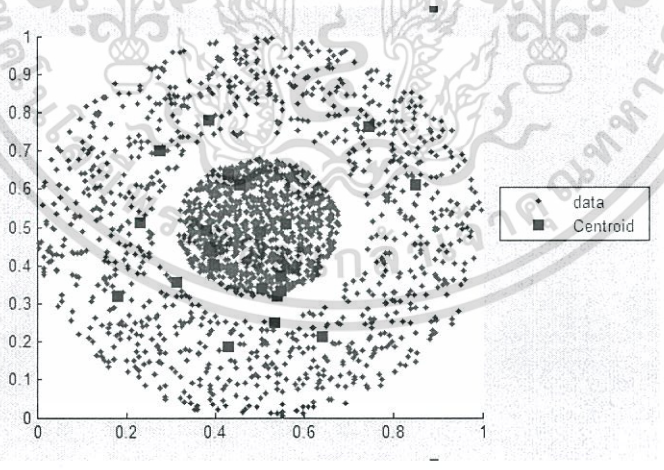
ตารางที่ 4.4(ต่อ) แสดงผลการทดลองของอัลกอริทึมที่สองโดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
5	Pima Indians diabetes	0.7734	0.8606	0.8233
6	Vehicle silhouettes	0.7820	3.1252	0.8270
7	Wine recognition	0.9722	1.0825	0.9700

2. ชุดข้อมูลที่สร้างขึ้น

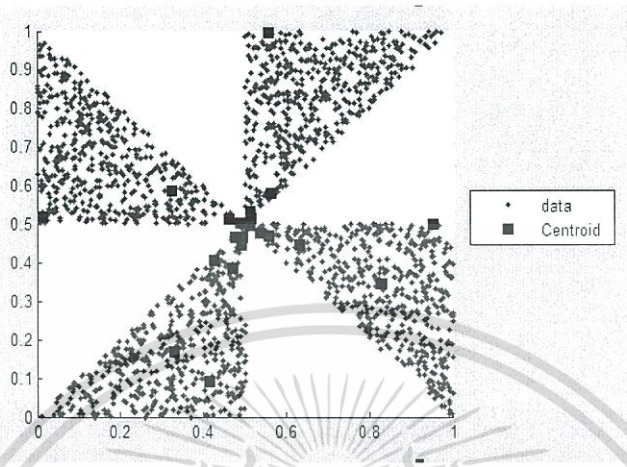
ตารางที่ 4.5 แสดงผลการทดลองของอัลกอริทึมที่สองโดยใช้ชุดข้อมูลที่สร้างขึ้น

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Donut	0.9980	0.5334	1
2	Fan	1	1.7737	1
3	Flower1	0.9850	1.9157	0.9940
4	Flower 2	0.9980	1.7609	1



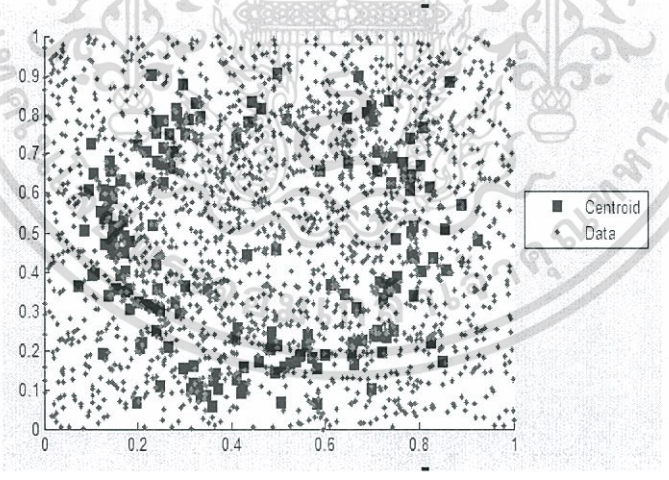
รูปที่ 4.9 แสดงผลการจำแนกประเภทของข้อมูลรปโดนัทจากอัลกอริทึมที่สอง

จากรูปที่ 4.9 แสดงผลการจำแนกประเภทข้อมูลรูปโดนัทจากอัลกอริทึมที่สอง ให้ค่าความถูกต้อง 0.9980 และจำนวน 22 กลุ่ม



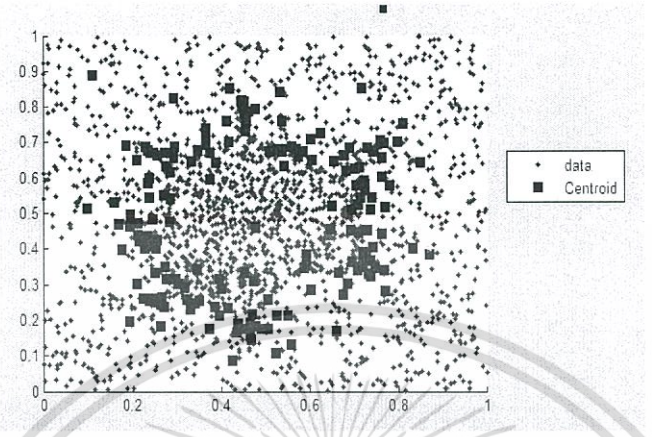
รูปที่ 4.10 แสดงผลการจำแนกประเภทของข้อมูลรปใบพัดจากอัลกอริทึมที่สอง

จากรูปที่ 4.10 แสดงผลการจำแนกประเภทข้อมูลรูปใบพัดจากอัลกอริทึมที่สอง ให้ค่าความถูกต้อง 1 และจำนวน 23 กลุ่ม



รูปที่ 4.11 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่สอง

จากรูปที่ 4.11 แสดงผลการจำแนกประเภทข้อมูลรูปดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่สอง ให้ค่าความถูกต้อง 0.9850 และจำนวน 190 กลุ่ม



รูปที่ 4.12 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่สองจากอัลกอริทึมที่สอง

จากรูปที่ 4.12 แสดงผลการจำแนกประเภทข้อมูลรูปดอกไม้แบบที่สองจากอัลกอริทึมที่สอง ให้ค่าความถูกต้อง 0.9980 และจำนวน 194 กลุ่ม

4.2.1.3 ผลการทดลองของอัลกอริทึมที่สาม

1. ชุดข้อมูลมาตรฐาน

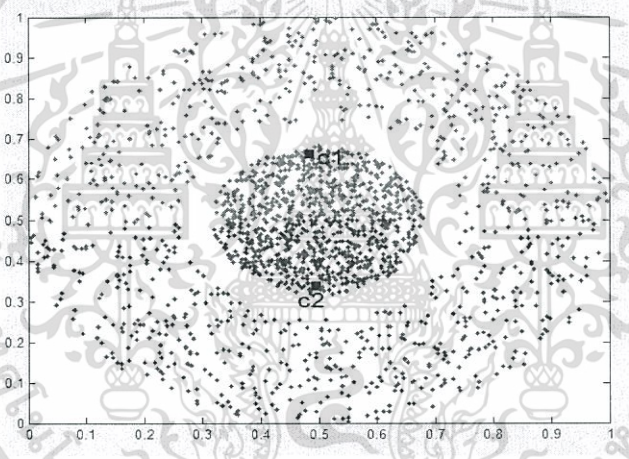
ตารางที่ 4.6 แสดงผลการทดลองของอัลกอริทึมที่สามโดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Balance scale weight and distance	0.7051	0.0413	0.7220
2	Haber man's survival	0.6393	0.0567	0.6390
3	Heart disease	0.7778	0.1722	0.7779
4	Iris	0.9667	0.1478	0.9696
5	Pima Indians diabetes	0.6693	0.4449	0.6700
6	Vehicle silhouettes	0.391	1.9224	0.4000
7	Wine recognition	0.75	0.0413	0.7520

2. ชุดข้อมูลที่สร้างขึ้น

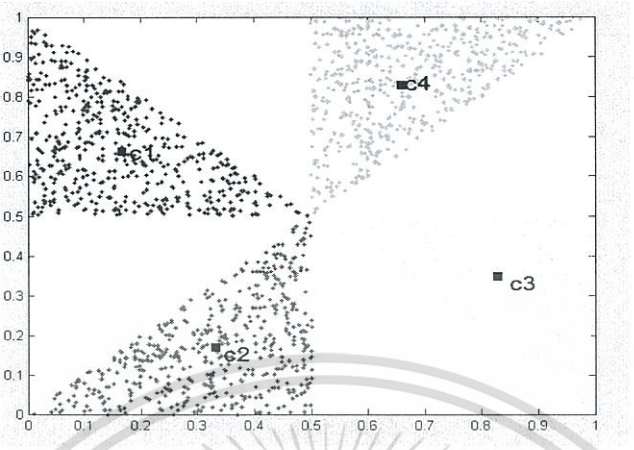
ตารางที่ 4.7 แสดงผลการทดลองของอัลกอริทึมที่สามโดยใช้ชุดข้อมูลที่สร้างขึ้น

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Donut	0.5045	0.6148	0.5040
2	Fan	1	1.8163	1
3	Flower1	0.5545	0.6828	0.5640
4	Flower 2	0.5585	0.5935	0.5583



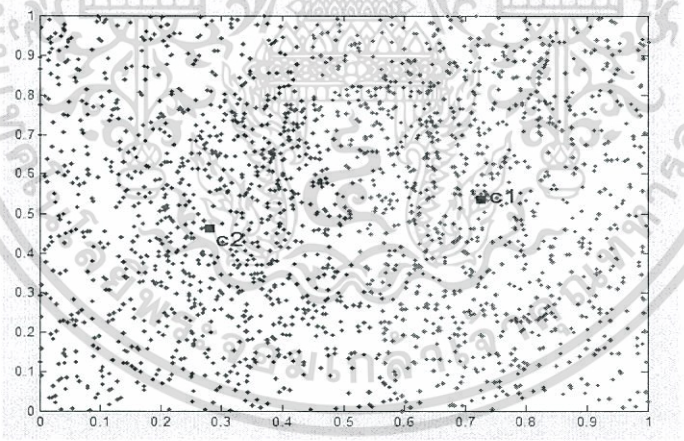
รูปที่ 4.13 แสดงผลการจำแนกประเภทของข้อมูลรูปโดนัทจากอัลกอริทึมที่สาม

จากรูปที่ 4.13 แสดงผลการจำแนกประเภทข้อมูลรูปโดนัทจากอัลกอริทึมที่สาม ให้ค่าความถูกต้อง 0.5045 และจำนวน 2 กลุ่ม



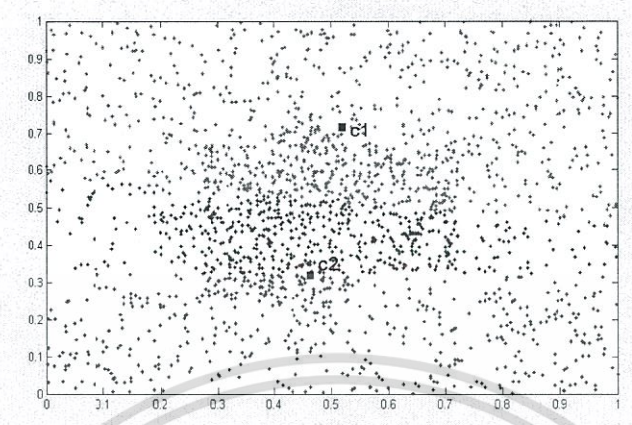
รูปที่ 4.14 แสดงผลการจำแนกประเภทของข้อมูลรูปใบพัดจากอัลกอริทึมที่สาม

จากรูปที่ 4.14 แสดงผลการจำแนกประเภทข้อมูลรูปใบพัดจากอัลกอริทึมที่สาม ให้ค่าความถูกต้อง 1 และจำนวน 4 กลุ่ม



รูปที่ 4.15 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่สาม

จากรูปที่ 4.15 แสดงผลการจำแนกประเภทข้อมูลดอกไม้แบบที่หนึ่งจากอัลกอริทึมที่สาม ให้ค่าความถูกต้อง 0.5545 และจำนวน 2 กลุ่ม



รูปที่ 4.16 แสดงผลการจำแนกประเภทของข้อมูลรูปดอกไม้แบบที่สองจากอัลกอริทึมที่สาม

จากรูปที่ 4.16 แสดงผลการจำแนกประเภทข้อมูลดอกไม้แบบที่สองจากอัลกอริทึมที่สาม ให้ค่าความถูกต้อง 0.5585 และจำนวน 2 กลุ่ม

4.2.2 ผลการทดลองของอัลกอริทึมที่นำมาเปรียบเทียบ

4.2.2.1 ผลการทดลองของอัลกอริทึม C4.5

1. ชุดข้อมูลมาตรฐาน

ตารางที่ 4.8 แสดงผลการทดลองของอัลกอริทึม C4.5 โดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Balance scale weight and distance	0.8237	0.13675	0.8250
2	Haber man's survival	0.6393	0.1641	0.6667
3	Heart disease	0.8296	0.1092	0.8500
4	Iris	0.9667	0.06855	0.9667
5	Pima Indians diabetes	0.6510	0.4303	0.6650
6	Vehicle silhouettes	0.8436	1.5626	0.8440
7	Wine recognition	0.8611	0.54125	0.8686

2. ชุดข้อมูลที่สร้างขึ้น

ตารางที่ 4.9 แสดงผลการทดลองของอัลกอริทึม C4.5 โดยใช้ชุดข้อมูลที่สร้างขึ้น

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Donut	0.8510	0.1334	0.8680
2	Fan	0.8040	0.4434	0.8100
3	Flower1	0.8420	0.4789	0.8500
4	Flower2	0.8455	0.4402	0.8100

4.2.2.2 ผลการทดลองของอัลกอริทึม LVQ

1. ชุดข้อมูลมาตรฐาน

ตารางที่ 4.10 ตารางแสดงผลการทดลองของอัลกอริทึม LVQ โดยใช้ชุดข้อมูลมาตรฐาน

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Balance scale weight and distance	0.8590	0.1413	0.8600
2	Haber man's survival	0.7619	0.0467	0.7800
3	Heart disease	0.8444	0.1622	0.8450
4	Iris	1	0.1470	1
5	Pima Indians diabetes	0.7995	0.500	0.8000
6	Vehicle silhouettes	0.6850	1.5224	0.6800
7	Wine recognition	0.9545	0.0419	0.9555

2. ชุดข้อมูลที่สร้างขึ้น

ตารางที่ 4.11 ตารางแสดงผลการทดลองของอัลกอริทึม LVQ โดยใช้ชุดข้อมูลที่สร้างขึ้น

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง	เวลา (วินาที)	10 Fold Cross Validation
1	Donut	0.8778	0.500	0.8870
2	Fan	1	1.6737	1
3	Flower1	0.6120	1.0157	0.6210
4	Flower2	0.6035	1.7600	0.7100

ตารางที่ 4.12 ตารางแสดงผลการทดลองของอัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบของข้อมูลทั้งหมด

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง (%)				
	ชุดข้อมูลมาตรฐาน	อัลกอริทึมที่นำเสนอ			อัลกอริทึมที่นำมาเปรียบเทียบ	
		อัลกอริทึมที่หนึ่ง	อัลกอริทึมที่สอง	อัลกอริทึมที่สาม	อัลกอริทึม C4.5	อัลกอริทึม LVQ
1	Balance scale weight and distance	85.00	85.58	70.51	82.37	85.90
2	Haber man's survival	83.45	90.16	63.93	63.93	76.19
3	Heart disease	78.00	82.22	77.78	82.96	84.44
4	Iris	98.00	100	96.67	96.67	100
5	Pima Indians diabetes	75.86	77.34	66.93	65.10	79.95
6	Vehicle silhouettes	78.06	78.2	39.10	84.36	68.50
7	Wine recognition	91.67	97.22	75.00	86.11	95.45

ตารางที่ 4.12(ต่อ) ตารางแสดงผลการทดลองของอัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบของข้อมูลทั้งหมด

ลำดับ	ชุดข้อมูล	ค่าความถูกต้อง (%)				
		อัลกอริทึมที่นำเสนอ			อัลกอริทึมที่นำมาเปรียบเทียบ	
	ข้อมูลที่สร้างขึ้น	อัลกอริทึมที่หนึ่ง	อัลกอริทึมที่สอง	อัลกอริทึมที่สาม	อัลกอริทึม C4.5	อัลกอริทึม LVQ
1	Donut	99.69	99.8	50.45	85.1	87.78
2	Fan	99.96	100	100	80.4	100
3	Flower1	98.51	98.5	55.45	84.2	61.2
4	Flower2	99.97	99.8	55.85	84.55	60.35

4.3 สรุปผลการทดลอง

4.3.1 สรุปผลการทดลองของชุดข้อมาตรฐาน

1. ข้อมูล Balance scale weight and distance

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูลอัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุดคือ อัลกอริทึม LVQ 85.90 % รองลงมา คือ อัลกอริทึมที่สอง ของอัลกอริทึมที่เสนอ 85.58 % อันดับต่อมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่เสนอ 85.00% อันดับต่อมาคือ อัลกอริทึม C4.5 82.37 % สุดท้ายคือ อัลกอริทึมที่สาม 70.50%

2. ข้อมูล Haber man's survival

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูลอัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่าค่าความถูกต้องมีความแตกต่างกัน กล่าวคือ อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุดคือ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 90.16 % รองมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่เสนอ 83.45% อันดับต่อมาคือ อัลกอริทึม LVQ 83.45 % และสุดท้ายคือ อัลกอริทึม C4.5 และอัลกอริทึมที่สามของอัลกอริทึมของอัลกอริทึมที่นำเสนอให้ให้ค่าความถูกต้องเท่ากันคือ 63.93 %

3. ข้อมูล Heart disease

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึม LVQ 84.44 % รองลงมาคืออัลกอริทึม C4.5 82.96% อันดับต่อมาคือ อัลกอริทึมริทึมที่สองของอัลกอริทึมที่นำเสนอ 82.22 % อันดับต่อมาคือ อัลกอริทึมที่หนึ่งของอัลกอิทึมที่นำเสนอ 78.00 % สุดท้ายคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ 77.78%

4. ข้อมูล Iris

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกกลุ่มข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึม LVQ และ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ คือ 100 % รองลงมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 98.00 % และอันดับสุดท้าย คือ อัลกอริทึม C4.5 และอัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอให้ค่าความถูกต้องเท่ากันคือ 96.97 %

5. ข้อมูล Pima Indians diabetes

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึม LVQ 79.95 % รองลงมาคืออัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 77.34 % อันดับต่อมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 75.86 % อันดับต่อมาคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ 66.93% และอันดับสุดท้ายคือ อัลกอริทึม C4.5 65.10 %

6. ข้อมูล Vehicle silhouettes

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึม C4.5 84.36 % รองลงมา คือ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 78.2 % อันดับต่อมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่เสนอ 78.2 % อันดับต่อมาคือ อัลกอริทึม LVQ คือ 68.50 % และสุดท้ายคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ คือ 39.10%

7. ข้อมูล Wine recognition

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึมที่สองของอัลกอริทึมที่เสนอ 97.22 % รองลงมาคือ อัลกอริทึม LVQ 95.45 % อันดับ

ต่อมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 91.67 % อันดับต่อมาคือ อัลกอริทึม C4.5 86.11 % และอันดับสุดท้ายคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ คือ 75%

4.3.2 สรุปผลการทดลองของชุดข้อมูลที่สร้างขึ้น

1. ข้อมูล Donut

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 99.80 % รองลงมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 99.69 % อันดับต่อมาคือ อัลกอริทึม LVQ 87.78 % อันดับต่อมาคือ อัลกอริทึม C4.5 85.10 % และสุดท้ายคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ คือ 50.45%

2. ข้อมูล Fan

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึมที่สอง อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอและอัลกอริทึม LVQ คือ 100 % รองลงมาคือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 99.96 % และอันดับสุดท้ายคือ อัลกอริทึมC4.5 80.40 %

3. ข้อมูล Flower1

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 98.51 % รองลงมาคือ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 98.50 % อันดับต่อมาคือ อัลกอริทึม C4.5 84.20 % อันดับต่อมา อัลกอริทึม LVQ 61.20 %และอันดับสุดท้ายคือ อัลกอริทึมที่สามของอัลกอริทึมที่นำเสนอ 55.45%

4. ข้อมูล Flower2

จากการทดลองโดยวัดประสิทธิภาพความถูกต้อง ของการจำแนกประเภทข้อมูล อัลกอริทึมที่นำเสนอและอัลกอริทึมที่นำมาเปรียบเทียบ พบว่า อัลกอริทึมที่ให้ค่าความถูกต้องมากที่สุด คือ อัลกอริทึมที่หนึ่งของอัลกอริทึมที่นำเสนอ 99.97 % รองลงมาคือ อัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ 99.80 % อันดับต่อมาคือ อัลกอริทึม C4.5 84.55 % อันดับต่อมาคือ อัลกอริทึม LVQ 60.35 % และอันดับสุดท้ายคืออัลริทึมที่สามของอัลกอริทึมที่นำเสนอ 55.85 %

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการจำแนกประเภทข้อมูล (Classification) แบบใหม่ที่ใช้หลักการทำงานของอัลกอริทึมเคมีนและค่าเอนโทรปีเป็นโครงสร้างหลักในการสร้างแบบจำลองการจำแนกประเภท

หลักการทำงานของงานวิจัยที่นำเสนอ กล่าวคืองานวิจัยนี้ได้เสนออัลกอริทึมสำหรับการสร้างแบบจำลองการจำแนกประเภทข้อมูล 3 อัลกอริทึม ดังนี้

1. อัลกอริทึมที่หนึ่ง หลักการทำงานคือ เริ่มจากสุ่มเลือกจุดศูนย์กลางเริ่มต้น จำนวนเท่ากับจำนวนกลุ่มของชุดข้อมูลที่ใช้สำหรับการทดลอง แล้วนำจุดเริ่มต้นดังกล่าวไปวัดระยะทาง ระหว่างจุดศูนย์กลางเริ่มต้นกับข้อมูลทั้งหมดในชุดข้อมูล จากนั้นจัดกลุ่มโดยพิจารณาจากระยะทางใกล้ เมื่อได้กลุ่มเริ่มต้นแล้ว นำแต่ละกลุ่มที่ได้มาวัดค่าเอนโทรปี เพื่อทำการแบ่งกลุ่ม การแบ่งกลุ่มจะเกิดขึ้นก็ต่อเมื่อค่าเอนโทรปีของกลุ่มนั้นๆ มีค่ามากกว่าค่าที่กำหนดไว้ วิธีการแบ่งกลุ่มคือแบ่งเท่ากับจำนวนคลาสที่อยู่ในกลุ่มนั้นๆ โดยคลาสเหมือนกันจัดให้อยู่กลุ่มเดียวกัน กลุ่มที่แบ่งออกจะมีตัวแทนของกลุ่มใหม่คือค่าเฉลี่ยของสมาชิกในกลุ่มนั้นๆ จากนั้นนำตัวแทนกลุ่มทั้งหมดที่ได้มาคำนวณหาระยะทางระหว่างตัวแทนของกลุ่มกับข้อมูลทั้งหมด พร้อมทั้งวัดค่าเอนโทรปีของกลุ่มใหม่ ทำซ้ำขั้นตอนการแบ่งกลุ่ม และคำนวณค่าเอนโทรปี จนกระทั่งกลุ่มทั้งหมดมีค่าเอนโทรปีน้อยกว่าค่าเอนโทรปีที่กำหนดไว้ สุดท้ายนำตัวแทนของทุกกลุ่มไปทำนายกลุ่มกับข้อมูลทดสอบระบบ และคำนวณค่าความถูกต้อง

2. อัลกอริทึมที่สอง หลักการทำงานคือ เริ่มจากการเลือกจุดศูนย์กลางเริ่มต้น โดยมีขั้นตอนการเลือกจุดศูนย์กลางเริ่มต้นมีดังนี้ เริ่มทำทีละแอตทริบิวต์ของข้อมูล คือนำข้อมูลทั้งหมดในแอตทริบิวต์ที่พิจารณา มาแบ่งช่วงข้อมูลและหาความถี่ จากนั้นเรียงความถี่จากมากไปหาน้อย แล้วพิจารณา สามอันดับแรกที่มีความถี่สูงสุด จากนั้นนำช่วงข้อมูลที่ถูกเลือกมาหาค่าความเป็นสมาชิกของคลาสทุกคลาสในชุดข้อมูล แล้วจึงนำค่าความเป็นสมาชิกมาหาสัดส่วน โดยพิจารณาจากจากคลาสที่หนึ่งก่อนและคลาสอื่นๆที่เหลือตามลำดับ จะทำการเลือกตัวแทนของคลาสจากค่าสัดส่วนที่มาก ทำซ้ำขั้นตอน การหาความถี่ หาค่าสัดส่วนเพื่อหาตัวแทนของแต่ละคลาสในแต่ละแอตทรบิวต์จนครบทุกแอตทริบิวต์ของชุดข้อมูลจะได้จุดศูนย์กลางเริ่มต้น จากนั้นนำจุดศูนย์กลางเริ่มต้นที่ได้ไปหาระยะทาง ระหว่างจุดศูนย์กลางเริ่มต้นกับข้อมูลทั้งหมดในชุดข้อมูล เพื่อจัดกลุ่มซึ่งจัดกลุ่มโดยพิจารณาจากระยะที่ใกล้

จากนั้นนำกลุ่มที่ได้มาหาคำนวณหาค่าเอนโทรปีเพื่อแบ่งกลุ่ม ซึ่งจะแบ่งกลุ่มก็ต่อเมื่อค่าเอนโทรปีของกลุ่มนั้นๆมีค่ามากกว่าค่าที่กำหนดไว้ โดยจะแบ่งเท่ากับคลาสในกลุ่มนั้นซึ่งตัวแทนของกลุ่มใหม่เกิดจากค่าเฉลี่ยของสมาชิกกลุ่มใหม่นั้นๆ จากนั้นทำการวัดระยะทางระหว่าง ตัวแทนของทุกกลุ่มเพื่อจุดศูนย์กลางของกลุ่มใหม่และคำนวณหาค่าเอนโทรปีของกลุ่ม ทำซ้ำขั้นตอนการจัดกลุ่มและแบ่งกลุ่มใหม่จนกระทั่งค่าเอนโทรปีของทุกกลุ่มมีค่าน้อยกว่าหรือเท่ากับค่าที่กำหนดไว้ ขั้นตอนสุดท้ายคือนำจุดศูนย์กลางของทุกกลุ่มไปทำนายกลุ่มกับข้อมูลทดสอบระบบและคำนวณค่าความถูกต้อง

3.อัลกอริทึมที่สาม หลักการทำงานคือ เริ่มจากการเลือกจุดเริ่มต้นโดยใช้วิธีการเลือกจุดเริ่มต้นที่นำเสนอ ซึ่งเป็นวิธีเดียวกับการเลือกจุดเริ่มต้นของอัลกอริทึมที่สอง จากนั้นนำจุดศูนย์กลางที่ได้ไปจัดกลุ่มโดยใช้อัลกอริทึมเคมีน ซึ่งจะทำการ จัดกลุ่มจนกระทั่งจุดศูนย์กลางไม่เปลี่ยน สุดท้ายนำจุดศูนย์กลางนั้นไปทำนายกลุ่มกับข้อมูลทดสอบระบบและคำนวณค่าความถูกต้อง

ชุดข้อมูลที่นำมาทดลองทั้งหมดจำนวน 11 ชุด เป็นชุดข้อมูลมาตรฐานจำนวน 7 ชุดและชุดข้อมูลที่สร้างขึ้นจำนวน 4 ชุด

5.2 ข้อดีของงานวิจัย

1. จากการทดลองของอัลกอริทึมที่นำเสนอ ในส่วนของการทดลองอัลกอริทึมที่หนึ่งหลักการทำงานส่วนใหญ่คล้ายกับการทำงานของอัลกอริทึมเคมีนแบบดั้งเดิม นั่นคือการเลือกจุดเริ่มต้นโดยใช้วิธีการสุ่มเลือก การคำนวณจึงไม่ซับซ้อน มีผลทำให้การทำงานอัลกอริทึมใช้เวลาสำหรับการทำงานน้อย

2. จากการทดลองของอัลกอริทึมที่นำเสนอ ในส่วนของอัลกอริทึมที่สอง จุดเด่นของอัลกอริทึมนี้คือ การเลือกจุดเริ่มต้นให้กับอัลกอริทึมเคมีน ซึ่งงานวิจัยนี้เสนอวิธีการเลือกจุดเริ่มต้นรายละเอียดในหัวข้อ 3.3 มีผลทำให้ผลการทดลองเสถียร นั่นคือ ได้ผลการทดลองเหมือนเดิมทุกครั้งของการรันอัลกอริทึม

5.3 ปัญหาที่พบในงานวิจัย

1. จากการทดลองของอัลกอริทึมที่นำเสนอ ในส่วนของอัลกอริทึมที่หนึ่ง ดังที่กล่าวข้างต้นว่าหลักการทำงานส่วนใหญ่คล้ายกับอัลกอริทึมเคมีน ทำให้ผลการทดลองแปรผันไปตามจุดเริ่ม มีผลทำให้การทดลอง ต้องทำการทดลองซ้ำหลายรอบ สำหรับงานวิจัยนี้ทำการทดลอง อย่างน้อย 10 ครั้ง และการทำงานของอัลกอริทึมในส่วนของการวนลูป ต้องวนซ้ำหลายรอบ มากกว่าการทำงานของอัลกอริทึมที่สองของอัลกอริทึมที่นำเสนอ

2. การกำหนดพารามิเตอร์ในส่วนของค่าเอนโทรปีที่กำหนดไว้ สำหรับการแบ่งกลุ่มยังไม่มีกฎเกณฑ์ที่แน่นอน

5.4 แนวทางการพัฒนาในอนาคต

1. พัฒนาวิธีการกำหนดพารามิเตอร์ในส่วนของค่าเอนโทรปีที่กำหนดไว้ ให้มีกฎเกณฑ์ที่แน่นอน
2. พัฒนาอัลกอริทึมของการจัดกลุ่ม(clustering) อัลกอริทึมอื่นๆ ให้เป็นการเรียนรู้แบบมีเป้าหมาย (Supervised Learning)โดยใช้เทคนิคของจำแนกประเภทข้อมูล

บรรณานุกรม

[1] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means Clustering Algorithm," **in Proceedings of the World Congress on Engineering, London**, UK, 2009.

[2] K. A. Abdul Nazeer, S. D. Madhu Kumar, and M. P. Sebastian,"Enhancing the k-means clustering algorithm by using a O(n logn) heuristic method for finding better initial centroids," **in Proceedings of the Second International Conference on Emerging Applications of Information Technology**, pp. 261-264, 2011.

[3] R. Evans, B. Pfahringer, and G. Holmes, "Clustering for classification," **in Proceedings of the 7th International Conference on IT in Asia**, 2011.

[4] Muhammad A. Khan, Muhammad Nazir, Arfan Jaffar and Anwar M. Mirza, " Fuzzy Clustering and Fuzzy Entropy based Classification Model ," **in Processdings of the 6th International Conference on Emerging Technologies (ICET)**, 2010.

[5] A. Kumar, R. Sinha, V. Bhattacherjee, D. S. Verma, and S. Singh, "Modeling using K-means clustering algorithm " **in Proceedings of the 1st International Conference on Recent Advances in Information Technology**, 2012.

[6] A. Frank and A. Asuncion, **"UCI Machine Learn inRepository."**[http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science, 2010.

[7] จิตราภรณ์ มูลวงค์. **"การจัดกลุ่มข้อมูลโดยใช้วิวัฒนาการทางด้านพฤติกรรมทางสังคมของมนุษย์."** วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2550.

ภาคผนวก

ภาคผนวก ก.

ผลงานวิจัยที่ได้รับการตีพิมพ์เผยแพร่

Arit Thammano and Pannee Kesisung."Enhanceing K-means Algorithm for Solving Classification Problems."**Proceeding of 2013 IEEE International Conference on Mechatronics and Automation.** August 4-7, Takamatsu, Japan.

IEEE

Co-sponsors: IEEE Robotics and Automation Society, Kagawa University

Technical Co-sponsors: RSJ, JSME, SICE, JSPE, HIT, BIT, IEEJ, URC, UESTC, CUST, BJUT, TJUT

Call for Papers

Index

The 2013 IEEE International Conference on Mechatronics and Automation (ICMA 2013) will take place in Takamatsu, Kagawa, Japan from August 4 to August 7, 2013. Takamatsu is a small city located at Shikoku which is the smallest island in 4 main islands of Japan. Shikoku contains a lot of temples including Zentsu-ji, where one of the most famous Buddhists, Kukai, was born.

As the host city of ICMA 2013, Takamatsu not only provides the attendees with a great venue for this event, but also an unparalled experience in the Japanese history through several historical architectures. You are cordially invited to join us at IEEE ICMA 2013 in Takamatsu. The objective of ICMA 2013 is to provide a forum for researchers, educators, engineers, and government officials involved in the general areas of mechatronics, robotics, automation and sensors to disseminate their latest research results and exchange views on the future research directions of these fields.

The topics of interest include, but not limited to the following:

- Intelligent mechatronics, robotics, biomimetics, automation, control systems,
- Opto-electronic elements and Materials, laser technology and laser processing
- Elements, structures, mechanisms, and applications of micro and nano systems
- Teleoperation, telerobotics, haptics, and teleoperated semi-autonomous systems
- Sensor design, multi-sensor data fusion algorithms and wireless sensor networks
- Biomedical and rehabilitation engineering, prosthetics and artificial organs
- Control system modeling and simulation techniques and methodologies
- AI, intelligent control, neuro-control, fuzzy control and their applications
- Industrial automation, process control, manufacturing process and automation

Contributed Papers: All papers must be submitted in PDF format prepared strictly following the IEEE PDF Requirements for Creating PDF Documents for IEEE Xplore. The standard number of pages is 6 and the maximum page limit is 8 pages with extra payment for the two extra pages. See detailed instructions in the conference web site. All papers accepted by IEEE ICMA 2013 will be indexed by EI and included in IEEE Xplore®. Extensions of selected papers will be published in a regular or a special issue of the journals of IJMA and JRM.

Organized Sessions: Proposals with the title, the organizers, and a brief statement of purpose of the session must be submitted to an OS Chair by March 20, 2013.
Tutorials & Workshops: Proposals for tutorials and workshops that address related topics must be submitted to one of the Tutorial/Workshop Chairs by May 1, 2013.

Important Dates:

March 20, 2013	**Full papers and organized session proposals**
May 1, 2013	**Proposals for tutorials and workshops**
May 15, 2013	Notification of paper and session acceptance
June 1, 2013	Submission of final papers in IEEE PDF format

For detailed up-to-date information, please visit the IEEE ICMA conference website at:

http://2013.ieee-icma.org

Proceedings of 2013 IEEE
International Conference on Mechatronics and Automation
August 4 - 7, Takamatsu, Japan

Enhancing K-means Algorithm for Solving Classification Problems

Arit Thammano
Computational Intelligence Laboratory
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520 Thailand
arit@it.kmitl.ac.th

Pannee Kesisung
Computational Intelligence Laboratory
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520 Thailand
s2560404@kmitl.ac.th

Abstract **— K-means is the most popular clustering algorithm because of its efficiency and superior performance. However, the performance of K-means algorithm depends heavily on the selection of initial centroids. This paper proposes an extension to the original K-means algorithm enabling it to solve classification problems. First, the entropy concept is employed to adapt the traditional K-means algorithm to be used as a classification technique. Then, to improve the performance of K-means algorithm, a new scheme to select the initial cluster centers is proposed. The proposed models are tested on seven benchmark data sets from the UCI machine learning repository. Experimental results have shown that the proposed models outperform the learning vector quantization network in most of the tested data sets.**

Index Terms **– *Data mining, K-means algorithm, Classification, Entropy***

I. Introduction

Data classification is one of the fundamental problems in data mining. Classification, as described by [1], is a process of finding a model that describes and distinguishes data classes, for the purpose of being able to use the model to predict the class of objects which class label is unknown. There are many classification techniques that have been used thus far such as Decision tree, Neural networks, Support vector machines, and Bayesian networks [2]. This paper focuses on a type of classification model that is based on K-means clustering algorithm. K-means is the most popular clustering algorithm. It is very efficient and very easy to implement. Besides being used as a clustering technique, K-means has also been adapted for data classification. Two main problems of K-means algorithm are that (i) the number of clusters is needed to be specified before running the algorithm, and (ii) the quality of the resulting clusters depends heavily on the selection of initial centroids. Many researchers, such as [3][4], have tried to overcome the above problems by introducing efficient methods for selecting the initial cluster centers. The experimental results show that their proposed algorithms produce better clusters in less computation time than the original K-means algorithm.

For classification, Kumar et al. [5] proposes a model for predicting the probability of the outcome of each class by using K-means clustering algorithm. Evans et al. [6] proposes a framework, called Cluster classifier. In Cluster classifier, K-means or other clustering algorithms are used to summarize a large data set by using the cluster centroids to create a more compact representation of the data set. Then the classification technique is applied to classify the summarized data set.

In this study, two new classification algorithms, which are based on the concept of K-means clustering algorithm and the concept of entropy, are proposed. The classification performance of the proposed methods is evaluated against the learning vector quantization network, which is one of the most powerful neural networks.

The rest of the paper is organized as follows: section II briefly describes the original K-means algorithm and the learning vector quantization network. The proposed algorithms are presented in section III. Section IV describes the tested benchmark problems and discusses the experimental results. Finally, section V is the conclusions.

II. Background

A. K-means Algorithm

K-means is the most famous clustering algorithm. It groups a set of n data points into K clusters using Euclidean distance as the similarity measure. The steps of the K-means algorithm are described as follows:

1. Arbitrarily select K initial cluster centers ($z_1, z_2, \ldots, z_K$) from the input data.
2. Assign each data pattern $\mathbf{x}_i$ to the cluster C_j to which it is the most similar.
3. When all input data has been assigned, update the cluster centers as follows:

$$z_j^* = \frac{1}{n_j} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i \qquad (1)$$

where n_j is the number of data belonging to cluster C_j.

4. If the stopping criterion is satisfied, terminate the loop. If not, go to step 2.

B. Learning Vector Quantization Network

Learning vector quantization (LVQ) network [7] is a special case of the Kohonen self-organizing maps. LVQ is a supervised competitive neural network model in which each output node represents a particular class. During training, the output node whose weight vector most closely matches the training input pattern is chosen as the winner. If the winning node has the correct class label, its weight vector will be moved toward the input pattern. However, if it belongs to the

978-1-4673-5560-5/13/$31.00 ©2013 IEEE

wrong class, its weight vector will be moved away from the input pattern.

III. Proposed Methodology

This paper introduces two new classification algorithms. Both classification algorithms are based on the concept of K-means clustering described in Section II. Besides the K-means clustering concept, the first algorithm employs the concept of entropy to allow K-means to be used as the classification algorithm.

Since the performance of K-means depends heavily on initial selection of the cluster centers, the method of arbitrarily selecting the initial cluster centers will not likely lead to the desired results. Therefore, the second algorithm proposes a new way of selecting the initial centers that produces more promising results.

The details of the two proposed algorithms are described in next subsections.

A. Algorithm 1

The following is the procedure used in training the proposed Algorithm 1.

1. Select the initial centers

Define the initial number of clusters (K). In this research, K is chosen to be equal to the number of data classes (m). Then randomly select K initial cluster centers ($z_1, z_2, \ldots, z_K$) from the input data.

2. K-means clustering

Assign each input data x_i to the cluster C_j to which it is the closest. The distance between the input data x_i and the center of the cluster C_j is calculated by using (3).

$$J = \arg\min_{j=1}^{K} \left[d(\mathbf{x}_i, \mathbf{z}_j) \right] \quad (2)$$

$$d(\mathbf{x}_i, \mathbf{z}_j) = \sqrt{\sum_{q=1}^{Q} (x_{iq} - z_{jq})^2} \quad (3)$$

where $\mathbf{z}_j$ is the center of the cluster C_j, Q is the dimension of the input data.

After assigning the data points to the clusters, K-means clustering is performed on the data set to obtain K disjoint clusters.

3. Measure the purity of the clusters

The purity of each cluster is determined by its entropy, which can be calculated by

$$H(C_j) = -\sum_{i=1}^{m} P_i \log_2 P_i \quad (4)$$

where P_i is the probability that an arbitrary data point in the cluster C_j belongs to the class i. The cluster whose entropy is greater than E_T is split into smaller clusters. The center of each smaller cluster is calculated as the mean of all the input data belonging to each data class.

4. Termination criteria

Steps 2 and 3 are repeated until one of the termination criteria is met.

5. Measure the accuracy

The classification accuracy is calculated as the ratio of the number of correctly classified instances to the total number of instances in the test set.

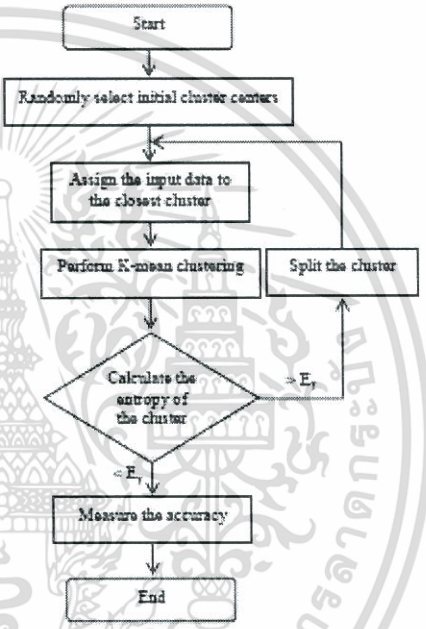


Fig. 1 Flowchart of Algorithm 1.

B. Algorithm 2

Algorithm 2 is an improvement over Algorithm 1, which randomly selects the initial cluster centers from the input data. The more efficient method for selecting the initial cluster centers is introduced in this algorithm. The followings are the details of Algorithm 2.

1. Select the initial centers

The first task is to select the initial centers, which is done as follows:

a) Define the initial number of clusters (K). In this research, K is chosen to be equal to the number of data classes.
b) Normalize the input data to a value between [0, 1].
c) Divide the normalized input data range into N segments $(s_1, s_2, \ldots, s_n, \ldots, s_N)$.
d) Starting with the first attribute, count the number of data points which lie in each segment. Then the segments are ranked in descending order of the number of members.
e) Select the top three ranked segments whose at least one members belongs to class 1.
f) For each of the three selected segments, calculate the ratio of the number of members in class 1 to the number of members in other classes.

$$r = \frac{n_1}{\sum_{i=2}^{m} n_i} \qquad (5)$$

where n_i is the number of data belonging to class i. Then the segment with the largest ratio value is selected. Next, the value of the first attribute of the first cluster center is defined as the average of members of the selected segment.
g) Remove the selected segment in step f from the ranked list.
h) Continue determining the values of the first attribute of the clusters 2, 3, ..., K by repeating steps e through g.
i) Repeat steps d through h for all the other attributes.

After finishing this step, the center of each cluster will be identified.

2. Assign the data points to the clusters

Assign each input data x_i to the closest cluster by using (2). The distance between the input data x_i and the center of the cluster C_j is calculated by using (3).

3. Measure the purity of the clusters

To measure the purity of each cluster, calculate the entropy of each cluster according to (4). The cluster whose entropy is greater than E_T is split into smaller clusters. The center of each smaller cluster is calculated as the mean of all the input data belonging to each data class.

4. Termination criteria

Steps 2 and 3 are repeated until one of the termination criteria is satisfied. The termination criteria used in this work are the specified maximum number of iterations and the acceptable entropy value.

5. Measure the accuracy

The classification accuracy is calculated as the ratio of the number of correctly classified instances to the total number of instances in the test set.

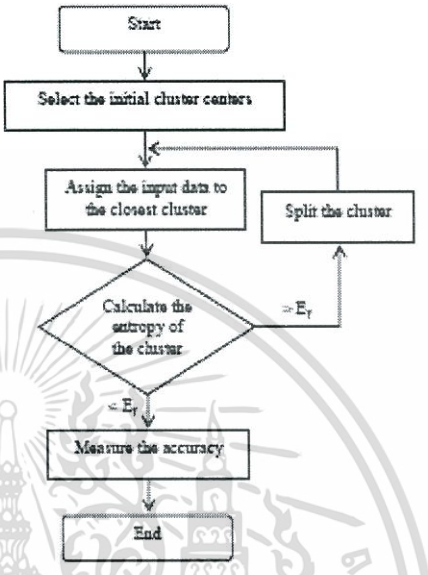


Fig. 2 Flowchart of Algorithm 2

IV. Experimental Results

In this research, the proposed algorithms 1 and 2 were benchmarked against the learning vector quantization network (LVQ). The experiments were conducted on 7 data sets from UCI machine learning repository [8] namely Iris, Wine recognition, Haberman's survival, Heart disease, Balance scale weight and distance, Vehicle silhouettes, and Pima Indians diabetes. The brief descriptions of the above data sets are as follows:

1. The first data set is the well-known Iris data. The sepal length, sepal width, petal length, and petal width of 150 Iris flowers from 3 species (Iris-setosa, Iris-versicolor, and Iris-virginica) are measured in centimeters and are used as the input of the problem. The training set contains 120 records while the testing set contains 30 records.
2. The second data set is the Wine recognition data. This data is the result of a chemical analysis of wines grown in the same region but from three different cultivars. Thirteen continuous attributes are used to determine the type of wine (class 1, 2, or 3). In this paper, the 178 instances in the database were randomly divided into a training set of 90 instances and a testing set of 88 instances.
3. The third data set is the Haberman's survival data. This data set contains 306 cases from a study that was conducted between 1958 and 1970 at the University of Chicago's

Billings Hospital on the survival of patients who had undergone surgery for breast cancer. Three numerical attributes are used to predict the output class (1 or 2). In this paper, the data was divided into a training set of 214 examples and a testing set of 92 examples.

4. The fourth data set is the Heart disease problem. The problem concerns the prediction of the absence (class 1) or presence (class 2) of heart disease given the results of various medical tests carried out on a patient. This data set contains 13 attributes and 270 records. In this paper, the 270 records in the database were randomly divided into a training set of 135 records and a testing set of 135 records.
5. The fifth data set is the Balance Scale Weight and Distance database. This data set was generated to model psychological experiments. Each example is classified as having the balance scale tipping to the right, tipping to the left, or being balanced. The input data consists of 4 numerical attributes: Left-weight, Left-distance, Right-weight, and Right-distance. In this paper, the training set contains 313 examples, while the testing set contains 312 examples.
6. The sixth is the Vehicle Silhouettes data set. The purpose is to classify a given silhouette as one of four types of vehicle (a double-decker bus, Chevrolet van, Saab 9000, and Opel Manta 400). Eighteen features were extracted from each of the silhouette. There are a total of 846 patterns in the data set. In this paper, the data was randomly divided into a training set of 424 examples and a testing set of 422 examples.
7. The seventh data set is the Pima Indians diabetes database. The problem is to predict whether a patient would test positive (class 1) or negative (class 0) for diabetes according to World Health Organization criteria. This database contains 768 examples. Each example is described by 8 numerical attributes. In this paper, the 768 examples in the database were randomly divided into a training set of 384 examples and a testing set of 384 examples.

In this paper, we conducted 2 sets of experiments. The first set emphasizes on revealing the performance of Algorithm 1 while the second set focuses on Algorithm 2. For each data set, Algorithm 1 is run 10 times. Each run starts with a different initial set of centers. The experimental results of Algorithm 1 in classifying the above 7 data sets are shown in Table I. It can be seen from Table I that the performance of Algorithm 1 varies with the initial set of centers. This is to be expected as Algorithm 1 is significantly based on K-means algorithm. Despite of the above issue, the average performance on the 7 test sets in terms of the classification accuracy is quite promising.

Table II illustrates the best classification accuracy on the tested data sets obtained from Algorithm 1, Algorithm 2, and LVQ. For Iris data, all three algorithms are able to obtain the perfect classification accuracy. For Vehicle silhouettes data and Pima Indians diabetes, Algorithm 1 comes first in the competition. The performance of Algorithm 1 is better than the performance of Algorithm 2 by 1.66 and 4.95% for Vehicle silhouettes data and Pima Indians diabetes, respectively. The performance of Algorithm 1 is better than the performance of LVQ by 11.38 and 2.34% for Vehicle silhouettes data and Pima Indians diabetes, respectively. For Wine recognition, Haberman's survival, and Balance scale weight and distance, Algorithm 2 comes out to be the best among the compared algorithms. The performance of Algorithm 2 is better than the performance of Algorithm 1 by 5.55, 6.55, and 1.60% for Wine recognition, Haberman's survival, and Balance scale weight and distance, respectively. The performance of Algorithm 2 is better than the performance of LVQ by 1.77, 15.71, and 0.32% for Wine recognition, Haberman's survival, and Balance scale weight and distance, respectively. The Heart disease problem is the only data for which LVQ attains the best classification accuracy while Algorithm 2 comes second.

TABLE I
EXPERIMENTAL RESULTS OF ALGORITHM 1

Experiments	Accuracy						
	Iris	Wine recognition	Haberman's survival	Heart disease	Balance scale weight and distance	Vehicle silhouettes	Pima Indians diabetes
1	0.9667	0.9167	0.8361	0.8074	0.8591	0.7844	0.7656
2	1.0000	0.9167	0.8361	0.8074	0.8109	0.7870	0.7422
3	0.9667	0.9167	0.8197	0.8074	0.8397	0.7725	0.8229
4	0.9667	0.9167	0.8525	0.8074	0.8813	0.7589	0.7422
5	1.0000	0.9167	0.8525	0.8148	0.8462	0.7986	0.7422
6	0.9667	0.9167	0.8361	0.8148	0.8213	0.7464	0.7604
7	1.0000	0.9167	0.8361	0.8148	0.8462	0.7891	0.7552
8	0.9667	0.9167	0.8033	0.8148	0.7855	0.7986	0.7760
9	1.0000	0.9167	0.8361	0.8148	0.8462	0.7938	0.7422
10	0.9667	0.9167	0.8361	0.8148	0.8173	0.7749	0.7422
Maximum	1.0000	0.9167	0.8525	0.8148	0.8462	0.7986	0.8229
Minimum	0.9667	0.9167	0.8033	0.8074	0.7855	0.7464	0.7422
Mean	0.980002	0.9167	0.834426	0.811184	0.832501	0.77172	0.753639
S.D.	0.017196	0.0000	0.01436	0.003821	0.021536	0.018299	0.025386

TABLE II
COMPARATIVE RESULTS OF ALGORITHM 1, ALGORITHM 2, AND LVQ

Data sets	Accuracy		
	Algorithm 1	Algorithm 2	LVQ
Iris	1	1	1
Wine recognition	0.9167	0.9722	0.9545
Haberman's survival	0.8825	0.9150	0.7659
Heart disease	0.8148	0.8222	0.8444
Balance scale weight and distance	0.8462	0.8627	0.8540
Vehicle silhouettes	0.7500	0.7820	0.6848
Pima Indians diabetes	0.8229	0.7734	0.7995

V. CONCLUSIONS

This paper presents two new learning models for classification tasks. The proposed models are based on K-means algorithm and the concept of entropy. In the proposed models, the entropy concept is employed to adapt the traditional K-means algorithm to be used as a classification technique. Moreover, to improve the performance of K-means algorithm, a new scheme to select the initial cluster centers is also proposed. The performance of the two proposed algorithms is measured in terms of the classification accuracy. The experimental results show that the proposed approaches perform better than the learning vector quantization network in most of the tested data sets.

REFERENCES

[1] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2006.

[2] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

[3] K. A. Abdul Nazeer, S. D. Madhu Kumar, and M. P. Sebastian, "Enhancing the k-means clustering algorithm by using a o(n log) heuristic method for finding better initial centroids" in Proceedings of the Second International Conference on Emerging Applications of Information Technology, 2011, pp. 261-264.

[4] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means Clustering Algorithm," in Proceedings of the World Congress on Engineering, London, UK., 2009.

[5] A. Kumar, R. Sinha, V. Bhattacherjee, D. S. Verma, and S. Singh, "Modeling using K-means clustering algorithm " in Proceedings of the 1st International Conference on Recent Advances in Information Technology, 2012.

[6] R. Evans, B. Habegger, and O. Hahmi, "Clustering for classification," in Proceedings of the 1st International Conference on IT in Asia, 2011.

[7] T. Kohonen, "The self-organizing maps," Proceedings of the IEEE, vol. 78, no. 9, pp. 1464-1480, 1990.

[8] A. Frank and A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science, 2010.

ประวัติผู้เขียน

ชื่อ-นามสกุล	พรรณี เกษีสังข์
วัน-เดือน-ปี เกิด	7 – เมษายน – 2529 ที่สุรินทร์
ที่อยู่	56 หมู่ 3 ตำบล กุดขาคีม อำเภอ รัตนบุรี จังหวัด สุรินทร์ 32130
ประวัติการศึกษา	2552 ปริญญาตรีวิทยาศาสตร์บัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเกษตรศาสตร์