การศึกษากระบวนการแบ่งกลุ่มเมล็ดพันธุ์ด้วยข้อมูลโครงสร้างของเมล็ด

STUDY OF THE SEED CLUSTERING PROCESS USING STRUCTURAL DATA

สุทธิพงศ์ ผ่องแผ้ว

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2561

การศึกษากระบวนการแบ่งกลุ่มเมล็ดพันธุ์ด้วยข้อมูลโครงสร้างของเมล็ด

สุทธิพงศ์ ผ่องแผ้ว

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2561
ลิขสิทธิ์ของมหาวิทยาลัยศรีนครินทรวิโรฒ

STUDY OF THE SEED CLUSTERING PROCESS USING STRUCTURAL DATA

SUTTIPONG PONGPAW

A Master's Project Submitted in partial Fulfillment of Requirements

for MASTER OF SCIENCE (Information Technology)

Faculty of Science Srinakharinwirot University

2018

Copyright of Srinakharinwirot University

สารนิพนธ์

เรื่อง

การศึกษากระบวนการแบ่งกลุ่มเมล็ดพันธุ์ด้วยข้อมูลโครงสร้างของเมล็ด

ของ

สุทธิพงศ์ ผ่องแผ้ว

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

---------- คณบดีบัณฑิตวิทยาลัย

(ผู้ช่วยศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก
(อาจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

..... ประธาน
(อาจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.จันตรี ผลประเสริฐ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ประจักษ์ เฉิดโฉม)

ชื่อเรื่อง	การศึกษากระบวนการแบ่งกลุ่มเมล็ดพันธุ์ด้วยข้อมูลโครงสร้างของเมล็ด
ผู้วิจัย	สุทธิพงศ์ ผ่องแผ้ว
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2561
อาจารย์ที่ปรึกษา	อาจารย์ ดร. วีรยุทธ เจริญเรืองกิจ

การคัดแยกเมล็ดพันธุ์จะสำเร็จได้ต้องมีผู้เชี่ยวชาญที่มีทักษะและประสบการณ์เฉพาะในการจำแนกเมล็ดพันธุ์ในการตรวจสอบตัวอย่างทีละตัวอย่าง ซึ่งเป็นกระบวนการที่ต้องใช้เวลามากสำหรับการแบ่งออกเป็นกลุ่ม เป้าหมายคือการสร้างกลุ่มของเมล็ดพันธุ์ที่มีลักษณะทางกายภาพภายในกลุ่มที่คล้ายกัน โดยการจัดกลุ่มเมล็ดแบบอัตโนมัติสามารถช่วยลดเวลาและปรับปรุงประสิทธิภาพได้ ซึ่งความท้าทายของการจัดกลุ่มข้อมูลคือการกำหนดจำนวนกลุ่มข้อมูล และผลความถูกต้องในการจัดกลุ่ม งานวิจัยนี้มีจุดมุ่งหมายเพื่อศึกษาอัลกอริทึมการจัดกลุ่มและแนวทางในการเลือกจำนวนของกลุ่มรวมทั้งการประเมินค่าด้วย performance vectors ที่ประกอบด้วย 2 วิธี คือ Davies-Bouldin และ Average distance within centroid เพื่อประเมินความถูกต้องจากการทดลองข้อมูลเมล็ดพันธุ์ที่ใช้ในการทดลองได้จากฐานข้อมูล UCI

การทดลองโดยใช้เทคนิคทำเหมืองข้อมูลด้วย Rapid Miner Studio7.6 แสดงให้เห็นว่า performance vector สามารถใช้เพื่อประเมินความถูกต้องของอัลกอริทึมการจัดกลุ่ม แต่ไม่ถูกต้องเสมอไป ดังแสดงด้วยการประเมินผลอัลกอริทึมการจัดกลุ่ม หลังจากเลือก attribute ด้วยค่า Coefficient of variation มีค่า performance vectors ที่ดีขึ้น แต่ความถูกต้องในการแบ่งกลุ่มลดลง และพบว่าอัลกอริทึม K-means เป็นอัลกอริทึมการจัดกลุ่มที่ดีกว่าการแบ่งกลุ่มด้วยอัลกอริทึม DBSCAN สำหรับชุดข้อมูลเมล็ด UCI

คำสำคัญ : การแบ่งกลุ่มข้อมูล, K-means, Data Mining, Feature selection

Title	STUDY OF THE SEED CLUSTERING PROCESS USING STRUCTURAL DATA
Author	SUTTIPONG PONGPAW
Degree	MASTER OF SCIENCE
Academic Year	2018
Thesis Advisor	Lecturer Werayuth Charoenruengkit , Ph.D.

Seed clustering can be a time consuming process and requires specific human skills and experience to place a large number of seeds into several groups. The goal is to establish certain groups of seeds with similar physical charateristics in a group. It was more effective to cluster seeds automatically to reduce time and improve efficiency. The challenge of establishing a clustering algorithm to determine the number of clusters and computational efficiency to cluster data into the correct groups. This research aimed to study the clustering algorithms and approaches to select the number of clusters as well as the evaluation of the feature selections. Davies-Bouldin and Average distance within centroid were the performance vectors chosen to evaluate the accuracy from the experments. The seed data used in the experiement were obtained from the UCI database.

The experiment using the data mining modules from Rapid Miner Studio 7.6 showed that the performance vector can be used to evaluate the accuracy of the clustering algorithm, but was not always correct in terms of an evaluation of clustering algorithms as it was falsely shown that the coefficient of variation could improve the clustering accuracy performance. The kmean algorithm is the best clustering algorithm in comparision to the DBscan algorithms for the UCI seed dataset.

Keyword : Segmentation, K-means, Data Mining, Feature selection

กิตติกรรมประกาศ

สารนิพนธ์นี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอกราบขอบพระคุณท่านคณาจารย์ทุกท่าน ที่ได้กรุณาให้คำแนะนำในการศึกษา และ อาจารย์ ดร.วีรยุทธ เจริญเรืองกิจ อาจารย์ที่ปรึกษาสารนิพนธ์ช่วยเหลือทั้งในด้านการดำเนินงานวิจัย และด้านวิชาการ ที่ให้คำแนะนำในกระบวนการวิจัย เพื่อเขียนสารนิพันธ์ฉบับนี้ ขอขอบคุณเพื่อนๆ บัณฑิตทุกท่าน ที่ให้คำปรึกษา และให้ความช่วยเหลืออย่างดีเสมอมา

สุดท้ายนี้ ผู้วิจัยขอขอบคุณ คณาจารย์ทุกท่านที่ได้ประสิทธิ์ประสาทวิชาความรู้ต่างๆ ทั้งในอดีต และปัจจุบัน และขอกราบขอบพระคุณบิดา มารดา ที่ให้กำลังใจ รวมถึงเป็นแรงผลักดันที่ยิ่งใหญ่แก่ผู้วิจัยและส่งเสริมการศึกษาเป็นอย่างดีมาโดยตลอด

สุทธิพงศ์ ผ่องแผ้ว

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการทำวิจัย

ในประเทศไทยเกษตรกรมักจัดจำแนกเมล็ดตามคุณภาพขนาดและคุณสมบัติอื่น ๆ เพื่อให้สามารถบรรจุหรือขายได้ตามชนิดที่จัดไว้ ขั้นตอนการจำแนกและวิเคราะห์เมล็ดพันธุ์เหล่านี้ต้องใช้กระบวนการคัดแยกด้วยตนเองซึ่งต้องใช้เวลามาก

การจำแนกประเภทและการวิเคราะห์เมล็ดสามารถช่วยควบคุมคุณภาพเมล็ดพันธุ์เพื่อป้องกันการปนเปื้อนของเมล็ดพันธุ์ โดยปัญหาการคัดแยกเมล็ดพันธุ์จะสำเร็จได้ต้องมีผู้เชี่ยวชาญในการตรวจสอบตัวอย่างทีละตัวอย่าง ซึ่งเป็นงานใช้เวลานาน จึงใช้เทคโนโลยีมาประยุกต์ใช้ในการแบ่งกลุ่มเพื่อตรวจสอบสายพันธุ์ของเมล็ดข้าวเพื่อให้ได้ช่วยลดระยะเวลาในการทำงาน และเพิ่มความแม่นยำของการแบ่งกลุ่มเมล็ดพันธุ์ข้าว[1]

จากที่มาและความสำคัญดังกล่าว ผู้วิจัยจึงเห็นความสำคัญของเมล็ดพันธุ์ข้าว จึงได้เลือกทำการศึกษาข้อมูลทางด้านเมล็ดพันธุ์ข้าวจาก UCI (University of California Irvine) ปี 2012 เกี่ยวกับ เมล็ดข้าวสาลี 3 สายพันธุ์มีข้อมูล 210 รายการ มี 7 แอตทริบิวต์ด้วยเทคนิคทาง K-means โดยการสร้างแบบจำลองทาง Data Mining เพื่อเพิ่มความแม่นยำในการแบ่งกลุ่มเมล็ดพันธุ์ข้าว เพื่อที่จะสามารถนำไปประยุกต์ใช้ทางด้านอุตสาหกรรมเกี่ยวกับเมล็ดพันธุ์ข้าวได้ และมีวัตถุประสงค์ของการวิจัย ดังนี้ เพื่อศึกษาการเลือก feature selection จากข้อมูล เพื่อเพิ่มความแม่นยำในการแบ่งกลุ่ม และ เพื่อศึกษาการใช้เทคนิคทาง K-means Clustering ในการแบ่งกลุ่มข้อมูล

1.2 วัตถุประสงค์ของการวิจัย

1.เพื่อศึกษาและเปรียบเทียบผลการใช้เทคนิค K-means Clustering และ DBSCAN Clustering ในการแบ่งกลุ่มข้อมูล

2. เพื่อศึกษาวิธีการเลือกค่าจำนวนกลุ่มที่เหมาะสมในการแบ่งกลุ่มข้อมูล

3. เพื่อศึกษาการเลือก feature selection โดยใช้สัมประสิทธิ์ของการแปรผัน

4. เพื่อศึกษาการใช้ Rapid Miner Studio7.6 เป็นเครื่องมือในการวิเคราะห์การแบ่งกลุ่มข้อมูล

1.3 ขอบเขตของการวิจัย

การวิจัยนี้เป็นการศึกษาด้วยเทคนิคทาง K-means โดยการสร้างแบบจำลองทาง Data Mining เพื่อเพิ่มความแม่นยำในการแบ่งกลุ่ม

1. ศึกษาทฤษฎีและเทคนิคจากงานวิจัยในรูปแบบต่าง ๆ ที่มีความเกี่ยวข้องกับการแบ่งกลุ่มข้อมูลทาง Data Mining เพื่อนำมาประยุกต์ใช้ในการแบ่งกลุ่ม

2. ศึกษาการเลือก feature selection ด้วยสัมประสิทธิ์ของการแปรผันจากข้อมูล

3. ข้อมูลที่นำมาใช้ในการศึกษาการวิจัย เป็นข้อมูล จาก UCI ปี 2012 เกี่ยวกับ เมล็ดข้าวสาลี 3 สายพันธุ์ มีข้อมูล 210 รายการ มี 8 แอตทริบิวต์

4. ใช้อัลกอริทึมที่มีให้ใน Rapid Miner Studio7.6 เป็นเครื่องมือในการวิเคราะห์การแบ่งกลุ่มข้อมูล โดยจะใช้ k-means clustering และ DBSCAN ในการศึกษา

1.4 วิธีดำเนินการวิจัย

1. ศึกษาผลงานนิพนธ์: Literature Review ทบทวนวรรณกรรมที่เกี่ยวข้อง โดยต้องศึกษาทฤษฎีและศึกษางานวิจัยที่เกี่ยวกับเทคนิคทาง Data Mining รวมถึงการแบ่งกลุ่มทางด้าน K-means Clustering ในการวิเคราะห์การแบ่งกลุ่มข้อมูล เพื่อใช้ประกอบในงานวิจัย

2. กำหนดขอบเขตของการศึกษา ประกอบด้วย

- ข้อมูล คือ เตรียมข้อมูล, รวบรวมข้อมูล, ความต้องการต่าง ๆ ที่จะนำมาวิเคราะห์
- ในการวิจัย ในการวิจัยนี้ทำการเลือกใช้ Data Mining ที่มี Rapid Miner Studio7.6 เป็นโปรแกรมในการประมวลผลข้อมูลในการหาค่าทางสถิติ

3. ศึกษาทฤษฎีที่เกี่ยวข้อง ทางด้าน Data Mining เทคนิคการแบ่งกลุ่มแบบ K-means Clustering และ การเตรียมข้อมูล data preparation อีกทั้ง การใช้งานโปรแกรม Rapid Miner Studio7.6 รวมไปถึง การวิเคราะห์กลุ่ม เพื่อที่จะสามารถใช้อัลกอริทึมในการแบ่งกลุ่มสำหรับแยกเมล็ดพันธุ์ เพื่อใช้ในการแก้ไขปัญหาของงานวิจัยในครั้งนี้

4. ใช้โปรแกรม Rapid Miner Studio7.6 เพื่อทำการวิเคราะห์ข้อมูล ด้วยเทคนิค K-means Clustering และ DBSCAN โดยการแบ่งกลุ่มของข้อมูล และตรวจสอบความถูกต้อง

5. ทำการสรุป และ อภิปรายผลของงานวิจัย

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สามารถศึกษาการเลือก feature selection ที่เหมาะสมในการแบ่งกลุ่มได้

2. สามารถศึกษาการใช้เทคนิคทาง K-means Clustering ในการแบ่งกลุ่มข้อมูลได้

3. สามารถศึกษาการใช้เทคนิค DBSCAN Clustering ในการแบ่งกลุ่มข้อมูลได้

4. สามารถศึกษาการใช้ Rapid Miner Studio7.6 เป็นเครื่องมือวิเคราะห์การแบ่งกลุ่มข้อมูลได้

5. สามารถเป็นแนวทางการวิเคราะห์เพื่อจำแนกเมล็ดพันธุ์ได้

บทที่ 2
แนวคิด ทฤษฎี เทคโนโลยี และระบบงานที่เกี่ยวข้อง

ในการศึกษาของงานวิจัยเรื่อง การศึกษาอัลกอริทึมการแบ่งกลุ่มโดยใช้ K-mean สำหรับการแยกเมล็ดพันธุ์ ในครั้งนี้ มี 2 ประเด็น ประเด็นแรกพิจารณาจำนวนกลุ่ม (k) สำหรับการจัดกลุ่ม K-means โดยสมมติว่าไม่มีความรู้ labels ของข้อมูล โดยการทดสอบจะพิจารณาค่าที่ดีที่สุดของ k ตามดัชนีการประเมินเท่านั้น และ ความถูกต้องได้รับการยืนยันโดยการประเมินความถูกต้องของการจัดกลุ่มด้วย k ที่ดีที่สุด ประเด็นที่สองการทำ feature selection โดยใช้สัมประสิทธิ์ของการแปรผัน และวัดผลของข้อมูลโดย Performance vector อีกทั้งใช้เทคนิค DBSCAN เพื่อเปรียบเทียบความถูกต้องในการจำแนกข้อมูลเมล็ดพันธุ์

จึงจำเป็นที่จะต้องมีความรู้ความเข้าใจ แนวคิด ทฤษฎี เทคโนโลยี และระบบงานที่เกี่ยวข้องจากการศึกษาทฤษฎีที่เกี่ยวข้อง ทางด้านสถิติ รวมไปถึง การวิเคราะห์ข้อมูลในแบบต่าง ๆ เพื่อที่จะสามารถนำไปใช้ในการแก้ไขปัญหาของงานวิจัยในครั้งนี้ ประกอบด้วยดังนี้

2.1 ทฤษฎีพื้นฐานเกี่ยวกับการ Data Mining

Data Mining เป็นเทคนิคในการวิเคราะห์ข้อมูลอย่างหนึ่ง ซึ่งมาจากคำว่า เหมืองข้อมูล เป็นการค้นหาสิ่งที่มีประโยชน์จากฐานข้อมูลที่มีขนาดใหญ่ โดยเป็นกระบวนการสกัด หรือค้นหาสารสนเทศ เพื่อให้ได้ความรู้ หรือสารสนเทศบางมุมที่ซ่อนเร้นอยู่ในฐานข้อมูลขนาดใหญ่ ซึ่งต้องใช้ข้อมูลในอดีตเป็นจำนวนมาก เพื่อนำความรู้ที่ได้ไปช่วยในการวิเคราะห์ และประกอบการตัดสินใจของผู้บริหารในธุรกิจด้านการวิเคราะห์ยอดการสั่งซื้อของลูกค้าที่สามารถช่วยแบ่งกลุ่มและวิเคราะห์ยอดการสั่งซื้อเพื่อนำไปผลิตและขายสินค้าได้ตรงตามกลุ่มเป้าหมายแต่ละกลุ่ม [2]

โดยข้อมูลสามารถแบ่งออกเป็น 2 ประเภทคือ ซึ่งเป็น Supervised Learning และ Unsupervised Learning โดย Supervised Learning เป็นกระบวนการแยกข้อมูลที่อิงกับข้อมูลในอดีตที่ได้รับการระบุไว้ กระบวนการเรียนรู้สร้างแบบจำลองหรือเรียนรู้เกี่ยวกับคุณสมบัติของข้อมูล และ และ label ที่เกี่ยวข้องเพื่อทำนาย label จากข้อมูลที่มองไม่เห็น ตัวอย่างของเทคนิค Supervised Learning มีรายละเอียดดังนี้

Supervised Learning คือกระบวนการเรียนรู้ ที่มีการสอน โดยจะสอน Machine ว่า Inputs ของข้อมูล มีลักษณะอย่างไร แล้วจะได้ Output ของข้อมูลแบบใด อาทิเช่น เราต้องการให้

Machine เรียนรู้เรื่อง เด็กนักเรียนว่าคนไหนสอบผ่าน หรือ สอบตก โดยหน้าตา Data Set จะมีลักษณะดังตารางที่ 1 [3]

ตาราง 1 Input สำหรับ Supervises Learning

ชื่อ	A	B	C	D	E	F
คะแนน	90	25	73	51	40	85
Label	Pass	Fail	Pass	Pass	Fail	?

โดยในตารางที่ 1 Label คือการแสดง Machine ว่าOutput จะมีลักษณะอย่างไร และนำตารางนี้เข้าไปใน Machine จะทำให้สามารถเรียนรู้ได้ว่าระดับคะแนนที่เท่าไหร่คือผ่าน หรือคะแนนที่เท่าไหร่คือตก ดังนั้น Machineสามารถคาดเดาได้ว่า นักเรียนชื่อ F มีความน่าจะเป็นที่จะได้ Pass [3] ซึ่งเป็นผลจากข้อมูลที่ได้เรียนรู้มาโดยไม่จำเป็นต้องรู้ว่าคะแนนเท่าไรคือผ่าน

2. Unsupervised Learning คือกระบวนการเรียนรู้แบบไม่มีการสอน โดยการเรียนรู้วิธีการจะประกอบด้วย Inputs ของข้อมูล และทำให้ Machine ทราบว่าต้องการอะไรดังตัวอย่างเช่นการแบ่งกลุ่ม ในตารางที่ 2 ดังนี้ [3]

ตาราง 2 Input สำหรับ Unsupervised Learning

ชื่อ	A	B	C	D	E	F
คะแนน	90	25	73	51	40	85

จะเห็นได้ว่าการแบ่งเป็น 2 กลุ่ม ผลลัพธ์ที่ออกมาจะไม่สามารถทราบได้ว่า 2 กลุ่มนั้นคืออะไร คอมพิวเตอร์จะทำหน้าที่เพียงแบ่งกลุ่ม หากแบ่งเป็น 2 กลุ่ม จะได้ผลลัพธ์ดังตารางที่ 3

ตาราง 3 ผลลัพธ์จาก Unsupervised Learning

ชื่อ	A	B	C	D	E	F
คะแนน	90	25	73	51	40	85
Output	1	2	1	1	2	1

2.2 การวิเคราะห์กลุ่ม (Cluster Analysis)

การวิเคราะห์กลุ่ม (Cluster Analysis) เป็นเทคนิคในการแบ่งกลุ่มหน่วยข้อมูล หรือเป็นการแบ่งคน สัตว์ สิ่งของ องค์กร ฯลฯ ออกเป็นกลุ่มย่อยอย่างน้อย 2 กลุ่ม โดยมีหลักเกณฑ์ในการแบ่งดังนี้ “ให้หน่วยที่อยู่ในกลุ่มเดียวกันมีลักษณะที่สนใจเหมือนกันหรือคล้ายกัน แต่หน่วยที่อยู่ต่างกลุ่มกันจะมีลักษณะที่สนใจต่างกัน” [4]

การแบ่งกลุ่มข้อมูล (Clustering) เมื่อฐานข้อมูลมีขนาดใหญ่ขึ้น และมีความหลากหลายมากขึ้น จะมีความจำเป็นต้องแบ่งข้อมูลออกเป็นส่วนๆ โดยข้อมูลในแต่ละส่วนมีความเกี่ยวพันกันอย่างเหมาะสม ก่อนที่จะใช้วิธีปฏิบัติการอื่นต่อไป ทั้งนี้ เพื่อความสะดวกในการทำเหมืองข้อมูลรวมทั้งการทำให้การสรุปสาระในแต่ละส่วนมีความหมาย และใช้ประโยชน์ได้ตรงกับความต้องการ การแบ่งข้อมูลนี้ อาจใช้การเป็นวิธีการกำหนดกลุ่ม หรือประเภทที่มีความแตกต่างกันจำนวนหนึ่ง [5] ตัวอย่างเช่น การทำ k-means และการทำ DBSCAN ใน Rapid Miner Studio 7.6

2.3 เทคนิค Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) เป็นสิ่งที่สำคัญขั้นตอนหนึ่ง ที่ควรทำในกระบวนการ Data Science โดยอาศัยเทคนิคการทำ data visualization และ quantitative methods ต่าง ๆ ในการสรุป และทำความเข้าใจใน data-set ที่มี เพื่อให้เข้าใจข้อมูลก่อนทำการวิเคราะห์ นอกจากนี้ยังใช้ประโยชน์จากหลายวิธีเชิงปริมาณเพื่ออธิบายข้อมูล คือ ช่วยตรวจสอบข้อผิดพลาด, ช่วยเลือกแบบจำลองที่เหมาะสมเบื้องต้น, ช่วยกำหนดความสัมพันธ์ระหว่างตัวแปรอธิบาย รวมถึงช่วยประเมินทิศทางและความสัมพันธ์ที่หยาบของความสัมพันธ์ระหว่างคำอธิบาย ซึ่ง John W.Tukey ได้เขียนหนังสือ เกี่ยวกับExploratory Data Analysis ในปี 2520 ได้กล่าวว่าการทำ EDA ใช้ในการทำความเข้าใจ และสรุปเนื้อหาของชุดข้อมูล โดยปกติเทคนิคการทำ EDA ที่เป็นที่นิยม คือ การทำ clustering เช่น Hierarchical Clustering, K-Means และ DBSCAN ซึ่งมีลักษณะดังภาพประกอบที่ 1

ภาพประกอบ 1 ลักษณะของกราฟ Exploratory Data Analysis

ที่มา : https://svds.com/value-exploratory-data-analysis/

จากภาพประกอบที่ 1 จะเห็นได้ว่าลักษณะการจัดกลุ่มของข้อมูลที่คล้ายกันจะอยู่รวมกันเป็นกลุ่มก้อน และในชุดข้อมูลจะมีข้อมูลที่แตกต่างกันซึ่งจะเป็นจุดข้อมูลเล็ก ๆ ทำให้สามารถระบุได้ง่ายขึ้นว่าข้อมูลมี outlier มากน้อยเพียงใด แต่ข้อมูลที่จะทำการ Cluster ไม่มีสีในการแยกจึงจะใช้ EDA ในการสังเกตสิ่งต่อไปนี้

1. เพื่อสำรวจลักษณะข้อมูลในมุมต่าง ๆ ในทุก ๆตัวแปร หรือเปรียบเทียบกันระหว่าง Attribute

2. เพื่อทำให้ข้อมูลในเชิงปริมาณดูเข้าใจง่าย ทำให้เห็นภาพรวมของข้อมูลได้ชัดเจน ซึ่งง่ายต่อการนำเสนอ และ การวิเคราะห์

2.4 การแบ่งกลุ่มข้อมูลแบบเคมีน (K-means clustering Analysis)

เทคนิคการแบ่งกลุ่มแบบ K-means Clustering, K-means หรือ การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis)คือ อัลกอริทึมชนิดหนึ่งที่เป็นเทคนิค

ในการตัดแบ่ง (Partition) วัตถุออก เป็น K กลุ่ม ซึ่งในแต่ละกลุ่มแทนค่าด้วยค่าเฉลี่ยของกลุ่ม และเป็นจุดศูนย์กลาง (centroid) ของกลุ่มที่ใช้ในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน โดยเลือกกลุ่มที่แบ่งนั้นจะมีระยะห่างจากค่ากลางของกลุ่มที่น้อยที่สุด แล้วทำการประมวลหาค่ากลางของกลุ่มใหม่ ทำเช่นนี้จนกระทั่งค่ากลางของกลุ่มเปลี่ยนแปลงน้อยกว่าค่าที่กำหนดไว้ หรือครบจำนวนรอบที่กำหนดไว้ [6]

2.5 วิธีการทำงานของ K-Means Cluster Analysis

ขั้นตอนในการทำงานของแบ่งกลุ่ม ของ k-means [5]

1) กำหนดจำนวนกลุ่ม K กลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด

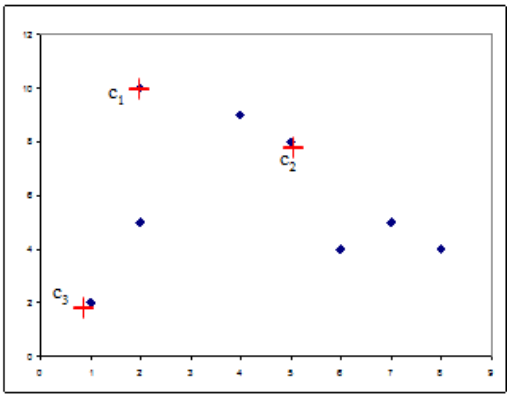
2) นำชุดข้อมูลทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้วัตถุนั้นมากที่สุด ซึ่งคำนวณจากการวัดระยะห่าง ระหว่างจุดที่น้อยที่สุด

3) คำนวณจุดศูนย์กลาง K จุดใหม่ โดยหาค่าเฉลี่ยทุกวัตถุที่อยู่ในกลุ่ม

4) ทำซ้ำในข้อที่ 2) จนกระทั้งจุดศูนย์กลางเปลี่ยนแปลงน้อยกว่าค่าที่ตั้งไว้

2.5.1 การกำหนดจุดเริ่มต้นของ centroid กำหนดอย่างไร

ในการสุ่มเลือกจุดศูนย์กลางเริ่มต้นของกลุ่มจำนวน k จุด โดยกำหนดจุดด้วยวิธี random หรือใช้ค่าเฉลี่ยของข้อมูล นอกจากนี้ยังมีวิธีการเลือกจุดศูนย์กลางเพื่อเพิ่มความถูกต้องในการจัดกลุ่ม โดยใช้ k-means++ [4]



ภาพประกอบ 2 กราฟแสดงข้อมูล และการกำหนดจุดเริ่มต้นของ centroid ในการทำ k-means

ที่มา : https://wipawanblog.files.wordpress.com/2014/06/chapter-8-clustering-k-means.pdf

จากภาพประกอบที่ 2 จะสุ่มเลือกจุดศูนย์กลางเริ่มต้นของกลุ่มจำนวน 3 จุด ได้แก่ c1(2,10), c2(5,8) และc3(1,2)

2.5.2 X-Means Clustering

X-Means Clustering เป็นอัลกอริทึมการขยาย K-Means ซึ่งจะพยายามกำหนดจำนวนของกลุ่มโดยอัตโนมัติ ซึ่งใช้ Bayesian Information Criterion เริ่มต้นด้วยกลุ่มเดียว อัลกอริทึม X-Means จะเข้าสู่การทำงานหลังจากการประมวลผลของ K-Means แต่ละครั้ง ทำให้การตัดสินใจในกลุ่มที่เกี่ยวกับเซตย่อยของ centroids ในปัจจุบันควรแบ่งออกเพื่อให้พอดีกับข้อมูลได้ดียิ่งขึ้น การตัดสินใจแบ่งแยกจะกระทำโดยการคำนวณเกณฑ์ข้อมูลเบย์ [6]โดย BIC จะแสดงถึงโมเดลไหนดีกว่า และจะพยายามเลือกโมเดลที่เป็นจริง ซึ่งมีลักษณะการกำหนดกลุ่มที่แท้จริงภายในชุดของข้อมูลที่ไม่มีป้ายกำกับ และเป็นวิธีที่รวดเร็วที่มีประสิทธิภาพในการจัดกลุ่มข้อมูล

2.5.3. K-Means++

K-Means++ เป็นอัลกอริทึมในการหาจุดเริ่มต้นที่ดีที่สุด ที่กำหนดจุดศูนย์กลางของ cluster ด้วยวิธีการสุ่ม โดยค่าแรกกำหนดจากระยะทางที่สั้นที่สุดของจุดข้อมูลไปยังจุดศูนย์กลางที่อยู่ใกล้ที่สุด เพื่อกำหนดจุดศูนย์กลางของ cluster และในจุดที่สอง กำหนดโดยระยะทางไกลที่สุดของจุดศูนย์กลางของ cluster แรกกับข้อมูลที่มี และทำซ้ำต่อ ๆไป กระทั้งจุดศูนย์กลางเปลี่ยนแปลงน้อยกว่าค่าที่ตั้งไว้

2.6 เทคนิค Hierarchical Cluster Analysis

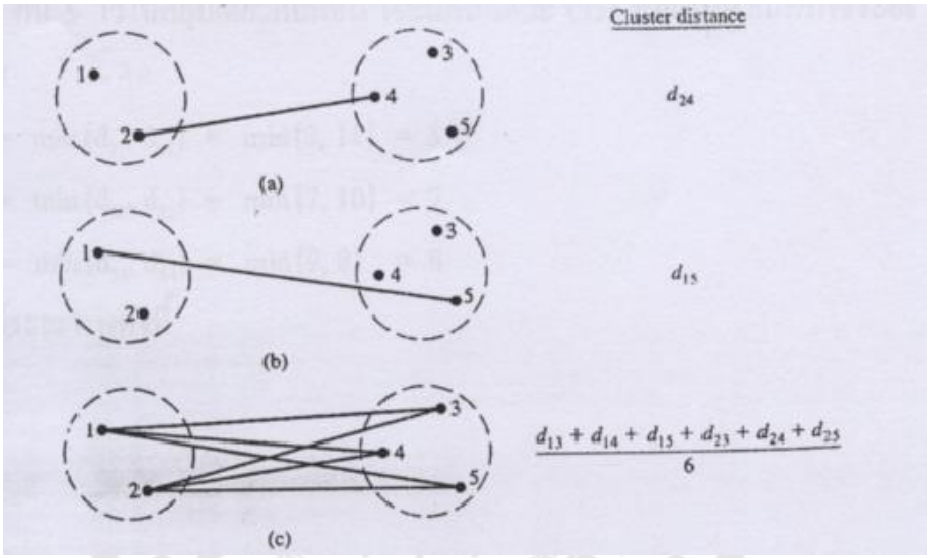
เทคนิค Hierarchical Cluster Analysis หรือ การจัดกลุ่มแบบลำดับชั้น เป็นเทคนิควิธีการจัดกลุ่มตามความคล้ายกันของข้อมูล ด้วยวิธีการวัดความคล้าย หรือความต่าง เช่น Euclidean distance, Cosine similarity, Manhattan distance เป็นต้น ซึ่งลักษณะการแสดงผลของ Hierarchical Clustering จะแสดงในรูปแบบของต้นไม้ และเทคนิคนี้สามารถแบ่งวิธีการสร้างต้นไม้ได้เป็น 2 ประเภท คือ Agglomerative (Bottom-Up) และ Divisive (Top-Down) โดยในแต่ละ class node จะประกอบไปด้วย child nodes [8]

2.6.1 Agglomerative Hierarchical Cluster Analysis [9]

Agglomerative Hierarchical Cluster Analysis เป็นกระบวนการที่เริ่มโดยสมมติให้มี n กลุ่มย่อยของจุดข้อมูลที่คล้ายกันมากที่สุด หรือมีระยะสั้นที่สุดจึงรวมกันเป็นกลุ่มก่อน จะได้ n-1 กลุ่มย่อย โดยที่ n คือ จำนวนข้อมูล

จากนั้นหาความคล้ายหรือคำนวณระยะทางจาก n – 1 กลุ่มย่อยใหม่ และสังเกตลักษณะกลุ่มย่อยใดคล้ายกันมากที่สุด หรือมีระยะทางสั้นที่สุด ให้รวมกลุ่มย่อยนั้นเข้าด้วยกัน ทำเช่นนี้ต่อ ๆ ไป สุดท้ายจะมีเพียง 1 กลุ่ม ซึ่งประกอบด้วยสิ่งของ n สิ่ง

Agglomerative Hierarchical Cluster Analysis โดยทั่วไปประกอบด้วย 3 วิธีการ คือ 1.single linkage (nearest neighbor) 2.complete linkage (furthest neighbor) และ 3.average linkage (average distance) ซึ่งแนวคิดทั้ง 3 วิธี มีลักษณะดังภาพประกอบที่ 3



ภาพประกอบ 3 ลักษณะการจัดกลุ่มของ Agglomerative Hierarchical Cluster Analysis

ที่มา : http://www.slideshare.net/khuwawa2513/12-cluster-analysis

จากภาพประกอบที่ 3 จะเห็นว่า (a) คือ กลุ่มที่ได้มาจากระยะทางที่สั้นที่สุดเป็นการจัดกลุ่มโดยวิธี single linkage, (b) คือ กลุ่มที่ได้มาจากระยะทางที่ยาวที่สุดระหว่างจุดข้อมูลด้วยวิธี complete linkage และ(c) คือ กลุ่มที่ได้จากระยะทางเฉลี่ยระหว่างจุดข้อมูลทั้งหมด นั้นคือวิธี average distance

ขั้นตอนของเทคนิค Hierarchical Cluster สำหรับการจัดกลุ่มลักษณะของข้อมูล

กัลยา วานิชย์บัญชา ได้นำเทคนิค Hierarchical Cluster สำหรับการจัดกลุ่มตามลักษณะของข้อมูล ไว้ 3 ขั้นตอนด้วยกันซึ่งมีรายและเอียดดังนี้ [9]

ขั้นที่ 1 เลือกปัจจัยหรือตัวแปรที่คาดว่ามีอิทธิพลที่ทำให้ลักษณะของข้อมูลต่างกัน นั่นคือ จะทำให้สามารถจัดกลุ่มลักษณะของข้อมูลตัวแปรนั้นได้ชัดเจน

ขั้นที่ 2 เลือกวิธีการคำนวณเพื่อวัดค่าความคล้าย หรือระยะห่างระหว่างลักษณะของข้อมูลแต่ละคู่

ขั้นที่ 3 เลือกหลักเกณฑ์ในการรวมกลุ่ม หรือรวม Cluster

2.6.2 Divisive Hierarchical Cluster Analysis

คือ วิธีการที่สมมติว่ามีกลุ่มของข้อมูล จำนวน n โดยที่ n คือจำนวนของกลุ่มข้อมูล จากนั้นก็จะแบ่งแยกออกเป็น 2 กลุ่ม โดยหาจากระยะทางที่ไกลสุดของกลุ่มของข้อมูล เพื่อวัดระยะห่างระหว่างชุดของแต่ละกลุ่มข้อมูลซึ่งมีการจำแนก 3 ประเภท ดังนี้ 1.Binary, 2.Count Data และ 3.Interval scale

จะเห็นได้ว่าการทำ Agglomerative Hierarchical Cluster Analysis เป็นการรวมเข้าของข้อมูลที่อยู่ใกล้กันเป็นกลุ่ม แต่การทำ Divisive Hierarchical Cluster Analysis เป็นการแบ่งข้อมูลแยกออกเป็นกลุ่ม โดยวัดจากระยะห่างของข้อมูลที่ไกลเพื่อแบ่งกลุ่มของข้อมูล

2.7 การแบ่งกลุ่ม DBSCAN คือ

DBSCAN (Density-based spatial clustering of applications with noise) เป็น algorithm หนึ่งในการทำ clustering โดยไม่ต้องกำหนดจำนวน cluster ที่ต้องการแบ่งเหมือนกับ K-means จึงทำให้ DBSCAN เหมาะสำหรับข้อมูลที่ไม่เป็นกลุ่มก้อน หรือกระจายตัวกันแบบมี pattern ที่มีลักษณะเป็นรูปทรงต่าง ๆ ที่ k-means ไม่สามารถจัดกลุ่มได้ อีกทั้งเหมาะสำหรับการตัด Outliner หรือ Noise ออกไป [10]

Core point : เป็นจุดที่มีลักษณะข้อมูลอยู่ใกล้กันเป็นจำนวนมาก ถ้า core รวมกันได้ก็จะ merge เป็นกลุ่มเดียวกัน

โดย DBSCAN เป็นการหาบริเวณที่ข้อมูลเกาะกลุ่มกัน ซึ่งสามารถคำนวณได้จากจุดข้อมูล ที่อยู่ในบริเวณรอบ ๆ ในรัศมีที่กำหนด ซึ่งการที่จะใช้ DBSCAN ได้ จำเป็นต้องมี 2 parameter คือ[10]

1. Epsilon คือ รัศมีจากจุดศูนย์กลาง
2. Min Points คือ จำนวนจุดข้อมูลขั้นต่ำสำหรับการกำหนด center

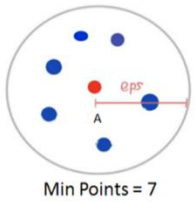
ในการกำหนดจุดแรกของ DBSCAN จะพิจารณาจากความหนาแน่นที่สามารถจัดกลุ่มได้ ด้วยการเปลี่ยนจุด Centroid ไปใกล้ และหาจุดใหม่ โดยสังเกตว่าจุดที่จะกำหนดนั้นคือ มีค่ามากกว่า minpoint จึงนับจุดแรก

ลักษณะของ Algorithm DBSCAN [10]

1. ในแต่ละจุดข้อมูลจะคำนวณหา จุดข้อมูลที่อยู่ใกล้กันทั้งหมดในรัศมี Epsilon ถ้าจุดข้อมูลไหนมี จุดข้อมูลที่อยู่ใกล้กันมากกว่า หรือเท่ากับ Min Points ให้จุดข้อมูลนั้นเป็น core point และสร้างเป็น cluster ใหม่

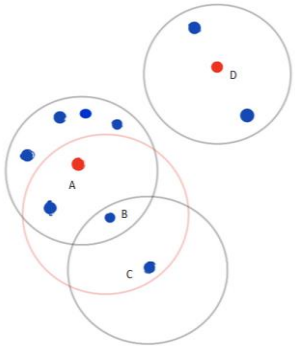
2. ในแต่ละ core point ถ้ามีจุดข้อมูลที่อยู่ใกล้กันที่เชื่อมต่อกับอีก core point ได้ ให้รวมเป็น cluster เดียวกัน

3. ถ้าจุดข้อมูลใดไม่เชื่อมต่อกับ core point ก็จะให้จุดข้อมูลนั้นเป็น Outliner ซึ่งไม่อยู่ใน cluster ใด ๆ



ภาพประกอบ 4 การกำหนด Core point สำหรับการทำ DBSCAN clustering

จากภาพประกอบที่ 4 ตำแหน่ง A จะเรียกว่า Core point เพราะในรัศมีจาก A นั้นมีจุดข้อมูลที่อยู่ใกล้กันอย่างน้อย 6 จุด



ภาพประกอบ 5 การกำหนดจำนวนการจัดกลุ่มข้อมูลในการทำ DBSCAN clustering

จากภาพประกอบที่ 5 จะแสดงลักษณะในการกำหนดตำแหน่งต่าง ๆ เพื่อใช้ในการจัดกลุ่มของ cluster ดังนี้

ตำแหน่ง A จะเรียกว่า Core point เพราะ มีจุดข้อมูลที่อยู่ใกล้กันอย่างน้อย 7

ตำแหน่ง B จะเรียกว่า Border เพราะ B มีจุดข้อมูลที่อยู่ใกล้กันไม่ถึง 7 แต่อยู่ในรัศมีของ core point A

ตำแหน่ง C จะเรียกว่า Border เพราะ C มีจุดข้อมูลที่อยู่ใกล้กันไม่ถึง 7 แต่อยู่ในรัศมีของ B ซึ่ง B ก็อยู่ในรัศมีของ Core point A ทำให้ C ถือว่าอยู่ใน cluster เดียวกันกับ A และ B

ตำแหน่ง D จะเรียกว่า Outlier หรือ Noise เพราะจุดนั้นไม่ได้อยู่ในรัศมีของ Core point ใด ๆเลย ซึ่งNoise นั้นจะเป็นข้อมูลที่เราต้องการตัดออกไป และไม่รวมอยู่ใน Cluster [10]

2.8 การวัดความคล้าย (Similarity Measure)

การวัดความคล้าย คือ การวัดความคล้ายกันของกรณีการจัดกลุ่มของข้อมูล ทีละคู่ในบางครั้งที่เป็นกรณีการจัดกลุ่มของข้อมูล หรือก็คือหลักเกณฑ์ของเทคนิค Cluster ที่จะใช้ในกรณีการจัดกลุ่มของข้อมูลที่คล้ายกันไว้ในกลุ่มเดียวกัน เพื่อใช้ในการจัดกลุ่มตัวแปรที่สัมพันธ์กันไว้ในกลุ่มเดียวกัน

ในส่วนของการจัดกลุ่มตัวแปรการวัดความคล้าย จะเป็นการวัดความคล้ายของตัวแปรแต่ละคู่ ซึ่งหมายความว่า ถ้าระยะห่างระหว่างข้อมูลคู่ใดต่ำ แสดงว่า ระยะห่างระหว่างข้อมูลคู่นั้นอยู่ใกล้กัน หรือมีความคล้ายกัน ควรจะจัดให้อยู่ในกลุ่ม หรือ Cluster เดียวกัน โดยจะต้องหาความคล้ายของกรณีการจัดกลุ่มของข้อมูล ซึ่งการคำนวณจะขึ้นอยู่กับชนิดของข้อมูล มี 3 ประเภท คือ [9]

1. Binary คือ ข้อมูลที่เป็นเลขฐานสอง มีได้ 2 ค่า คือ 0 กับ 1
2. Count Data คือ ข้อมูลที่อยู่ในรูปความถี่เป็นจำนวนนับ และมีค่าเป็นบวก
3. Interval scale คือ ข้อมูลที่แบ่งออกเป็นกลุ่ม โดยมีระยะห่างเท่าๆ กัน สามารถวัดค่าได้ถึงแม้ว่าการเริ่มต้นของ 0 อาจจะไม่เท่ากัน

การวัดความคล้ายด้วยของกรณีในการจัดกลุ่มข้อมูล คือ ถ้าค่าความคล้ายของ กรณีในการจัดกลุ่มข้อมูล คู่ใดมีค่ามากแสดงว่า กรณีในการจัดกลุ่มข้อมูล คู่นั้นคล้ายกันมาก จึงควรจัดให้อยู่ในกลุ่มเดียวกัน การคำนวณค่าความคล้ายจะแตกต่างกัน ถ้าชนิดของข้อมูลแตกต่างกัน

การวัดความคล้ายของตัวแปรด้วยค่าสัมประสิทธิ์สหสัมพันธ์ คือ หากตัวแปรคู่ใด มีค่าสัมประสิทธิ์สหสัมพันธ์มาก แสดงว่าคู่นั้นสัมพันธ์กันมากควรจัดไว้ในกลุ่มเดียวกัน

2.8.1. Euclidean distance

เป็นวิธีการที่ได้มาจากทฤษฎีพีทาโกรัส โดยวัดระยะห่างปกติระหว่างจุด 2 จุดในแนวเส้นตรง ซึ่งอาจวัดได้ด้วยไม้บรรทัด ระยะห่างยูคลิเดียน ระหว่างจุด p และ จุด q แสดงด้วย d(p,q) คำนวณได้ดังสมการ [7]

$$\sqrt{(p_i - q_i)^2 + (p_i - q_i)^2 + \cdots + (p_i - q_i)^2} \quad (1)$$

$$\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2)$$

โดยที่กำหนด p = {p1, p2, p3, …, pn} และ q = {q1, q2, q3, …, qn} คือ จุด 2 จุดที่ต้องการคำนวณระยะห่าง ซึ่งค่า d(p,q) มีค่ามาก แสดงว่า 2 จุดนี้มีความหากกัน หรือแตกต่างกันมาก และหาก 2 จุด p และ q น้อย แสดงว่า มีความใกล้เคียงกันมาก และหากมีค่าเป็น 0 จะหมายถึงทั้ง 2 จุดนั้น คือจุดเดียวกัน

2.8.2 Cosine similarity

เป็นวิธีการวัดมุมโคไซน์ของเวคเตอร์ทั้งสอง โดยสามารถวัดได้จากความคล้ายคลึงระหว่าง 2 เวคเตอร์ ซึ่งคำนวณได้จากสมการ [7]

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{|A||B|} \quad (3)$$

$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum^{n} (A)_i^{\,2}} \times \sqrt{\sum^{n} (B)_i^{\,2}}} \quad (4)$$

โดยกำหนดให้ A = {a1, a2, a3, … , an} และ B = {b1, b2, b3, …, bn} ซึ่งเป็น 2 เวคเตอร์ที่จะนำมาเปรียบเทียบ และค่าสหสัมพันธ์โคไซน์จะมีค่าอยู่ระหว่าง -1 ถึง 1 โดยมีความหมายดังนี้

ถ้าทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางเดียวกัน เมื่อมี ค่าเข้าใกล้ 1

ถ้าทั้ง 2 เวคเตอร์มีความสัมพันธ์กันมากไปในทิศทางตรงข้ามกัน เมื่อค่าเข้าใกล้ -1

ถ้าทั้ง 2 เวคเตอร์ไม่มีความสัมพันธ์กัน เมื่อมีค่าเข้าใกล้ 0

2.8.3. Manhattan distance

เป็นวิธีการวัดระยะทางระหว่างจุดสองจุดตามแกนวัดมุมขวา โดยชื่อลอกเลียนมาจากตารางเค้าโครงของถนนในแมนฮัตตัน ซึ่งทำให้สามารถใช้เส้นทางที่สั้นที่สุดระหว่างจุดสองจุดในเมือง คำนวณได้ดังสมการ [7]

$$d = \sum_{i}^{n} |x_i - y_i| \quad (5)$$

โดยที่ x_i และ y_i คือจุด 2 จุดที่ต้องการคำนวณระยะห่าง

2.9 เทคนิค Principal Component Analysis (PCA)

Principal Component Analysis (PCA) คือ เทคนิคที่นิยมในการทำ decomposition ข้อมูลสำหรับการคลีนข้อมูล และกระบวนการเตรียมข้อมูล เพื่อลด features redundancy ซึ่งช่วยลด feature space ให้มีขนาดเล็กลง โดยตัดตัวแปรที่มีความสำคัญน้อยออกไป และไม่ส่งผลให้สูญเสียข้อมูลสำคัญ หรือ information ไป [11]

PCA เป็นการสร้าง component ใหม่ โดยการหา linear combination จากตัวแปรตั้งต้น เพื่ออธิบาย information ให้ได้มากที่สุด โดยเป็นเทคนิคที่ทำ dimension reduction ประเภท Feature Extraction ทำให้ได้จำนวนตัวแปรใหม่ที่มีขนาดลดลง โดยการทำ PCA มีประโยชน์ดังนี้

1.ช่วยลดปัญหาตัวแปรมีความสัมพันธ์กัน (Multicollinearity) เนื่องจาก component แต่ละตัวจะไม่มีความสัมพันธ์กัน [11]

2.ในกรณีที่ไม่อยากจะเสีย information สามารถช่วยลดจำนวนของตัวแปรลง โดยที่ไม่อยากจะตัดตัวแปร แต่อาจส่งผลให้ตีความผลการวิเคราะห์หรืออธิบายได้ยากขึ้น เนื่องจากการลดจำนวนตัวแปรส่งผลให้ที่ได้จากการวิเคราะห์ในแต่ละ Attribute อาจมีความแตกต่างกันน้อย

การทำงานของ PCA จะทำการสร้างตัวแปรขึ้นมาใหม่เรียกว่า component ซึ่ง component ตัวแรกจะมีค่า variance สูงที่สุด โดยในแต่ละ component จะไม่มีความสัมพันธ์กัน ซึ่งจะอธิบาย information ได้มากที่สุด อีกทั้งในตัวต่อ ๆ ไป จะมีค่าความแปรปรวนที่ลดหลั่นไปตามลำดับ โดยปกติจำนวน component ที่เหมาะสมที่ถูกเลือกมาใช้จะครอบคลุมค่าความแปรปรวนประมาณ 80-90%

วิธีการหา component จะหาได้จากสมการ Y = WX

Y คือ Matrix ของ component ขนาด p x 1

W คือ Matrix ค่าสัมประสิทธิ์ของตัวแปร X ขนาด p x p และ

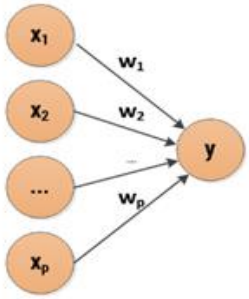
X คือ Matrix ของตัวแปรตั้งต้น ขนาด p x 1

$$y_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p$$

$$y_2 = w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p$$

$$\vdots$$

$$y_p = w_{p1}x_1 + w_{p2}X_2 + \cdots + w_{pp}x_p$$



ภาพประกอบ 6 ลักษณะการกระจายของ linear combination ในการทำ PCA

ที่มา : https://medium.com/ingenio/principal-components-analysis-pca-ต่างจาก-factor-analysis-fa-ยังไง-ตอนที่-1c395e55bdc3

จากภาพประกอบที่ 6 จะเห็นว่าการทำ PCA จะเป็นการสร้าง Component ใหม่ขึ้นมาโดยทำการรวม information ของตัวแปร X ทั้งหมด โดยไม่ตัดตัวแปรใด ๆ จึงทำให้ไม่เสีย information ซึ่งแตกต่างกับการทำ Feature Selection ที่เสีย information (ตัวแปร) ไปบางส่วน

2.10 เทคนิค Coefficient of variation (CV)

ค่าสัมประสิทธิ์ของการแปรผัน (coefficient of variation) ตัวย่อ(C.V.) เป็นอัตราส่วนระหว่างส่วนเบี่ยงเบนมาตรฐานกับค่าเฉลี่ยเลขคณิตของข้อมูลชุดนั้น [12]

$$\text{สูตร} \quad c.v. = \frac{S}{\overline{x}} \tag{6}$$

เมื่อกำหนดให้ C.V. คือ สัมประสิทธิ์ของการแปรผัน

S คือ ส่วนเบี่ยงเบนมาตรฐาน

$\overline{x}$ คือ ค่าเฉลี่ยเลขคณิต

โดยที่ค่าเบี่ยงเบนมาตรฐานแสดงให้เห็นว่า ข้อมูลมีการกระจายจากค่าเฉลี่ยของข้อมูล และค่าเฉลี่ยทำหน้าที่ในการปรับค่าเบี่ยงเบนมาตรฐานของแต่ละ Attribute ให้อยู่ในช่วงเดียวกัน ทำให้สามารถเปรียบเทียบการแปรผันของข้อมูลได้ ดังนั้น Coefficient of variation (CV) จึงเป็นตัวกำหนดความสำคัญของแต่ละ Attribute ซึ่งใช้เป็นเกณฑ์ในการทำ feature selection ด้วยสมมุติฐานที่ว่า ถ้าค่า Attribute ของข้อมูลมีลักษณะที่กระจายมากก็จะสามารถแบ่งกลุ่มได้ง่าย

2.11 ตัววัดประสิทธิภาพโมเดล

ตัววัดประสิทธิภาพโมเดลเป็นวิธีการต่าง ๆ ที่ใช้ในการประเมินคุณภาพของการจัดกลุ่ม รวมไปถึง การประเมินประสิทธิภาพของโมเดลภายในและภายนอก

2.11.1. Davies Bouldin

Davies-Bouldin index เป็นเกณฑ์ที่ใช้ในการวัดคุณภาพการจัดกลุ่มที่ใช้ในการวิเคราะห์ เพื่อการแบ่งกลุ่มข้อมูล โดย Davies-Bouldin index คือ เป็นอัตราส่วนระหว่างผลรวมของการกระจายตัวของข้อมูลในกลุ่ม และ ระยะห่างระหว่างกลุ่ม จะเห็นว่าการแบ่งกลุ่มที่ดีนั้น การกระจายตัวในกลุ่มจะต้องน้อย และ ระยะห่างระหว่างแต่ละกลุ่มจะต้องมาก ซึ่งสำหรับการคำนวณค่าของ Davies-Bouldin index เป็นดังสมการต่อไปนี้ [13]

$$\frac{1}{K}\sum_{k=1}^{K}\max_{k\neq 1}\left\{\frac{S(U_k)+S(U_1)}{d(U_k,SU_1)}\right\} \tag{7}$$

โดย ค่า S(Uk)+S(Ul) เป็นระยะห่างของข้อมูลภายในกลุ่ม k และกลุ่ม l

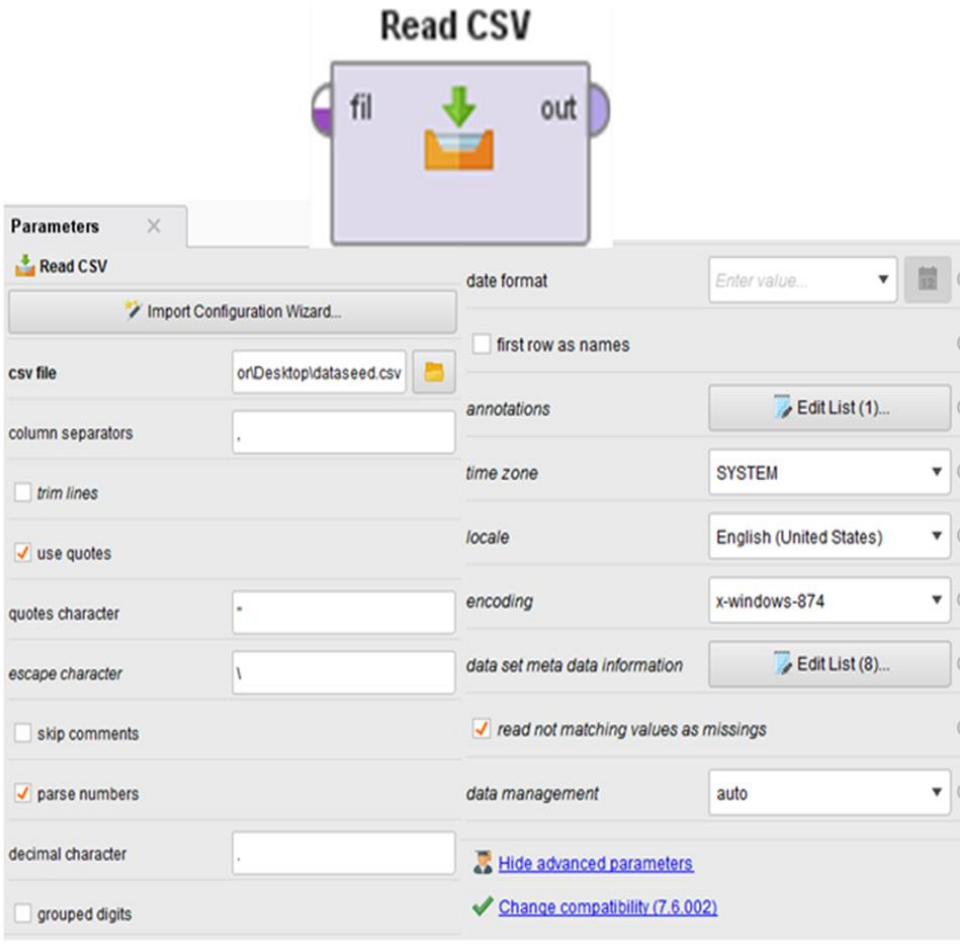
ค่า d(Uk, Ul) เป็นระยะห่างระหว่างจุดกึ่งกลางกลุ่ม k กับกลุ่ม l

ดังนั้น หากค่าของ Davies-Bouldin index มีค่าเล็กที่สุดจะทำให้ได้การแบ่งแยกของกลุ่มดีที่สุด

2.12 Rapid miner studio7.6

Rapid miner studio7.6 เป็นเครื่องมือ ที่ช่วยในการคำนวณวิเคราะห์ประมวลผลโดยมี Operators ของ Rapid Miner Studio 7.6 ที่ใช้ในการทดลองประกอบไปด้วยดังนี้

2.12.1 Read CSV



ภาพประกอบ 7 Operator Read CSV

จากภาพประกอบที่ 7 เป็นตัวที่ใช้ในการ import ชุดข้อมูลสกุล.CSV (comma separated values) ซึ่งเป็นไฟล์ที่มีเครื่องหมาย comma ที่เป็นไฟล์ excel เข้ามาในโปรแกรม Rapid miner studio 7.6 โดยมี parameter ที่สำคัญและใช้ในการวิเคราะห์ คือ

- CSV file คือ รายละเอียดที่อยู่ของไฟล์ที่ import ชุดข้อมูลสกุล.CSV

• Column separator คือพารามิเตอร์ที่ใช้ในการเลือกรูปแบบของตัวคั่น เช่น comma, semicolon, space และ tab เป็นต้น

• trim lines คือ พารามิเตอร์นี้ใช้ลบช่องว่างที่ตำแหน่งเริ่มต้น และส่วนท้าย ก่อนที่จะมีการแบ่งคอลัมน์

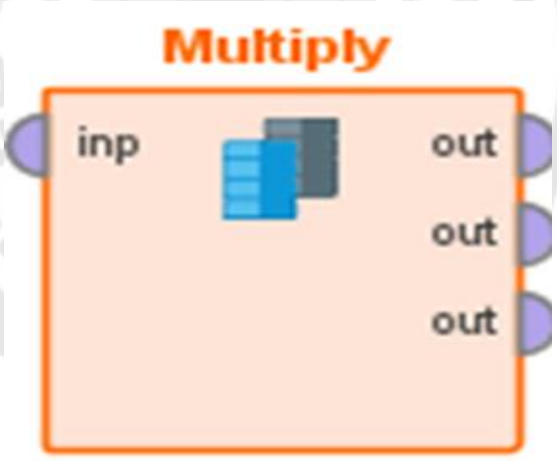
• use quotes คือ พารามิเตอร์ที่ใช้สำหรับคั่นคอลัมน์

• first row as names คือ การกำหนดให้แถวแรกของตารางเป็นชื่อในแต่ละ Attribute

• data set meta data information คือการกำหนด Attribute และ Label ที่ใช้ในการสร้างตารางเพื่อใช้ในการวิเคราะห์

• read not matching values as missings คือพารามิเตอร์ที่กำหนดค่าว่างเป็นเครื่องหมาย '?' ตัวอย่างเช่น ถ้า 'abc' เขียนในคอลัมน์จำนวนเต็มจะถือว่าเป็นค่าที่ขาดหายไป ในไฟล์ CSV เครื่องหมายคำถาม (?) จะถูกอ่านเป็นค่าที่ขาดหายไป

2.12.2. Multiply



ภาพประกอบ 8 Operator Multiply

จากภาพประกอบที่ 8 เป็น Operator Multiply ที่ใช้ในการสร้างสำเนาของข้อมูล โดย output ที่ได้จะสร้างขึ้นได้เหมือนกับต้นฉบับ คือ ข้อมูล input เพื่อนำไปใช้ในการวิเคราะห์ต่อไป ซึ่งในการทดลองนี้ได้ใช้ Operator Multiply ในการคัดลอกตารางข้อมูลที่ได้จากการทำ clustering เพื่อไปทดสอบกับ Operator Aggregate ต่อไป ทำให้สามารถได้สำเนาของข้อมูลหลายสำเนาเพื่อนำไปใช้ในการวิเคราะห์ที่แตกต่างกัน

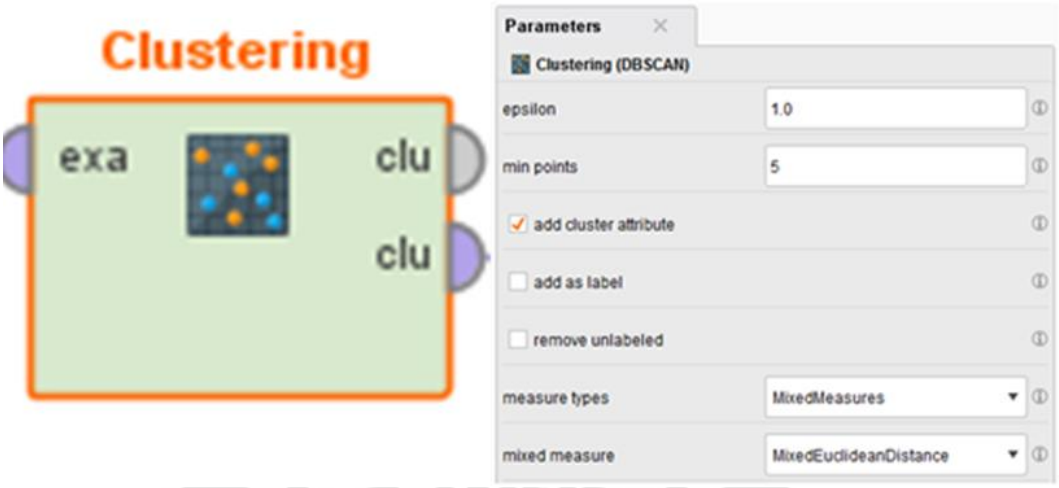
2.12.3 Clustering (k-means)

ภาพประกอบ 9 Operator Clustering (k-means)

จากภาพประกอบที่ 9 เป็นตัวที่ใช้ในการสร้าง cluster สำหรับการแบ่งกลุ่มข้อมูล โดยไม่ต้องกำหนดจำนวนกลุ่ม โดยมี parameter ที่สำคัญและใช้ในการวิเคราะห์ ดังนี้

- K คือ การกำหนดจำนวนกลุ่ม ที่ใช้ในการสร้าง cluster
- max runs คือ พารามิเตอร์ที่ระบุจำนวนสูงสุดของการรันของ k-means ด้วยการเริ่มต้นแบบสุ่มจุด centroidsที่ทำขึ้นเพื่อใช้ในการวิเคราะห์
- determine good start values คือ พารามิเตอร์ที่กำหนดให้จุด centroids ตัวแรกของclusterเป็นตัวที่ดีสุด โดยใช้อัลกอริทึม k-means++
- measure types คือ พารามิเตอร์ที่ใช้สำหรับเลือกชนิดของการวัดที่จะใช้ในการวัดระยะห่างระหว่างจุด ตัวอย่างเช่น
- numerical measure คือ พารามิเตอร์ที่ใช้เลือกการวัดผลในรูปแบบของตัวเลขสำหรับการวิเคราะห์
- max optimization steps คือ พารามิเตอร์ที่ระบุจำนวนสูงสุดของการทำซ้ำสำหรับการการประมวลผลหนึ่งครั้งของ k-means

2.12.4 Clustering (DBSCAN)



ภาพประกอบ 10 Operator Clustering (DBSCAN)

จากภาพประกอบที่ 10 เป็นตัวที่ใช้ในการสร้าง cluster สำหรับการแบ่งกลุ่มข้อมูลโดยไม่ต้องกำหนดจำนวนกลุ่ม ซึ่งสามารถคำนวณได้จาก data point ที่อยู่รอบ ๆ ในรัศมีที่กำหนด ชุดข้อมูลสกุล.CSV ที่เป็นไฟล์ excel เข้ามาในโปรแกรม Rapid miner studio 7.6 โดยมี parameter ที่สำคัญและใช้ในการวิเคราะห์ ดังนี้

- epsilon คือ พารามิเตอร์ที่ระบุขนาดของรัศมีจากจุดศูนย์กลางเพื่อระบุขนาดของพื้นที่ใกล้เคียงโดยรอบ
- min points คือ จำนวน data point ขั้นต่ำสำหรับการกำหนด center ที่สร้างกลุ่ม
- measure types คือ พารามิเตอร์ที่ใช้สำหรับเลือกชนิดของการวัดที่จะใช้ในการวัดระยะห่างระหว่างจุด เช่น Mix Measures, Nominal Measures และNumerical Measures
- mixed measure คือ พารามิเตอร์นี้จะใช้ได้เมื่อชนิดพารามิเตอร์ของการวัดถูกตั้งค่าเป็น 'mixed measures' และมีตัวเลือกเดียวที่มีให้คือ 'Mixed Euclidean Distance'

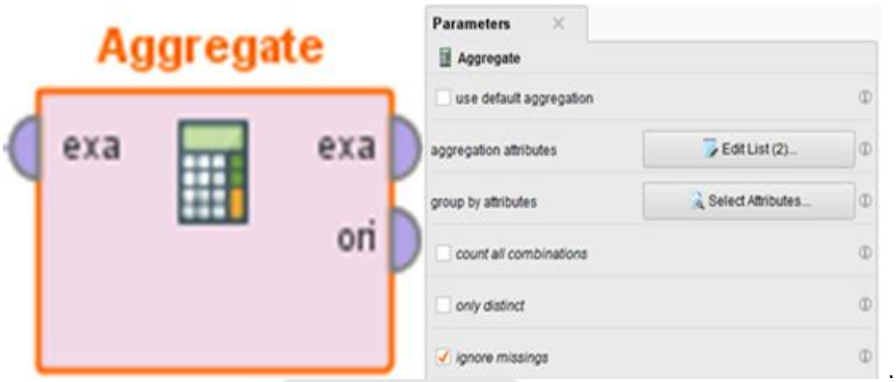
2.12.5 Performance

ภาพประกอบ 11 Operator Performance

จากภาพประกอบที่ 11 เป็น Operator Performance ที่ใช้เพื่อประเมินประสิทธิภาพของวิธีการจัดกลุ่มของเซนทรอยด์ และ แสดงรายการค่าของเกณฑ์ประสิทธิภาพของเซนทรอยด์ในแต่ละคลัสเตอร์ที่ใช้ในการวิเคราะห์ โดยมี main criterion เป็น parameter ที่สำคัญคือ

- main criterion คือ ใช้ในการระบุเกณฑ์หลักที่จะใช้สำหรับการประเมินผล ซึ่งมีเกณฑ์ในการประเมิน คือ Avg. within centroid distance และ Davies Bouldin
- main criterion only คือ การระบุว่าพารามิเตอร์นี้ใช้เกณฑ์หลักในการประเมินผลเพียงชนิดเดียว
- normalize คือ พารามิเตอร์นี้จะระบุว่าให้ผลลัพธ์เป็นแบบ normalize คือการลดความซ้ำซ้อนในข้อมูล โดยผลลัพธ์ที่ได้ในการประเมินจะมีค่าน้อยลง เนื่องจากมีกระบวนที่ดำเนินการอย่างเป็นลำดับเพื่อลดปัญหาการซ้ำซ้อนของข้อมูล
- maximize คือ ผลลัพธ์ที่ได้จากการวัดประสิทธิภาพจะมีค่าเป็นบวก

2.12.6 Aggregate



ภาพประกอบ 12 Operator Aggregate

จากภาพประกอบที่ 12 เป็นตัวที่ใช้ในการสรุปค่าต่าง ๆ ของ Example Set ใหม่จากอินพุต Example Set ที่แสดงผลลัพธ์ของฟังก์ชันการรวมที่เลือก มีฟังก์ชั่นการรวมหลายรายการ รวมถึง SUM, COUNT, MIN, MAX, AVERAGE และฟังก์ชันที่คล้ายคลึงกันอื่น ๆ ที่รู้จักกันในชื่อ SQL ฟังก์ชันการทำงานของประโยค GROUP BY ของ SQL สามารถเลียนแบบได้โดยการใช้กลุ่มตามพารามิเตอร์ attributes ซึ่งใช้ในการสรุปผลการแบ่งกลุ่มของเมล็ดพันธุ์ให้ทราบถึงจำนวนของ cluster และในแต่ละ cluster มีข้อมูลเมล็ดพันธุ์จำนวนกี่เมล็ด รวมถึงมีชนิดใดอยู่บ้าง

งานวิจัยที่เกี่ยวข้อง

M. Emre Celebi และคณะ [14]ได้นำเสนอการเปรียบเทียบวิธีการกำหนดจุดเริ่มต้นของ centroid ของการทำ k-means clustering และนำเสนอเกณฑ์ประสิทธิภาพต่าง ๆ ที่ได้จากการทดลองกับชุดข้อมูล (dataset) ที่แตกต่างกัน

Ahmad และ Reza [15] ทำระบบจำแนกกลุ่มของเมล็ดพันธุ์ข้าวสาลี จาก UCI ด้วยอัลกอริทึม K-Means ด้วยโปรแกรม MATLAB ในการประมวลผล โดยเริ่มจากกำหนด k (จำนวนกลุ่ม) เพื่อหา centroid k จากนั้นคำนวณระยะทางไปยัง centroid และกำหนดกลุ่มตามระยะห่างของจุดศูนย์กลาง คลัสเตอร์ อีกทั้งระบบสามารถเปลี่ยนค่า centroid โดยใช้สูตรEuclidean distance, โคไซน์ และความสัมพันธ์ตัววัดระยะทาง ทำให้จัดกลุ่มเมล็ดทั้งหมดได้อย่างถูกต้อง

Jia-zhou ZHENG [16] ได้ทำการศึกษาค้นคว้าในเรื่อง การแบ่งข้อมูลผลิตภัณฑ์จากเสื้อผ้าด้วยการวิเคราะห์องค์ประกอบหลัก ด้วย K-means algorithm โดยมีการแบ่งกลุ่มผลิตภัณฑ์ตามต้นทุนการจัดเก็บ และตัวชี้วัดต้นทุนการขาดทุน ที่มีตัวชี้วัด 12 ตัว คือค่า KPI ซึ่งเป็นปัจจัยที่เกี่ยวข้องที่สามารถบ่งชี้ข้อมูลระยะเวลาการขายได้ทั้งหมด และใช้ซอฟต์แวร์คือ Clementine 12.0 ที่กำหนด K ไว้ที่ 3 แบ่งตามต้นทุนในการจัดเก็บ คือของที่ขายไม่ดี เป็น 3 กลุ่ม จากนั้นทำการทดสอบความถูกต้องของการแบ่งกลุ่มเปรียบเทียบวิธีการต่าง ๆ โดยมี PCA เพื่อแยกความสัมพันธ์ระหว่างระยะเวลาการขายที่อยู่ติดกัน และมีการใช้อัลกอริทึม K-means ที่ถ่วงน้ำหนักไว้ จากผลการศึกษาพบว่าวิธีการแบ่งกลุ่มถ่วงน้ำหนักแบบ 3 กลุ่ม สามารถเชื่อถือได้

วาทินี นุ้ยเพียร และคณะ [17]ได้ทำการศึกษาเทคนิคการทำเหมืองข้อมูล เพื่อจัดกลุ่มโอโซน โดยใช้ข้อมูลปริมาณโอโซนจาก UCI ปี 2551 โดยมีโมเดลต่าง ๆ ที่ทำการวิเคราะห์ ด้วยซอฟต์แวร์ Weka ซึ่งต้องมีการเตรียมข้อมูล เช่น กำจัด Missing Value จากนั้นทำ Feature Selection 2 แบบ คือ แบบ CfsSubsetEval มี 6 แอททริบิวต์ และ แบบ พิจารณาจากผู้วิจัย มี 24 แอททริบิวต์ เพื่อทำการเปรียบเทียบผลการทดสอบ ซึ่งแบ่งชุดข้อมูล training set กับ test set ซึ่งสรุปผลลัพธ์ได้ว่า K-Means และ XMeans มีค่าดีกว่าโมเดลอื่น ๆ

เรวดี ศักดิ์ดุลยธรรม [5] ได้ทำการศึกษาการใช้เทคนิคดาต้าไมน์นิ่งมาใช้ในการวิเคราะห์การเลือกวิธีในการรักษาโรคนิ้วล็อก โดยการนำข้อมูลส่วนตัว ข้อมูลการตรวจร่างกาย และข้อมูลการรักษาโรคนิ้วล็อกของผู้ป่วย มาทำการวิเคราะห์ปัจจัยที่มีผลต่อความสำเร็จในการรักษาโรคนิ้วล็อกในแบบต่าง ๆ ด้วยการแบ่งกลุ่มของข้อมูลแบบ K-means Clustering เป็น 3 กลุ่ม และสร้างกฎความสัมพันธ์ของข้อมูล (Association Rules) ระหว่างลักษณะอาการของผู้ป่วย กับวิธีการรักษานิ้วล็อก เพื่อใช้เป็นแนวทางประกอบการรักษาโรคนิ้วล็อกได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

วีระยุทธ พิมพาภรณ และ พยุง มีสัจ [8]ได้ทำการศึกษา และนำเสนอผลการทดสอบประสิทธิภาพ ของอัลกอริทึม 2 อัลกอริทึม คือ Hierarchical Clustering กับ K - Means Clustering โดยทั้งสองจะมีกระบวนการสร้างการ (Feature Selection) โดยนำชุดข้อมูล (Data Set) ซึ่งเป็นข้อมูลทางด้านผลการเรียนบางสร้างแบบจำลอง ซึ่งผลการเปรียบเทียบประสิทธิภาพในด้าน ความถูกต้องของการจัดกลุ่ม Hierarchical Clustering มีค่าความถูกต้องดีกว่า K - Means Clustering แต่การใช้เทคนิค Hierarchical Clustering มีข้อเสียคือ มีความซับซ้อนในการวิเคราะห์และอาจจะต้องมีการปรับเพื่อให้ใช้ได้กับข้อมูลขนาดใหญ่

บทที่ 3
วิธีดำเนินการวิจัย

ขั้นตอนในการดำเนินงานวิจัยเริ่มต้นจากกรอบแนวคิดในการศึกษาอัลกอริทึมการแบ่งกลุ่มโดยใช้ k-mean สำหรับการแยกเมล็ดพันธุ์มีดังนี้

1.ศึกษาวิธีการ

- ศึกษาข้อมูลเมล็ดพันธุ์
- ศึกษาการใช้งานโปรแกรม Rapid Miner Studio7.6
- ศึกษาทฤษฎีการแบ่งกลุ่ม
- ศึกษาเทคนิคการเลือกฟีเจอร์

2. ออกแบบกระบวนการวิเคราะห์

- กำหนดขั้นตอนในการทำงานของโปรแกรม

3. ทดสอบระบบ

3.1 เตรียมข้อมูล, ศึกษารายละเอียดข้อมูล ที่จะนำมาวิเคราะห์

ภาพประกอบ 13 รายละเอียดของชุดข้อมูลเมล็ดพันธุ์ข้าวสาลี

ที่มา: https://archive.ics.uci.edu/ml/datasets/seeds

จากภาพประกอบที่ 13 จะแสดงลักษณะของกลุ่มข้อมูล geometric(ลักษณะทางเรขาคณิต)ของเมล็ดพันธุ์ข้าวสาลี 3 สายพันธุ์ ได้แก่ Kama, Rosa และ Canadian สายพันธุ์ล่ะ70 ข้อมูล โดยเป็นข้อมูล จาก UCI ปี 2012 ประกอบไปด้วย 7 Attribute ดังภาพประกอบที่14

Attribute	ความหมาย
1. area A	คือ ค่าขนาดพื้นที่ของเมล็ดพันธุ์
2. perimeter P	คือ ค่าความยาวเส้นรอบวงของเมล็ดพันธุ์
3. Compactness	คือ ค่าความแน่นของเมล็ดพันธุ์
4. length of kernel	คือ ค่าความยาวของเมล็ดพันธุ์
5. width of kernel	คือ ค่าความกว้างของเมล็ดพันธุ์
6. asymmetry coefficient	คือ ค่าสัมประสิทธิ์ความสมส่วนของเมล็ดพันธุ์
7. length of kernel groove	คือ ค่าความยาวร่องเคอร์เนลของเมล็ดพันธุ์

ภาพประกอบ 14 ความหมายของแต่ละ Attribute ของชุดข้อมูลเมล็ดพันธุ์ข้าวสาลี

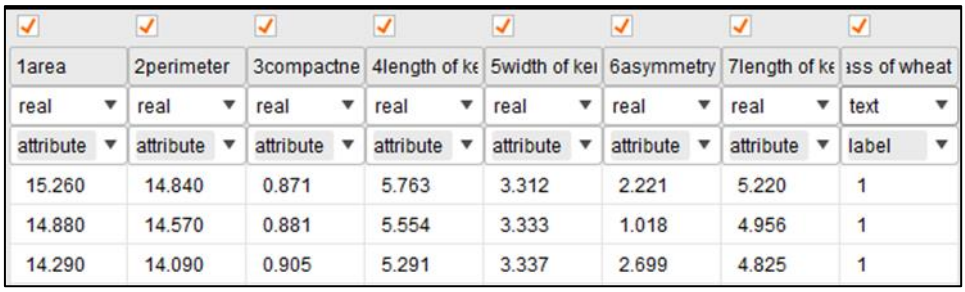
	A	B	C	D	E	F	G	H
1	1area	2perimeter	3compactness	4length of kernel	5width of kernel	6asymmetry coefficient	7length of kernel groove	8class of wheat
2	15.26	14.84	0.871	5.763	3.312	2.221	5.22	1
3	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	1
4	14.29	14.09	0.905	5.291	3.337	2.699	4.825	1
5	13.84	13.94	0.8955	5.324	3.379	2.259	4.805	1
6	16.14	14.99	0.9034	5.658	3.562	1.355	5.175	1

ภาพประกอบ 15 การทำ Data Transformation เป็นสกุล.csv

จากภาพประกอบที่ 15 ลักษณะของชุดข้อมูลที่แปลงข้อมูลจาก csv ที่ได้จากฐานข้อมูล UCI ซึ่งมีการเพิ่มชื่อของคอลัมน์ก่อนที่จะถูกนำเข้าไปยัง Rapid Miner Studio 7.6

3.2 การวิเคราะห์ข้อมูลใน Rapid Miner Studio 7.6

3.2.1 ทำการ import ข้อมูลคุณลักษณะต่าง ๆ ของเมล็ด เข้าใน Rapid Miner Studio 7.6 เพื่อใช้ในการวิเคราะห์และคำนวณหาค่าต่าง ๆ เพื่อใช้ในการแบ่งกลุ่ม



1area	2perimeter	3compactne	4length of ke	5width of kei	6asymmetry	7length of ke	ass of wheat
real	real	real	real	real	real	real	text
attribute	attribute	attribute	attribute	attribute	attribute	attribute	label
15.260	14.840	0.871	5.763	3.312	2.221	5.220	1
14.880	14.570	0.881	5.554	3.333	1.018	4.956	1
14.290	14.090	0.905	5.291	3.337	2.699	4.825	1

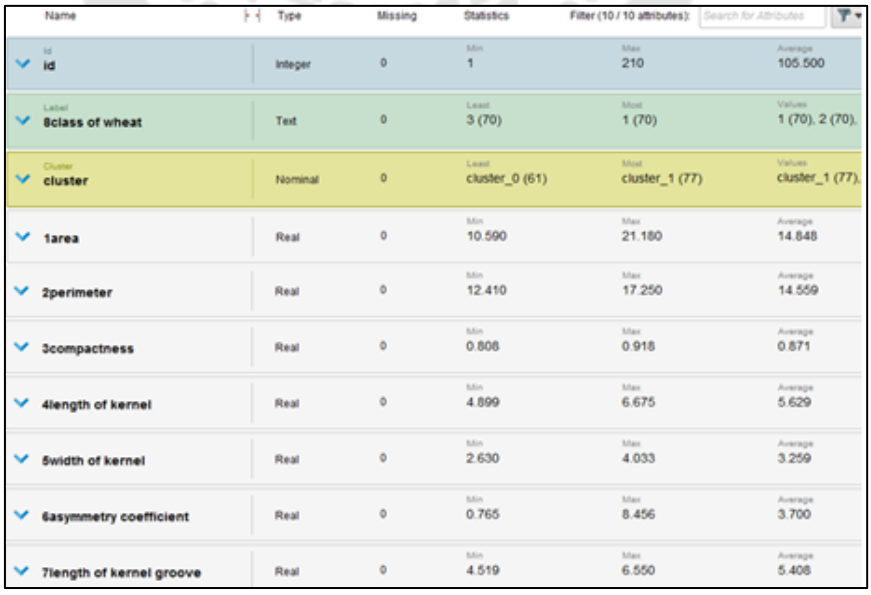
ภาพประกอบ 16 การกำหนดชนิดของข้อมูลในแต่ละ Attribute ที่จะนำมาวิเคราะห์

จากภาพประกอบที่ 16 เป็นการนำเข้าข้อมูลเพื่อนำมาวิเคราะห์ โดยกำหนด Attribute และ Label ให้เหมาะสมกับการวิเคราะห์การแบ่งกลุ่ม

การสำรวจข้อมูล ประกอบด้วยแถว 210 แถวแต่ละแถวจะมีป้ายกำกับว่า 1,2 หรือ 3 เพื่อระบุว่าแถวข้อมูลเป็น Kama, Rosa และ Canadian ตามลำดับ ซึ่งแต่ละคอลัมน์แสดงถึงคุณลักษณะเฉพาะของข้อมูล ซึ่งคุณลักษณะอยู่ 7 Attribute ที่แสดงในภาพประกอบที่ 16

3.2.2 ทำการวิเคราะห์ data exploration

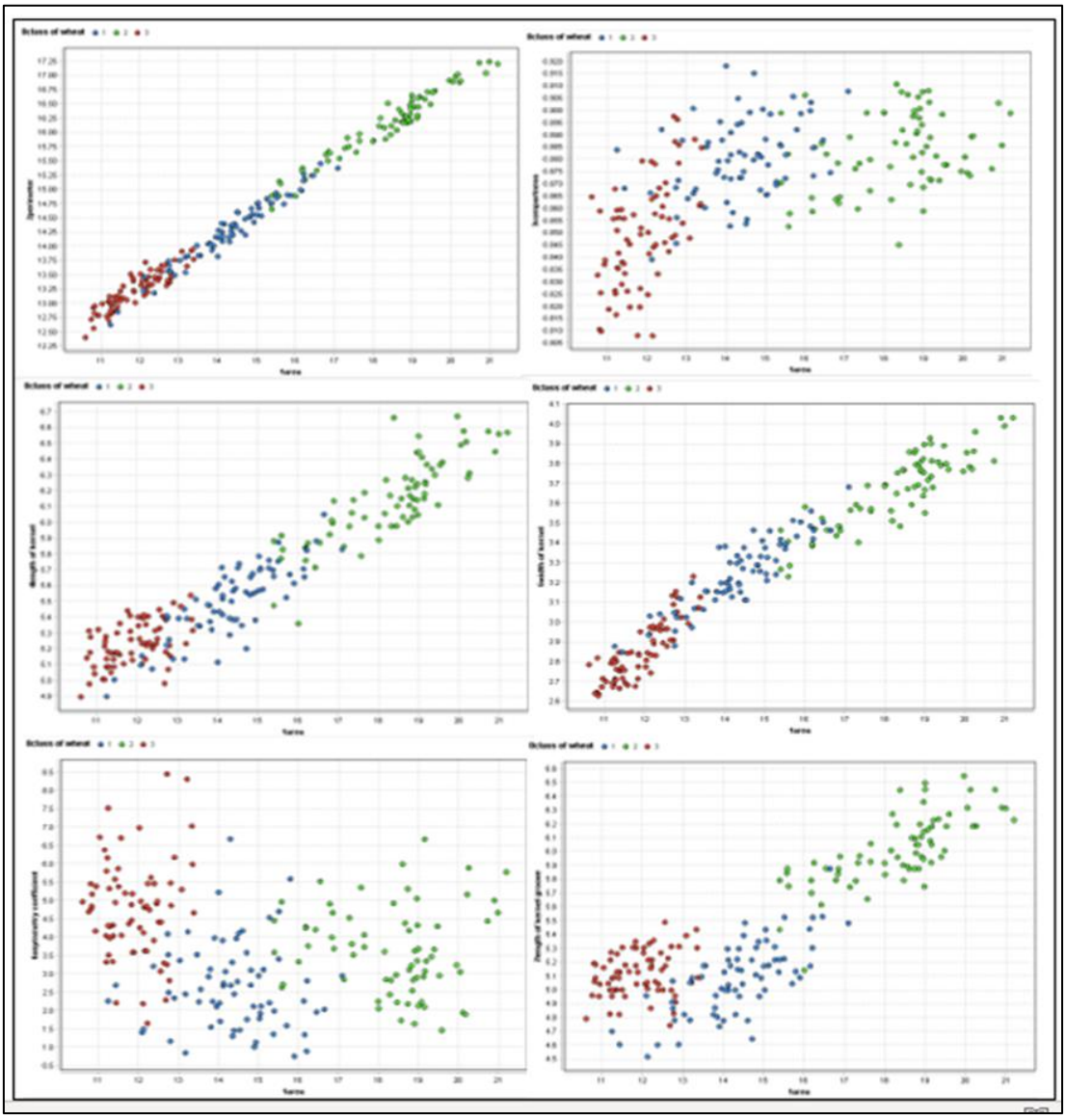
เพื่อทำความเข้าใจว่าอะไรอยู่ในชุดข้อมูลและลักษณะของข้อมูล และคำนวณหาค่าต่าง ๆ เช่น ค่า mean, ค่า distribution ของ data เพื่อสังเกตลักษณะของชุดข้อมูลเป็นอย่างไร เพื่อใช้ประกอบในการวิเคราะห์การแบ่งกลุ่ม



Name	Type	Missing	Statistics	Filter (10 / 10 attributes):	
id	Integer	0	Min 1	Max 210	Average 105.500
class of wheat	Text	0	Least 3 (70)	Most 1 (70)	Values 1 (70), 2 (70),
cluster	Nominal	0	Least cluster_0 (61)	Most cluster_1 (77)	Values cluster_1 (77),
1area	Real	0	Min 10.590	Max 21.180	Average 14.848
2perimeter	Real	0	Min 12.410	Max 17.250	Average 14.559
3compactness	Real	0	Min 0.808	Max 0.918	Average 0.871
4length of kernel	Real	0	Min 4.899	Max 6.675	Average 5.629
5width of kernel	Real	0	Min 2.630	Max 4.033	Average 3.259
6asymmetry coefficient	Real	0	Min 0.765	Max 8.456	Average 3.700
7length of kernel groove	Real	0	Min 4.519	Max 6.550	Average 5.408

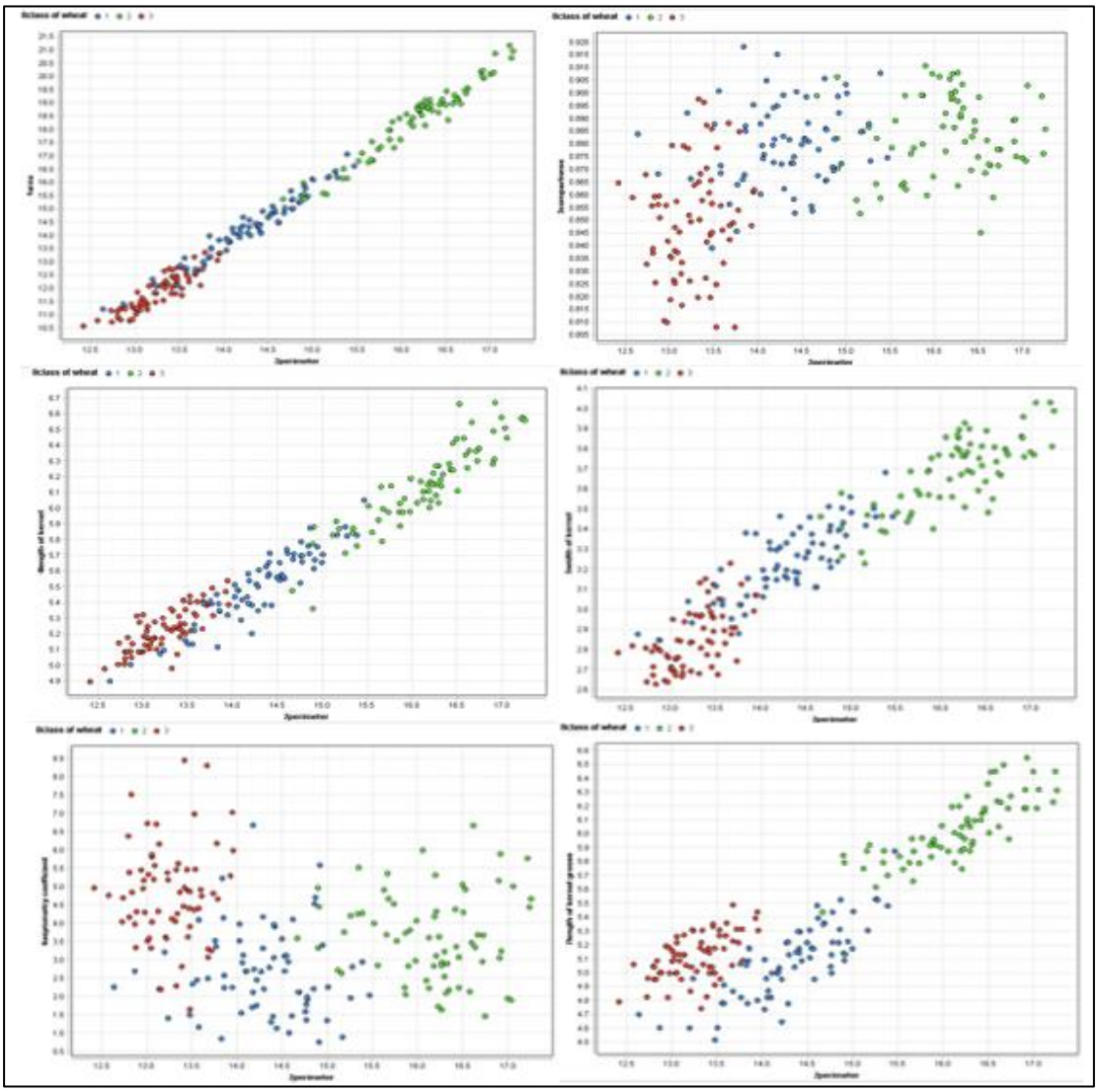
ภาพประกอบ 17 ค่าสถิติของชุดข้อมูลในแต่ละ Attribute

จากภาพประกอบที่ 17 เป็นการค่าทางสถิติของชุดข้อมูลในแต่ละ Attribute เพื่อใช้ประกอบการวิเคราะห์ โดยกำหนด Attribute และ Label ให้เหมาะสมกับการวิเคราะห์การแบ่งกลุ่ม ซึ่งค่า Mean, Max และ Min จะสามารถวิเคราะห์ได้ว่า Attribute นั้น ไม่พบข้อมูลที่เป็น outliner โดยการทำ Exploratory Data Analysis (EDA) เพื่อสังเกตลักษณะของข้อมูลที่ได้จากการพล็อตกราฟ รวมถึงสามารถตรวจสอบในชุดข้อมูลว่าครบถ้วนหรือไม่ จะเห็นได้ว่า การทำ data exploration เป็นสิ่งที่สำคัญที่จะส่งผลต่อการวิเคราะห์ได้



ภาพประกอบ 18 การทำ EDA เปรียบเทียบความสัมพันธ์ Attribute ต่าง ๆ กับ “Area”

จากภาพประกอบที่ 18 เป็นการแสดงกราฟแผนภาพการกระจาย (scatter plot) เพื่อสังเกตความสัมพันธ์ของ Attribute ชื่อ“Perimeter”, “Compactness”, “Length of kernel”, “Width of kernel”, “Asymmetry coefficient” และ “Length of kernel groove” ระหว่าง “Area” ของเมล็ดพันธุ์ข้าวสาลีทั้ง 3 สายพันธุ์ และสังเกตความสัมพันธ์ของ “Perimeter” ได้ดังภาพประกอบ 19



ภาพประกอบ 19 การทำ EDA เปรียบเทียบความสัมพันธ์ Attribute ต่าง ๆ กับ “Perimeter”

จากภาพประกอบที่ 19 แสดงกราฟแผนภาพการกระจาย (scatter plot) ความสัมพันธ์ของ Attribute ชื่อ“Area”, “Compactness”, “Length of kernel”, “Width of kernel”, “Asymmetry

coefficient” และ “Length of kernel groove” ระหว่าง “Perimeter” ของเมล็ดพันธุ์ข้าวสาลีทั้ง 3 สายพันธุ์ จะเห็นได้ว่า ความสัมพันธ์ระหว่าง “Compactness” กับ“Perimeter” และ “Asymmetry coefficient” กับ “Perimeter” มีลักษณะของจุดข้อมูลที่อยู่ห่างกัน และมีจุดข้อมูลมีการกระจายตัว ซึ่งจากภาพประกอบที่ 18 และ19 เป็นการทำ Exploratory Data Analysis ของชุดข้อมูลโดยไม่มี label ที่ใช้เป็นเกณฑ์ในการจำแนกจำนวนกลุ่มของข้อมูล จึงไม่สามารถระบุได้ว่าจุดข้อมูลเกาะกลุ่มกันหรือไม่ แต่สามารถสังเกตได้ว่ามีจุดข้อมูลใดที่เป็น outlier

บทที่ 4
ผลการศึกษา

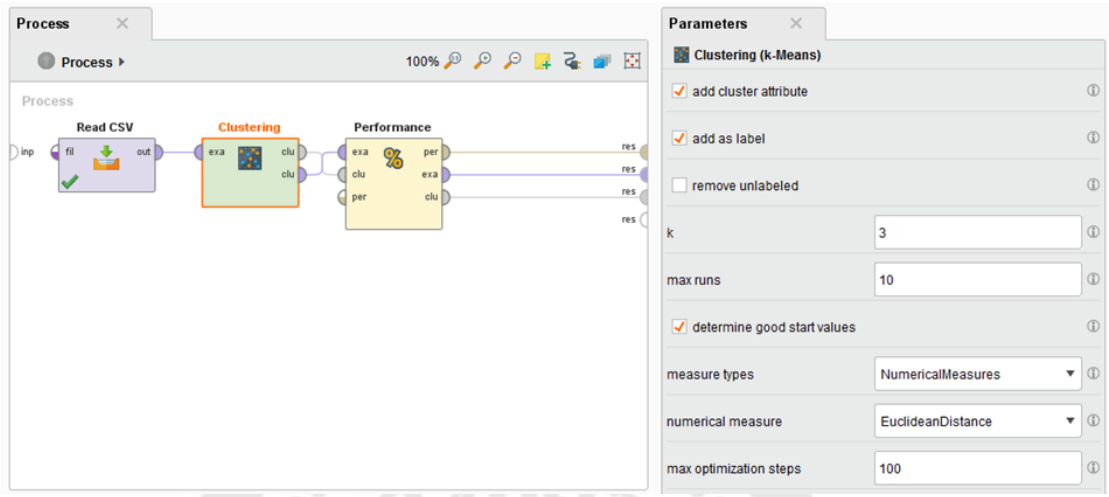
ผลการศึกษาอัลกอริทึมการแบ่งกลุ่มโดยใช้ k-mean สำหรับการแยกเมล็ดพันธุ์ ได้แบ่งออกเป็น 3 การทดลองที่เกี่ยวข้องกัน โดยการทดลองที่ 1 คือ การใช้ Performance vector เพื่อกำหนดจำนวนกลุ่ม ซึ่งกลุ่มที่เกิดขึ้นจะได้รับการตรวจสอบความถูกต้อง โดยเปรียบเทียบกับ label ของข้อมูล การทดลองที่ 2 ประกอบด้วย 2 วิธีการ วิธีการแรกคือ การใช้ CV เพื่อกำหนดความสำคัญของแต่ละ Attribute ซึ่งเป็นประโยชน์สำหรับการจัดกลุ่มที่มีจำนวน Attribute มาก จึงต้องคำนวณข้อมูลคลัสเตอร์มากขึ้น โดยมีการประเมินผล 2 วิธี วิธีแรกใช้ Performance vector เป็นตัวประเมินการประเมิน Attribute ก่อนทำfeature selection และเปรียบเทียบกับหลังการทำ feature selection ที่มีการตัด Attribute ว่าสามารถช่วยปรับปรุงค่า Performance vector index ในการแบ่งกลุ่มได้ดีขึ้นหรือไม่ วิธีที่สองคือการเปรียบเทียบผลการแบ่งกลุ่มกับ label ของข้อมูล และการทดลองสุดท้าย คือการแบ่งกลุ่มข้อมูลโดยการใช้เทคนิค DBSCAN และนำผลมาวิเคราะห์เปรียบเทียบระหว่างผลของเทคนิคกับ K-means

4.1 การทดลองการแบ่งกลุ่ม ทำ K-Means Clustering

การทำ Clustering จะเป็นการแบ่งกลุ่มข้อมูลที่เลือกค่า k ที่ดีที่สุดโดยพิจารณาจากค่า Performance Vector ซึ่งได้มาจากการคำนวณหาความห่างระหว่างข้อมูลกับจุด centroid โดยใช้ Euclidean distance เป็นการวัดความห่าง ในการทดลองได้ทำ clustering ที่ k = 2,3,4,5,6 และเมื่อได้ค่า k ที่เหมาะสม

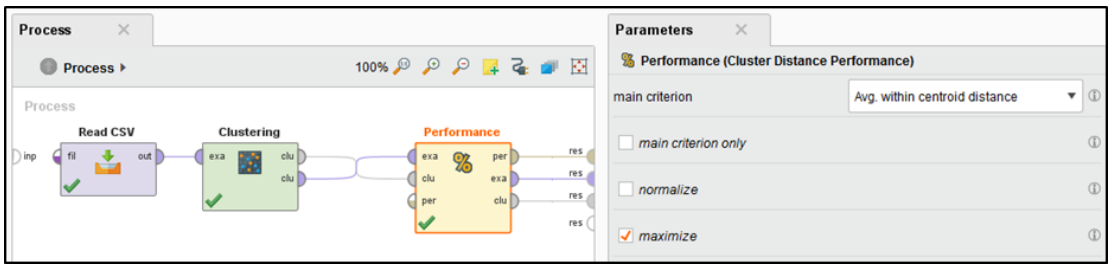
การทดสอบดำเนินการโดยสมมติว่าไม่มีความรู้เกี่ยวกับจำนวนกลุ่มข้อมูล อัลกอริทึม k-means ถูกเลือกเป็นวิธีการจัดกลุ่มข้อมูลเหล่านี้ให้อยู่ในกลุ่มที่แตกต่างกัน โดยที่ k ระบุจำนวนกลุ่ม ค่า k ถูกเลือกจาก k = 2,3,4,5,6 ค่าที่ดีที่สุดของ k ถูกเลือกขึ้นอยู่กับเทคนิคการประเมินที่แตกต่างกัน เมื่อเลือกจำนวนที่ดีที่สุดของ k จะเปรียบเทียบกับจำนวนจริงของกลุ่มเพื่อหาข้อผิดพลาด

4.1.1. การเลือกค่า K จาก Davies-Bouldin และ Avg_within_centroid_distance



ภาพประกอบ 20 การทำ k-means Clustering ใน Rapid Miner Studio7.6

จากภาพประกอบที่ 20 การทำ Clustering จะเป็นการแบ่งกลุ่มข้อมูลจากการสุ่มค่า k โดยมีการคำนวณหาค่า distance ซึ่งค่า distance ได้มาจากค่าความห่างระหว่างข้อมูลกับจุด centroid จากสูตร Euclidean distance และในกระบวนการนี้ทำการสุ่มค่า k=3 หรือก็คือแบ่งกลุ่มข้อมูลออกเป็น 3 กลุ่ม จะได้ cluster 3จำนวน อีกทั้งเลือกพารามิเตอร์ determine good start values เพื่อกำหนดให้จุด centroids ตัวแรกของ cluster เป็นตัวที่ดีสุด เนื่องจากการเลือก determine good start values เป็นการใช้อัลกอริทึม k-means++ รวมถึงกำหนดพารามิเตอร์ max runs สำหรับสุ่มจุด centroids เท่ากับ10 และ พารามิเตอร์ max optimization steps ที่ระบุจำนวนของการทำซ้ำสำหรับการประมวลผลหนึ่งครั้งของ k-means เท่ากับ 100 เนื่องจากเป็นค่าเริ่มต้นของพารามิเตอร์ดังกล่าว

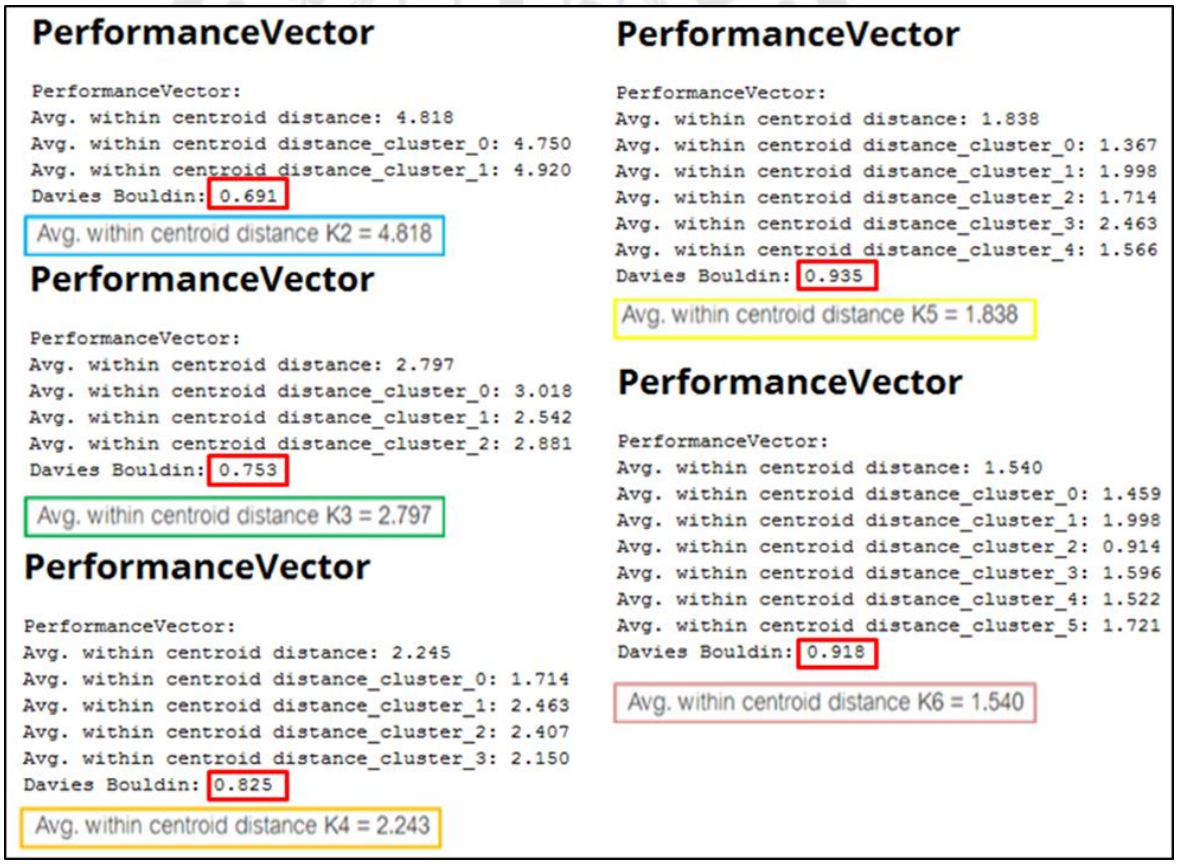


ภาพประกอบ 21 การทำ Performance cluster distance ใน Rapid Miner Studio7.6

จากภาพประกอบที่ 21 แสดงผลลัพธ์ของอัลกอริทึม k-mean ที่ค่าที่แตกต่างกันของ k การทำ Performance cluster distance จะมีการนำผลค่าเฉลี่ยภายในระยะคลัสเตอร์คำนวณโดยใช้ ค่าเฉลี่ยระยะห่างระหว่างข้อมูลแต่ละจุด และจุดเซนทรอยด์เป็นเกณฑ์ในการกำหนด ที่จะหยุดการทำซ้ำ ๆของอัลกอริทึม k-mean หลังจาก วนครบรอบสูงสุดที่ 100 ลูป ผลลัพธ์จะแสดงในภาพประกอบที่ 20 โดยมีการรายงานหลายค่าดังนี้

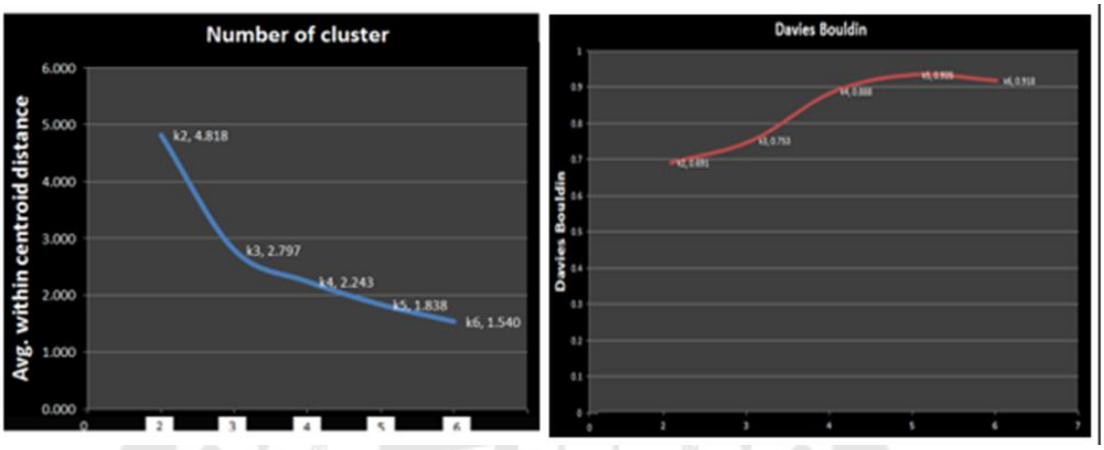
1. Avg_within_centroid_distance คือ การหาค่าเฉลี่ยภายในระยะคลัสเตอร์ คำนวณโดยใช้ระยะห่างเฉลี่ยระหว่างจุด centroid ของคลัสเตอร์ทั้งหมด

2. Davies-Bouldin index คือเกณฑ์ที่ใช้ในการวัดคุณภาพการจัดกลุ่มที่ใช้ในการวิเคราะห์ เพื่อการแบ่งกลุ่มข้อมูลการคำนวณค่าของ Davies-Bouldin index คือ เป็นอัตราส่วนระหว่างผลรวมของการกระจายตัวของข้อมูลในกลุ่ม และ ระยะห่างระหว่างกลุ่ม จะเห็นว่าการแบ่งกลุ่มที่ดีนั้น การกระจายตัวในกลุ่มจะต้องน้อย และ ระยะห่างระหว่างแต่ละกลุ่มจะต้องมาก



ภาพประกอบ 22 การเปรียบเทียบผลการทดลอง Performance Vector ของค่า k = 2,3,4,5,6

จากภาพประกอบที่ 22 จะแสดงลักษณะของข้อมูลที่มีค่า เฉลี่ยระยะห่างจาก centroid กับข้อมูลในแต่ละ cluster และ ค่า Avg. within Centroids Distance ของแต่ละกลุ่ม ที่ k=2, k=3, k=4, k=5, k=6 โดย ค่า Davies-Bouldin เป็นเกณฑ์ที่ใช้ในการวัดคุณภาพการจัดกลุ่มที่ใช้ในการวิเคราะห์ การแบ่งกลุ่มข้อมูล จะเห็นว่าการแบ่งกลุ่มที่ดีนั้น ค่าของ Davies-Bouldin index มีค่าน้อยเท่าไรจะแสดงให้เห็นการแบ่งแยกกลุ่มดีที่สุด



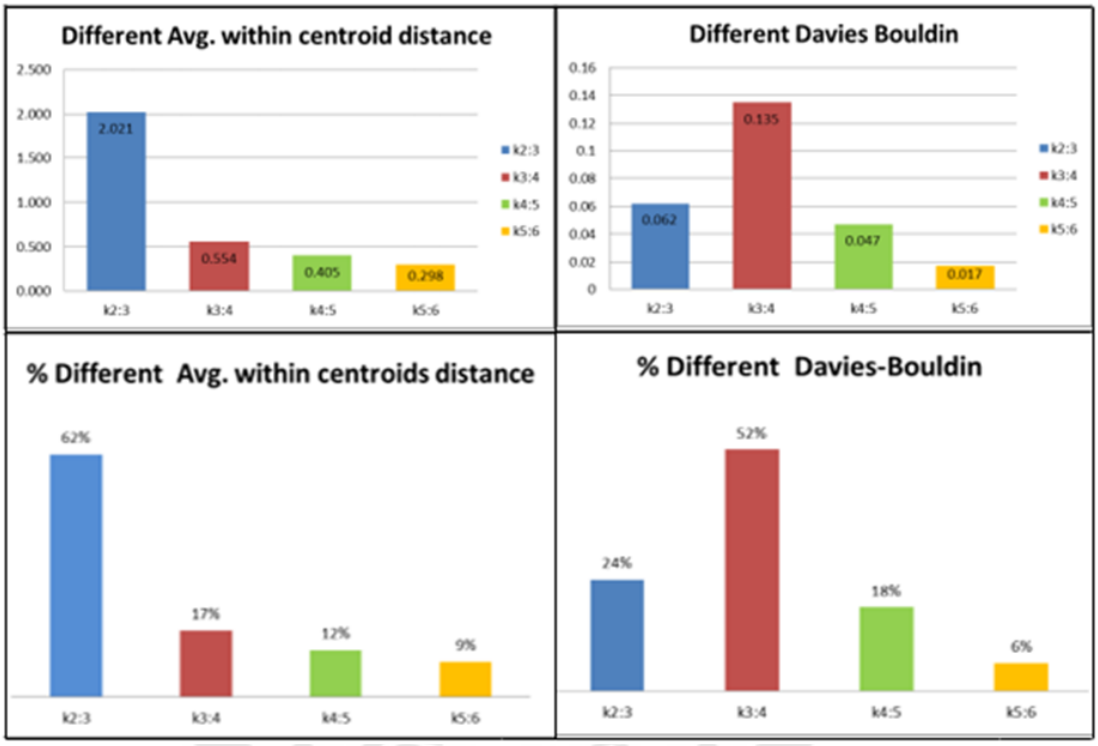
ภาพประกอบ 23 กราฟแสดงลักษณะ Avg. within centroids distance และค่า Davies-Bouldin

จากภาพประกอบที่ 23 ทางซ้ายจะแสดงลักษณะของข้อมูล Avg. within centroids distance ที่มีค่าเฉลี่ยระยะห่างจาก centroid กับข้อมูลในแต่ละกลุ่ม ที่ k = 2, 3, 4, 5 และ 6 โดยนำมาแสดงในรูปแบบกราฟเพื่อดูลักษณะของข้อมูล และทางขวาจะแสดงลักษณะของข้อมูลที่มีค่า Davies-Bouldin ที่ k = 2, 3, 4, 5 และ 6

โดยภาพด้านซ้ายจะเห็นว่า Avg. within centroids distance มีขนาดเล็กลงเมื่อจำนวนกลุ่มเพิ่มขึ้น ในอีกด้านหนึ่งทางขวา ตัวเลขค่าของ David Bouldn เพิ่มขึ้นต่อเนื่องเมื่อจำนวน k เพิ่มขึ้น ซึ่งผลลัพธ์ของทั้ง Avg. within centroids distance และ Davies-Bouldin มีความขัดแย้งกันจึงไม่สามารถที่จะสรุปได้ว่าจำนวนกลุ่มที่เหมาะสมในการทำ k-means และ แสดงให้เห็นว่าไม่มีตัวบ่งชี้ใดที่สามารถกำหนดจำนวนกลุ่มได้ จึงทำการวิเคราะห์ตรวจสอบเพิ่มเติม โดยพิจารณาความแตกต่างระหว่าง Performance vector ของ k ที่มีค่าแตกต่างกัน ดังผลการทดลองแสดงไว้ในภาพประกอบที่ 23 ที่สังเกตได้จากความลาดชันของกราฟ (อัตราการเปลี่ยนแปลงของ k ที่แตกต่างกัน) ที่รูปด้านซ้ายความลาดชันเปลี่ยนแปลงอย่างมากจาก k = 2 ถึง k = 3 ในขณะที่การเปลี่ยนแปลงของ David Bouldn จาก k = 3 เป็น k = 4 มีค่ามากที่สุด ดังแสดงในภาพประกอบที่ 24

หากไม่มีรู้เกี่ยวกับจำนวนกลุ่มสามารถเลือก k = 3 หรือ k = 4 เป็นจำนวนในการกำหนดกลุ่มข้อมูลที่ดีที่สุด เมื่อพิจารณาระยะห่างของแต่ละค่าจาก David Bouldn ที่ k = 4 อาจเป็นทางเลือกที่ดีที่สุด

อย่างไรก็ตามข้อมูลชุดนี้ระบุว่ามี label เพียง 3 กลุ่มเท่านั้น ซึ่งค่า Avg. within centroids distance อาจเป็นตัวชี้วัดที่ดีที่สุดสำหรับชุดข้อมูลนี้



ภาพประกอบ 24 การเปรียบเทียบค่า Different Performance Vector index

จากภาพประกอบที่ 24 จะแสดงข้อมูลผลต่างระหว่างกลุ่ม เพื่อใช้ประกอบการวิเคราะห์ในการกำหนดจำนวนในการแบ่งกลุ่ม โดยทางซ้ายจะแสดงค่า Different Avg. within centroids distance ของกลุ่มที่ดีที่สุด คือ สีฟ้า ช่วง k=2 กับ k=3 โดยมีค่าผลต่างเท่ากับ 2.021 หรือ 62% และ ส่วนทางขวาจะแสดงค่า Different Davies-Bouldin ซึ่งกลุ่มที่ดีที่สุด คือ สีแดง ช่วง k=3 กับ k=4 มีค่าผลต่างเท่ากับ 0.135 หรือ 52%

จะเห็นได้ว่า ค่าความแตกต่างของ Avg. within centroids distance ที่ k=2 กับ k=3 มีค่ามากที่สุดจำนวนกลุ่มที่มีความเหมาะสมที่จะนำมาใช้ในการแบ่งกลุ่ม คือ k=3 ในทางเดียวกันเมื่อดูที่ Davies-Bouldin มีค่าน้อยที่สุด จะทำให้ได้การแบ่งแยกของกลุ่มดีที่สุด และพบว่า ค่าความ

แตกต่างของ Davies-Bouldin มีค่าเพิ่มมากขึ้นที่ k=3 กับ k=4 ทำให้ค่า k ที่เหมาะสมจึงเป็นค่า k=3 เช่นเดียวกัน

4.1.2. เปรียบเทียบความถูกต้องของการแบ่งกลุ่มที่ K ต่าง ๆ

การทดลองนี้ดำเนินการ เพื่อยืนยันผลที่ได้จากการเลือกค่า k ที่เหมาะสมโดยวัดจากค่าความถูกต้องในการแบ่งกลุ่ม เพื่อตรวจว่าสามารถจัดกลุ่มได้อย่างถูกต้องหรือไม่ โดยเปรียบเทียบค่า k = 2,3,4,5,6

k=2

cluster	kmeans	TRUE	error
0	126	70	56
1	84	69	15
sum	210	139	71

k=3

cluster	kmeans	TRUE	error
0	61	60	1
1	77	68	9
2	72	60	12
sum	210	188	22

k=4

cluster	kmeans	TRUE	error
0	60	44	16
1	56	54	2
2	54	54	0
3	40	16	24
sum	210	168	42

k=5

cluster	kmeans	TRUE	error
0	43	30	13
1	42	40	2
2	31	22	9
3	48	48	0
4	46	46	0
sum	210	186	24

k=6

cluster	kmeans	TRUE	error
0	44	30	14
1	42	40	2
2	33	33	0
3	49	47	2
4	27	20	7
5	15	15	0
sum	210	185	25

ภาพประกอบ 25 การทำ K- Means Clustering มีค่า k=2, k=3, k=4, k=5, k=6

จากภาพประกอบที่ 25 มีการแบ่งกลุ่ม K- Means Clustering ที่มีค่า k=2, k=3, k=4, k=5, k=6 จะเห็นได้ว่า เลข Cluster label อาจไม่ตรงกับเลข Cluster ที่โปรแกรมรายงาน ดังนั้น การตัดสินใจว่า ความถูกต้องของการจัดกลุ่ม จึงดูที่จำนวนเมล็ดที่มี Cluster label ไปอยู่ในกลุ่มใดมากกว่าจะถือว่า เลข Cluster label นั้นตรงกับเลข Cluster ที่โปรแกรมรายงาน โดยคาดหวังว่าหากเพิ่มค่า k แล้ว เมล็ดควรอยู่แค่ 3 กลุ่ม และ มีเมล็ดไปปนในกลุ่มอื่นเล็กน้อย

ในการแบ่งกลุ่ม K- Means Clustering ได้แสดงผลข้อมูลดังนี้

- ที่ k=2 ได้จัดกลุ่ม K- Means Clustering เป็น 2 cluster โดย

ในกลุ่มที่ 1 cluster_0 มี 126 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 70 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 56 เมล็ด

ในกลุ่มที่ 2 cluster_1 มี 84 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 69 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 15 เมล็ด

- ที่ k=3 ได้จัดกลุ่ม K- Means Clustering เป็น 3 cluster โดย

ในกลุ่มที่ 1 cluster_0 มี 61 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 60 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 1 เมล็ด

ในกลุ่มที่ 2 cluster_1 มี 77 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 68 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 9 เมล็ด

ในกลุ่มที่ 3 cluster_2 มี 72 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 60 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 12 เมล็ด

- ที่ k=4 ได้จัดกลุ่ม K- Means Clustering เป็น 4 cluster โดย

ในกลุ่มที่ 1 cluster_0 มี 60 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 44 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 16 เมล็ด

ในกลุ่มที่ 2 cluster_1 มี 56 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 54 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 2 เมล็ด

ในกลุ่มที่ 3 cluster_2 มี 54 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 54 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 0 เมล็ด

ในกลุ่มที่ 4 cluster_3 มี 40 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 16 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_3 อยู่ 24 เมล็ด

- ที่ k=5 ได้จัดกลุ่ม K- Means Clustering เป็น 5 cluster โดย

ในกลุ่มที่ 1 cluster_0 มี 43 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 30 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 13 เมล็ด

ในกลุ่มที่ 2 cluster_1 มี 42 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 40 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 2 เมล็ด

ในกลุ่มที่ 3 cluster_2 มี 31 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 22 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 9 เมล็ด

ในกลุ่มที่ 4 cluster_3 มี 48 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 48 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_3 อยู่ 0 เมล็ด

ในกลุ่มที่ 5 cluster_4 มี 46 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 46 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_4 อยู่ 0 เมล็ด

- ที่ k=6 ได้จัดกลุ่ม K- Means Clustering เป็น 6 cluster โดย

ในกลุ่มที่ 1 cluster_0 มี 44 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 30 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 14 เมล็ด

ในกลุ่มที่ 2 cluster_1 มี 42 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 40 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 2 เมล็ด

ในกลุ่มที่ 3 cluster_2 มี 33 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 33 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 0 เมล็ด

ในกลุ่มที่ 4 cluster_3 มี 49 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 47 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_3 อยู่ 2 เมล็ด

ในกลุ่มที่ 5 cluster_4 มี 27 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 20 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_4 อยู่ 7 เมล็ด

ในกลุ่มที่ 6 cluster_5 มี 15 เมล็ด โดยมีเมล็ดพันธุ์ที่มีความถูกต้อง 15 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_5 อยู่ 0 เมล็ด

เมื่อทำการแบ่งกลุ่มด้วยค่า k=2, k=3, k=4, k=5, k=6 จะเห็นได้ว่า ค่า k=3 มีเมล็ดพันธุ์ชนิดอื่นมาปนค่าอยู่น้อยที่สุด จึงมีความเหมาะสมที่จะใช้ในการวิเคราะห์ข้อมูล

4.2 การทำ feature selection

หลังจากได้ทำการแบ่งกลุ่มด้วย k-mean และเลือก k=3 เป็นค่าที่ดีที่สุด จะมาทำการศึกษาความแตกต่างของ Attribute ต่าง ๆ ที่ k=3 เพื่อดูว่า Attribute ใดน่าจะเป็นตัวบ่งชี้หลักในการแบ่งกลุ่ม ซึ่งอาจทำให้สามารถตัด Attribute บางตัวออกจากการทำ clustering ซึ่งทำให้ความซับซ้อนของการคำนวณลดลง

4.2.1 การใช้สัมประสิทธิ์ของการแปรผัน (coefficient of variation)

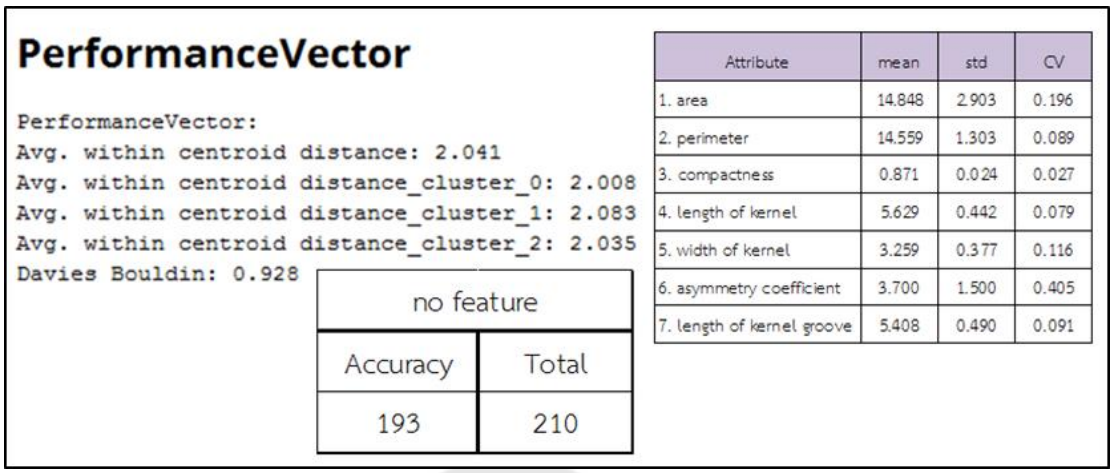
การใช้ค่า CV (Coefficient of variation) เพื่อเปรียบเทียบการกระจายระหว่างข้อมูลที่มากกว่าหนึ่งชุด โดยใช้การเปรียบเทียบการกระจายของข้อมูลระหว่างชุด เพื่อใช้เป็นเกณฑ์ในการเลือก Attribute เนื่องจากเป็นค่าที่ชี้ให้เห็นถึงความแตกต่างของค่าระยะห่างระหว่างกลุ่มข้อมูลของในแต่ละ Attribute ซึ่งค่า CV (Coefficient of variation) ที่มีค่าน้อยจะแสดงให้เห็นว่า ลักษณะของข้อมูลในแต่ละ Attribute มีความแตกต่างกันน้อย จึงไม่เหมาะที่จะนำ Attribute นั้นมาใช้เป็นเกณฑ์ในการแบ่งกลุ่ม

Attribute	Mean	Std	CV
1. area	14.848	2.903	0.196
2. perimeter	14.559	1.303	0.089
3. compactness	0.871	0.024	0.027
4. length of kernel	5.629	0.442	0.079
5. width of kernel	3.259	0.377	0.116
6. asymmetry coefficient	3.700	1.500	0.405
7. length of kernel groove	5.408	0.490	0.091

ภาพประกอบ 26 ค่าของ Coefficient of variation ในแต่ละ Attribute

จากภาพประกอบที่ 26 แสดงค่าของ Mean, Standard Deviation และ Coefficient of variation ในแต่ละ Attribute เพื่อใช้ในการวิเคราะห์ลักษณะข้อมูลทางด้าน (feature selection)

จะเห็นได้ว่า Attribute ที่ 3 ชื่อ “compactness” มีแนวโน้มที่จะตัดออกเป็นอันดับแรก เนื่องจากมีค่าของ Coefficient of variation เท่ากับ 0.027 ต่ำที่สุดซึ่งผลลัพธ์ที่ได้มีค่าน้อยที่สุดเมื่อเทียบกับ Attribute อื่น



PerformanceVector

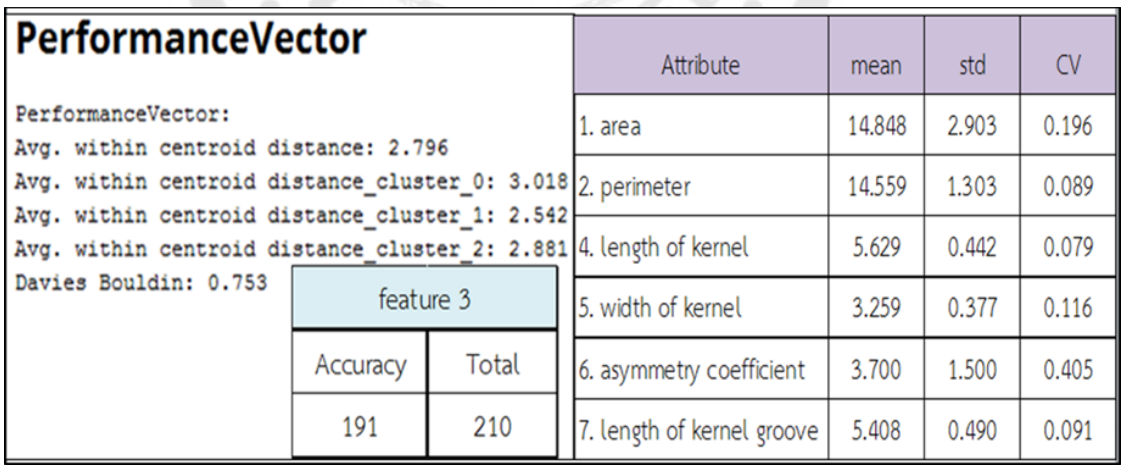
```
PerformanceVector:
Avg. within centroid distance: 2.041
Avg. within centroid distance_cluster_0: 2.008
Avg. within centroid distance_cluster_1: 2.083
Avg. within centroid distance_cluster_2: 2.035
Davies Bouldin: 0.928
```

no feature	
Accuracy	Total
193	210

Attribute	mean	std	CV
1. area	14.848	2.903	0.196
2. perimeter	14.559	1.303	0.089
3. compactness	0.871	0.024	0.027
4. length of kernel	5.629	0.442	0.079
5. width of kernel	3.259	0.377	0.116
6. asymmetry coefficient	3.700	1.500	0.405
7. length of kernel groove	5.408	0.490	0.091

ภาพประกอบ 27 การไม่ทำ feature selection

จากภาพประกอบที่ 27 แสดงให้เห็นถึง ค่า coefficient of variation ที่ควรตัดออกนั้น คือ Attribute ชื่อว่า “compactness” เนื่องจาก มีค่า CV เท่ากับ 0.027 เมื่อดูค่า Accuracy มีลักษณะการแบ่งกลุ่มของข้อมูลถูกต้อง193 จากเมล็ดพันธุ์ทั้งหมด 210 เมล็ด ซึ่งค่า Davies Bouldin มีค่าเท่ากับ 0.928 เมื่อตัด Attribute ออกแล้วจึงทำการวิเคราะห์การแบ่งกลุ่ม โดยไม่ใช้ Attribute “compactness” ทำให้ได้ผลการวิเคราะห์ ดังแสดงในภาพประกอบที่ 26



PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 2.796
Avg. within centroid distance_cluster_0: 3.018
Avg. within centroid distance_cluster_1: 2.542
Avg. within centroid distance_cluster_2: 2.881
Davies Bouldin: 0.753
```

feature 3	
Accuracy	Total
191	210

Attribute	mean	std	CV
1. area	14.848	2.903	0.196
2. perimeter	14.559	1.303	0.089
4. length of kernel	5.629	0.442	0.079
5. width of kernel	3.259	0.377	0.116
6. asymmetry coefficient	3.700	1.500	0.405
7. length of kernel groove	5.408	0.490	0.091

ภาพประกอบ 28 การทำ feature selection ตัด “compactness”

จากภาพประกอบที่ 28 แสดงผลการทำ feature selection ที่ทำการตัด Attribute ชื่อว่า “compactness” เมื่อดูลักษณะการแบ่งกลุ่มของ โดย ดูจากค่า Avg. within centroid distance

มีค่าลดลง ซึ่งสอดคล้องกับ ค่า Davies Bouldin ที่ลดลงเช่นเดียวกัน แสดงให้เห็นว่า การทำ feature selection ส่งผลให้การแบ่งกลุ่มของข้อมูลได้มีความชัดเจนมากยิ่งขึ้น แต่พบว่าค่า Performance vector มีค่าที่ดีขึ้น ถึงแม้ค่า Accuracy ของการแบ่งกลุ่มของข้อมูลถูกต้อง 191 จาก เมล็ดพันธุ์ทั้งหมด 210 เมล็ดก็ตาม

หลังจากนั้นทำการทดลองการทำ feature selection ด้วยการตัด Attribute อื่น ๆ โดยตัดเพิ่มขึ้นครั้งละ 1 Attribute ซึ่งเรียงลำดับในการตัดจากค่า Coefficient of variation ที่มีค่าน้อยที่สุด ดังภาพประกอบที่27 เพื่อเปรียบเทียบผลที่ได้จาการแบ่งกลุ่ม รวมถึงค่าดัชนีการประเมินของ Davies-Bouldin และ Avg. within centroid distance

4.3 ดัชนีการประเมินผลของ Davies-Bouldin และ Avg_within_centroid_distance

Performance Vector

Avg. centroid distance	Avg. centroid distance	Avg. centroid distance	Avg. centroid distance	Avg. centroid distance	Avg. centroid distance
2.041	1.475	1.269	1.112	0.886	0.219
Davies Bouldin	**Davies Bouldin**	**Davies Bouldin**	**Davies Bouldin**	**Davies Bouldin**	**Davies Bouldin**
0.928	0.845	0.837	0.830	0.811	0.775

No Cut feature		feature 3		feature 3,4		feature 3,4,2		feature 3,4,2,7		feature 3,4,2,7,5	
Accuracy	Total	Accuracy	Total	Accuracy	Total	Accuracy	Total	Accuracy	Total	Accuracy	Total
193	210	191	210	191	210	190	210	189	210	184	210

Cluster	Error	Cluster	Error	Cluster	Error	Cluster	Error	Cluster	Error	Cluster	Error
0	4	0	10	0	10	0	8	0	9	0	13
1	5	1	4	1	5	1	8	1	8	1	8
2	8	2	5	2	4	2	4	2	4	2	5

ภาพประกอบ 29 การเปรียบเทียบผลการทำ feature selection ด้วย Coefficient of variation

จากภาพประกอบที่ 29 แสดงการเปรียบเทียบผลการทำ feature selection จะเห็นได้ว่า มีจำนวน Error ใน cluster_0 เพิ่มขึ้นจาก 4 จำนวน เป็น 13 จำนวน, cluster_1 ลดลงจาก 5 จำนวน เหลือเป็น 4 และ cluster_2 ลดลงจาก 8 จำนวน เหลือเป็น 5 จำนวน เมื่อตัด Attribute ชื่อ" compactness", "length of kernel", "perimeter", "length of kernel groove" และ "width of kernel" ตามลำดับ จะได้ค่า Davies Bouldin ที่มีค่าน้อยลงเมื่อทำ feature selection จึงส่งผลให้

การแบ่งกลุ่มของข้อมูลมีความแม่นยำมากขึ้น โดยการทำ Coefficient of variation จะเป็นตัววัดที่ดีสำหรับการตัดสินใจเกี่ยวกับการเลือก Attribute มีดังนี้

ครั้งที่ 1 ทำการตัด Attribute ชื่อ “compactness”

เนื่องจากมีค่าเป็น 0 ทำให้ค่า Davies Bouldin มีค่าเท่ากับ 0.845 ซึ่งดีกว่าเดิม 0.083

ครั้งที่ 2 ทำการตัด Attribute ชื่อ “length of kernel”

เนื่องจากมีค่าเป็น 0 ทำให้ค่า Davies Bouldin มีค่าเท่ากับ 0.837 ซึ่งมีค่าเท่าเดิม 0.008

ครั้งที่ 3 ทำการตัด Attribute ชื่อ “perimeter”

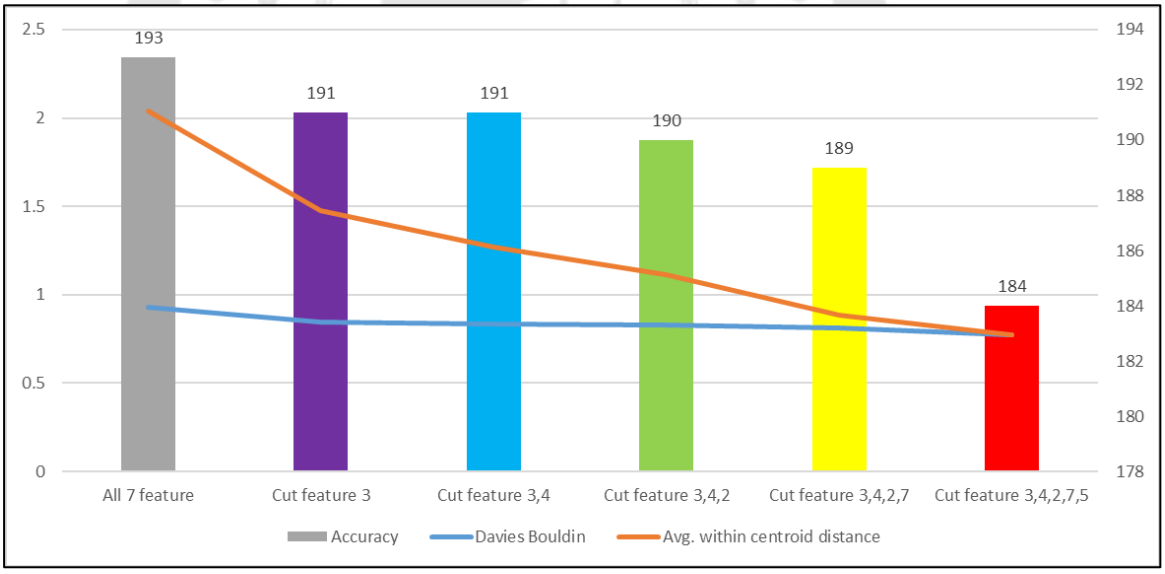
เนื่องจากมีค่าเป็น 0 ทำให้ค่า Davies Bouldin มีค่าเท่ากับ 0.830 ซึ่งดีกว่าเดิม 0.007

ครั้งที่ 4 ทำการตัด Attribute ชื่อ “length of kernel”

เนื่องจากมีค่าเป็น 0 ทำให้ค่า Davies Bouldin มีค่าเท่ากับ 0.811 ซึ่งดีกว่าเดิม 0.019

ครั้งที่ 5 ทำการตัด Attribute ชื่อ “width of kernel”

เนื่องจากมีค่าเป็น 0 ทำให้ค่า Davies Bouldin มีค่าเท่ากับ 0.775 ซึ่งดีกว่าเดิม 0.036



ภาพประกอบ 30 การเปรียบเทียบผลการไม่ทำ feature selection กับทำ feature selection

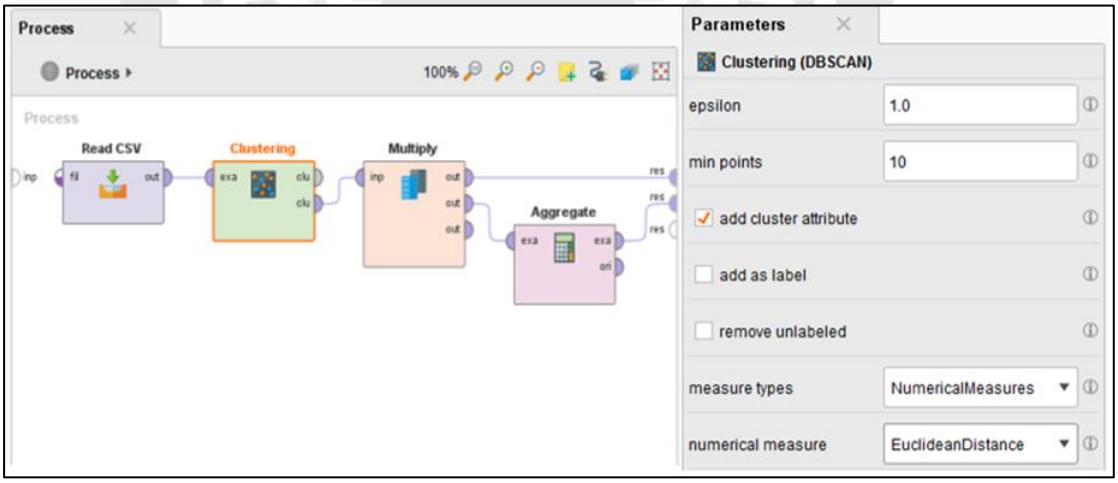
อีกทั้งทำการเปรียบเทียบ Performance vector ด้วยดัชนีการประเมินผลของ Davies-Bouldin, ค่า Avg. within centroid distance และเปรียบเทียบค่าความถูกต้อง หลังจากที่ทำ feature selection โดย การตัด Attribute ที่นำมาใช้ในการวิเคราะห์ทำให้ค่าของ Performance

vector มีค่าที่ดีขึ้นตามลำดับ ดังภาพประกอบที่ 30 ซึ่งตรงข้ามกับค่าความถูกต้องในการจำแนกเมล็ดพันธุ์ที่มีค่าลดน้อยลง

จะเห็นได้ว่าการทำ feature selection โดยใช้ Coefficient of variation ทำให้การแบ่งกลุ่มชัดเจนมากขึ้น โดย แสดงได้จากค่า Avg. within centroid distance และค่า Davies Bouldin ที่ลดลง แต่ความถูกต้องลดลง เนื่องจากการตัด Attribute ออกทำให้มีมีคุณลักษณ์ที่ใช้ในการจำแนกเมล็ดพันธุ์ข้าวได้ลดลง

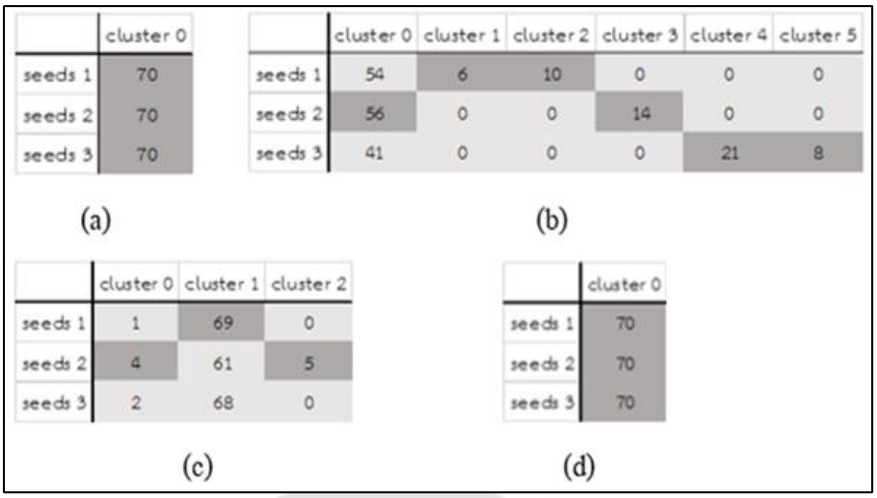
4.4 เทคนิค DBSCAN Clustering

จากนั้นทำการทดสอบชุดข้อมูล Seeds data set ด้วยเทคนิค DBSCAN เพื่อทำการเปรียบเทียบความแม่นยำในการแบ่งกลุ่มของข้อมูลเมล็ดพันธุ์ข้าวสาลี ทั้ง 3 สายพันธุ์ โดยเทคนิค DBSCAN Clustering เป็นเทคนิคในการทำ Clustering ซึ่งพิจารณาจำนวนข้อมูลที่อยู่ในรัศมีที่กำหนด โดยมีพารามิเตอร์ที่สำคัญดังนี้Epsilon (แอพสลอน) คือ การระบุรัศมีรอบข้อมูล และ min points คือ เป็นการระบุจำนวนข้อมูลในรัศมี ซึ่งมีลักษณะดังภาพประกอบที่ 31



ภาพประกอบ 31 การทำ DBSCAN Clustering ใน Rapid Miner Studio7.6

ในการทดลอง DBSCAN ได้ทำการกำหนดค่า parameters ต่าง ๆ เพื่อศึกษาผลกระทบของ parameter ต่อการจัดกลุ่ม ซึ่งทำการเลือก measure types เป็น Numerical Measures เนื่องจากลักษณะข้อมูลเป็นแบบตัวเลข โดยใช้การวัดความแตกต่างด้วย Euclidean distance และ Cosine Similarity ผลการทดลองดังแสดงในภาพประกอบที่ 31 และ 32



	cluster 0
seeds 1	70
seeds 2	70
seeds 3	70

(a)

	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
seeds 1	54	6	10	0	0	0
seeds 2	56	0	0	14	0	0
seeds 3	41	0	0	0	21	8

(b)

	cluster 0	cluster 1	cluster 2
seeds 1	1	69	0
seeds 2	4	61	5
seeds 3	2	68	0

(c)

	cluster 0
seeds 1	70
seeds 2	70
seeds 3	70

(d)

ภาพประกอบ 32 การเปรียบเทียบ DBSCAN ด้วย “Euclidean distance” ที่ Epsilon 0, 0.5, 1 และ 2 โดย minpoint เท่ากับ 5

จากภาพประกอบที่ 32 ผลการจัดกลุ่มด้วย DBSCAN โดยใช้ Euclidean Distance เมื่อค่า Epsilon, minpoint set เป็น (a) 0,5 (b) 0.5,5 (c) 1,5 (d) 2,5 ตามลำดับ จะเห็นได้ว่า ในการแบ่งกลุ่ม DBSCAN Clustering ได้แสดงผล Confusion matrix ของข้อมูลดังนี้

- เมื่อ (a) ที่ Epsilon 0, minpoint 5 ได้จัดกลุ่ม DBSCAN เป็น 1 cluster โดย ในกลุ่มที่ 1 cluster_0 มีเมล็ดพันธุ์ที่ 1, 2 และ 3 อยู่ทั้งหมด 210 เมล็ด โดยมีเมล็ดพันธุ์ละ 70 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 140 เมล็ด
- เมื่อ (b) ที่ Epsilon 0.5, minpoint 5 ได้จัดกลุ่ม DBSCAN เป็น 6 cluster โดย ในกลุ่มที่ 1 cluster_0 มีเมล็ดพันธุ์อยู่ทั้งหมด 151 เมล็ด โดยมีเมล็ดพันธุ์ที่ 2 อยู่มากที่สุด จำนวน56 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 140 เมล็ด
ในกลุ่มที่ 2 cluster_1 มีเมล็ดพันธุ์อยู่ทั้งหมด 6 เมล็ด โดยมีเมล็ดพันธุ์ที่ 1 อยู่มากที่สุด จำนวน 6 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 0 เมล็ด
ในกลุ่มที่ 3 cluster_2 มีเมล็ดพันธุ์อยู่ทั้งหมด 10 เมล็ด โดยมีเมล็ดพันธุ์ที่ 1 อยู่มากที่สุด จำนวน 10 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 0 เมล็ด
ในกลุ่มที่ 4 cluster_3 มีเมล็ดพันธุ์อยู่ทั้งหมด 14 เมล็ด โดยมีเมล็ดพันธุ์ที่ 2 อยู่มากที่สุด จำนวน 14 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_3 อยู่ 0 เมล็ด
ในกลุ่มที่ 5 cluster_4 มีเมล็ดพันธุ์อยู่ทั้งหมด 21 เมล็ด โดยมีเมล็ดพันธุ์ที่ 3 อยู่มากที่สุด จำนวน 21 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_4 อยู่ 0 เมล็ด

ในกลุ่มที่ 6 cluster_5 มีเมล็ดพันธุ์อยู่ทั้งหมด 8 เมล็ด โดยมีเมล็ดพันธุ์ที่ 3 อยู่มากที่สุด จำนวน 8 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_5 อยู่ 0 เมล็ด

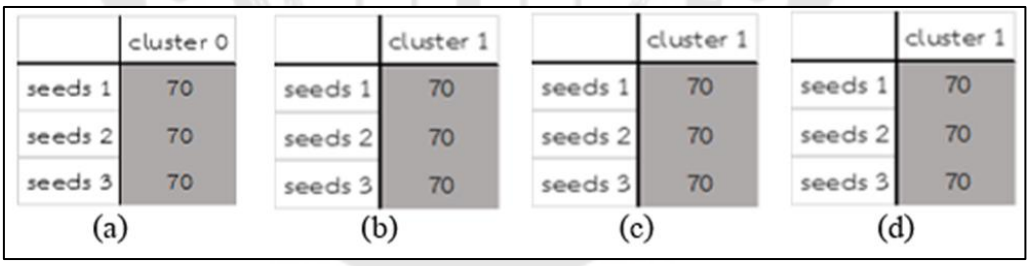
- เมื่อ (c) ที่ Epsilon 1, minpoint 5 ได้จัดกลุ่ม DBSCAN เป็น 3 cluster โดย

ในกลุ่มที่ 1 cluster_0 มีเมล็ดพันธุ์อยู่ทั้งหมด 7 เมล็ด โดยมีเมล็ดพันธุ์ที่ 2 อยู่มากที่สุด จำนวน 4 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 3 เมล็ด
ในกลุ่มที่ 2 cluster_1 มีเมล็ดพันธุ์อยู่ทั้งหมด 198 เมล็ด โดยมีเมล็ดพันธุ์ที่ 1 อยู่มากที่สุด จำนวน 69 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_1 อยู่ 129 เมล็ด
ในกลุ่มที่ 3 cluster_2 มีเมล็ดพันธุ์อยู่ทั้งหมด 5 เมล็ด โดยมีเมล็ดพันธุ์ที่ 2 อยู่มากที่สุด จำนวน 5 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_2 อยู่ 0 เมล็ด

- เมื่อ (d) ที่ Epsilon 2, minpoint 5 ได้จัดกลุ่ม DBSCAN เป็น 1 cluster โดย

ในกลุ่มที่ 1 cluster_0 มีเมล็ดพันธุ์ที่ 1, 2 และ 3 อยู่ทั้งหมด 210 เมล็ด โดยมีเมล็ดพันธุ์ละ 70 เมล็ด ส่งผลให้มีเมล็ดพันธุ์ชนิดอื่นมาปน (ค่า error) อยู่ใน cluster_0 อยู่ 140 เมล็ด

อีกทั้งทำการทดสอบด้วยการเลือกพารามิเตอร์ measure type เป็น Cosine Similarity จึงได้ผลลัพธ์ ดังภาพประกอบที่ 33



ภาพประกอบ 33 การเปรียบเทียบ DBSCAN ด้วย “Cosine Similarity” ที่ Epsilon 0, 0.5, 1 และ 2 โดย minpoint เท่ากับ 5

จากภาพประกอบที่ 33 ได้กำหนดพบว่า Numerical Measures เป็น “Cosine Similarity” โดย Epsilon เป็น 0.0, 0.5, 1.0 และ2.0 และ กำหนด min pints 5 จุด ซึ่งผลลัพธ์ที่ได้จากการทดสอบ มีจำนวน cluster เพียง 1 กลุ่ม จึงทำให้ทราบว่า พารามิเตอร์ Numerical Measures แบบ “Euclidean distance” ซึ่งเป็นการคำนวณการวัดระยะห่างระหว่างจุดที่ได้ผลลัพธ์ในการแบ่งกลุ่ม cluster ได้ดีกว่าแบบ “Cosine Similarity” ที่ความแตกต่างที่กำหนดได้จากการคำนวณการวัดความคล้ายคลึงของมุมโคไซน์ระหว่าง 2 เวกเตอร์

	cluster 0	cluster 1	cluster 2
seeds 1	1	69	0
seeds 2	4	61	5
seeds 3	2	68	0

ภาพประกอบ 34 ผลการทำ DBSCAN Clustering ที่ min points เป็น 5, Epsilon เป็น 1.0

จากภาพประกอบที่ 34 จะแสดงถึง การแบ่งกลุ่มของข้อมูลเป็น 3 กลุ่ม ได้ 3 cluster คือ 1.cluster_0 , 2.cluster_1 และ 3.cluster_2 ซึ่งสามารถวิเคราะห์การจำแนกเมล็ดพันธุ์ใน cluster ต่าง ๆได้ดังนี้

ใน cluster_0 มีเมล็ดพันธุ์ชนิดที่ 1 อยู่ 1 เมล็ด, เมล็ดพันธุ์ชนิดที่ 2 อยู่ 4 เมล็ด และเมล็ดพันธุ์ชนิดที่ 3 อยู่ 2 เมล็ด ดังนั้น class of wheat ของ cluster_0 ควรเป็น เมล็ดพันธุ์ชนิดที่ 2 ซึ่งมากกว่าเมล็ดพันธุ์ชนิดที่ 1 และ 3 จึงมี จำนวน Error ของเมล็ดสายพันธุ์ที่ 1 มาปนใน cluster_0 อยู่ 1 เมล็ด และ เมล็ดสายพันธุ์ที่ 3 มาปนใน cluster_0 อยู่ 2 เมล็ด ทำให้มีความแม่นยำในการจัดกลุ่มเมล็ดพันธุ์ชนิดที่ 2 ถูกต้อง 4 เมล็ด

ใน cluster_1 มีเมล็ดพันธุ์ชนิดที่ 1 อยู่ 69 เมล็ด, เมล็ดพันธุ์ชนิดที่ 2 อยู่ 61 เมล็ด และเมล็ดพันธุ์ชนิดที่ 3 อยู่ 68 เมล็ด ดังนั้น class of wheat ของ cluster_1 ควรเป็น เมล็ดพันธุ์ชนิดที่ 1 ซึ่งมากกว่าเมล็ดพันธุ์ชนิดที่ 2 และ 3 จึงมี จำนวน Error ของเมล็ดสายพันธุ์ที่ 2 มาปนใน cluster_1 อยู่ 61 เมล็ด และ เมล็ดสายพันธุ์ที่ 3 มาปนใน cluster_1 อยู่ 68 เมล็ด ทำให้มีความแม่นยำในการจัดกลุ่มเมล็ดพันธุ์ชนิดที่ 1 ถูกต้อง 69 เมล็ด

ใน cluster_2 มีเมล็ดพันธุ์ชนิดที่ 2 เพียงชนิดเดียว อยู่ 5 เมล็ด ทำให้มีความแม่นยำในการจัดกลุ่มเมล็ดพันธุ์ชนิดที่ 2 ถูกต้อง 5 เมล็ด

จะเห็นได้ว่า การทำ DBSCAN Clustering มีความแม่นยำในการจำแนกเมล็ดพันธุ์เพียง 74 เมล็ด จากทั้งหมด 210 เมล็ด ดังภาพประกอยที่ 35

เทคนิค	ทั้งหมด	ความแม่นยำ	ผิดพลาด
การทำ k-means no feature selection	210	193	17
การทำ DBSCAN	210	74	136

ภาพประกอบ 35 การเปรียบเทียบผลระหว่างเทคนิค feature selection และ เทคนิค DBSCAN

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากการศึกษาผลการวิเคราะห์ การพัฒนาระบบการแบ่งกลุ่มเมล็ดพันธุ์ ด้วยการทำเหมืองข้อมูลในการแบ่งกลุ่ม K-means Clustering ทางด้านการศึกษาการเลือก feature selection ที่เหมาะสมในการแบ่งกลุ่ม, ศึกษาการใช้เทคนิคทาง K-means Clustering ในการแบ่งกลุ่มข้อมูล, ศึกษาการใช้เทคนิค DBSCAN Clustering ในการแบ่งกลุ่มข้อมูลและ การศึกษาการใช้ Rapid Miner Studio7.6 เป็นเครื่องมือในการวิเคราะห์การแบ่งกลุ่มข้อมูล สามารถนำมาประยุกต์ใช้ร่วมกันในการทดสอบได้ จะเห็นได้ว่า การแบ่งกลุ่มด้วยค่า k=3 มีค่า error น้อยที่สุด จึงมีความเหมาะสมที่จะใช้ในการวิเคราะห์ข้อมูล โดยใช้ Performance Vector เป็นตัววัดค่าความถูกต้องของการแบ่งกลุ่มข้อมูลที่จะแสดงค่า Davies Bouldin และ Avg. within centroid distance แต่เนื่องจากค่า Performance vector ที่ได้จาก Davies Bouldin และ Avg. within centroid distance ที่ได้มีค่าผันผวนกัน จึงใช้ค่า Different Performance vector เพื่อกำหนดค่า k ได้ถูกต้อง ถึงแม้ค่าที่ได้จากการทำ feature selection ด้วย k-means ที่มีค่าที่ดีขึ้นนั้น ไม่ได้เพิ่มค่าความถูกต้อง ส่งผลให้ค่า Performance vector ไม่ได้แสดงถึงความสามารถในการแบ่งกลุ่มที่ถูกต้องมากขึ้น จึงทำให้การ clustering ควรเลือกใช้ Performance vector ที่มีดัชนีประเมินผลที่หลากหลายชนิดร่วมกัน อีกทั้งทำการเปรียบเทียบผลการทดลองระหว่างการทำ k-means no feature selection กับเทคนิค DBSCAN พบว่า การทำ k-means no feature selection สามารถจำแนกเมล็ดพันธุ์ได้ความถูกต้องที่ดีกว่า ดังนั้นการจัดกลุ่มข้อมูลที่มีค่า k=3 เป็นการแบ่งกลุ่มที่ดีที่สุด และเทคนิคการทำ K-means Clustering แบบ no feature selection ได้ค่าความแม่นยำในการจำแนกเมล็ดพันธุ์ ที่ถูกต้องถึง 193 เมล็ด และมีค่า Error ในการจำแนกเมล็ดพันธุ์น้อยที่สุด คือ 17 เมล็ด ซึ่งเป็นผลลัพธ์ที่ดีกว่าเทคนิคการทำ DBSCAN Clustering ที่มีค่าความแม่นยำในการจำแนกเมล็ดพันธุ์ ที่ถูกต้องเพียง 74 เมล็ด และมีค่า Error ในการจำแนกเมล็ดพันธุ์ คือ 136 เมล็ด โดยการแบ่งกลุ่มของข้อมูลด้วยเทคนิค feature selection จะมีความแม่นยำในการจำแนกเมล็ดพันธุ์ ที่ถูกต้องเพียง 184 เมล็ด และมีค่า Error ในการจำแนกเมล็ดพันธุ์ คือ 26 เมล็ด ซึ่งมีรายละเอียดดังนี้

1. การศึกษาการพัฒนาระบบการแบ่งกลุ่มเมล็ดพันธุ์ ด้วยการทำเหมืองข้อมูลในการแบ่งกลุ่มโดยใช้ โปรแกรม Rapid Miner Studio7.6 เป็นเครื่องมือที่ใช้ในการประมวลผลวิเคราะห์การแบ่งกลุ่มข้อมูล ที่ประโยชน์ต่อการศึกษาครั้งนี้เป็นอย่างมาก เนื่องจากเป็นมีความรวดเร็วและง่ายต่อผู้ใช้งาน เพียงเข้าใจถึงกระบวนการที่จะใช้สำหรับการวิเคราะห์การแบ่งกลุ่มด้วย k-means

2. การศึกษาการกำหนดจำนวนที่ใช้ในแบ่งกลุ่มใช้ Performance Vector เป็นตัววัดค่าความถูกต้องของการแบ่งกลุ่มข้อมูลที่จะแสดงค่า Davies Bouldin และ Avg. within centroid distance ด้วยค่า k=2, k=3, k=4, k=5 และ k=6 ซึ่งค่า Performance vector ที่ได้จาก Davies Bouldin และ Avg. within centroid distance ที่ได้มีค่าผันผวนกัน จึงใช้ค่า Different Performance vector เพื่อกำหนดค่า k ได้ถูกต้อง โดยจำนวนที่ใช้ในแบ่งกลุ่มที่ k=3 มีค่า error น้อยที่สุด จึงมีความเหมาะสมที่จะใช้ในการวิเคราะห์ข้อมูล

3. การศึกษาเทคนิคการเลือก ฟีเจอร์ (feature selection) เพื่อเพิ่มความแม่นยำในการแบ่งกลุ่ม ด้วยการใช้ค่า CV (Coefficient of variation) เพื่อเปรียบเทียบการกระจายระหว่างข้อมูล ในแต่ละ Attribute เพื่อใช้เป็นเกณฑ์ในการทำ feature selection นั้นคือ การตัด Attribute ชื่อ "compactness", "length of kernel", "perimeter", "length of kernel groove" และ "width of kernel" ตามลำดับ จะได้ค่า Performance vector ที่ใช้ดัชนีการประเมินผลของ Davies Bouldin และ Avg. within centroid distance มีค่าที่ดีขึ้นเมื่อทำ feature selection ซึ่งทำให้จำนวน Error ลด cluster_1 ลดลงจาก 5 จำนวน เหลือเป็น 4 และ cluster_2 ลดลงจาก 8 จำนวน เหลือเป็น 5 จำนวน แต่ว่าจำนวน Error ในcluster_0 เพิ่มขึ้นจาก 4 จำนวน เป็น 13 ส่งผลให้จำนวน Error โดยรวมเพิ่มขึ้นจาก 17 เป็น 26 จำนวน

4. การศึกษาเทคนิค DBSCAN Clustering เพื่อทำการเปรียบเทียบความแม่นยำในการแบ่งกลุ่มของข้อมูลเมล็ดพันธุ์ข้าวสาลี ทั้ง 3 สายพันธุ์ โดยสามารถจำแนกเมล็ดพันธุ์ได้ถูกต้องเพียง 74 เมล็ดและมีการแบ่งกลุ่มที่ผิดพลาดถึง 136 เมล็ด ซึ่ง เทคนิค k-means no feature selection จำแนกเมล็ดพันธุ์ข้าวสาลีทั้ง 3 ชนิด ได้ถูกต้องแม่นยำกว่า เทคนิคการทำ DBSCAN Clustering

5.2 ข้อเสนอแนะ

จะทำการศึกษาและประยุกต์ใช้กับข้อมูลทางด้านอื่นที่ต้องการแบ่งกลุ่มโดยไม่มีการระบุชนิดของข้อมูล (Unsupervised learning) ซึ่งยังเป็นสิ่งที่ท้าทายในทางปฏิบัติเนื่องจากความแตกต่างของคุณลักษณะ (attribute) ของข้อมูลที่ต่างกัน

ในการเลือก feature พบว่า Performance vector ที่ดีขึ้นไม่ได้แสดงถึงความสามารถในการแบ่งกลุ่มที่ถูกต้องมากขึ้น จึงแนะนำว่าการทำ clustering ควรเลือกใช้ Performance vector ที่หลากหลายชนิดร่วมกัน

บรรณานุกรม

1. การปฏิรูประบบการเกษตร การพัฒนาศูนย์กลางการผลิตเมล็ดพันธุ์ [ออนไลน์]. เข้าถึงได้จาก : http://www.library.senate.go.th/document/Ext10567/10567739_0002.PDF(วันที่ค้นข้อมูล : 27 ธันวาคม 2560).

2. ภัทธิรา สุวรรณโค; นิศาชล จานงศรี; และ จิติมนต์ อั่งสกุล. (2017). แบบจำลองการพยากรณ์ความเสี่ยงในการเกิดอุบัติเหตุทางถนนในเทศกาลปีใหม่. วารสารวิทยาการและเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีมหานคร (JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY). 2(7): 10-19.

3. ศิริกาญจนา พิลาบุตร. (2016). การหาความสัมพันธ์ที่มีผลต่อการสำเร็จการศึกษาด้วยเทคนิคการจำแนกประเภทข้อมูล กรณีศึกษาวิทยาลัยนครราชสีมา. การประชุมวิชาการและนำเสนอผลงานวิจัยระดับชาติ ราชธานีวิชาการ. 1(1): 1854-1862.

4. กัลยา วานิชย์บัญชา. (2552). การวิเคราะห์ข้อมูลหลายตัวแปร. กรุงเทพมหานคร : ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย. พิมพ์ครั้งที่ 4

5. เรวดี ศักดิ์ดุลยธรรม. (2554). การใช้เทคนิคดาต้าไมน์นิงในการวิเคราะห์ปัจจัยที่มีผลต่อความสำเร็จในการรักษาโรคนิ้วล็อกในแบบต่างๆ ของคณะแพทยศาสตร์วชิรพยาบาลมหาวิทยาลัยกรุงเทพมหานคร : มหาวิทยาลัยราชพฤกษ์.

6. Dan Pelleg and Andrew W Moore. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. International Conference on Machine Learning : 727-734.

7. สุพจน์ เฮงพระพรหม และ ชนาธิป หมั่นเพียรสุข. (2558). ประสิทธิภาพของฟังก์ชันความเหมือนต่อขั้นตอนวิธีเพื่อนบ้านใกล้ที่สุดเค สำหรับการจำแนกประเภทข้อมูล. นครปฐม :มหาวิทยาลัยราชภัฏ.

8. วีระยุทธ พิมพาภรณ และ พยุง มีสัจ. (2014). การเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลโดยวิธีการเลือก ลักษณะสำคัญแบบพลวัตเพื่อเพิ่มประสิทธิภาพของอัลกอริทึม การจัดกลุ่มบนปริภูมิย่อย. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ. วารสารเทคโนโลยีสารสนเทศ (Information Technology Journal). 2(10). 43-51.

9. จันทร์จิรา พิลาแดง. (2558). การจัดกลุ่มแบบสองด้านโดยขั้นตอนวิธีเชิงพันธุกรรมเพื่อแบ่งกลุ่มระดับความเข้มแข็งของครอบครัวไทย. วิทยานิพนธ์. วิทยาศาสตรมหาบัณฑิต.ปทุมธานี: มหาวิทยาลัยธรรมศาสตร์.ถ่ายเอกสาร

10. ธรรมศักดิ์ เธียรนิเวศน์. (2548). การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่. วิทยานิพนธ์ วศ.บ. (คอม), นครราชสีมา: มหาวิทยาลัยเทคโนโลยีสุรนารี.ถ่ายเอกสาร

11. สุเมธ พิลึก. (2559). การวิเคราะห์องค์ประกอบสมรรถภาพทางวิชาชีพของผู้สำเร็จการศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจตามความต้องการของตลาดแรงงาน. ในสมาคมถาบันอุดมศึกษาเอกชนแห่งประเทศไทยในพระราชูปถัมภ์ สมเด็จพร ะเทพรัตนราชสุดาฯ สยามบรมราชกุมารี.22(1). 96-109.

12. จินดา สวัสดิ์ทวี. (2554). สัมประสิทธิ์การแปรผันของสองตัวแปรที่สัมพันธ์กัน. วารสารวิทยาศาสตร์ลาดกระบัง. 20(2). 72-80.

13. วีรศักดิ์ ช่องงูเหลือม. (2555). การจัดกลุ่มข้อมูลด้วยเทคนิคราฟเคมีดอยส์แบบขนาน บนหน่วยประมวลผลกลางแบบหลายแกนหลัก. วิทยานิพนธ์ วศ.บ. (คอม), นครราชสีมา: มหาวิทยาลัยเทคโนโลยีสุรนารี.ถ่ายเอกสาร

14. M . E. Celebi, H. A. Kingravi and P. A. Vela. (2013). A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. Expert Systems with Applications, 40(1). 200–210.

15. A. RezaParnian and R. Javidan. (2014). Autonomous Wheat Seed Type Classifier System. International Journal of Computer Applications. 96(12). 14–17.

16. Jia-zhou ZHENG (2016). Classification of Clothing Products Based on Principal Component Analysis and Clustering Algorithm. International Conference on Economics and Management.

17. วาทินี นุ้ยเพียร, ภรัณยา อำมฤครัตน์, เดช ธรรมศิริ, ณรงค์ โพธิ และพยุง มีสัจ. (2552). การเปรียบเทียบอัลกอริธึมการจัดกลุ่มข้อมูล Ozone day โดยใช้เทคนิคการทำเหมืองข้อมูล. National Conference on Computing and Information Technology. 5. 911-916.

ประวัติผู้เขียน

ชื่อ-สกุล	สุทธิพงศ์ ผ่องแผ้ว
วัน เดือน ปี เกิด	4 ธันวาคม 2535
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	ปริญญาตรี
ที่อยู่ปัจจุบัน	44 ม.9 ซ.ติวานนท์24แยก1/6 ถ.ติวานนท์ ต.บางกระสอ อ.เมือง จ.นนทบุรี 11000
ผลงานตีพิมพ์	-
รางวัลที่ได้รับ	-