Cluster analysis

Cluster analysis (CA)เป็นการวิเคราะห์เชิงสำรวจที่พยายามจะจัด case ที่มีลักษณะเหมือนหรือคล้ายคลึงกัน (homogeneous) เข้าอยู่ในกลุ่มเดียวกัน โดยcase ที่เป็นสมาชิกของกลุ่มเดียวกันจะมีความเหมือนหรือมีลักษณะที่ใกล้เคียงกัน และมีความแตกต่างจาก case ที่เป็นสมาชิกของกลุ่มอื่น ๆ

CA อาจมีชื่ออื่นได้แก่ segmentation analysis / taxonomy analysis/classification analysis หรือ numerical taxonomy

Case ที่จะใช้ CA ในการจัดกลุ่มอาจอยู่ในรูปผลที่ได้จากการสังเกต/ทดลอง ผู้ที่เข้ามามีส่วนร่วมผู้ตอบแบบสอบถาม คน สัตว์ สิ่งของ ทั้งนี้ case แต่ละ case จะประกอบด้วยตัวแปรแสดงลักษณะที่มีจำนวนตั้งแต่สองตัวแปรขึ้นไป โดยตัวแปรแสดงลักษณะแต่ละตัวจะบ่งบอกมิติใดมิติหนึ่งที่มีการใช้วัด ตัวอย่างเช่น

- ในกรณีของการจัดกลุ่มรถยนต์ case แต่ละcase จะประกอบด้วยตัวแปรบอกลักษณะประกอบด้วยประเภทของรถ (รถเก๋ง/รถบรรทุก/รถตู้/รถSUV) ราคา ขนาดของเครื่องยนต์ กำลังแรงม้า ฐานล้อ ความกว้าง ความยาวของตัวรถ น้ำหนักรถเปล่า ความจุน้ำมัน จำนวนไมล์ต่อแกลลอน เป็นต้น ฯ

- ในกรณีของการจัดกลุ่มผู้บริโภค case แต่ละ case จะประกอบด้วยตัวแปรบอกลักษณะ ประกอบด้วยเพศ อายุ การศึกษา รายได้ ลักษณะบ้านที่อยู่อาศัย ขนาดของครอบครัว จำนวนเงินที่ซื้อในแต่ละครั้ง ความถี่ในการจับจ่ายใช้สอย เป็นต้น ฯ

ตัวแปรที่แสดงลักษณะอาจเป็นตัวแปรที่มีค่าเพียงสองค่า (binary variable) ข้อมูลนามบัญญัติ (nominal data) หรือข้อมูลเรียงลำดับ (ordinal data)

การจัดกลุ่มโดยใช้ CA จะเริ่มจากฐานข้อมูลที่ประกอบด้วย case ต่าง ๆปะปนกัน โดยเราจะไม่ทราบเลยว่า case ใดสังกัดอยู่กลุ่มใด ที่สำคัญก็คือบ่อยครั้งผู้วิเคราะห์มักจะไม่ทราบจำนวนกลุ่มล่วงหน้า และอาจไม่สามารถกำหนดชื่อกลุ่มที่สามารถสื่อความหมายให้ชัดเจนได้ อาจทำได้เพียงแค่แยกแยะ case ที่มีมากมายทั้งหมด ออกเป็นกลุ่มหนึ่ง กลุ่มสอง กลุ่มสาม กลุ่มสี่เท่านั้น ฯ

CAไม่ได้ใช้วิธีการหรือแบบจำลองทางสถิติใด ๆเป็นการเฉพาะ และไม่ได้มีข้อสมมติฐานเกี่ยวกับการกระจายค่า (distribution) ของตัวแปรใด ๆเป็นการเฉพาะ

CA ถูกนำไปใช้ในหลากหลายวงการ ต่อไปนี้เป็นตัวอย่างของการนำ CA ไปใช้

วงการการตลาด

นักการตลาดใช้ CA ในการจัดกลุ่มลูกค้าออกเป็นกลุ่มหรือ segment โดยจะมีกลุ่มวัยรุ่นผู้นำเทรนด์ที่มีความหิวกระหายแนวโน้มแฟชั่นใหม่ล่าสุดตลอดเวลา กลุ่มคนวัยกลางคนที่ชื่นชอบกับดีไซน์ที่ไม่มีวันตกยุค และอื่น ๆ ทั้งนี้ก็เพื่อกำหนดกลยุทธ์ทางการตลาดโดยอาศัยข้อมูลเกี่ยวกับลักษณะการจับจ่ายใช้สอย อายุ ข้อมูลประชากรศาสตร์ และความชื่นชอบในสไตล์

วงการการให้ความบันเทิงด้านดิจิทัล

ธุรกิจที่ให้บริการภาพยนตร์ทางอินเตอร์เน็ตสามารถจัดแบ่งลูกค้าออกเป็นกลุ่มที่ชื่นชอบภาพยนตร์แนวโรมันติคแสนหวาน ในขณะที่อีกกลุ่มมีความชื่นชอบในภาพยนตร์ประเภทสยองขวัญ จากนั้นสามารถรณรงค์การโฆษณาที่มุ่งเน้นเสนอบริการหรือสินค้าที่ปลุกเร้าความต้องการให้เหมาะกับสิ่งที่ลูกค้าแต่ละกลุ่มชื่นชอบ

วงการสาธารณสุข

โรงพยาบาลสามารถใช้ CA ในการจัดกลุ่มผู้ป่วยออกเป็นกลุ่มตามความหนักเบาของอาการป่วยและการตอบสนองต่อการรักษา ทั้งนี้เพื่อให้สามารถนำเสนอรูปแบบแผนการรักษาที่มุ่งสนองความต้องการตลอดจนเหมาะสมกับสถานะของผู้ป่วยแต่ละท่าน

วงการสาธารณสุขที่เฝ้าระวังการอุบัติของโรคติดต่อ

นักวิจัยสามารถใช้ CA ในการกำหนดพื้นที่หรือกลุ่มประชากรที่มีความถี่ของโรคบางอย่างสูงเช่น ไข้หวัดใหญ่ ทั้งนี้เพื่อผู้รับผิดชอบงานด้านสาธารณสุขจะได้จัดสรรทรัพยากรและลงพื้นที่ปฏิบัติงานได้อย่างมีประสิทธิภาพ ให้ความทุ่มเทในพื้นที่ ๆมีความต้องการมากที่สุด

วงการสื่อสังคมออนไลน์

แพลตฟอร์มสื่อสังคมออนไลน์สามารถใช้ CA ในการจัดแบ่งผู้ใช้ออกเป็นกลุ่มๆตามระดับกิจกรรมไล่เรียงตั้งแต่กลุ่มแฟนพันธ์แท้ไปจนถึงพวกผู้ใช้ที่แวะเวียนเข้ามาบ้างบางครั้ง และใช้กลยุทธ์สำหรับแต่ละกลุ่ม โดยกลุ่มแรกอาจต้องใช้ proactive approach มีการแจ้งเตือนในเวลาที่ถูกที่ควรซึ่งจะช่วยดึงดูดความสนใจ ในขณะที่กลุ่มหลังอาจให้ข้อแนะนำหรือการโต้ตอบที่รวดเร็วเพื่อเป็นตัวกระตุ้นให้มีการเข้าร่วมในแพลตฟอร์ม

การวางแผนงานด้านสาธารณูปโภค

พนักงานเทศบาลสามารถใช้ CA ช่วยในการจัดสรรทรัพยากรให้เหมาะกับชุมชนต่าง ๆ ตัวอย่างเช่นชุมชนคนรุ่นหนุ่มสาวซึ่งเสาะแสวงหาความเพลิดเพลินและการเรียนรู้จะมีความต้องการสนามเด็กเล่น สนามกีฬา โรงเรียนที่มีกิจกรรมหลากหลาย ในขณะที่ชุมชนผู้สูงวัยชอบสภาพแวดล้อมที่เงียบสงบ มีห้องสมุดที่พวกเขาสามารถพักผ่อนหย่อนใจ

วงการอีคอมเมิร์ซ

บริษัทที่ทำธุรกิจด้าน e-commerce ใช้ CA ในการจัดกลุ่มลูกค้าที่มักจะซื้อสินค้าประเภทที่ผลิตจากสารอินทรีย์ และทำการส่งเสริมการขายในสินค้าที่ผลิตจากสารธรรมชาติเหล่านี้

วงการแพทย์

แพทย์สามารถใช้ CA ร่วมกับ scanner ในการตรวจสอบเนื้อเยื่อที่ส่อว่าจะกลายเป็นมะเร็งโดยดูจากขนาด รูปร่าง สีของเซลล์หรือเนื้อเยื่อจากภาพถ่ายทางการแพทย์

วงการให้ความบันเทิงทางด้านดิจิทัล

ธุรกิจเช่น Netflix , Spotify, You tube ใช้ CA ในการจัดกลุ่มลูกค้าจากพฤติกรรมการชมภาพยนตร์ ฟังเพลงและนำเสนอภาพยนต์ตามแนวที่ลูกค้าชื่นชอบ

วงการประกันสุขภาพ

นักคณิตศาสตร์ประกันภัยอาจจะเก็บข้อมูลของครัวเรือนเกี่ยวกับจำนวนครั้งที่ไปพบแพทย์ต่อปี ขนาดของครัวเรือน จำนวนสมาชิกในครัวเรือนที่มีอาการป่วยเรื้อรัง อายุโดยเฉลี่ยของสมาชิกในครัวเรือน จากนั้นใช้ CA ในการจัดกลุ่มครัวเรือนที่มีลักษณะเดียวกัน พร้อมกันนั้นสามารถกำหนดค่าเบี้ยประกันต่อเดือนโดยพิจารณาจากจำนวนครั้งที่คาดหวังว่าครัวเรือนในแต่ละกลุ่มจะใช้

วงการขายสินค้าปลีก

บริษัทค้าปลีกพยายามเก็บข้อมูลจากครัวเรือนในด้านรายได้ ขนาดของครัวเรือน อาชีพของหัวหน้าครัวเรือน ระยะห่างจากชุมชนเมืองที่อยู่ใกล้ที่สุด จากนั้นใช้ CA ในการกำหนดกลุ่มครัวเรือนออกเป็นกลุ่มครัวเรือนเล็กแต่จ่ายเงินหนัก กลุ่มครัวเรือนใหญ่และจ่ายเงินหนัก กลุ่มครัวเรือนเล็กไม่ค่อยจับจ่ายซื้อของ กลุ่มครัวเรือนใหญ่แต่ไม่ค่อยจับจ่ายซื้อของ แล้วจัดแคมเปญประชาสัมพันธ์ตลอดจนส่งจดหมายเชิญชวนไปยังครัวเรือนที่คาดว่าเป็นกลุ่มเป้าหมาย

ด้านการศึกษา

CA ช่วยให้สามารถกำหนดกลุ่มนักเรียนที่ครู/ผู้มีส่วนเกี่ยวข้อง อาจต้องให้ความใส่ใจดูแลเป็นพิเศษ โดยCA จะช่วยในการจัดนักเรียนที่มีลักษณะคล้ายๆกันเข้ากลุ่มแต่ละกลุ่ม (ตัวอย่างเช่นมีกลุ่มนักเรียนที่เป็นเลิศในทุกวิชา กลุ่มนักเรียนที่เป็นเลิศในบางวิชาแต่มีปัญหาในวิชาอื่น ๆบางวิชา กลุ่มนักเรียนที่มีปัญหาในแทบทุกวิชา) จากนั้นผู้วิเคราะห์สามารถใช้ discriminant analysis เพื่อให้ทราบว่า นักเรียนแต่ละกลุ่มมีปัจจัยด้านจิตวิทยา สภาพแวดล้อม ความสามารถ ทัศนคติเป็นอย่างไรที่แตกต่างจากของเด็กนักเรียนกลุ่มอื่น ๆ

ด้านมานุษยวิทยา

CA ใช้เป็นเครื่องมือในการกำหนดความเหมือนของวัตถุที่มนุษย์สมัยโบราณผลิตขึ้น (artifacts) ตลอดจนสิ่งของที่เป็นมรดกทางวัฒนธรรมที่ได้มีการขุดค้นพบ

ข้อสังเกต:- CAไม่มีกลไกที่ช่วยให้นักวิเคราะห์สามารถแยกแยะระหว่างตัวแปรที่เกี่ยวข้องกับตัวแปรที่ไม่เกี่ยวข้องกับการบอกลักษณะความเหมือนหรือความแตกต่างออกจากกัน ดังนั้นก่อนการวิเคราะห์โดย CA ผู้วิเคราะห์จะต้องมั่นใจว่าตัวแปรที่ใช้บอกลักษณะความเหมือนหรือความแตกต่างเป็นตัวแปรที่ได้กลั่นกรองว่าเกี่ยวข้องแล้วเท่านั้น ตัวอย่างเช่นในการศึกษาจัดกลุ่มคน ตัวแปรที่บ่งบอกความเหมือนหรือความแตกต่างอาจจะเป็น เพศ การศึกษา อายุ ศาสนา รายได้ อาชีพ ส่วนตัวแปรอื่น ๆเช่น ชื่อ นามสกุลอาจไม่ได้มีส่วนช่วยในการจัดแบ่งกลุ่มคนและเป็นตัวแปรที่ไม่เกี่ยวข้อง

สิ่งหนึ่งที่ต้องทำความเข้าใจกันก็คือ CA มักจะเป็นส่วนหนึ่งของการวิเคราะห์ลำดับขั้นตอน (sequence analysis) ที่มีลำดับขั้นตอนดังนี้

Factor analysis---→cluster analysis-----→discriminant analysis

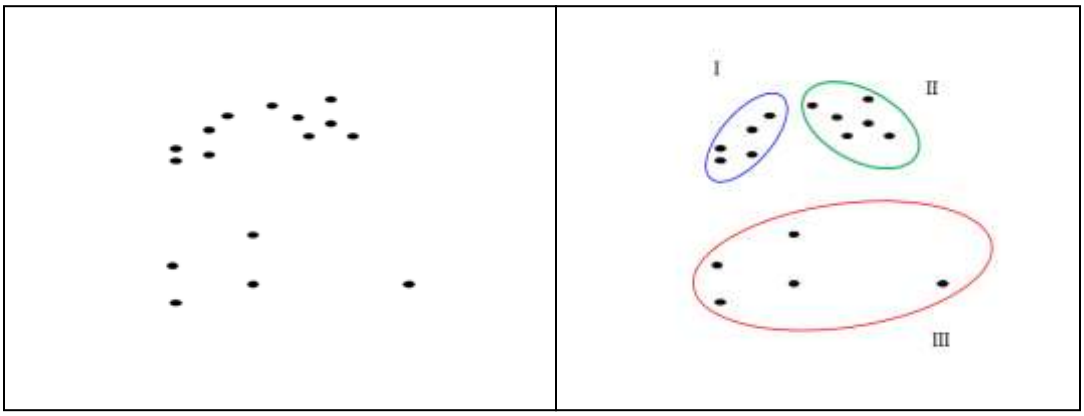
นั่นก็คือ กระบวนการวิเคราะห์ลำดับขั้นตอน (sequence analysis) จะเริ่มจากการวิเคราะห์ปัจจัย (factor analysis) เป็นลำดับแรก(ดู Statistics Talks# 12-15) ตามด้วย CA เป็นลำดับที่สองและตามด้วย discriminant analysis (ดู Statistics Talks# 20-22) เป็นลำดับที่สาม โดยการวิเคราะห์ปัจจัยจะช่วยในการลดมิติและ/หรือลดรูปข้อมูลให้เหลือกลุ่มตัวแปรที่ไม่ซ้ำซ้อนและขจัดตัวแปรที่มีความสัมพันธ์(multicollinearity) ออกไป ทำให้ง่ายต่อการทำ CA เพื่อแยกแยะกลุ่ม เมื่อเสร็จจากการทำ CA แล้ว discriminant analysis จะช่วยเข้ามาตรวจสอบความถูกต้องของแบบจำลอง (goodness of fit) และช่วยในการสร้างprofile ของกลุ่ม ในขั้นตอนนี้ CA พึ่งพา discriminant analysis ในการตรวจสอบการแบ่งกลุ่ม ตลอดจนตรวจสอบว่าตัวแปรใดที่มีความสำคัญ(ทางสถิติ)ในการ

แยกแยะความแตกต่างระหว่างกลุ่ม อย่างไรก็ตามนี่มิได้หมายความว่ากลุ่มต่าง ๆที่แยกจากกันแล้วจะมีความหมายเป็นที่เข้าใจในตัวเอง การตีความตลอดจนการเลือกกลุ่มให้ถูกต้องเป็นศิลปะที่ต้องการฯ ฝึกฝน

ลักษณะการจัดกลุ่ม

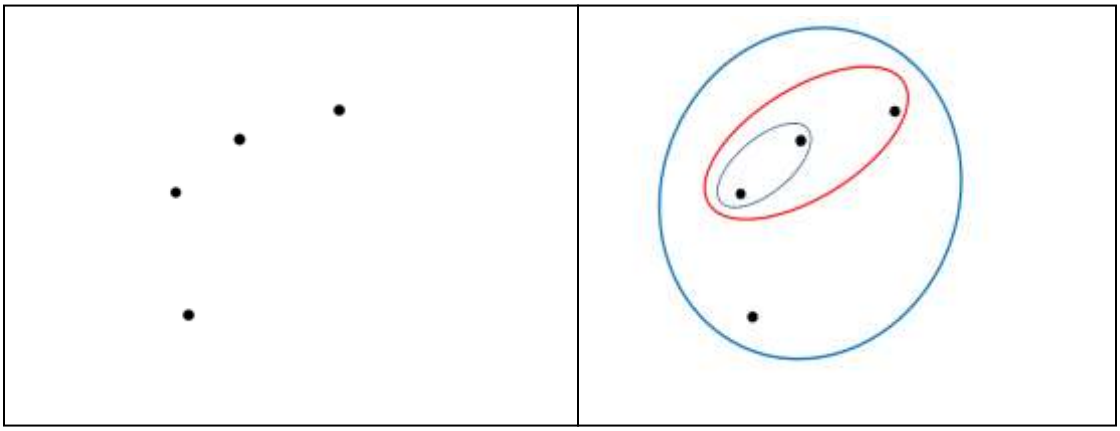
การจัดกลุ่มอาจอยู่ในรูปแบบสองแบบดังต่อไปนี้

1.Partitional clusters: กลุ่มที่จัดแบบนี้จะอยู่แยกจากกันโดยเด็ดขาด และไม่มีส่วนใดส่วนหนึ่งของ cluster ที่จะทับซ้อนกัน ภาพด้านล่างแสดงการจัดกลุ่มแบบนี้



ก่อนการจัดกลุ่ม **หลังจัดกลุ่มแล้ว**

2.Hierarchical cluster: กลุ่มที่จัดแบ่งแบบนี้จะมีลักษณะที่บางกลุ่มจะเป็น subset หรือสังกัด (nested) หรือทับซ้อน (overlap) กับกลุ่มอื่น โดยสมาชิกของกลุ่มหนึ่งก็จะเป็นสมาชิกของกลุ่มอื่นที่ใหญ่กว่า ภาพด้านล่างแสดงการจัดกลุ่มแบบนี้



ก่อนการจัดกลุ่ม **หลังจัดกลุ่มแล้ว**

ขั้นตอนในการจัดกลุ่ม

มี 2 ขั้นตอนได้แก่

1.เริ่มจากจำนวนcase จำนวนหนึ่งที่เราต้องการจัดcase เหล่านี้เข้ากลุ่มที่มีลักษณะคล้ายๆกัน โดยเรา เลือกตัวแปรที่ใช้วัดลักษณะที่บ่งบอกความเหมือนหรือความแตกต่าง

2 เลือกวิธีการจัดกลุ่มที่จะใช้

วิธีในการจัดกลุ่ม(Clustering Methods)

การจัดกลุ่มสามารถทำได้ 3 วิธีคือ

1. Hierarchical clustering:เป็นวิธีการที่ใช้กันมากที่สุด ทั้งนี้ตามวิธีนี้จะมีการจัดกลุ่มโดยเริ่มจาก case เพียงcase เดียวไปจนถึงการจัดกลุ่มที่มี case ทุกcase สามารถใช้ในการจัดกลุ่มตัวแปรได้ เหมือน factor analysis สามารถใช้ได้กับข้อมูลที่เป็น interval /count / binary ในเอกสารนี้เราจะกล่าวถึงเฉพาะส่วนที่เกี่ยวข้องกับการจัดกลุ่มตัวแปรที่มีลักษณะเป็น intervalเท่านั้น
   -การจัดกลุ่มแบบhierarchical clustering นี้สามารถทำได้สองรูปแบบคือ agglomerative clustering หรือ divisive clustering ในเอกสารฉบับนี้ เราจะกล่าวถึงการจัดกลุ่มที่เป็น agglomerative clustering เท่านั้น

การจัดกลุ่มแบบ agglomerative hierarchical clustering เป็นการจัดกลุ่มแบบ forward clustering โดยเริ่มจากcase แต่ละ case จะเป็น cluster ในตัวของมันเอง จากนั้น case 2 cases ที่มีระยะห่างจากกันสั้นที่สุดจะรวมกันเป็น cluster ก่อน จากนั้นพิจารณา case ที่ 3 ที่มี ระยะห่างจาก case 1 หรือ 2 แต่หาก case ที่ 3 อยู่ใกล้ case ที่4 มากกว่าก็จะรวมกันเข้าเป็น cluster ที่สอง กระบวนการนี้จะดำเนินไปเรื่อย ๆ จน case ทุก case รวมกันเข้าเป็นกลุ่มเดียว โดยcluster บางcluster จะซ้อนเป็น subset ของ cluster ที่ใหญ่กว่า พึงสังเกตว่า cluster สุดท้ายนี้จะมีเพียง cluster เดียวประกอบด้วยcase ทุก case แม้cluster สุดท้ายนี้ด้วยตัวมันเองไม่สามารถใช้ประโยชน์ได้มากนัก แต่โครงสร้างของ cluster ย่อย ๆที่มารวมกลุ่มกันอาจชี้ให้เห็นจำนวนcluster หลักๆ ที่สามารถแยกออกจากcluster อื่น และนำไปใช้ประโยชน์ต่อหรือเป็นข้อมูลในการกำหนดจำนวน cluster ที่เหมาะสมต่อไป

การจัดกลุ่มแบบ hierarchical divisive clustering จะกลับกันกับ hierarchical agglomerative clustering โดยเริ่มจากcluster ใหญ่ cluster เดียวที่ประกอบด้วยcase ทุก case รวมกันอยู่ จากนั้นจะมีการแตกแยกย่อยออกเป็น clusterเล็กและจบลงที่ case แต่ละ case เป็น cluster ในตัวของมันเอง

-การจัดกลุ่มแบบนี้เหมาะกับชุดข้อมูลที่มีจำนวน case (n) ขนาดเล็ก (n < 250) ถ้า n ใหญ่ คอมพิวเตอร์จะเสียเวลาในการคำนวณนาน

-ในการจัดกลุ่มแบบนี้ กลุ่มอาจจะอยู่ซ้อนกันเป็นชั้น ๆ(nested)แทนที่จะอยู่แยกจากกัน จากนั้นใช้ผลที่ได้จากการจัดกลุ่มนี้ในการกำหนดจำนวน cluster ที่เหมาะสม ก่อนที่จะใช้วิธีการจัดกลุ่มแบบอื่น(k-means clustering)เพื่อหา solution ในขั้นตอนต่อไป

-การจัดกลุ่มแบบนี้ ผู้ใช้ต้องกำหนดวิธีในการรวมกลุ่ม(agglomeration method) ซึ่งมีหลายวิธี ได้แก่ single linkage, complete linkage, average linkage , average linkage within groups , centroid method , median method และ Ward's method เพื่อความมั่นใจใน solution นักวิเคราะห์อาจทดลองใช้สองหรือสามวิธีข้างต้น เพื่อดูผลลัพธ์ที่ได้ว่ามีความความสอดคล้องกันหรือไม่

-การจัดกลุ่มแบบนี้ ผู้ใช้ต้องให้คำนิยามของความเหมือน (similarity) หรือ distance โดยเลือกสถิติที่ใช้วัดความห่าง/ความเหมือนของcase ที่แตกต่างกันสองcase

<u>วิธีรวมกลุ่ม</u>

Nearest neighbor / single linkage / SLINK cluster: ตามวิธีนี้ สอง cluster จะมารวมกันก็ต่อเมื่อระยะห่างระหว่างcase หนึ่งๆในcluster แรกกับ case ในcluster ที่สองมีค่าน้อยที่สุด

ข้อดี : เป็นวิธีการรวมกลุ่มที่ง่าย และเหมาะกับการรวมกลุ่มที่cluster เดิมไม่ได้มีลักษณะเป็นรูปทรงกลม(spherical) หรือรูปวงรี (elliptical)

ข้อเสีย: ไม่ได้คำนึงถึงโครงสร้างของcluster และอาจก่อให้เกิดปัญหาที่เรียกว่า "chaining" โดย cluster ที่ถูกจัดกลุ่มแล้วจะมีรูปร่างยาวและยืดไม่เป็นระเบียบ ที่สำคัญคือ solution ที่ได้อาจเปลี่ยนแปลงหากมีข้อมูลที่มีค่าสุดโต่ง(outlier) หรือ noise ในข้อมูล

Furthest neighbor/complete linkage/ CLINK: ตามวิธีนี้ cluster สอง cluster จะมารวมกันก็ต่อเมื่อระยะห่างระหว่างcase หนึ่งๆในcluster แรกกับ case ในcluster ที่สองมีระยะห่างไกลกันมากที่สุด วิธีนี้มีแนวโน้มที่จะทำให้เราได้ cluster ขนาดย่อมที่มีขนาดใกล้เคียงกัน แต่มีข้อเสียตรงที่ไม่ได้คำนึงถึงโครงสร้างของ cluster และมีแนวโน้มที่จะทำให้clusterใหญ่แตกแยกออกเป็นcluster ย่อย

ข้อดี:ไม่ sensitive กับการมี noise หรือ outliers

Between-group linkage/ average linkage/unweighted paired-group method (UPGMA linkage): ตามวิธีนี้ กลุ่มสองกลุ่มที่จะมารวมกันจะมีค่าเฉลี่ยของระยะห่างระหว่างคู่ของcase หนึ่งๆ ในกลุ่มแรกกับ case ที่อยู่ในกลุ่มที่สองต่ำที่สุด เป็นวิธีการรวมกลุ่มที่ได้รับความนิยมมากกว่าการรวมกลุ่มแบบ single and complete linkage

ข้อดี:ไม่sensitive กับการมี noise หรือ outliers

ข้อเสีย:มีแนวโน้มที่จะทำให้เกิด cluster รูปวงกลม

Within-group linkage/average linkage within groups: ตามวิธีนี้ cluster สอง clusterที่แต่แรกแยกกัน จะมารวมกันก็ต่อเมื่อระยะห่างเฉลี่ยระหว่างcase ต่าง ๆใน cluster ที่รวมกันแล้วมีค่าน้อยที่สุด สอดคล้องกับวัตถุประสงค์ที่จะหาความเหมือนกันภายใน cluster (homogeneity within clusters)

Centroid clustering/weighted pair-group method using centroid averages: ตามวิธีนี้ cluster สอง clusterที่จะมารวมกันก็ต่อเมื่อ centroids(centroidคือค่าเฉลี่ยของแต่ละตัวแปรในcluster หนึ่งๆ)ของทั้งสอง cluster อยู่ใกล้กันมากที่สุด centroidของ cluster ใหม่คือค่าเฉลี่ยถ่วงน้ำหนักของcentroid ของทั้งสอง clusterเดิม

Median clustering/ unweighted pair-group method using centroid averages:ตามวิธีนี้ cluster สองclusterที่จะมารวมกันก็ต่อเมื่อcentroids(centroidคือค่าเฉลี่ยของแต่ละตัวแปรใน cluster หนึ่งๆ)ของทั้งสอง cluster อยู่ใกล้กันมากที่สุด centroidของ cluster ใหม่คือค่าmedian ของ centroidของทั้งสองcluster เดิม เหมาะกับการรวมกลุ่มสองกลุ่มที่มีจำนวนcaseเท่า ๆกัน

Ward's method: ตามวิธีนี้ จะมีการคำนวณค่าเฉลี่ยของตัวแปรทุกตัวของแต่ละกลุ่ม จากนั้นจะมีการคำนวณ squared Euclidean distance จากค่าเฉลี่ยนี้ของ case แต่ละ case และมีการหาผลรวมของdistance นี้ของทุก ๆcase กลุ่มสองกลุ่มที่จะมีการมารวมกันจะต้องมีผลรวมของ squared distance ภายในกลุ่มเพิ่มขึ้นน้อยที่สุด หรือมีผลทำให้ pooled within-cluster variation มีค่าเพิ่มน้อยที่สุด วิธีนี้เป็นวิธีการรวมกลุ่มที่ค่อนข้างได้รับความนิยมเช่นเดียวกับวิธี average linkage มีแนวโน้มที่จะทำให้มี cluster ที่มีขนาดเท่า ๆกัน (ซึ่งบางครั้งก็ไม่เป็นผลดีเสมอไป) และผลลัพธ์ที่ได้ค่อนข้างขึ้นอยู่กับการมี outliers

สถิติที่ใช้วัดความเหมือน/ความแตกต่าง

Euclidean:	$\sqrt{(y-x)^2}$
Squared Euclidean:	$\Sigma (y-x)^2$
City-Block:	$\Sigma \lvert y-x \rvert$
Chebychev:	$\max \lvert y-x \rvert$
Cosine:	$\cos r_{xy}$
Minkowski:	$\{\Sigma \lvert y-x \rvert^m\}^{1/m}$
Pearson's correlation:	$1-r_{xy}$

-ในการจัดกลุ่มแบบ hierarchical clustering นี้ บางครั้ง case แต่ละ case อาจประกอบด้วยตัวแปรหลายตัวที่บ่งบอกลักษณะหรือคุณสมบัติ หรือมีมาตรวัดแตกต่างกัน ดังนั้นก่อนการจัดกลุ่ม จำเป็นที่จะต้องมีการ standardize ตัวแปรเสียก่อน

ข้อสังเกต: ในบางกรณี การ standardize ตัวแปรก็มีข้อเสียตรงที่อาจมีผลทำให้case ที่แตกต่างกัน (ก่อน standardized) มีความเหมือนกันมากขึ้นภายหลังstandardized แล้ว ทางแก้ก็คือบางทีเราอาจจำเป็นต้องทำ CA สองครั้ง ครั้งแรกโดยไม่มีการstandardize ข้อมูล และครั้งที่สองเมื่อมีการ standardize เพื่อพิจารณาดูว่าจะสร้างความแตกต่างในผลที่ได้ในการทำCA มากน้อยเพียงใด

ความหมายของข้อมูลทางสถิติเมื่อมีการวิเคราะห์โดยใช้ hierarchical clustering:

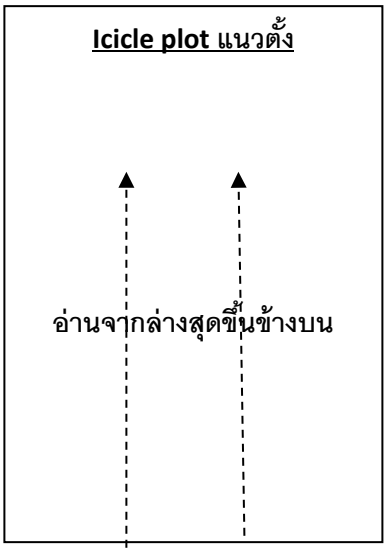
Proximity matrix: ให้ข้อมูลในรูปตารางเกี่ยวกับระยะห่างระหว่างcase ต่าง ๆกันที่มีในข้อมูล มักจะเป็นประโยชน์ในกรณีที่ขนาดของข้อมูลมีขนาดเล็กเพราะสามารถเปรียบเทียบความเหมือน/ความแตกต่างของ case ต่าง ๆกันได้ง่าย แต่ไม่เหมาะกับกรณีที่ข้อมูลมีขนาดใหญ่เพราะ matrix จะมีขนาดใหญ่มาก จนไม่อาจพิมพ์ลงในกระดาษแผ่นหนึ่งได้ และอาจสร้างความสับสนให้กับผู้วิเคราะห์เอง

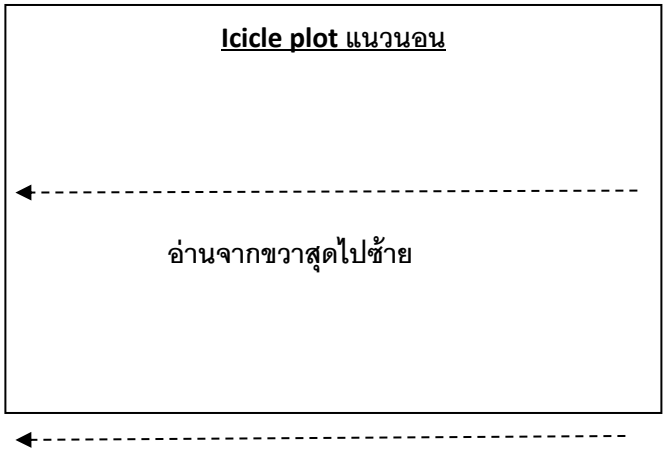
Agglomeration schedule : ให้ข้อมูลในรูปตารางเกี่ยวกับการรวมกันของ case ต่าง ๆ เข้าเป็น cluster ในแต่ละขั้นตอนของการจัดกลุ่ม มีการให้ข้อมูลเกี่ยวกับcase 2 case ที่นำมารวมกัน ระยะห่างระหว่าง case (ดูจาก coefficients) ขั้นตอนที่จะมีการupdate cluster โดยมี case ใหม่เข้ามารวมอยู่ใน cluster (ดูจาก Next Stage)

Icicle plot: มีสองแบบคือแบบแนวตั้ง(vertical icicle plot) และแบบแนวนอน (horizontal icicle plot) เป็นตารางแสดงให้เห็นกระบวนการรวมกันเข้าเป็น cluster ในแต่ละขั้นตอน ในการอ่านมีกฎเกณฑ์ดังนี้

- หากเป็นแบบแนวตั้ง แถวแต่ละแถวจะแสดงลำดับขั้นตอนที่มีการรวมกันเข้าเป็น cluster ส่วนคอลัมน์แต่ละคอลัมน์แสดง case แต่ละ case ในการอ่านความหมาย ผู้วิเคราะห์จะต้องอ่านจากล่างสุดขึ้นข้างบน โดยเริ่มจากแถวล่างสุด ซึ่งจะแสดงการรวมกันของ case ต่างกัน 2 case ขึ้นเป็น cluster แถวรองแถวล่างสุดจะแสดงการรวมกันของ case 2 case หรือระหว่าง cluster แรกกับ case อื่นขึ้นเป็น cluster ใหม่ ในแต่ละแถวต่อมาก็จะมีการรวมกันในลักษณะนี้จนกระทั่งถึงแถวบนสุดที่ทุก ๆcase จะรวมกันอยู่ใน cluster เดียว

- หากเป็นแบบแนวนอน แถวแต่ละแถวจะแสดง case แต่ละ case ส่วนคอลัมน์จะแสดงขั้นตอนที่มีการรวมกันเข้าเป็น cluster ในการอ่านความหมาย ผู้วิเคราะห์จะต้องอ่านจากขวาไปซ้าย โดยเริ่มจากคอลัมน์ด้านขวาสุด ซึ่งจะแสดงการรวมกันของ case ต่างกัน 2 case ขึ้นเป็น cluster คอลัมน์ต่อจากคอลัมน์ขวาสุดจะแสดงการรวมกันของ case 2 case หรือระหว่าง cluster แรกกับ case ขึ้นเป็น cluster ใหม่ ในแต่ละคอลัมน์ต่อมาก็จะมีการรวมกันในลักษณะนี้จนกระทั่งถึงคอลัมน์ที่อยู่ซ้ายสุดที่ทุก ๆcase จะรวมกันอยู่ใน cluster เดียว





Dendrogram: เป็นรูปกราฟที่แสดงให้เห็นผลของการจัดกลุ่มในขั้นสุดท้าย โดยแสดงระยะห่างระหว่าง case ต่าง ๆ ที่มารวมกันในรูปของเส้นกิ่งก้านที่แสดงในแนวนอน หากเส้นที่เชื่อมระหว่างcase ยาวมากแสดงว่าระยะห่างระหว่าง case มีสูง

บ่อยครั้ง dendrogram จะใช้เป็นเครื่องมือที่ชี้ให้ผู้วิเคราะห์ทราบในเบื้องต้นว่า จำนวนcluster ที่เหมาะสมควรจะเป็นเท่าใด แต่ในหลายกรณี dendrogram อาจไม่ช่วยให้ผู้วิเคราะห์สามารถกำหนดจำนวน cluster ที่เหมาะสมได้ และผู้วิเคราะห์อาจต้องทำการทดสอบความไว (sensitivity analysis) โดยเปลี่ยนวิธีการรวมกัน(clustering method) และ/หรือเปลี่ยนมาตรวัด(measure) ที่ใช้บอกระยะทางระหว่าง case 2 cases

จุดแข็งของ hierarchical clustering method; วิธีการจัดกลุ่มแบบนี้ ผู้วิเคราะห์ไม่จำเป็นต้องกำหนดจำนวน cluster ไว้ก่อน

2.**K-means clustering**: การจัดกลุ่มแบบนี้มีเงื่อนไขว่าเราต้องกำหนดจำนวนของกลุ่มไว้ก่อนล่วงหน้า จากนั้นโปรแกรมจะทำการเลือกจับ case เข้าcluster เอง วิธีนี้ใช้การคำนวณน้อยไม่หนักกับคอมพิวเตอร์ cluster ที่ได้ในขั้นตอนสุดท้ายจะเป็นกลุ่มที่แยกจากกัน (partitional cluster) ไม่ทับซ้อนกัน

-เป็นการรวมกลุ่มที่เหมาะกับกรณีที่มีข้อมูลขนาดใหญ่ (n >1,000) เนื่องจากไม่ต้องคำนวณหา proximity matrix หาระยะห่างซึ่งวัดความเหมือนหรือแยกแยะความแตกต่างเหมือน hierarchical clustering

- caseหนึ่งๆอาจจะเปลี่ยนแปลงออกจาก cluster ที่เคยอยู่ไปอยู่ยัง cluster ใหม่ได้ตลอดกระบวนการรวมกลุ่ม

-เนื่องจากcaseหนึ่งๆอาจเปลี่ยนจากกลุ่มหนึ่งไปอยู่กลุ่มอื่น ๆได้ในแต่ละขั้นตอนที่มีการคำนวณ จนกว่าจะสิ้นสุดกระบวนการรวมกลุ่มกันเป็นคำตอบขั้นสุดท้าย บางครั้งจึงจัดการรวมกลุ่มแบบนี้เข้าเป็นการจัดกลุ่มแบบโยกย้ายกลุ่ม(relocating clustering method)

-การคำนวณระยะห่างใน k-means clustering จะใช้ Euclidean distance โดยกระบวนการจะเริ่มจากค่าเฉลี่ยตั้งต้นชุดหนึ่งจากนั้นก็จะมีการจัดแต่ละกรณีเข้ากลุ่มโดยพิจารณาจากระยะห่าง จากศูนย์กลาง จากนั้นก็จะมีการคำนวณค่าเฉลี่ยของกลุ่มขึ้นอีก มีการโยกย้ายกรณีแต่ละกรณีโดยพิจารณาจากค่าเฉลี่ยใหม่ที่คำนวณนี้ ทำอย่างนี้ซ้ำอีกจนกระทั่งค่าเฉลี่ยของกลุ่มไม่เปลี่ยนแปลงไปจากขั้นตอนก่อนหน้าอย่างมีนัยสำคัญหรือจำนวนรอบ iteration ครบจำนวนรอบสูงสุดที่ได้กำหนดเอาไว้ ท้ายที่สุดก็จะมีการคำนวณค่าเฉลี่ยของ clusterอีกครั้งและจัดกรณีต่าง ๆ ไปอยู่ยังกลุ่มที่ถาวร ไม่มีการสลับหรือจัดกลุ่มใหม่อีกแล้ว

- K-means clustering จะสร้าง ANOVA tableแสดง mean-square errors จุดประสงค์สำคัญไม่ใช่เพื่อใช้ทดสอบสมมติฐานทางสถิติใด ๆ แต่เพื่อใช้พิจารณาประสิทธิภาพของตัวแปรแต่ละตัวในการแบ่งแยก case เข้ากลุ่ม

-K-means clustering มุ่งที่จะminimize ความแตกต่างภายในcluster และ maximize ความแตกต่างระหว่าง cluster

-K-means clustering เป็นประโยชน์กับผู้วิเคราะห์ในการทดสอบแบบจำลองหลายแบบที่มีจำนวน cluster แตกต่างกัน

คำเตือน:วิธีของK-means clusteringนี้ผลที่ได้จะขึ้นอยู่กับการเกิดกรณีที่ที่มีข้อมูลสุดโต่งแยกจากกลุ่ม (outliers) ค่อนข้างมาก ดังนั้นก่อนเริ่ม ให้พิจารณากลั่นกรองว่ามีกรณีที่มี outliers หรือไม่ และเอากรณีนั้นออกก่อนที่จะทำการวิเคราะห์ใด ๆ

Input ของ K-means clustering

-การหาk-means clustering จะต้องระบุข้อมูลส่วนที่เป็น inputs ต่อไปนี้

ก .ตัวแปรที่จะใช้ในการรวมกลุ่ม

ข.จำนวน cluster เป้าหมายที่ต้องการ

ค.วิธีการ(iterate หรือ iterate และ classify)

ง.วิธีการเลือก cluster center (จากalgorithm ในโปรแกรมเอง จากไฟล์ข้อมูลที่สร้างไว้ก่อนแล้ว)

ปุ่มiterate

-เราอาจเลือกปุ่ม iterate และระบุจำนวน รอบของiteration (จำนวนรอบสูงสุดต้องไม่เกิน 10)และเงื่อนไขการรวมกัน (convergence criterion) ซึ่งตั้งไว้ว่ากระบวนการ iteration จะหยุดลงเมื่อใดก็ตามที่ขนาดของการเปลี่ยนแปลงในค่าเฉลี่ยของ cluster น้อยกว่า 2% ของระยะห่างระหว่าง initial clusters หากผู้วิเคราะห์ต้องการกำหนดระยะห่างนี้เองก็สามารถทำได้โดยระบุตัวเลขที่มีค่ามากกว่า 0 แต่ไม่เกิน 1 ลงไป

ปุ่ม save

-เมื่อผู้วิเคราะห์ต้องการบันทึกหมายเลขที่ของcluster ที่ case แต่ละ case เป็นสมาชิกอยู่ โดยโปรแกรมจะบันทึกไว้เป็นข้อมูลในคอลัมน์สุดท้ายของไฟล์ข้อมูลโดยใช้ชื่อว่า QLC_1 หรือหากผู้วิเคราะห์ต้องการบันทึกข้อมูลเกี่ยวกับระยะห่างระหว่างcase แต่ละ case กับ cluster center ของมัน ก็สามารถทำได้โดยใช้ชื่อว่า QLC_2 โดยคลิ๊กเช็คที่ช่อง Distance from cluster center

ปุ่ม Options

Initial cluster centers: ถูกตั้งเป็น default อยู่แล้ว หากไม่ต้องการให้คลิ๊กออกไป ในส่วนนี้จะให้ข้อมูลของตัวแปรที่อยู่ใน cluster ตั้งต้น

Output ของ K-means clustering

Iteration history table : เป็นตารางแสดงการเปลี่ยนแปลงใน cluster mean ในแต่ละขั้นตอนที่มีการ iterate และจะสิ้นสุดลงต่อเมื่อcluster mean เปลี่ยนแปลงไปในอัตราที่ต่ำกว่า cut-off rate

ในส่วนของ Cluster centers หากเว้นว่างไว้ โปรแกรมจะไปสุ่มคำนวณหาค่าเฉลี่ยด้วยตัวมันเอง แต่หากผู้วิเคราะห์มีข้อมูลเก็บไว้ในรูปไฟล์ ก็สามารถเลือกระบุชื่อไฟล์ที่มีข้อมูลเกี่ยวกับ cluster mean ที่เคยคำนวณไว้ก่อนหน้านี้ได้

ในกรณีที่กระบวนการจัดกลุ่มสิ้นสุดลง หากผู้วิเคราะห์ต้องการเก็บข้อมูลเกี่ยวกับ cluster mean ก็สามารถกำหนดชื่อไฟล์ได้เพื่อประโยชน์ของการวิเคราะห์ในอนาคตกับกรณีมีข้อมูลตัวอย่างเข้ามาเพิ่ม

ANOVA table:จะให้ข้อมูลที่แสดงให้เห็นว่ามีตัวแปรใดบ้างที่มีระยะห่างจากค่าเฉลี่ยของ cluster แตกต่างอย่างมากจาก ตัวแปร อื่น ๆ (พิจารณาจากค่า Mean Square Error ของตัวแปรนั้น ๆ) ซึ่งมีความหมายว่า ตัวแปรนั้น ๆมีส่วนน้อยมากในการแยกแยะความแตกต่างระหว่าง cluster

ANOVA table นี้ไม่สามารถนำไปใช้ในการทดสอบสมมติฐานทางสถิติใด ๆได้ เพียงแต่อาศัยให้เห็นค่าของ Mean Square Error ของตัวแปรหนึ่งๆเท่านั้น

Cluster information for each case: จะระบุหมายเลขcluster ที่ case แต่ละ case เป็นสมาชิกอยู่ตลอดจน Euclidean distance ระหว่างcase กับ cluster center

Number of cases in cluster table: แสดงให้เห็นจำนวน case ทั้งหมดที่มีในแต่ละ cluster

ในกรณีที่จำนวนของ case ในแต่ละ cluster มีความไม่สมดุล ผู้วิเคราะห์อาจกำหนดจำนวน cluster ให้เพิ่มขึ้นหรือเรียงลำดับ case แต่ละ case ในข้อมูลใหม่ จากนั้นจึงทำการวิเคราะห์ K-means clustering อีกครั้ง

Cluster membership table:แสดงข้อมูลของแต่ละ case ว่าเป็นสมาชิกอยู่ใน cluster ใด และมีระยะห่างจาก centroid ของ cluster นั้นอย่างไร

3. **Two-step clustering**: การจัดกลุ่มแบบนี้มีความแตกต่างจากวิธีการจัดกลุ่มอื่น ๆ ตรงที่

1. สามารถจัดกลุ่มที่ใช้ตัวแปรที่เป็นได้ทั้งแบบแยกประเภท(categorical) และตัวแปรที่มีค่าต่อเนื่อง (continuous)
2. จะทำการคัดเลือกจำนวน cluster ด้วยตนเอง(automatic selection of cluster)
3. ใช้สำหรับวิเคราะห์ไฟล์ขนาดใหญ่ได้อย่างมีประสิทธิภาพ

-การจัดกลุ่มแบบนี้ใช้ likelihood distance measure เป็นมาตรวัดในการจัดกลุ่ม ในกรณีที่ตัวแปรมีค่าต่อเนื่องทั้งหมด สามารถเลือกใช้ Euclidean distance ได้

-ข้อสมมติฐานของการจัดกลุ่มวิธีนี้

ก. ตัวแปรที่ใช้ไม่ว่าจะเป็นแบบแยกประเภทหรือที่มีค่าต่อเนื่องต่างมีความเป็นอิสระซึ่งกันและกัน

ข. การกระจายของค่าตัวแปรที่มีค่าต่อเนื่องมีลักษณะเป็น Normal ในขณะที่การกระจายของตัวแปรที่เป็นแบบแยกประเภท (categorical variable) มีลักษณะเป็น multinomial

-.ในการจัดกลุ่มแบบนี้ caseต่าง ๆถูกจัดเข้าเป็น preclusters ก่อน หลังจากนั้นจาก preclusters ก็รวมกันเป็น clusters โดยใช้ agglomerative clustering algorithm ซึ่งแสดงจำนวนกลุ่มที่มีการรวมกันเป็น cluster เริ่มตั้งแต่การรวมกลุ่มขึ้นเป็นกลุ่มเดียว ไปจนถึงหลายกลุ่ม ในจำนวนกลุ่มที่มารวมกันจะมีการให้ข้อมูลค่าสถิติ Schwarz's Bayesian Information Criterion (BIC) และ/ หรือ Akaike's Information Criterion(AIC) โดยจำนวนCluster ที่ optimal จะมี BIC/AIC ต่ำที่สุดในขณะที่มี Ratio of Distance Measures สูงที่สุด

คำเตือน: solutionสุดท้ายที่ได้ จะขึ้นอยู่กับลำดับของcase ที่มีอยู่ในไฟล์ เพื่อลดปัญหานี้ควรจะจัดกลุ่มแบบสุ่ม

Input ของการจัดกลุ่มแบบนี้

Distance measure: Log-likelihood เป็น default หรือ Euclidean

Number of clusters: ให้เป็นหน้าที่ของ algorithm จะกำหนดเองโดยอัตโนมัติ อยู่ที่ 15 กลุ่ม (default) หรือจะเลือกระบุจำนวนcluster ที่ต้องการก็ได้

Clustering criterion: เลือกระหว่าง Schwarz's Bayesian Information Criterion(BIC) ซึ่งเป็น default หรือเลือก Akaike's Information Criterion

Plots: ให้เลือกระบุภาพแสดง Within cluster percentage chart cluster pie chart และ Rank of variable Importance

Output: เลือก Statistics เช่น Descriptives by cluster cluster frequencies และ Information criterion (AIC หรือ BIC)

Working data file และ XML files

Outputของ 2-stage clustering

Auto clustering table: เป็นตารางแสดงจำนวนของcluster ที่รวมกันพร้อมกับค่าสถิติ Bayesian Information Criterion (BIC) ที่ทำให้พิจารณาได้ว่า ควรจะจัดกลุ่มในจำนวนที่เท่าใด

Cluster Distribution Table:แสดงให้เห็นว่าในแต่ละ cluster ประกอบด้วย case ต่าง ๆมากน้อยเท่าใด

Cluster centroid table:แสดงให้เห็นจุดศูนย์กลางของแต่ละcluster จำแนกตามค่าของตัวแปรต่าง ๆ

Frequencies table:แสดงจำนวนของ case ในแต่ละ cluster จำแนกตามตัวแปรแบ่งแยกประเภท

AIC หรือ BIC เป็นดัชนีใช้วัดคุณภาพของ model รูปแบบต่าง ๆกัน แสดงการ trade-off ระหว่างความซับซ้อน(complexity) ที่เพิ่มมากขึ้น กับ goodness-of—fit ของ แบบจำลองที่ใช้ ในส่วนที่เป็น CA นี้ การเพิ่มจำนวน cluster ขึ้นทำให้มีความซับซ้อนมากขึ้นซึ่งไม่เป็นผลดี ในขณะเดียวกันจำนวน cluster ที่น้อยลงมีผลทำให้จำเป็นที่จะต้องมีการจัดกลุ่ม case บาง case เข้าอยู่เป็นสมาชิกของ cluster บาง cluster แทนที่จะแยกออกไปเป็น cluster ต่างหาก และมีผลทำให้มีการสูญเสียข้อมูลบางอย่างไป

Contribution this issue: ดร. ดนัย ปัตตพงศ์

งานวิจัยที่อ้างอิงบทความวิชาการนี้

ดารณี พิมพ์ช่างทอง “ การวิเคราะห์จัดกลุ่มเพื่อการรณรงค์ทางการตลาดด้วยการใช้เครือข่ายสังคมออนไลน์ “ . Global Business and Economics Review, ปีที่ 13 ฉบับที่หนึ่ง เดือนมิถุนายน 2561

อยากเรียนรู้การนำสถิติข้างต้นนี้ไปใช้ในการวิจัยระดับสารนิพนธ์ (independent study) วิทยานิพนธ์ (thesis) ดุษฎีนิพนธ์(dissertation) ปรึกษาได้ที่ dpattaphongse@gmail.com

* ผู้แต่ง MBA's Made Easy (160+ issues) เอกสารวิชาการด้านศาสตร์การบริหารธุรกิจที่ช่วยให้ธุรกิจสามารถยืนหยัดและอยู่รอดได้ในภาวะที่โลกเปลี่ยนแปลงอยู่ตลอดเวลา

* ผู้พัฒนา FINALYSIS… a dedicated software สำหรับให้บริการนักธุรกิจที่ต้องการวิเคราะห์ความเป็นไปได้ทางการเงินของโครงการพัฒนาอสังหาริมทรัพย์ (บ้านจัดสรร/จัดสรรที่ดินเพื่อการอุตสาหกรรม/อาคารชุด/อาคารสำนักงานให้เช่า) โรงแรม โรงพยาบาลเอกชน ห้างสรรพสินค้า โรงงานน้ำตาล โรงงานกระดาษ โรงไฟฟ้าชีวมวล ฯลฯ ได้เห็นตัวเลขก่อนโครงการเกิด หลีกเลี่ยงความผิดพลาดเป็นร้อยเป็นพันล้านหากเกิดการลงทุนจริง(กำหนด DEBUT 1 เมษายน 2569)

* ผู้แต่งหนังสือ"การวิเคราะห์ความเป็นไปได้ทางการเงินและการจัดวงเงินเครดิตของโครงการลงทุน"ประกอบด้วยตัวอย่างของธุรกิจจริงที่ไม่เปิดเผยชื่อนับ 100 บริษัท ครอบคลุมอุตสาหกรรม 24 อุตสาหกรรม

*Co-developer ซอฟแวร์ en◉gex◉cel® สำหรับใช้ทดสอบ/เรียนรู้ศัพท์(ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 90 บทครอบคลุมศัพท์ระดับ SAT/IELTS/TOEFL กว่า 12,000 คำ) และไวยากรณ์อังกฤษ (ประกอบด้วยแบบฝึกหัดและเฉลยกว่า 160 บทหรือกว่า 10,000 ข้อครอบคลุมเนื้อหาระดับอุดมศึกษาและTOEFL) มาพร้อมกับไฟล์เสียง/ไฟล์ข้อมูล/ฯลฯ อีกมาก(กำหนด DEBUT 1 เมษายน 2569)