

ปิยธิดา รุจะศิริ 2552: การเปรียบเทียบเทคนิคการแบ่งกลุ่มข้อมูลสำหรับการวิเคราะห์กลุ่ม ปริญญาวิทยาศาสตรมหาบัณฑิต (สถิติ) สาขาสถิติ ภาควิชาสถิติ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์บุญอ้อม โคมทิ, Ph.D. 185 หน้า

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบความสามารถในการแบ่งกลุ่มของเทคนิคการแบ่งกลุ่ม 5 วิธี ได้แก่ วิธีการจัดกลุ่มแบบลำดับชั้น อัลกอริทึมจัดกลุ่มแบบเคมีน วิธีจัดระบบด้วยตัวเองของโคโฮเนน อัลกอริทึมจัดกลุ่มแบบเคเมคอยด์ และการใช้อัลกอริทึมจัดกลุ่มแบบเคเมคอยด์ร่วมกับวิธีวัดระยะห่างแบบไดนามิกโทมวอร์ปปีง โดยทำการศึกษากับข้อมูลจริงและข้อมูลจำลอง ซึ่งมีการแจกแจงแบบปกติหลายตัวแปร ในการจำลองข้อมูลใช้เทคนิคมอนติคาร์โลสร้างข้อมูลจำนวน 1,000 ชุด สำหรับแต่ละชุดข้อมูลซึ่งมีขนาด 25, 100 และ 300 ตัวอย่าง และกำหนดจำนวนกลุ่มข้อมูลขนาด 2-5, 2-8 และ 2-15 กลุ่มตามลำดับ กำหนดจำนวนตัวแปรที่ศึกษาเป็น 3, 5 และ 7 ตัวแปร โดยพิจารณาจากเกณฑ์ค่าความแตกต่างของข้อมูลภายในกลุ่ม (RMSSTD) ซึ่งถ้าค่า RMSSTD ต่ำ แสดงว่าเทคนิคการแบ่งกลุ่มนั้นเป็นวิธีที่ดี และพิจารณาจากเกณฑ์ค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (RS) ซึ่งถ้าค่า RS สูง แสดงว่าเทคนิคการแบ่งกลุ่มนั้นเป็นวิธีที่ดี ผลการวิจัยพบว่า ทั้งการศึกษากการแบ่งกลุ่มจากข้อมูลจำลองและจากข้อมูลจริงให้ผลสอดคล้องกัน กล่าวคือ อัลกอริทึมจัดกลุ่มแบบเคมีนให้ค่า RMSSTD ต่ำที่สุดและค่า RS สูงที่สุดในเกือบทุกกรณี ซึ่งสรุปได้ว่า อัลกอริทึมจัดกลุ่มแบบเคมีนให้ผลการวิเคราะห์กลุ่มที่ดีที่สุด ยกเว้นกรณีที่แบ่งกลุ่มข้อมูลขนาด 25 ตัวอย่างเป็น 4 หรือ 5 กลุ่ม ซึ่งวิธีการจัดกลุ่มแบบลำดับชั้นด้วยวิธี Ward linkage ให้ผลการวิเคราะห์กลุ่มดีที่สุด นอกจากนี้เมื่อจำนวนกลุ่มเพิ่มขึ้นประสิทธิภาพในการแบ่งกลุ่มของทั้ง 5 เทคนิควิธี มีแนวโน้มเพิ่มขึ้น แต่จำนวนตัวแปรและขนาดของชุดข้อมูลที่เพิ่มขึ้นไม่มีผลต่อประสิทธิภาพในการแบ่งกลุ่มของเทคนิคการแบ่งกลุ่มทั้ง 5 วิธี

Piyatida Rujasiri 2009: The Comparison of Clustering Techniques for Cluster Analysis. Master of Science (Statistics), Major Field: Statistics, Department of Statistics. Thesis Advisor: Assistant Professor Boonorm Chomtee, Ph.D. 185 pages.

The objective of this research is to compare effectiveness of the 5 clustering techniques for multivariate data; these techniques consist of Hierarchical clustering method, K-means clustering algorithm, Kohonen's Self-Organizing Maps method (SOM), K-medoids method and K-medoids method integrated with Dynamic Time Warping distance measure (DTW). To evaluate these 5 techniques, Root Mean Square Standard Deviation (RMSSTD) and R Squared (RS) are used to be the criteria. For RMSSTD, the lower value is the better technique and for RS, the higher value is the better technique. These approaches are evaluated by using both real and simulated data which are multivariate normally distributed. Each datasets were generated by Monte Carlo technique with 25, 100 and 300 sample sizes and repeated 1,000 times for 3, 5 and 7 variables. In this research, 2-15 clusters are studied. For the results, both real and simulated datasets provide the same result. That is, K-means clustering method yields the lowest RMSSTD and highest RS for almost situations. Therefore, K-means is the most suitable algorithm of clustering techniques for cluster analysis, except categorizing 25-sample dataset to 4 or 5 clusters which Ward linkage gives the best clustering result. In addition, increasing the number of clusters trends to increase the efficiency of the 5 clustering methods. However, the number of variables and the size of dataset do not affect to the efficiency of the 5 clustering methods.