งานวิจัยนี้ เป็นการปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล ที่ใช้ได้ผลดีกับการตัดแบ่งชั้นสี (Color Quantization) หลักการตัดแบ่งข้อมูลคือการแบ่งข้อมูลตามแกนที่มีค่าความแปรปรวนสูงสุดให้ได้จำนวนกลุ่มตามที่ต้องการจัดกลุ่มข้อมูลด้วยอัลกอริทึมเคมีน (K-means) และใช้จุดศูนย์กลางของข้อมูลที่แบ่งแล้วเป็นจุดเริ่มต้นของการจัดกลุ่มด้วยอัลกอริทึมเคมีน การใช้จุดเริ่มต้นที่ดีจะลดข้อจำกัด และข้อเสียของการใช้ค่าเริ่มต้นแบบสุ่ม ที่ให้ผลการจัดกลุ่มที่ไม่แน่นอน และกลุ่มข้อมูลบางกลุ่มอาจไม่มีจำนวนสมาชิก

การทดสอบประสิทธิภาพของอัลกอริทึมที่นำเสนอได้ทำกับข้อมูลจาก UCI และ Web Access Log โดยเปรียบเทียบผลกับอัลกอริทึมเคมีนที่มีการกำหนดค่าเริ่มต้นแบบสุ่ม อีกทั้งยังใช้การเปรียบเทียบกับอัลกอริทึมที่ใช้ในการกำหนดค่าเริ่มต้นสำหรับการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วย Cluster Center Initialization Algorithm (CCIA) จากผลการทดสอบประสิทธิภาพของการจัดกลุ่มข้อมูล ถือได้ว่าอัลกอริทึมที่นำเสนอนี้มีประสิทธิภาพดีกว่าการใช้ค่าเริ่มต้นแบบสุ่ม และให้ประสิทธิภาพใกล้เคียงกับ CCIA ซึ่งมีวิธีการที่ซับซ้อนกว่า

196131

In this research, we propose an algorithm to compute initial cluster centers for K-means clustering. We use novel approach for color quantization that divides color spaces into small clusters or cells with intercluster distances as large as possible and intracluster distance as small as possible. In the proposed algorithm, data in a cell is partitioned using a cutting plane that divide cell in two small cells. The plane is perpendicular to the data axis with the highest variance and is designed to reduce the sum squared errors of the two cells as much as possible. Cells are partitioned one at a time until the number of cells reaches the desired number K. The centers of the K cells become the initial cluster centers for K-means.

We evaluated our method by clustering 10 UCI data sets (UCI Machine Learning Repository) and Web Access Log data set. We also present the experimental results on some datasets in comparison with CCIA algorithm. The experimental results reveal that the proposed algorithm computes initial cluster centers that help K-means converge to better clustering than the random initial cluster centers and almost guarantee every cluster has its data membership. The proposed algorithm also performs as good as CCIA algorithm which is more difficult to implement.