การจัดกลุ่มข้อมูล เป็นงานรวมกลุ่มข้อมูลด้วยการพิจารณาความคล้ายคลึงกัน อัลกอริทึม k-means เป็นอัลกอริทึมมาตรฐานที่ทำการรวมกลุ่มข้อมูล โดยเริ่มต้นกระบวนการด้วยการคำนวณระยะห่างเพื่อกำหนดข้อมูลแต่ละตัวเข้ากลุ่มที่อยู่ใกล้ที่สุด จากนั้นคำนวณค่าจุดกึ่งกลางของแต่ละกลุ่มโดยใช้ค่าเฉลี่ยของข้อมูลทั้งหมดในกลุ่ม อัลกอริทึมจะทำสองขั้นตอนนี้ซ้ำจนกระทั่งค่าจุดกึ่งกลางไม่มีการเปลี่ยนแปลงและข้อมูลไม่มีการเปลี่ยนกลุ่ม ผู้วิจัยได้เสนออัลกอริทึมชื่อ density-biased clustering เพื่อปรับปรุง k-means ให้ทำงานกับข้อมูลขนาดใหญ่ได้รวดเร็ว ความเร็วในการทำงานจะได้จากขั้นตอนการสุ่มด้วยเทคนิคการสุ่มตามความหนาแน่น เพื่อคัดเลือกข้อมูลตัวแทนที่มีความหนาแน่นสูง ค่าความหนาแน่นคำนวณจากจำนวนข้อมูลที่อยู่ใกล้เคียง จากนั้นนำข้อมูลตัวแทนที่คัดเลือกไว้เข้าสู่กระบวนการจัดกลุ่ม งานวิจัยนี้ยังได้ปรับปรุงการจัดกลุ่มข้อมูลให้เป็นการจัดกลุ่มแบบเพิ่มพูนเพื่อให้สามารถจัดกลุ่มข้อมูลขนาดใหญ่ที่มีขนาดของกลุ่มแตกต่างกันได้ ผลการทดลองได้แสดงว่าวิธีการจัดกลุ่มข้อมูลตามความหนาแน่นสามารถคัดเลือกข้อมูลที่เป็นตัวแทนได้ดี และอัลกอริทึมทำงานได้ผลดีกับข้อมูลขนาดใหญ่ที่มีจำนวนมิติปานกลาง (มีจำนวนมิติหรือแอททริบิวต์ 8 ถึง 23 แอททริบิวต์) การจัดกลุ่มกับข้อมูลแบบเพิ่มพูนกับข้อมูลที่ถูกคัดเลือกตามความหนาแน่นให้ผลการจัดกลุ่มที่ใกล้เคียงกับวิธี k-means แต่ใช้เวลาประมวลผลที่น้อยกว่า

229253

Clustering is a task of grouping data based on similarity. A standard k-means algorithm groups data by firstly assigning all data points to the closest clusters, then determining the new cluster mean based on the average values of its members. The algorithm repeats these two steps until it converges; that is until there is no change in cluster assignment among data points. We propose a variation called incremental density-biased clustering to improve the scalability of the clustering process. To speed up the clustering process, we develop the density-biased clustering algorithm as an efficient data clustering technique. Efficiency is due to the reduced data set. Density of each data point is calculated based on its surrounding neighbors. Only dense data are selected for further clustering. We also implement the incremental clustering algorithm to get the benefit from its advantage of efficient memory usage and extend it to deal with clustering large data of varying cluster sizes. Our experimental results reveal the effectiveness of drawing samples of high density from large data sets of moderate dimensions (approximately 8-23 attributes), then incrementally grouping these data into clusters. The clustering results produce almost the same results as the standard k-means, but our incremental algorithm takes less computational time.