K-MEANS CLUSTERING USING CORRELATED ATTRIBUTES

THIPPRAPA POPIYATRAKUL 4937921 ITCS/M

M.Sc.(COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE: SUDSANGUAN NGAMSURIYAROJ, Ph.D., SONGSRI TANGSRIPAIROJ, Ph.D., ANANTA SRISUPHAB, Ph.D.

ABSTRACT

K-Means clustering is one of the widely used knowledge discovery techniques. One disadvantage of K-means clustering used for a large data set is how to find an initial set of clustered data that are approximately close to the final set of clustered data so that it would not take a lot of time in clustering the final set when compared with the method that selects a data set in a random fashion.

This thesis proposed a new efficient way to do the K-means clustering when only a few correlated attributes are used. For each data set, the correlation among attributes is computed in order to determine which attributes are most related so that they could be the representatives of all attributes. Subsequently, the correlated attributes are selected for computing the clustering, and the results are compared with the clustering outputs obtained from using all attributes in the computation.

We evaluated our work using 10 datasets of UCI Machine Learning Repository including Iris, Ecoli, Yeat, Abalone, White Wine, Page Blocks, Magic Grammar Telescope, Breast Cancer, Waveform, and Letter Recognition. We used the tool called PML (Parallel Machine Learning) developed by IBM for performing the K-means clustering. Our results illustrate that the centroids of correlated attributes are close to the centroids of all attributes while having less computation time.

KEY WORDS: K-MEANS CLUSTERING / INITIAL CENTROID / CORRELATION

50 pages

การทำคลัสเตอร์ด้วยวิธี K-MEANS ที่ใช้ข้อมูลที่สัมพันธ์กัน

K-MEANS CLUSTERING USING CORRELATED ATTRIBUTES

ทิพย์ประภา โพธิ์ปิยตระกูล 4937921 ITCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์ : สุดสงวน งามสุริยโรจน์, Ph.D., ทรงศรี ตั้งศรีไพโรจน์, Ph.D., อนันต์ ศรีสุภาพ, Ph.D.

บทคัดย่อ

ในปัจจุบันการทำเหมืองข้อมูลได้รับความนิยมอย่างแพร่หลาย และอัลกอริทึ่มที่นิยมใช้คือ K-Mean clustering แต่อัลกอริทึ่มนี้มีข้อเสียหลายอย่างเช่น การหาจุดศูนย์กลางเริ่มต้นของ Cluster ซึ่งส่วนใหญ่นิยมใช้วิธีการสุ่มเพื่อหาจุดศูนย์กลางเริ่มต้นนั้น ทำให้เกิดปัญหาเช่น เมื่อเลือกโดยวิธีการสุ่มอาจได้จุดศูนย์กลางเริ่มต้นเป็น Outliner เมื่อทำการ Clustering อาจใช้เวลานานกว่าจะได้ผลลัพธ์ที่ต้องการ และปัญหาที่สำคัญอีกประการหนึ่งคือ การทำเหมืองข้อมูลนั้นจะใช้ข้อมูลที่มีขนาดใหญ่ซึ่งอาจเป็นข้อมูลที่สะสมมาเป็นเวลานานดังนั้นกว่าจะได้ผลลัพธ์ที่ต้องการจะทำให้เสียเวลามาก

ดังนั้นเพื่อเป็นการแก้ไขปัญหาข้างต้นที่กล่าวมางานวิจัยนี้จึงได้นำเสนอแนวคิดที่จะลดขนาดข้อมูลลงและยังสามารถหาจุดศูนย์กลางเริ่มต้นที่ใกล้เคียงกับจุดศูนย์กลางเริ่มต้นของผลลัพธ์อีกด้วย โดยใช้คุณลักษณะของข้อมูลที่มีความสัมพันธ์กัน ซึ่งวิธีการนี้จะเริ่มจากการหาค่าความสัมพันธ์กันระหว่าง Attribute ทุกคู่ เพื่อหากลุ่มของ Attribute ที่มีความสัมพันธ์ใกล้ชิดกันมากที่สุดเป็นตัวแทนของข้อมูลทั้งหมด ซึ่งใช้ในการ Clustering

จากวิธีการที่นำเสนอได้ทำการทดสอบกับชุดของข้อมูลทั้งหมด 10 ชุดข้อมูลจาก UCI Machine Learning Repository และใช้โปรแกรม PML (Parallel Machine Learning) ของบริษัท IBM ในการทำ Clustering ซึ่งทำการทดสอบโดยเปรียบเทียบผลลัพธ์ระหว่างการใช้ Attribute ทุกตัวของข้อมูลกับการใช้เพียง Attribute ที่มีความสัมพันธ์ใกล้ชิดกัน ผลลัพธ์ที่ได้นั้นจุดศูนย์กลางเริ่มต้นของการใช้ Attribute ที่มีความสัมพันธ์กันมีความใกล้เคียงกันกับการ Attribute ทั้งหมด ดังนั้นการใช้เพียง Attribute ที่มีความสัมพันธ์ใกล้ชิดกันสามารถหาจุดศูนย์กลางเริ่มต้นของ Cluster ได้

50 หน้า