การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน ด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล

สิริชัย ดีเลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)
คณะสถิติประยุกต์
สถาบันบัณฑิตพัฒนบริหารศาสตร์
2550

การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน
ด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล

สิริชัย ดีเลิศ
คณะสถิติประยุกต์

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาแล้วเห็นสมควรอนุมัติให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)

ผู้ช่วยศาสตราจารย์ .....ประธานกรรมการ
(ดร. โอม ศรนิล)

รองศาสตราจารย์ .....กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์
(ดร. สุรพงค์ เอื้อวัฒนามงคล)

รองศาสตราจารย์ ..... กรรมการ
(ดร. บุญธีร์ เครือตราชู)

ผู้ช่วยศาสตราจารย์ ..... กรรมการ
(ดร. ปรีชา วิจิตรธรรมรส)

รองศาสตราจารย์ ..... คณบดี
(ดร. สุรพงค์ เอื้อวัฒนามงคล)
วันที่ 26 เดือน กันยายน พ.ศ. 2550

บทคัดย่อ

ชื่อวิทยานิพนธ์	การปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีนด้วยการหาค่าเริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูล
ชื่อผู้เขียน	สิริชัย ดีเลิศ
ชื่อปริญญา	วิทยาศาสตรมหาบัณฑิต (ระบบสารสนเทศประยุกต์)
ปีการศึกษา	2550

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

ABSTRACT

Title of Thesis	Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance
Author	Mr. Sirichai Deelers
Degree	Master of Science (Applied Information System)
Year	2007

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.

กิตติกรรมประกาศ

วิทยานิพนธ์เรื่องการปรับปรุงประสิทธิภาพการจัดกลุ่มข้อมูลของอัลกอริทึมเคมีน ด้วยการหาค่นริ่มต้นโดยวิธีการตัดแบ่งกลุ่มข้อมูลนี้ สำเร็จลุล่วงได้ด้วยดี เนื่องมาจากบุคคลหลายท่านที่ได้กรุณาช่วยเหลือให้ข้อมูล ข้อเสนอแนะ คำปรึกษาแนะนำ ความคิดเห็นและกำลังใจ

ผู้เขียนขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. สุรพงค์ เอื้อวัฒนามงคล ที่ได้ให้คำแนะนำ ชี้แนะแนวทาง และตรวจสอบวิทยานิพนธ์ในทุกขั้นตอน และขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. โอม ศรนิล ผู้ช่วยศาสตราจารย์ ดร. ปรีชา วิจิตรธรรมรส และรองศาสตราจารย์ ดร.บุญธีร์ เครือตราชู ที่ได้ชี้แนะแนวทางในการศึกษา และเปรียบเทียบประสิทธิภาพของอัลกอริทึมในการจัดกลุ่มข้อมูล

ขอขอบพระคุณคณาจารย์ของคณะสถิติประยุกต์ทุกท่าน ที่ประสิทธิ์ประสาทวิชาและถ่ายทอดความรู้ให้แก่ผู้ศึกษา และขอขอบคุณเจ้าหน้าที่คณะสถิติประยุกต์ ที่ได้ให้ความช่วยเหลือในการติดต่อประสานงานเป็นอย่างดี

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วันชัย สุทธะนันท์ คณบดี คณะวิทยาการจัดการ มหาวิทยาลัยศิลปากร ที่ได้ให้โอกาสและสนับสนุนการศึกษาในทุกด้าน

ขอขอบคุณนักศึกษาร่วมรุ่น รุ่นพี่ และรุ่นน้องในคณะสถิติประยุกต์ที่ได้ให้ความช่วยเหลือและประสานงานตลอดช่วงเวลาที่ได้ศึกษาอยู่ที่คณะสถิติประยุกต์

ท้ายสุดนี้ ขอกราบขอบพระคุณบิดา มารดา และญาติพี่น้องที่ได้ช่วยส่งเสริมสนับสนุนและเป็นกำลังใจให้แก่ผู้จัดทำวิทยานิพนธ์ตลอดมา

สิริชัย ดีเลิศ

สิงหาคม 2550