ชื่อเรื่องการค้นคว้าแบบอิสระ การเปรียบเทียบประสิทธิภาพระหว่างวิธีระยะทางสูงสุดกับวิธีแบ่งแบบไบนารีในรุ่นปรับปรุงสำหรับอัลกอริทึมเคมีนส์

ผู้เขียน นายสัจจะ ตันจันทร์พงศ์

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

อาจารย์ที่ปรึกษาการค้นคว้าแบบอิสระ อาจารย์ ดร.สรรพวรรธน์ กันตะบุตร

บทคัดย่อ

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพระหว่างวิธีระยะทางสูงสุดกับวิธีแบ่งแบบไบนารีในรุ่นปรับปรุงในการหาโค๊ตบุ๊คตั้งต้นสำหรับอัลกอริทึมเคมีนส์ ซึ่งทั้งสองวิธีนี้เป็นวิธีที่จะสร้างโค๊ตบุ๊คตั้งต้นที่ทำให้จำนวนรอบของการคำนวณหาโค๊ตบุ๊คที่เหมาะสมนั้นมีจำนวนรอบน้อยที่สุด ในขณะเดียวกันก็พยายามที่จะให้มีค่าความบิดเบือนไปน้อยที่สุดด้วย โดยในการศึกษาได้นำข้อมูลมาทดสอบกับโปรแกรมที่สร้างขึ้น 2 กรณี คือ ข้อมูลที่ได้จากการสุ่มขึ้นมาเอง ซึ่งกำหนดให้ข้อมูลทดสอบมีตั้งแต่ 20 จุด 30 จุด 40 จุด และ 50 จุด มีมิติตั้งแต่ 2 มิติ 4 มิติ และ 6 มิติ และทำการแบ่งจำนวนกลุ่มข้อมูลตั้งแต่ 4 กลุ่ม 8 กลุ่ม 16 กลุ่มเช่นกัน และข้อมูลจริงที่ได้จากศูนย์รวมข้อมูลของกลุ่มห้องปฏิบัติการและวิจัยการเรียนรู้ของระบบจักรกล คณะสารสนเทศและวิทยาการคอมพิวเตอร์ มหาวิทยาลัยแคริฟลอเนีย ซึ่งมีจำนวน 4 ฐานข้อมูล ได้แก่ ข้อมูลเกี่ยวกับรถยนต์ ข้อมูลการวินิจฉัยมะเร็งเต้านม ข้อมูลเกี่ยวกับการระบุประเภทของแก้ว และข้อมูลเกี่ยวกับการระบุแหล่งที่ผลิตของไวน์ โดยผลลัพธ์ที่ได้จากการศึกษาในกรณีแรกนี้ ปรากฏว่า วิธีแบบสุ่ม วิธีหาระยะทางสูงสุด และวิธีการแบ่งแบบไบนารีในรุ่นปรับปรุงนั้นให้ผลลัพธ์ของจำนวนรอบของการหาโค๊ตบุ๊คที่เหมาะสมและค่าความบิดเบือนใกล้เคียงกันมาก ส่วนในกรณีที่สองซึ่งใช้ข้อมูลจากฐานข้อมูลที่มีการจัดเก็บจริงและทราบจำนวนกลุ่มล่วงหน้าอยู่แล้ว ปรากฏว่าวิธีการแบ่งแบบไบนารีในรุ่นปรับปรุงใช้จำนวนรอบในการหาโค๊ตบุ๊คที่เหมาะสมน้อยที่สุด ซึ่งน้อยกว่าวิธีการหาระยะทางสูงสุด วิธีการแบบสุ่ม ตามลำดับ แต่ทั้ง 3 วิธีก็ให้ผลลัพธ์ของค่าของความบิดเบือนไปใกล้เคียงกัน

Independent Study Title Efficiency Comparison of Maximum Distance Method and Modified Binary Splitting Method in K-Means Algorithm

Author Mr. Sajja Tanchanpong

Degree Master of science (Computer Science)

Independent Study Advisor Lecturer Dr. Sampawat Kantabutra

ABSTRACT

In this independent study, we compared the efficiency between two initialization methods for K-means algorithm: maximum distance and modified binary splitting methods. These two methods aim to reduce the number of iterations for the K-means algorithm while maintaining the quality of clustering. In our experiments, we use two kinds of data sets. The first is based on random data containing cases: 20 points, 30 points, 40 points, and 50 points; 2 dimensions, 4 dimensions, and 6 dimensions; 4 groups, 8 groups, and 16 groups. The second is based on real data obtained from UCI (University of California, Irvine) Machine Learning Repository in which there are 4 databases: Auto-Mpg Database, Wisconsin Breast Cancer Databases, Glass Identification Database, and Wine Recognition Database. In the random input case, the results are very similar in terms of both the number of iterations and the quality of clustering. In the case of real data, the results show that the modified binary splitting method is the best performer, based on the number of iterations, while producing similar quality of clustering when compared to the other two initialization methods.