รหัสโครงการ SUT7-705-46-12-37

รายงานการวิจัย

การศึกษาผลของการลดขนาดของข้อมูลในกระบวนการทำเหมืองข้อมูล

(The effect of data reduction in the process of data mining)

ผู้วิจัย

หัวหน้าโครงการ

ผู้ช่วยศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ พ.ศ. 2546

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

มิถุนายน 2548

บทคัดย่อภาษาไทย

การทำเหมืองข้อมูล หรือการค้นหาความรู้ เป็นกระบวนการคัดแยกข้อสนเทศที่มีประโยชน์และยังไม่ถูกค้นพบมาก่อนออกจากเซตหรือกลุ่มข้อมูลขนาดใหญ่ แต่เนื่องจากกระบวนการนี้ใช้เวลาในการประมวลผลมากโดยเฉพาะเมื่อข้อมูลมีขนาดใหญ่ การใช้กลุ่มตัวอย่างแทนที่จะใช้ข้อมูลทั้งหมดจะช่วยให้การประมวลผลรวดเร็วขึ้น แต่ทั้งนี้ผลลัพธ์ที่ได้จะต้องมีคุณภาพคงเดิม งานวิจัยนี้จึงมีจุดมุ่งหมายที่จะศึกษาพฤติกรรมของอัลกอริทึมค้นหาความรู้ เมื่อกลุ่มข้อมูลมีขนาดลดลงตามลำดับจากการสุ่มข้อมูล ทั้งนี้เพื่อค้นหาขนาดของกลุ่มข้อมูลที่ให้ผลลัพธ์ใกล้เคียงที่สุดกับผลลัพธ์ที่ได้จากข้อมูลประชากรทั้งหมด อัลกอริทึมที่ใช้ในการสังเคราะห์ความรู้ได้แก่ อัลกอริทึมเบย์อย่างง่าย และอัลกอริทึมสร้างต้นไม้ตัดสินใจ ซึ่งจัดอยู่ในอัลกอริทึมประเภทค้นหากฎที่สามารถจำแนกข้อมูลและสามารถอธิบายข้อมูลแต่ละประเภทหรือแต่ละคลาสได้ เทคนิคการสุ่มข้อมูลที่ใช้ในการศึกษานี้ได้แก่ การสุ่มอย่างง่าย การสุ่มแบบเป็นระบบ การสุ่มแบบแบ่งชั้น การสุ่มแบบก้าวหน้าเชิงเลขคณิต และการสุ่มแบบก้าวหน้าเชิงเรขาคณิต ข้อมูลที่ใช้ในการศึกษาวิจัยนี้มีจำนวน 5 ชุด แต่ละชุดแยกเป็นข้อมูลฝึกและข้อมูลทดสอบ โดยข้อมูลเหล่านี้เป็นข้อมูลมาตรฐานจากแหล่งข้อมูลมหาวิทยาลัยแคลิฟอร์เนียเมืองเออร์ไวน์ การทดสอบคุณภาพของกลุ่มตัวอย่างจะใช้วิธีทดสอบความแม่นตรงของโมเดลซึ่งเป็นผลลัพธ์ที่ได้จากอัลกอริทึมค้นหาความรู้

บทคัดย่อภาษาอังกฤษ

Data mining or knowledge discovery is the process of extracting useful and previously unknown information from the very large data set. However, extracting knowledge from a large data set is computationally inefficient. Using a sample from the original data can speed up the data mining process, but this is only acceptable if it does not reduce the quality of the induced information. We thus investigate the behavior of learning algorithms on decreasing sample sizes to decide which sample is sufficiently similar to the original data. We observe the accuracy of the induced classification rules extracted from training samples of various sizes and use these results to determine when a sample is sufficiently small, yet maintain the acceptable accuracy rate. We evaluate four sampling methods: simple random, systematic random, stratified random, arithmetic progressive, and geometric progressive. The five data sets to be sampled are taken from the UCI repository and the learning algorithms to induce knowledge from each sample are naive Bayes and decision tree induction. The performance of each sampling scheme is evaluated on the basis of the induced-model accuracy tested on the supplied test data.