29th Congress on Science and Technology of Thailand
October 20-22, 2003, Khon Kaen University, Thailand

การศึกษาเปรียบเทียบเทคนิคการจัดการข้อมูลสูญหายในการทำเหมืองข้อมูลประเภทงานจำแนก*

A Comparative Study of Techniques to Handle Missing Values in the Classification Task of Data Mining

นิตยา เกิดประสพ, กิตติศักดิ์ เกิดประสพ, ยอด สายแวว, ปรีชา พุ่มรุ่งเรือง

Nittaya Kerdprasop , Kittisak Kerdprasop, Yawd Saiveaw, Preecha Pumrungreong
School of Computer Engineering, Suranaree University of Technology, 111 University Ave., Muang District, Nakorn Ratchasima 30000, Thailand; e-mail address: nittaya@ccs.sut.ac.th

บทคัดย่อ: งานวิจัยนี้เป็นการศึกษาเปรียบเทียบเทคนิคต่างๆ ที่ใช้จัดการกับกรณีข้อมูลบางส่วนสูญหายในกระบวนการวิเคราะห์ข้อมูลอัตโนมัติด้วยเทคนิคการทำเหมืองข้อมูลโดยเน้นเฉพาะงานจำแนกประเภทข้อมูล ข้อมูลที่ใช้ในการทดลองนำมาจากแหล่งข้อมูลมาตรฐานของมหาวิทยาลัยแคลิฟอร์เนียที่เออร์ไวน์ โดยเลือกมาทั้งข้อมูลประเภทจำนวนเลขและประเภทข้อความ ข้อมูลมาตรฐานถูกจำลองให้มีบางส่วนสูญหาย จากนั้นใช้เทคนิคที่แตกต่างกันสี่เทคนิคเพื่อเติมส่วนที่สูญหาย การทดสอบประสิทธิภาพของเทคนิคการเติมข้อมูลสูญหายใช้อัลกอริทึมการจำแนกประเภทข้อมูลด้วยวิธีเบย์ส์อย่างง่าย วิธีสร้างต้นไม้ตัดสินใจเชิงอุปนัย และวิธีการคำนวณระยะห่างของข้อมูล ผลการทดลองชี้ว่าถ้าข้อมูลเป็นประเภทจำนวนเลขการตัดทิ้งเรคคอร์ดที่มีข้อมูลสูญหายจะให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ดีกว่า ในขณะที่กรณีข้อมูลประเภทข้อความการเติมข้อมูลสูญหายด้วยสัญลักษณ์ “?” จะให้ประสิทธิภาพการจำแนกประเภทข้อมูลที่ดีกว่า

Abstract: We study and review the techniques for dealing with missing attribute values in data mining. Then, we conduct the experiments to observe the performance of classification algorithms on each strategy of missing-value substitution. The algorithms we used are naïve Bays, tree-based and instance-based classifiers. Four approaches of handling missing values are introduced to the numeric and nominal data sets taken from the UCI repository. The experimental results reveal the superior suggestive choice of ignoring numerical data instances with missing values, whereas replacing the unknown values with the symbol “?” produces a better classification results for the nominal data set.