บทความนี้ได้นำเสนอขั้นตอนวิธีใหม่ในการจำแนกกลุ่มข้อมูล 2 ขั้นตอนวิธีด้วยกัน ขั้นตอนวิธีแรกเรียกชื่อว่า Fuzzy C-means Decision Tree(FMDT) และขั้นตอนที่สองเรียกชื่อว่า Naive Bayes Random Forest (NBRF) ซึ่งทั้งสองขั้นตอนวิธีมีหลักการทำงานเหมือนกันดังนี้คือ กระบวนการตัดสินใจเริ่มด้วยการสกัดแยกข้อมูลที่ตัดสินใจได้ยากออกมาก่อน จากนั้นจึงทำการสร้างตัวแบบเพื่อเรียนรู้เพื่อจำแนกข้อมูลส่วนนี้อีกครั้งหนึ่ง งานวิจัยนี้ใช้การจำแนกขั้นต้นด้วย 2 ขั้นตอนวิธีคือ Fuzzy C-means(FCM) ซึ่งทำงานขนานกับขั้นตอนวิธี Decision Tree(DT) ข้อมูลที่ทั้งสองขั้นตอนวิธีจำแนกแล้วได้คำตอบไม่ตรงกัน ซึ่งเป็นข้อมูลส่วนที่อยู่บริเวณรอยต่อของแต่ละกลุ่มจะถูกจัดประเภทเป็นข้อมูลที่ตัดสินใจยาก จากนั้นใช้ขั้นตอนวิธี DT ทำการจำแนกกลุ่มข้อมูลประเภทนี้อีกครั้งหนึ่ง สำหรับขั้นตอนวิธี NBRF เปลี่ยน Fuzzy C-means (FCM) เป็น Naive Bayes(NB) และเปลี่ยน Decision Tree(DT) เป็น Random Forest(RF) การทดลองการจำแนกกลุ่มข้อมูลจากข้อมูล 11 แบ่งเป็นข้อมูลสังเคราะห์ 6 ชุดข้อมูล ได้แก่ ข้อมูล Clus1000 Clus200 Rand1000 Rand200 Pat1 Pat2 และเป็นข้อมูลจริง 4 ชุดข้อมูลได้แก่ ข้อมูลการออกเสียงสระภาษาของชาวพื้นเมืองอินเดีย (Vowel) ข้อมูลคนไข้ที่ป่วยเป็นโรคตับ (Hepato) ข้อมูลดอกไม้ไอรีส (Iris) ข้อมูลคนไข้ที่ป่วยเป็นโรคคลิชมาเนียซีส (Kla-azar) นั้นแสดงให้เห็นว่าขั้นตอนวิธีใหม่มีประสิทธิภาพดีกว่าแบบเดิมคือ NBRF มีประสิทธิภาพการจำแนกสูงกว่า FCMRF(Fuzzy C-Mean Random Forest) 4.81% FCMDT(Fuzzy C-Mean Decision Tree) 13.57% RF(Random Forest) 3.91% NB(Naive Bayes) 17.15% และสูงกว่า FCM 35.54%

216959

This paper proposes two novel techniques for data classifying namely Fuzzy C-means Decision Tree(FMDT) and Naive Bayes Random Forest (NBRF). Two techniques have same algorithm. The decision process starts by extracting the difficult designate data, if any, and follows by a construction of learning model for re-classifying them. This research performed the first step by using two algorithms; the Fuzzy C-means paralleled with the Decision Tree. The data getting different answers from the two algorithms will be selected as the difficult designate data. These data are lying around the border of the different classes. For NBRF changed from FCM to NB and DT to RF. The experiments performed on 11 benchmarks; 6 sets are synthesized data including Clus1000, Clus200, Rand1000, Rand200, Pat1, and Pat2 data, and 4 sets are real world data including Vowel, Hepato, Iris and Kla-azar data. The experimental results showed that the NBRF technique outperformed the existing models. The average performance of NBRF is better than those of Fuzzy C-Mean Random Forest 3.90%, Fuzzy C-Mean Decision Tree 13.57%, Random Forest 4.81%, Naive Bayes 17.15% and FCM 35.54%.