

Assignment 4: การใช้โปรแกรม WEKA

ในการบ้านนี้ให้นักเรียนใช้ tool ที่ชื่อว่า WEKA (<http://www.cs.waikato.ac.nz>) ในการเรียนรู้ machine learning algorithm แบบต่างๆ ในการบ้านนี้ให้ใช้ข้อมูลจาก zoo.arff, zoo.train.arff และ zoo.test.arff (ซึ่ง zoo.arff มีข้อมูลของทั้ง zoo.train.arff และ zoo.test.arff รวมกันอยู่)

1. ให้นักเรียนเรียนรู้การใช้งาน WEKA โดยเริ่มที่ decision tree algorithm (เรียกว่า J48 ใน WEKA)
 - a. ให้ใช้ J48 ในการสร้าง decision tree จาก zoo.train.arff แล้วเขียนใน รายงานว่า อธิบาย decision tree ที่สร้างขึ้นจาก J48
 - b. ให้ใช้ decision tree ที่สร้างขึ้นหาจำนวนข้อมูลที่ถูก classified ได้อย่างถูกต้องและผิดพลาด รวมทั้งใส่ข้อมูลใน ทั้งสองในรายงานและ text files ที่ชื่อว่า DT_correct.txt, DT_incorrect.txt
 - c. ใน WEKA มี classifier ที่ชื่อว่า ZeroR ที่ใช้ประเมิน classification algorithms ให้ใช้ ZeroR ในการ classify ทั้ง zoo.train.arff และ zoo.test.arff แล้วหาจำนวนข้อมูลที่ถูก classified ได้อย่างถูกต้องและผิดพลาด รวมทั้งใส่ข้อมูลใน ทั้งสองในรายงานและ text files ที่ชื่อว่า ZeroR_testcorrect.txt, ZeroR_testincorrect.txt, ZeroR_traincorrect.txt, ZeroR_trainincorrect.txt พร้อมทั้งอธิบายว่าตัวเลขที่ได้หมายความว่าอย่างไร และเปรียบเทียบกับ ตัวเลขที่ได้จาก decision tree ในกรณีของ test data
 - d. ถ้าคุณใช้ WEKA classifier กับ training set โดยไม่ได้ให้ test set ไปด้วย WEKA จะทำ 10-fold cross-validation บน training set โดยอัตโนมัติ ให้ใช้ zoo.arff ในข้อนี้แล้วรายงาน จำนวนข้อมูลที่ถูก classified ได้อย่างถูกต้องและผิดพลาด รวมทั้งใส่ข้อมูลใน ทั้งสองในรายงานและ text files ที่ชื่อว่า CrossValidation_correct.txt, CrossValidation_incorrect.txt พร้อมทั้ง อธิบายว่า 10-Fold Cross Validation คืออะไร และมีจุดประสงค์เพื่ออะไร
2. ให้ใช้ WEKA ในการเปรียบเทียบ machine learning algorithms ต่อไปนี้
 - K nearest neighbor (called IBk in WEKA)
 - Decision Trees (called J48 in WEKA)
 - Naïve Bayes (called Naïve Bayes in WEKA)
 - a. จาก data sets ข้างต้นให้วิเคราะห์ว่า machine learning algorithms ตัวใดเหมาะสมที่สุดในการ classification
 - b. ให้ใช้ WEKA ในการศึกษา machine learning algorithms แต่ละตัว แล้ววิเคราะห์ข้อมูล เปรียบเทียบกับวิเคราะห์เบื้องต้นที่คุณทำในข้อ 2(a)

กำหนดส่งการบ้านนี้คือวันที่ 11 กันยายน 2552 ให้ส่งมาทาง email ของอาจารย์