บทคัดย่อ

ชื่อรายงานการวิจัย : การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ

ชื่อผู้วิจัย : นายนิเวศ จิระวิชิตชัย

ปีที่ทำการวิจัย : 2553

.....

งานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคด้านเหมืองข้อมูล เพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติทดสอบประสิทธิภาพในการจำแนก (Classification) สำหรับข้อมูลทางการแพทย์ โดยทดลองกับ 7 อัลกอริทึม ซึ่งประกอบด้วย Naïve Bayes,Multilayer Perceptron,Radial Basis Function Network, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Ripper ทำการศึกษาเปรียบเทียบวิธีลดคุณลักษณะที่เหมาะสมด้วยวิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) รวมถึงทดสอบอัลกอริทึมประเภท Single learning และ Multiple learning และทำการเพิ่มประสิทธิภาพการจำแนกด้วยวิธี Bagging และ Boosting

ผลจากการวิจัยพบว่าทุกโมเดลที่สร้างขึ้น มีประสิทธิภาพในการจำแนกประเภทของโรคในระดับ 80 % ขึ้นไป เมื่อไม่ลดคุณลักษณะ และเมื่อเรียงค่าความถูกต้อง (Accuracy) แยกตามประเภทของข้อมูลพบว่า กลุ่มข้อมูล Hypothyroid การสร้างโมเดลด้วยอัลกอริทึม Decision Tree ให้ประสิทธิภาพดีที่สุด 99.57% กลุ่มข้อมูล Leukemia การสร้างโมเดลด้วยอัลกอริทึม Naive Bayes กับ Support Vector Machine ให้ประสิทธิภาพดีที่สุด 98.61% กลุ่มข้อมูล Breast-w การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 96.99% กลุ่มข้อมูล Lymphography การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 86.48% กลุ่มข้อมูล Hepatitis การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ K-Nearest Neighbor ให้ประสิทธิภาพดีที่สุด 85.80% กลุ่มข้อมูล Heart-c การสร้างโมเดลด้วยอัลกอริทึม Support Vector Machine ให้ประสิทธิภาพดีที่สุด 84.15% กลุ่มข้อมูล Heart-statlog การสร้างโมเดลด้วยอัลกอริทึม Radial Basis Function กับ Support

Vector Machine ให้ประสิทธิภาพดีที่สุด 84.07% ตามลำดับ ซึ่งการสร้างโมเดลดังกล่าวมีคุณภาพในระดับที่ยอมรับได้ และสามารถนำไปพัฒนาเป็นซอฟต์แวร์ในการวินิจฉัยโรคอัตโนมัติได้

การเพิ่มประสิทธิภาพด้วยวิธี Multiple Learning ด้วยอัลกอริทึม Bagging และ Boosting ส่งผลให้ค่าความถูกต้องเพิ่มขึ้นเฉพาะบางกลุ่มข้อมูลเท่านั้น โดยมีข้อสังเกตว่าสัดส่วนของกลุ่มตัวอย่างในแต่ละคลาสจะต้องมีปริมาณใกล้เคียงกันหรือเท่ากัน กรณีที่การกระจายของของกลุ่มตัวอย่างในแต่ละคลาส มีสัดส่วนที่แตกต่างกันมาก ส่งผลให้เทคนิค Bagging และ Boosting ไม่ช่วยเพิ่มประสิทธิภาพการจำแนกข้อมูล

การลดคุณลักษณะด้วยวิธี Correlation-based Feature Subset Selection (CFS) และวิธี Feature selection method based on correlation measure and relevance & redundancy analysis (FCBF) ส่งผลให้ประสิทธิภาพความถูกต้อง (Accuracy) ในการจำแนกประเภทของโรคใกล้เคียงกับการไม่ลดคุณลักษณะ แต่การลดมิติของข้อมูลดังกล่าวทำให้ประหยัดทรัพยากรของระบบคอมพิวเตอร์และระยะเวลาในการเรียนรู้เพื่อสร้างโมเดลได้เป็นอย่างดี