โมเดลการเรียนรู้แบบกลุ่มเพื่อการจำแนกข้อมูลความผิดปกติของทารกในครรภ์

Ensemble Learning Model for Cardiotocography Classification

วรกร พิมพาคุณ (Warrakorn Pimpakun)* ปัญญาพล หอระตะ (Punyaphol Horata)**

บทคัดย่อ

บทความนี้ได้เสนอการเรียนรู้แบบกลุ่มที่มีชนิดของตัวจำแนกต่างชนิดกัน เพื่อเพิ่มความแม่นยำการจำแนกข้อมูลความผิดปกติของทารกในครรภ์ ซึ่งเป็นข้อมูลที่ประกอบด้วยอัตราการเต้นของหัวใจของทารกในครรภ์และการบีบตัวของมดลูกมารดา ที่ได้จากกระบวนการ Cardiotocography (CTG) เพื่อใช้วิเคราะห์ความผิดปกติของทารกในครรภ์ โดยการวิเคราะห์ด้วยโมเดลการเรียนรู้แบบกลุ่มที่ประกอบด้วยสมาชิกโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชั่น (RBF) C5.0 และ ซัพพอร์ทเว็กเตอร์แมชชีน (SVM) ในการเรียนรู้รูปแบบความผิดปกติใช้ 70 % ของข้อมูล 2,126 รายการ และใช้ Majority voting กับสมาชิก ผลการทดลองแสดงให้เห็นว่าโมเดลการเรียนรู้แบบกลุ่มที่มีชนิดตัวจำแนกต่างชนิดกันซึ่งประกอบด้วยสมาชิกโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชั่น (RBF) จำนวน 7 ตัว C5.0 จำนวน 2 ตัว และ ซัพพอร์ทเว็กเตอร์แมชชีน (SVM) จำนวน 1 ตัว สามารถจำแนกได้แม่นยำ 99.81% เมื่อเทียบกับโมเดลการเรียนรู้แบบกลุ่มที่มีสมาชิกชนิดเดียวที่ประกอบด้วยสมาชิกโครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชั่น (RBF) จำนวน 7 ตัว ได้แม่นยำ 99.24%

ABSTRACT

This paper demonstrates the use of heterogeneous classifier ensemble learning for improving classification accuracy in cardiotocography (CTG) diagnosis. The CTG is consisting of fetal heart rate and tocographic signals, is a method used for intrapartal evaluation of the wellbeing of the fetus. The members of heterogeneous classifier ensemble is consisting of seven RBFs, two C5.0 and a SVM that trained on 70% of 2,126 records of CTG and using majority voting of the member. The results showed that the heterogeneous classifier ensemble correctly identified 99.81% compared to the homogeneous classifier ensemble that consist seven RBFs correctly classified 99.24%.

คำสำคัญ: การเรียนรู้แบบกลุ่ม การจำแนกข้อมูล

Key Words: Ensemble learning, Classification

** นักศึกษา หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

*** ผู้ช่วยศาสตราจารย์ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น*

บทนำ

ในการวินิจฉัยทางการแพทย์ ความแม่นยำและถูกต้องเป็นสิ่งที่สำคัญที่สุด เนื่องจากส่งผลถึงชีวิตของผู้เข้ารับการตรวจรักษา แนวทางหนึ่งที่เป็นทางเลือกในการวินิจฉัยอาการของโรคหรือการเกิดโรค คือ การวิเคราะห์ข้อมูลที่ส่งผลกระทบหรือปัจจัยแวดล้อมที่แสดงถึงอาการหรือมีความสัมพันธ์กับการเกิดโรคเหล่านั้น ผู้เชี่ยวชาญทางการแพทย์เมื่อวิเคราะห์ข้อมูลเหล่านี้ สามารถจำแนกผลลัพธ์เป็นกลุ่มได้ เช่น ผลลัพธ์คือ เกิด หรือ ไม่เกิดโรคหรือ ระดับความรุนแรงของโรค ซึ่งเมื่อได้ข้อมูลเหล่านี้สามารถวิเคราะห์โดยหลักการทางสถิติได้ โดยปัจจัยต่างๆล้วนมีผลต่อการเกิดโรคทั้งสิ้นแต่มีผลมากน้อยแตกต่างกัน การวิเคราะห์ข้อมูลสามารถใช้ข้อแตกต่างนี้ในการคำนวณผลลัพธ์ที่ถูกต้องแนวทางนี้ให้ผลลัพธ์เป็นที่น่าพอใจและช่วยในการตัดสินใจวินิจฉัยทางการแพทย์ได้

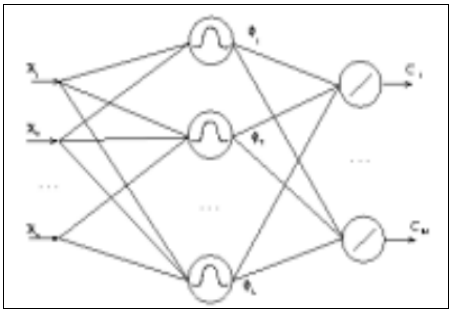
ปัจจุบันมีการพัฒนาวิธีการใหม่ๆ มาช่วยในการตัดสินใจวินิจฉัยทางการแพทย์ได้ เช่น อัลกอริทึม C5.0 เบย์เซียนเน็ตเวิร์ค โครงข่ายประสาทเทียม (ธนกร, 2554) เป็นเทคโนโลยีอย่างหนึ่งที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence: AI) เรียกว่า เทคนิคการเรียนรู้ของเครื่องจักร (Machine Learning Technique) เป็นการสร้างขั้นตอนวิธี (Algorithms) หรือโปรแกรมคอมพิวเตอร์ จากการให้ข้อมูลฝึก (Training data) สำหรับสอนให้คอมพิวเตอร์เรียนรู้ เพื่อให้ได้สมมติฐาน (Hypothesis) ในการนำมาใช้แยกแยะและพยากรณ์ ซึ่งเป็นวิธีการที่ถูกประยุกต์ใช้อย่างแพร่หลายในการวิเคราะห์รูปแบบของข้อมูล วิเคราะห์แนวโน้มของข้อมูล และศึกษาองค์ความรู้จากข้อมูล มีการประยุกต์ใช้ในการวิเคราะห์ข้อมูลอย่างแพร่หลาย และเพื่อเพิ่มประสิทธิภาพในการแยกแยะและพยากรณ์ การเรียนรู้แบบกลุ่ม (Ensemble Learning) จึงถูกคิดค้นขึ้นและได้มีการนำไปใช้ในการช่วยตัดสินใจทางการแพทย์ในหลายๆ ด้าน ซึ่งจะมีใช้ตัวจำแนก (Classifier) มากกว่าหนึ่งตัวในการเรียนรู้ แต่ละตัวจำแนกมีกระบวนการทำงานของตัวเอง และทุกตัวจำแนกกระทำกับข้อมูลเดียวกัน เมื่อได้ผลการจำแนกของแต่ละตัวจำแนกแล้วก็จะนำผลเหล่านั้นมาผ่านวิธีการรวบรวม (Combination, Integration หรือ Vote) และตัดสินสุดท้าย (กาญจนา, 2548)

บทความนี้ได้เสนอการจำแนกข้อมูลที่ประกอบด้วยอัตราการเต้นของหัวใจของทารกในครรภ์และการบีบตัวของมดลูกมารดา ที่ได้จากกระบวนการ Cardiotocography เพื่อใช้วิเคราะห์ความผิดปกติของทารกในครรภ์ โดยการใช้โครงข่ายประสาทเทียมแบบเรเดียลเบสิสฟังก์ชั่น (RBF) C5.0 และ SVM (Support Vector Machine) มาใช้วิธีการเรียนรู้แบบกลุ่มในการเพิ่มประสิทธิภาพในการจำแนกข้อมูล ผู้วิจัยคาดว่าการใช้วิธีการเรียนรู้แบบกลุ่มจะช่วยเพิ่มประสิทธิภาพในการจำแนกข้อมูลและได้ผลการจำแนกที่มีความถูกต้องมากขึ้น

เทคนิคที่ใช้ในการทดลอง

Radial basis function networks (RBF networks)

RBF Networks มีลักษณะเป็นโครงข่ายประสาทเทียมแบบไปข้างหน้า (feed forward networks) โดยประกอบด้วยชั้นของ Input และ Output โดยระหว่างทั้งสองชั้นจะมี Hidden layer อยู่โดยที่ Hidden unit แต่ละตัวใน Hidden layer ประกอบด้วย Radial basis function โครงสร้างของเครือข่ายจะแตกต่างกันตามงานที่จะวิเคราะห์ เมื่อต้องการจำแนกข้อมูลเพื่อหารูปแบบของข้อมูลจะใช้โครงสร้างดังภาพที่ 1



ภาพที่ 1 แบบจำลอง RBF-NN สำหรับจำแนกข้อมูล
ที่มา: Bors, Introduction of the Radial basis function(RBF) networks)

ในการวิเคราะห์หารูปแบบข้อมูลโดยการจำแนกข้อมูลจะใช้ Gaussian function ส่วนของ Output หาได้จากสมการการรวมของค่าน้ำหนักของ Outputของ Hidden units ดังสมการที่ (1)

$$\psi_k(\mathbf{X}) = \sum_{j=1}^{L} \lambda_{jk} \varphi_j(\mathbf{X}) \quad (1)$$

สำหรับ k=1,...,M เมื่อ คือค่าน้ำหนักของ Output โดยแต่ละค่าเป็นการเชื่อมโยงของ Hidden units และ Output Units และ M แทนลำดับของ Output unit ในการจำแนกข้อมูลเพื่อหารูปแบบของข้อมูล Output ของ Radial basis function จะถูกจำกัดที่ค่า 0,1 โดย sigmoid function

อัลกอริทึม C5.0

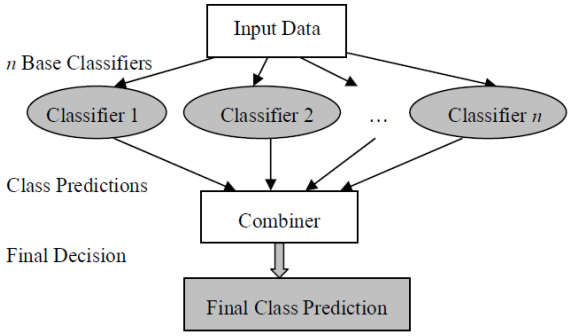
เป็นอัลกอริทึมที่พัฒนาโดยมีพื้นฐานจากอัลกอริทึม C4.5 ของ (Quinlan, J.R., 1999) ซึ่งอาศัยหลักพื้นฐานมาจากต้นไม้การตัดสินใจ สำหรับการสร้างต้นไม้การตัดสินใจจะใช้หลักการของทฤษฎีข่าวสาร (Information Gain) ค่าที่วัดได้สูงสุดจะถูกนำไปใช้เป็นเกณฑ์ในการแบ่งข้อมูล เมื่อคุณสมบัติของข้อมูล (Attribute) ใดให้ค่า Gain สูงสุดจะถูกเลือกให้เป็นโหนด จากนั่นค่าคุณสมบัติอื่นๆในแต่ละกิ่งเพื่อเลือกโหนดต่อไป หรือเป็นรูปแบบหนึ่งของการใช้ Decision Tree ที่พัฒนาเพิ่มเติม โดยไม่ได้ใช้ Gain เป็นตัวแบ่ง แต่ใช้ Gain ratio เป็นตัวแบ่งของ tree ในการทำงานขั้นตอนแรกคล้ายกับการทำงานด้วย ID3 คือต้องหา Info และ Gain ออกมาก่อน ซึ่ง C5.0 ถูกพัฒนาให้อยู่ในซอฟต์แวร์เหมืองข้อมูล เพื่อใช้ประโยชน์ในงานเชิงพาณิชย์มากขึ้นโดยสามารถสกัดหากฎการจำแนกจากต้นไม้การตัดสินใจได้ และยังรองรับการทำงานแบบ boosting ซึ่งเป็นแนวความคิดจาก (Freund & Schapire, 1999) สิ่งที่ C5.0 ต้องทำเพิ่มเติมคือ การหาค่า Gain ratio

SVM (Support Vector Machine)

ถูกนำเสนอขึ้นโดย Vapnik V. ในปี 1960 (Joachims, 1998) หลักการของ SVM คือการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นหรือระนาบที่เหมาะสมในการจำแนกประเภทข้อมูล โดย SVM จะทำการแมพ (Map) เวกเตอร์ใน Input space ให้เข้าสู่ Feature space โดยใช้ฟังก์ชันหรือเรียกว่าเคอร์เนล (Kernel) ชนิดต่างๆ เช่น โพลิโนเมียล (Polynomial) ซิกมอยด์ (Sigmoid) เป็นต้น

ตัวจำแนกที่ทำงานร่วมกัน หรือ เอ็นเซมเบิล (Ensemble)

มีความหมายเดียวกับ Committee, Classifier fusion และ Combination โดยเป็นวิธีการที่อาศัยตัวจำแนกมากกว่าหนึ่งตัว (Base Classifiers) ซึ่งแต่ละตัวจำแนกมีกระบวนการทำงานของตัวเอง และทุกตัวจำแนกกระทำกับข้อมูลเดียวกัน เมื่อได้ผลการจำแนกของแต่ละตัวจำแนกแล้วก็จะนำผลเหล่านั้นมาผ่านวิธีการรวบรวม (Combination, Integration หรือ Vote) และตัดสินสุดท้าย (Final Decision) เพื่อให้ได้เพียงผลการจำแนกเดียวเท่านั้น



ภาพที่ 2 ตัวจำแนกที่ทำงานร่วมกัน

ประเภทของวิธีการด้านตัวจำแนกที่ทำงานร่วมกัน (Categorization Scheme of Classifier Ensemble) ตัวจำแนกที่ทำงานร่วมกันถูกนำไปใช้และศึกษาวิจัยเพื่อปรับปรุงประสิทธิภาพแล้วหลายวิธี ขึ้นอยู่กับความเหมาะสมและข้อจำกัดของงาน ในที่นี้แบ่งเป็น 4 ประเภทตามงานวิจัยของ Shixin (2003) เป็นหลัก ดังนี้

ก. ตัวจำแนกที่ทำงานร่วมกันโดยจัดการข้อมูลฝึก (Classifier Ensemble Manipulation Training Sample)

ข. ตัวจำแนกที่ทำงานร่วมกันกับลักษณะเด่นของชุดข้อมูลนำเข้าย่อยที่แตกต่างกัน (Classifier Ensemble with Difference Input Feature Subsets) คือพิจารณาลักษณะเด่นของชุดข้อมูลฝึกที่แตกต่างกัน เพื่อสร้างแต่ละตัวจำแนก

ค. ตัวจำแนกที่ทำงานร่วมกันโดยมีชนิดของตัวจำแนกที่แตกต่างกัน (Heterogeneous Classifier Ensemble) คือตัวจำแนกแต่ละตัวมีขั้นตอนการฝึกและการจำแนกต่างกัน

ง. ตัวจำแนกที่ทำงานร่วมกันโดยมีชนิดของตัวจำแนกที่เหมือนกัน (Homogeneous Classifier Ensemble) คือ แต่ละตัวจำแนกมีวิธีการฝึกและการจำแนกเหมือนกัน

วิธีการเรียนรู้ (Learning Algorithms) ที่ประสบความสำเร็จและเป็นที่รู้จักมากที่สุด จนถือได้ว่าเป็นตัวแทนของการพัฒนาในด้านตัวจำแนกที่ทำงานร่วมกันจนถึงปัจจุบันนี้ ก็คือวิธีการที่เรียกว่า แบ็กกิ้ง (Bagging) และ บูสต์ติง (Boosting)

วิธีการจำแนกข้อมูลด้วยการใช้แบ็กกิ้ง (Breiman, 1996) อาศัยชุดของการฝึกที่แตกต่างกันซึ่งถูกสร้างโดยการสุ่มและการแทนที่จากชุดข้อมูลดั้งเดิมสำหรับการฝึก (Bootstraps Replacement) และนำแต่ละชุดของการการฝึก ทำการฝึกเพื่อให้ได้ตัวจำแนกต่าง ๆ สำหรับการจำแนกข้อมูล แต่ละตัวจำแนกให้ผลลัพธ์ของแต่ละตัว ต่อมาจะนำผลลัพธ์เหล่านี้ไปผสมกันโดยวิธีการลงคะแนน

วิธีการจำแนกข้อมูลด้วยการใช้บูสต์ติง (Freund & Schapire, 1996) จะทำการปรับน้ำหนักให้กับข้อมูลฝึกและสร้างตัวจำแนกต่าง ๆ ขึ้นมาจากข้อมูลที่ถูกปรับน้ำหนักนั้นสุดท้ายอาศัยเทคนิคการลงคะแนนต่าง ๆ วิธีบูสต์ติง ถูกพิสูจน์ว่า แต่ละขั้นตอนของวิธีการเรียนรู้ที่อ่อนแอ (Weak Learning Algorithm) อาจจะถูกปรับปรุง (Boosted) ให้มีความเสถียรขึ้นได้วิธีการจำแนกข้อมูลโดยการใช้บูสต์ติง ที่มีประสิทธิภาพเป็นที่รู้จักโดยทั่วไปเช่น เอด้าบูสต์ (AdaBoost)

Cardiotocography Data Set

ข้อมูลที่ใช้ในการทดลองได้จาก UC Irvine Machine Learning Repository ซึ่งเป็นข้อมูลที่ประกอบด้วยอัตราการเต้นของหัวใจของทารกในครรภ์และการบีบตัวของมดลูกมารดา ที่ได้จากกระบวนการ Cardiotocography เพื่อใช้วิเคราะห์ความผิดปกติของทารกในครรภ์ได้จำแนกข้อมูลออกเป็นกลุ่ม 3 กลุ่ม N=ปกติ S=น่าสงสัย P=ผิดปกติและข้อมูลการจำแนกกลุ่มได้มีการตรวจสอบความถูกต้องโดยผู้เชี่ยวชาญเฉพาะทาง

ตารางที่ 1 คุณลักษณะของชุดข้อมูล Cardiotocography

Data Set Characteristics:	Multivariate
Attribute Characteristics:	Real
Associated Tasks:	Classification
Number of Instances:	2126
Number of Attributes:	23

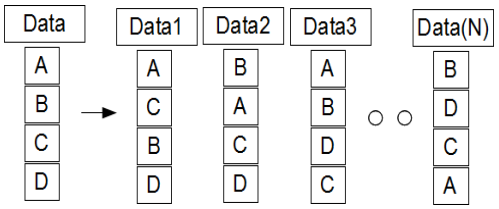
การจัดการชุดข้อมูลก่อนการทดลอง

การกำหนดจำนวนข้อมูลเรียนรู้และชุดทดสอบ

ชุดข้อมูลที่ใช้ในการทดลองถูกแบ่งเป็น 2 ส่วนคือ ข้อมูลที่ใช้ในการเรียนรู้ของตัวจำแนก (Training data) และข้อมูลที่ใช้ตรวจสอบความถูกต้องของตัวจำแนก (Testing data) การแบ่งข้อมูลผู้วิจัยได้กำหนดอัตราส่วนระหว่าง training data และ testing data เป็น 70:30 ตามงานวิจัยของธนกร ญาณกาย (ธนกร, 2554) ซึ่งให้ผลดีที่สุด

การเตรียมข้อมูลเพื่อใช้ในการเรียนรู้แบบกลุ่ม

การทดลองโดยใช้การเรียนรู้แบบกลุ่มต้องมีการเพิ่มจำนวนชุดข้อมูลที่ใช้ในการวิจัย ตามจำนวนสมาชิกของการเรียนรู้แบบกลุ่ม โดยการเพิ่มข้อมูลใช้หลักการ bootstrap จำนวนชุดข้อมูลที่เพิ่มขึ้นมาจะมีรายละเอียดภายในคล้ายกันแตกต่างกันเพียงลำดับการเรียงข้อมูล



ภาพที่ 3 การเตรียมข้อมูลโดยหลักการ bootstrap เพื่อใช้ในการเรียนรู้แบบกลุ่ม

การวัดผลการทดลอง

การวิเคราะห์ประสิทธิภาพของการจำแนกข้อมูลของตัวจำแนกโครงข่าย RBF, C5.0 และ SVM วัดจากความถูกต้องในการจำแนกข้อมูลเปรียบเทียบกับชุดข้อมูลจริง ส่วนการวัดประสิทธิภาพของการเอ็นเซมเบิลค่าที่ได้คือการจำแนกข้อมูลร่วมกันของแต่ละสมาชิกและเปรียบเทียบกับชุดข้อมูลจริง

วิธีการวิจัย

การจำแนกข้อมูลแบ่งเป็น 4 ส่วนคือ การจำแนกข้อมูลโดยโครงข่ายประสาทเทียมแบบ RBF, C5.0, SVM และการจำแนกข้อมูลโดยการเอ็นเซมเบิลด้วยวิธีการแบ็กกิ้ง และตัดสินผลการจำแนกข้อมูลจากทุก ๆ ตัวจำแนก ด้วยวิธีการ Voting

1. จากงานวิจัยของ ธนกร ญาณกาย (ธนกร, 2554) การจำแนกข้อมูลโดยการเอ็นเซมเบิลโครงข่ายประสาทเทียมแบบ RBF มีการกำหนดค่าข้อมูลที่ใช้ในการเรียนรู้เป็น 70 % ของข้อมูลทั้งหมด กำหนดค่า RBF cluster 200 และใช้ตัวจำแนกจำนวน 7 ตัว ดังตารางที่ 2 ซึ่งให้ประสิทธิภาพดีที่สุด

ตารางที่ 2 ผลการจำแนกข้อมูลโดยการเอ็นเซมเบิลโครงข่ายประสาทเทียมแบบ RBF

Members	RBF1 (%)	RBF2 (%)	RBF3 (%)	RBF4 (%)	RBF5 (%)	RBF6 (%)	RBF7 (%)	Voting Correct (%)
3	98.35	98.65	98.65					99.09
4	98.35	98.65	98.65	98.05				99.09
5	98.35	98.65	98.65	98.05	98.50			99.24
6	98.35	98.65	98.65	98.05	98.50	98.65		99.24
7	98.35	98.65	98.65	98.05	98.50	98.65	98.50	99.24

2. การจำแนกข้อมูลของอัลกอริทึม C5.0

ตารางที่ 3 ความถูกต้องของการจำแนกข้อมูลของอัลกอริทึม C5.0

Boosting	Cross-validate	Correct (%)
10	10	97.89
10	20	97.89
10	30	97.89
20	10	97.89
20	20	97.89
20	30	97.89
30	10	97.89
30	20	97.89
30	30	97.89
50	10	97.89
50	20	97.89
50	30	97.89
100	10	97.89
100	20	97.89
100	30	97.89

3. การจำแนกข้อมูลของ SVM โดยใช้เคอร์เนลโพลิโนเมียล (Kernel Polynomial)

ตารางที่ 4 ความถูกต้องของการจำแนกข้อมูลของ SVM

C	Degree	Correct (%)
10	2	78.20
50	2	78.20
100	2	78.20
10	3	78.20
50	3	78.20
100	3	78.20
10	4	78.20
50	4	78.20
100	4	78.20
10	5	78.20
50	5	78.20
100	5	78.20

4. การจำแนกข้อมูลโดยการเอ็นเซมเบิลโครงข่าย RBF, C5.0 และ SVM

ตารางที่ 5 ความถูกต้องของการจำแนกข้อมูลโดยการเอ็นเซมเบิลโครงข่าย RBF, C5.0 และ SVM

RBF1 (%)	RBF2 (%)	RBF3 (%)	RBF4 (%)	RBF5 (%)	RBF6 (%)	RBF7 (%)	C5.0 (1)	C5.0 (2)	SVM1	Voting
										Correct (%)
98.35	98.65	98.65	98.05	98.50	98.65	98.50				99.24
98.35	98.65	98.65	98.05	98.50	98.65	98.50	98.20			99.39
98.35	98.65	98.65	98.05	98.50	98.65	98.50	98.20	97.74		99.39
98.35	98.65	98.65	98.05	98.50	98.65	98.50	98.20	97.74	78.20	99.81

ผลการวิจัย

จากการทดลองพบว่าการจำแนกข้อมูลโดยใช้อัลกอริทึม C5.0 จากตารางที่ 3 การปรับค่า Boosting และ Cross-Validate ไม่มีผลต่อประสิทธิภาพการจำแนก ซึ่งให้ค่าความถูกต้องเท่ากับ 97.89 และการจำแนกข้อมูลของ SVM โดยใช้เคอร์เนลโพลิโนเมียล การปรับค่า C และ Degree ก็ไม่มีผลต่อประสิทธิภาพการจำแนกเช่นกัน ซึ่งให้ค่าความถูกต้อง 78.20 ดังตารางที่ 4 จากงานวิจัยของธนกร ญาณกาย(ธนกร, 2554) การจำแนกข้อมูลโดยการเอ็นเซมเบิลโครงข่ายประสาทเทียมแบบ RBF มีการกำหนดค่าข้อมูลที่ใช้ในการเรียนรู้เป็น 70 % ของข้อมูลทั้งหมด กำหนดค่า RBF cluster 200 และใช้ตัวจำแนกชนิดเดียวกันจำนวน 7 ตัว ได้ค่าความถูกต้อง 99.24 ซึ่งให้ประสิทธิภาพดีที่สุดดังตารางที่ 2 นั้น พบว่าสามารถเพิ่มค่าความถูกต้องขึ้นไปเป็น 99.81 โดยเพิ่มจำนวนสมาชิกของการเอ็นเซมเบิลด้วยอัลกอริทึม C5.0 และ SVM ไปอีก 2 ตัว และ 1 ตัว ตามลำดับ ซึ่งเป็นตัวจำแนกต่างชนิดกันดังตารางที่ 5

การอภิปรายผลและสรุปผล

บทความนี้ได้นำเสนอการนำทฤษฎีการเอ็นเซมเบิลที่ใช้ตัวจำแนกที่ทำงานร่วมกันโดยมีชนิดของตัวจำแนกที่แตกต่างกัน ซึ่งประกอบด้วยโครงข่าย RBF, C5.0 และ SVM สามารถเพิ่มประสิทธิภาพการทำงานของการคัดแยกข้อมูลทางการแพทย์ที่ต้องอาศัยการวินิจฉัยที่มีความถูกต้องแม่นยำสูง สามารถวิเคราะห์ได้ผลมีความถูกต้องแม่นยำมากขึ้นกว่าการเอ็นเซมเบิลโครงข่าย RBF จำนวน 7 ตัว ซึ่งเป็นตัวจำแนกที่ทำงานร่วมกันโดยมีชนิดของตัวจำแนกที่เหมือนกัน นอกจากให้ความแม่นยำแล้วนั้น C5.0 และ SVM ยังสามารถวิเคราะห์ได้อย่างรวดเร็ว ซึ่งจากการค้นพบครั้งนี้สามารถนำไปประยุกต์ใช้ในงานอื่นๆ ได้

เอกสารอ้างอิง

กาญจนา แสงทองพัฒนา . 2548. การจำแนกประเภทเว็บเพจโดยใช้ลักษณะเด่นบนตัวจำแนกแบบเบย์สอย่างง่ายที่ทำงานร่วมกัน. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์.

ธนกร ญาณกาย. 2554. กลไลเรเดียลเบสิสฟังก์ชั่นแบบคณะกรรมการเพื่อการจำแนกข้อมูลความผิดปกติของทารกในครรภ์. การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาประจำปี ครั้งที่ 12. มหาวิทยาลัยขอนแก่น.

Breiman, L. 1996. Bagging predictors. Machine Learning. 24(2): 123-140.

Bors, A. n.d. Introduction of the radial basis function. Department of computer science. UK.: University of York.

Freund, Y. and R.E. Schapire. 1996. Experiments with a new boosting algorithm, pp. 148-156. Proceedings of the Thirteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann.

_____. 1999. A Short Introduction to Boosting, Journal of Japanese Society for Artificial Intelligence. 14(5): 771-780.

Joachims, T. 1998. “Text categorization with support vector machines : Learning with many relevant features”. In European Conference on Machine Learning (F.CML).

Quinlan, J.R. 1999. Simplifying Decision Trees. Technology Square. Cambridge. MA.

Shixin, Y. 2003. Feature selection and classifier ensembles: A Study on Hyperspectral Remote Sensing Data. Ph.D. thesis, University of Antwerp (CMI).