การวิเคราะห์ความทนทานของการจำแนกแบบรวมกลุ่ม

A Robust Hybrid Classification Method Analysis

กาญจน์ ณ ศรีธะ[1*], จารี ทองคำ[2], วาทินี สุขมาก[3]

Karn Na Sritha[1*],JareeThongkam[2], VatineeSukmak[3]

Received: 10 January 2014 ; Accepted: 5 March 2014

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ความทนทานของเทคนิคสำหรับการจำแนกแบบรวมกลุ่ม ด้วยเทคนิคแบ็กกิ้ง และ เอดาบูท โดยใช้การเรียนรู้ของเทคนิคต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และซัพพอร์ทเวกเตอร์แมชชีน เพื่อการจำแนกกลุ่มในชุดข้อมูล Colon Tumor, Pima Diabetes และ Molecular Biology Promoter ในการทดลองได้ทำการเพิ่มข้อมูลรบกวนในชุดข้อมูลที่ระดับ 0%,10%, 20%, 30%, 40% และ 50% จากนั้นใช้ 10-Fold Cross Validation สำหรับการแบ่งชุดข้อมูลสอน และชุดข้อมูลทดสอบ ใช้ค่า Precision, Recall และ F-Measure ในการแสดงค่าประสิทธิภาพของเทคนิคสำหรับการจำแนก ผลการทดลองพบว่าเทคนิคAD+C4.5 มีความทนทานดีที่สุดกับชุดข้อมูล Colon Tumorส่วนเทคนิคAD+ANN มีความทนทานที่ดีกับชุดข้อมูล Pima Diabetes และเทคนิคBG+SVM มีความทนทานที่ดีกว่า AD+C4.5 และ AD+ANN กับชุดข้อมูล Molecular Biology Promoter

คำสำคัญ: ข้อมูลรบกวน ความทนทาน เทคนิคการจำแนกแบบรวมกลุ่ม

Abstract

This paper aims at analyzing the robustness of hybrid classification techniques including Bagging(BG) and Adaboost(AD) with base learners Decision Trees (C4.5), Artificial Neural Network (ANN) and Support Vector Machine (SVM) for predicting groups in The Colon Tumor, Pima Diabetes and Molecular Biology Promoter data sets. The experiments were conducted adding 0%, 10%, 20%, 30%, 40% and 50% noise. Ten-fold cross validation was utilized to divide the data into training and testing datasets. Precision, Recall and F-Measure were used to evaluate the performance of the models. The experimental results showed that the technique AD+C4.5 has the greatest robustness in the colon tumor data set while AD+ANN has the top robustness in the pima diabetes data set. BG+SVM is better than AD+C4.5 and AD+ANN in molecular biology promoters data set.

Keywords : noisy data, robust, hybrid classification

บทนำ

เทคนิคสำหรับการจำแนกที่นิยมในปัจจุบันแบ่งได้เป็น 2 ลักษณะตามการทำงาน คือ แบบเดี่ยว และ แบบรวมกลุ่ม เทคนิคสำหรับการจำแนกแบบเดี่ยวมีขั้นตอนการทำงานโดยการรับข้อมูลเข้ามาเรียนรู้แล้วทำการจำแนกข้อมูลตามที่เทคนิคได้เรียนรู้มา ส่วนในเทคนิคสำหรับการจำแนกแบบรวมกลุ่มจะมีเทคนิคส่งเสริมสำหรับช่วยในการจัดกลุ่มข้อมูลเพื่อส่งข้อมูลให้กับเทคนิคสำหรับการจำแนกข้อมูลอีกทีหนึ่ง ทำให้การเรียนรู้ของเทคนิคสำหรับการจำแนกมีประสิทธิภาพสูงขึ้น ซึ่งโดยส่วนมากเทคนิควิธีแบบรวมกลุ่มจะมีประสิทธิภาพสำหรับการจำแนกข้อมูลที่ดีกว่าเทคนิควิธีแบบเดี่ยว[1, 2]

[1] *นิสิตปริญญาโท, [2]ผู้ช่วยศาสตราจารย์, คณะวิทยาการสารสนเทศ, [3]รองศาสตราจารย์, คณะพยาบาลศาสตร์ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150*

[1] *Master degree student, [2]Assist. Prof., Faculty of Informatics, Mahasarakham University, Kantharawichai District, Mahasarakham 44150, Thailand.*

[3] *Assoc. Prof., Faculty of Nursing, Mahasarakham University, Kantharawichai District, Mahasarakham 44150, Thailand.*

* *Corresponding author; Karn Na Sritha, Faculty of Informatics, Mahasarakham University, Kantharawichai District, Mahasarakham 44150, Thailand.karn.n@windowslive.com*

การที่เทคนิคแบบรวมกลุ่มส่วนมากมีประสิทธิภาพที่ดีกว่าเทคนิคแบบเดี่ยว จึงมีนักวิจัยให้ความสนใจ นำเทคนิคแบบรวมกลุ่มไปสร้างแบบจำลองสำหรับพยากรณ์ข้อมูลของตน แต่ในบางข้อมูลเมื่อทำการพยากรณ์ออกมาแล้วปรากฏว่าประสิทธิภาพที่ได้ไม่เป็นที่น่าพอใจ สืบเนื่องมากจากข้อมูลจริงที่ใช้นั้นมีข้อมูลรบกวนปะปนอยู่ ซึ่งข้อมูลรบกวนเหล่านี้เป็นข้อมูลที่ทำการกำจัดได้ยาก เนื่องจากเป็นข้อมูลที่มีอยู่จริงแต่ค่าที่ระบุไว้นั้นเป็นค่าที่ผิดไปจากค่าที่เป็นจริง จึงทำให้ยากต่อการกำจัดด้วยเทคนิคสำหรับการจำแนกได้ ผู้ใช้จึงจำเป็นต้องค้นหาเทคนิคที่มีความเหมาะสม และมีความทนทานต่อข้อมูลรบกวนมาใช้สำหรับการจำแนกข้อมูลที่มีข้อมูลรบกวนอยู่[3, 4] การที่ในข้อมูลจริงมีข้อมูลรบกวนปะปนอยู่ จะส่งผลให้แบบจำลองที่สร้างขึ้นจากข้อมูลชุดที่มีข้อมูลรบกวน มีประสิทธิภาพในการพยากรณ์ที่ต่ำ จนไม่สามารถนำแบบจำลองที่ได้ไปใช้ในการพยากรณ์ข้อมูลในอนาคตได้ ซึ่งโดยส่วนใหญ่ในงานวิจัยจะให้ความสนใจกับข้อมูลรบกวนที่คลาสในการวิเคราะห์ความทนทานของแบบจำลองเพื่อการพยากรณ์[5, 6]

จากปัญหาที่ในชุดข้อมูลมีข้อมูลรบกวนปะปนอยู่ จึงมีนักวิจัยได้ทำการทดสอบความทนทานของเทคนิคในการจำแนกแบบรวมกลุ่ม เช่น Hong Hu และคณะ[7]ใช้เทคนิคC4.5, Bagging+C4.5, Adaboost+C4.5, Random Forest, CS4 และ MDMT กับชุดข้อมูล Microarray Cancer ทำการเพิ่มข้อมูลรบกวนที่ระดับ 0%, 20% และ 60% ผลการทดลองพบว่า เทคนิค MDMT มีความทนทานต่อข้อมูลรบกวนดี Thomas Dietterichและคณะ[8]ได้ทดสอบความทนของเทคนิคในกลุ่มต้นไม้ตัดสินใจ โดยใช้เทคนิค Bagging+C4.5, Adaboost+C4.5 และ Randomization+C4.5กับชุดข้อมูลพนักงานจำนวน 33 คน เพิ่มข้อมูลรบกวนที่ระดับ 0%, 5%, 10% และ 20% ผลการทดลองพบว่า เทคนิค Bagging+C4.5 และ Randomization+C4.5 มีความทนต่อข้อมูลรบกวนดี นิตยา เกิดประสพ และคณะ[9]ได้ทดสอบความทนของเทคนิคต้นไม้ตัดสินใจเชิง โดยใช้เทคนิค ID3 และ Robust Tree กับชุดข้อมูล Monk, Audiology, Breast Cancer และ Vote ทำการเพิ่มข้อมูลรบกวนที่ระดับ 0%, 1%, 5%, 10%, 20%, 25% และ 30% จากการวิจัยพบว่า Robust Tree มีความทนมากกว่า ID3จากงานวิจัยเหล่านี้จะพบว่า เทคนิค Bagging และ Adaboostซึ่งเป็นเทคนิคส่งเสริม และ C4.5 ซึ่งเป็นเทคนิคสำหรับการจำแนก เมื่อนำมารวมกลุ่มกันจะมีประสิทธิภาพที่ดีขึ้น และมีความทนทานต่อข้อมูลรบกวนที่ดีขึ้นด้วย แต่ในปัจจุบันมีเทคนิคสำหรับการจำแนกเพิ่มมากขึ้นและเป็นที่นิยมใช้อย่างแพร่หลาย เช่น SVM และ ANN ซึ่งมีประสิทธิภาพที่ดีสำหรับการจำแนกข้อมูล จึงมีนักวิจัยนำเทคนิคเหล่านี้มารวมกลุ่มเพื่อใช้ในการจำแนกข้อมูลขึ้น[1, 2, 10-12]ซึ่งในงานวิจัยเหล่านี้เป็นการทดสอบประสิทธิภาพของเทคนิคสำหรับการจำแนกเท่านั้น ยังขาดในส่วนของการทดสอบความทนทานของเทคนิคสำหรับการจำแนกข้อมูลที่มีข้อมูลรบกวนอยู่

ดังนั้นในงานวิจัยนี้จึงขอเสนอ การวิเคราะห์ความทนทานของการจำแนกแบบรวมกลุ่ม โดยใช้เทคนิคสำหรับการจำแนก คือ C4.5, Support Vector Machine และ Artificial Neural Network รวมกลุ่มกับเทคนิคแบบรวมกลุ่ม คือ Bagging และ Adaboostทำการทดสอบกับชุดข้อมูลมาตรฐาน 3 ชุด คือ Colon Tumor, Pima Diabetes และ Molecular Biology Promoters เพื่อให้ทราบ และเป็นการเปรียบเทียบความทนทานของเทคนิคแบบรวมกลุ่มสำหรับการจำแนกข้อมูลที่มีข้อมูลรบกวนในระดับต่าง ๆ กัน

ทฤษฎีที่เกี่ยวข้อง

ข้อมูลรบกวน

ข้อมูลรบกวน (Noise Data)[3-9]คือ ข้อมูลในความเป็นจริงที่มีความบกพร่อง เช่น ข้อมูลมีค่าที่ผิดพลาดแบบสุ่ม (Error) หรือมีค่าผิดปกติคลาดเคลื่อน (Outliers) สาเหตุของความผิดพลาดสืบเนื่องมาจากอุปกรณ์เก็บรวบรวมข้อมูลทำงานผิดพลาด ปัญหาการบันทึกหรือป้อนค่าข้อมูลผิดพลาด ปัญหาการส่งข้อมูล (Data Transmission) ผิดพลาด และข้อจำกัดทางเทคโนโลยี เช่น ข้อจำกัดของขนาดบัฟเฟอร์ และขนาดพื้นที่จัดเก็บข้อมูล ข้อมูลรบกวนมี 2 ประเภท คือ ข้อมูลรบกวนที่คลาส (Class Noise) และ ข้อมูลรบกวนที่ตัวแปร (Attribute Noise)

1) ข้อมูลรบกวนที่คลาส คือ ข้อมูลในคลาสที่มีความคลาดเคลื่อนของข้อมูล หรือมีค่าที่ผิดปกติไปจากความเป็นจริงซึ่งเกิดจากความผิดพลาดในการบันทึก และความผิดพลาดของเครื่องมือในการบันทึกข้อมูล

2) ข้อมูลรบกวนที่ตัวแปร คือ ข้อมูลในตัวแปรที่มีค่าคลาดเคลื่อนไป หรือผิดปกติของข้อมูลไปจากความเป็นจริงเกิดขึ้นได้จากการเก็บบันทึกข้อมูล

เทคนิคซัพพอร์ทเวกเตอร์แมทชีน

ซัพพอร์ทเวกเตอร์แมทชีน(Support Vector Machine : SVM)[13]เป็นเทคนิคที่มีกระบวนการปรับรูปแบบข้อมูล จากข้อมูลที่มีมิติต่ำ (Low Dimension Dataset) บนพื้นที่ข้อมูลนำเข้า (Input Space) ให้อยู่ในรูปแบบของข้อมูลที่มีมิติสูง (High Dimension Dataset) บนพื้นที่ข้อมูลคุณลักษณะ (Feature Space) โดยใช้ฟังก์ชันในการปรับรูปแบบข้อมูลที่มีโครงสร้างแบบเส้นตรง (Linear Classifier) และ สามารถสร้างพื้นที่ระยะห่างระหว่างตัวจัดประเภทข้อมูลเองกับค่าที่ใกล้ที่สุดของแต่ละ

กลุ่มข้อมูลได้มากที่สุด ซึ่งเส้นที่เหมาะสมดังกล่าวถูกเรียกว่า ระนาบแบ่งเขตข้อมูลที่เหมาะสม (The Optimal Separating Hyperplane)

เทคนิคโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Network : ANN)[14]เป็นเทคนิคที่มีโครงสร้างเป็นชั้น ๆ ข้อมูลที่นำเข้ามาเรียนรู้จะถูกนำเข้าไปในชั้นแรกสุดซึ่งเป็นชั้นรับข้อมูล (Input Layer) และเมื่อผ่านการคำนวณจากชั้นแรกแล้ว ผลลัพธ์จะถูกส่งต่อไปยังชั้นกลางหรือชั้นซ่อนของโครงข่าย (Hidden Layer) ซึ่งในแต่ละหน่วยของชั้นนี้จะรับข้อมูลจากทุกหน่วยในชั้นก่อนหน้ามาคำนวณ แล้วส่งต่อไปยังชั้นถัดไปและเมื่อข้อมูลถูกส่งต่อกันมาจนถึงชั้นสุดท้าย (Output Layer) จะได้ผลลัพธ์ออกมาการส่งผ่านข้อมูลต่อ ๆ กันไปแบบนี้เป็นการส่งต่อข้อมูลแบบการป้อนไปข้างหน้า (Feed Forward) หากความคลาดเคลื่อนจากเป้าหมายมากเกินไประบบจะนำค่าความคลาดเคลื่อนนี้ไปปรับน้ำหนักการเรียนรู้ใหม่ (Weight) แบบแพร่ย้อนกลับ (Back-Propagation) ซึ่งเป็นการปรับน้ำหนักความคลาดเคลื่อนจากชั้นผลลัพธ์ไปยังชั้นก่อนหน้าและทำการปรับน้ำหนักไปเรื่อย ๆ จนถึงชั้นรับข้อมูล

เทคนิคต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจ (Decision Tree : C4.5)[15]เป็นเทคนิคสำหรับการพยากรณ์ข้อมูลที่นิยมใช้ในเหมืองข้อมูล โดยมีรูปแบบการทำงานแบบต้นไม้ โดยการเรียนรู้จากชุดข้อมูลสอนแล้วสร้างแบบจำลองขึ้น ซึ่งเลือกปัจจัยที่มีค่า Gain Ratio สูงที่สุดเป็นโหนดราก รูปแบบของต้นไม้จะประกอบไปด้วย โหนดแรก เรียกว่า โหนดราก (Root Node) โหนดถัดลงมาจะถูกเรียกว่า โหนดลูก (Children Node) และ โหนดสุดท้ายเรียกว่า โหนดปลาย (Leaf Node)

เทคนิคแบ็กกิ้ง

แบ็กกิ้ง(Bagging : BG)[16]คือเทคนิคแบบรวมกลุ่มสำหรับช่วยเพิ่มประสิทธิภาพในการสร้างแบบจำลอง มีหลักการทำงานคือ นำชุดข้อมูลมาสร้างชุดข้อมูลสอน (Training Data) หลาย ๆ ชุดด้วยวิธีสุ่มเลือกด้วยวิธีการBootstrap นำชุดข้อมูลสอนเหล่านี้มาสร้างแบบจำลอง จากนั้นทำการคัดเลือกแบบจำลองดีที่สุดมาเป็นแบบจำลองต้นแบบ

เทคนิคเอดาบูท

เอดาบูท (Adaboost : AD)[17]เป็นเทคนิคแบบรวมกลุ่มสำหรับช่วยเพิ่มประสิทธิภาพในการสร้างแบบจำลอง ด้วยการสร้างชุดข้อมูลสอนโดยกำหนดน้ำหนัก (Weight) ให้กับข้อมูลสอนแต่ละตัว แล้วนำข้อมูลสอนแต่ละตัวไปสร้างแบบจำลอง ในแต่ละรอบที่มีการสร้างแบบจำลองใหม่ค่าน้ำหนักจะถูกเปลี่ยนไปตามความผิดพลาดของผลลัพธ์ที่แบบจำลองในรอบนั้น ๆ กระทำต่อข้อมูล ถ้าแบบจำลองตอบผลลัพธ์ถูกสำหรับข้อมูลสอนตัวใด ข้อมูลสอนตัวนั้นก็จะถูกลดค่าน้ำหนัก และในทางตรงข้ามถ้าแบบจำลองตอบผลลัพธ์ผิดสำหรับข้อมูลสอนตัวใด ข้อมูลสอนตัวนั้นก็จะถูกเพิ่มค่าน้ำหนัก โดยค่าน้ำหนักที่ได้จะใช้สำหรับกำหนดความน่าจะเป็นที่ข้อมูลตัวนั้นจะถูกเลือกให้อยู่ในชุดข้อมูลสอนชุดต่อ ๆ ไป ซึ่งในตอนแรกข้อมูลทุกตัวจะถูกกำหนดให้มีค่าความน่าจะเป็นที่ถูกเลือกเท่ากัน

ขั้นตอนการวิจัย

ในงานวิจัยนี้เป็นการทดสอบความทนต่อข้อมูลรบกวนของแบบจำลองโดยมีขั้นตอนวิธีวิจัยดังนี้ 1.การเตรียมชุดข้อมูล 2.การเพิ่มข้อมูลรบกวน 3.การแบ่งชุดข้อมูลสอนชุดข้อมูลทดสอบและ 4.การทดสอบประสิทธิภาพของแบบจำลอง Figure 1

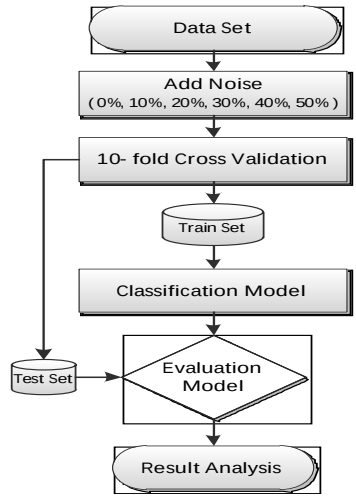

Figure 1 Research Method

ชุดข้อมูลที่ใช้ในงานวิจัย

งานวิจัยนี้ได้นำชุดข้อมูล 3 ชุด คือชุดข้อมูลจาก The Kent Ridge Biological Data Set Repository จำนวน 1 ชุด คือ Colon Tumor[18]และ ชุดข้อมูลจาก UCI 2 ชุด ประกอบไปด้วย Pima Diabetes[19]และ Molecular Biology Promoters[20] ซึ่งเป็นข้อมูลที่มีความแตกต่างกันในลักษณะ Attribute น้อย แต่ Instance มาก กับ ข้อมูลที่มีจำนวน Attribute มาก แต่ Instance น้อย ลักษณะของข้อมูลมีทั้งแบบ Numeric และ Nominal เพื่อให้มีความแตกต่างกันในชุดข้อมูลดัง Table1

Table 1 Description of the data sets

Data set	Data Type	Instance	Attributes	Class
Molecular Biology Promoters	Nominal	106	59	2
Pima Diabetes	Numeric	768	9	2
Colon Tumor	Numeric	62	2001	2

การเพิ่มข้อมูลรบกวน

ในงานวิจัยนี้เพิ่มข้อมูลรบกวน โดยการสุ่มข้อมูลรบกวนที่ชั้นคลาสในระดับ 0%, 10%, 20%, 30%, 40%และ 50% เพื่อให้ทราบถึงความทนของแบบจำลองต่อข้อมูลรบกวนที่ระดับต่าง ๆ กัน โดยใช้โปรแกรม Weka3.7.10[21]

การวัดประสิทธิภาพ

งานวิจัยนี้วัดประสิทธิภาพของเทคนิคโดยใช้ 10-fold cross validation[22] สำหรับการแบ่งชุดข้อมูลสอน (Training Data) กับชุดข้อมูลทดสอบ (Testing Data) โดยข้อมูลจะถูกแบ่งออกเป็น 10 ชุดเท่าๆกัน ใช้ 9 ชุดเป็นชุดข้อมูลสอน และที่เหลือ 1 ชุด เป็นชุดข้อมูลทดสอบ ทำแบบเปลี่ยนกลุ่มไปจนครบ 10 ชุด ใช้ค่าความแม่นยำ (Precision) ดังสมการที่ 1ค่าการเรียกคืน (Recall) ดังสมการที่ 2ค่า F-Measure ดังสมการที่ 3 ในการวัดประสิทธิภาพของแต่ละแบบจำลองที่มีข้อมูลรบกวนในระดับ 0%, 10%, 20%, 30%, 40%และ 50%

$$\Pr ecision = \frac{tp}{tp + fp} \quad (1)$$

$$\mathrm{Re}call = \frac{tp}{tp + fn} \quad (2)$$

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (3)$$

โดยที่

คือ จำนวน True Positive

คือ จำนวนTrue Negative

คือ จำนวน False Positive

คือ จำนวน False Negative

ผลการทดลอง

ในการทดลองนี้ผู้วิจัยได้ใช้เครื่องคอมพิวเตอร์ส่วนบุคคล ความเร็วCPUIntel I7 8602.8 GHz. หน่วยความหลัก 8 Gb. หน่วยความจำสำรอง500 Gb. โปรแกรม Wekaversion 3.7.10 เพิ่มข้อมูลรบกวนโดยการสุ่มข้อมูลรบกวนที่คลาส ในระดับ 0%, 10%, 20%, 30%, 40%และ 50% ใช้เทคนิค 10-fold cross-validation ในการทดสอบแบบจำลอง ใช้ค่าความแม่นยำ (Precision)ค่าการเรียกคืน (Recall) และ ค่า F-Measure ในการวัดค่าประสิทธิภาพของแบบจำลอง

การเปรียบเทียบค่าความแม่นยำ

ค่าความแม่นยำ (Precision)[23] เป็นค่าที่แสดงถึงประสิทธิภาพของแบบจำลองในการขจัดข้อมูลที่ไม่เกี่ยวข้องได้ถูกต้อง ในการพยากรณ์กลุ่มของชุดข้อมูล Pima Diabetes, Molecular Biology Promoters และ Colon Tumor ที่มีข้อมูลรบกวนสามารถแสดงค่าความแม่นยำของการพยากรณ์ได้ดังตารางที่ 2

จากตารางที่ 2แสดงการเปรียบเทียบค่าความแม่นยำที่แบบจำลองสามารถพยากรณ์ได้ จากการทดลองพบว่า ในข้อมูล Colon Tumor ที่ข้อมูลรบกวนระดับ 0% แบบจำลอง C4.5 มีค่าความแม่นยำสูงสุด รองลงมาคือ BG+C4.5 และ AD+C4.5 เมื่อเพิ่มข้อมูลรบกวนถึง 50% พบว่าแบบจำลอง AD+ANN มีค่าความแม่นยำสูงที่สุด รองลงมาคือ AD+C4.5 และ ANN ในข้อมูล Pima Diabetes ในระดับไม่มีข้อมูลรบกวน พบว่าแบบจำลอง ANN มีค่าความแม่นยำสูงสุดรองลงมาคือ AD+ANN และ C4.5 เมื่อเพิ่มข้อมูลรบกวนที่ 50% พบว่าแบบจำลอง ANN ยังมีค่าความแม่นยำสูงสุดอยู่รองลงมาคือ AD+ANN และ BG+C4.5ในข้อมูล Molecular Biology Promoters พบว่าแบบจำลอง SVM มีค่าความแม่นยำสูงที่สุดจากระดับที่ไม่มีข้อมูลรบกวน ถึงระดับ 30% รองลงมาคือ BG+SVM และ AD+SVM แต่ในระดับข้อมูลรบกวน

Table 2 Precision score of the data sets

Noise	SVM	ANN	C4.5	BG+SVM	AD+SVM	BG+ANN	AD+ANN	BG+C4.5	AD+C4.5
					Precisionx 100 (%)				
					Colon Tumor				
0%	42.09	68.64	**82.92**	42.09	42.09	44.57	77.27	82.55	80.95
10%	38.20	69.26	68.21	38.20	38.20	45.50	**75.35**	69.32	71.21
20%	34.31	59.25	68.29	34.31	33.98	37.36	**69.41**	65.44	66.81
30%	32.36	50.36	65.05	32.03	31.36	36.79	55.38	62.18	**68.82**
40%	28.48	51.52	53.10	26.57	26.86	35.46	**58.50**	58.07	56.07
50%	23.67	49.81	49.25	24.53	25.39	32.45	**57.83**	48.09	51.18
					Pima Diabetes				
0%	42.39	**76.40**	74.65	42.39	43.81	42.39	76.15	74.46	71.59
10%	38.58	**68.39**	67.55	38.56	38.52	38.58	68.06	66.76	65.95
20%	34.92	**62.50**	61.25	34.93	34.90	34.95	62.27	59.35	59.84
30%	31.32	54.34	51.78	35.34	37.97	31.83	**56.04**	53.04	52.77
40%	35.83	47.98	28.52	33.68	34.78	35.01	**49.78**	47.97	28.52
50%	44.30	**53.28**	31.74	37.51	30.78	32.60	52.84	50.22	31.74
					Molecular Biology Promoters				
0%	**98.65**	92.87	80.49	98.52	98.47	92.95	92.87	86.53	87.60
10%	**90.27**	76.30	73.05	89.73	89.14	79.00	76.30	76.79	74.52
20%	**79.50**	55.93	67.87	79.01	77.03	59.25	55.46	41.77	67.87
30%	**63.86**	49.23	60.32	62.58	59.67	50.69	48.55	25.96	60.15
40%	51.74	54.95	**63.12**	54.13	51.44	55.51	54.59	40.42	**63.12**
50%	51.96	48.63	51.18	**54.06**	53.63	51.34	48.25	25.65	51.18

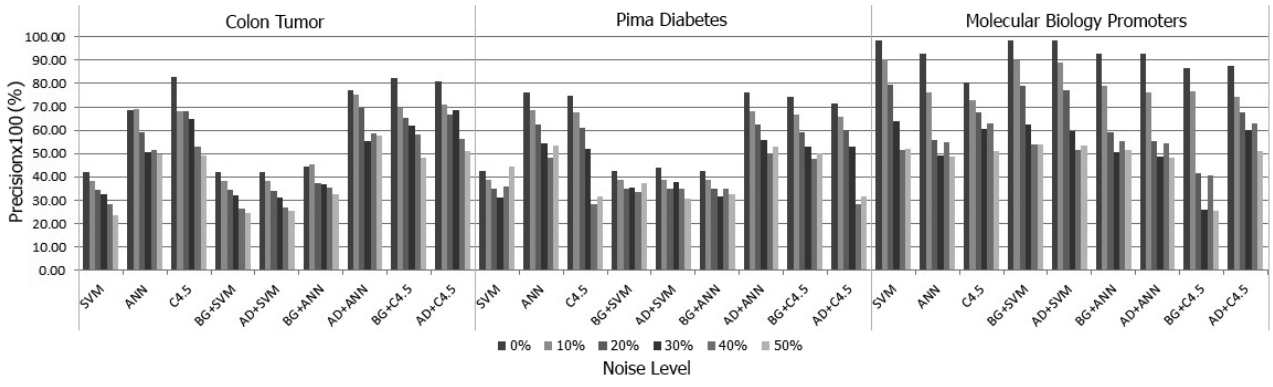

Figure 2 Precision results

50% แบบจำลอง BG+SVM มีค่าความแม่นยำสูงสุดรองลงมาคือ AD+SVM และ SVM ตามลำดับดังรูปที่ 2

การเปรียบเทียบค่าการเรียกคืน

ค่าการเรียกคืน (Recall)[23]เป็นค่าที่แสดงถึงประสิทธิภาพของแบบจำลองที่สามารถดึงข้อมูลที่เกี่ยวข้องได้ถูกต้องในการพยากรณ์กลุ่มของชุดข้อมูล Pima Diabetes, Molecular Biology Promoters และ Colon Tumor ที่มีข้อมูลรบกวนสามารถแสดงค่าการเรียกคืนของการพยากรณ์ได้ดังตารางที่ 3

จากตารางที่ 3 แสดงการเปรียบเทียบค่าการเรียกคืนที่แบบจำลองสามารถพยากรณ์ได้ จากการทดลองพบว่าข้อมูล Colon Tumor ที่ข้อมูลรบกวนระดับ 0% แบบจำลอง BG+C4.5 มีค่าการเรียกคืนสูงสุด รองลงมาคือ C4.5 และ AD+C4.5 เมื่อเพิ่มข้อมูลรบกวนที่ 50% แบบจำลอง AD+ANN มีค่าการเรียกคืนสูงที่สุด รองลงมาคือ ANN และ AD+C4.5 ในข้อมูล Pima Diabetes ที่ข้อมูลรบกวนระดับ 0% แบบจำลอง ANN มีค่าการเรียกคืนสูงสุด รองลงมาคือ AD+ANN และ BG+C4.5 เมื่อเพิ่มข้อมูลรบกวนที่ 50% แบบจำลองที่มีค่าการเรียกคืนสูงสุดคือ ANN, AD+ANN

Table 3 Recall score of the data sets

Recallx 100 (%)									
Noise	SVM	ANN	C4.5	BG+SVM	AD+SVM	BG+ANN	AD+ANN	BG+C4.5	AD+C4.5
Colon Tumor									
0%	64.76	74.45	81.95	64.76	64.76	61.21	77.93	**82.45**	80.79
10%	61.43	69.38	67.29	61.43	61.43	59.81	**74.52**	68.43	69.69
20%	58.10	60.31	66.71	58.10	57.76	54.79	**70.05**	64.64	65.52
30%	56.43	56.64	63.45	56.10	55.43	54.05	58.50	60.17	**66.52**
40%	53.10	57.07	51.90	51.19	51.48	52.33	**59.00**	57.19	55.48
50%	48.57	55.02	51.69	49.43	50.29	50.98	**58.19**	48.76	52.12
Pima Diabetes									
0%	65.11	**76.46**	74.49	65.11	65.16	65.11	76.25	74.66	71.69
10%	62.11	**68.71**	67.73	62.04	61.86	62.11	68.45	67.31	66.37
20%	59.00	**63.01**	61.34	59.02	58.91	59.11	62.88	60.22	60.08
30%	55.87	55.72	54.46	56.03	55.99	55.99	**56.54**	53.86	54.81
40%	**52.99**	50.74	52.70	52.84	52.39	52.16	51.10	48.33	52.70
50%	48.32	**53.02**	48.11	48.65	49.97	50.10	52.76	50.21	48.11
Molecular Biology Promoters									
0%	**98.40**	91.66	79.04	98.21	98.13	91.97	91.66	84.56	85.83
10%	**88.96**	73.79	71.12	88.29	87.36	76.94	73.79	74.75	72.58
20%	**77.73**	55.00	65.79	77.15	74.66	58.35	54.43	53.94	65.79
30%	**62.66**	49.17	59.17	60.07	58.35	49.97	48.87	48.79	58.99
40%	51.58	54.15	**61.35**	53.57	52.03	54.64	53.95	53.85	**61.35**
50%	51.01	48.40	50.45	51.85	**52.94**	50.65	48.29	49.66	50.45

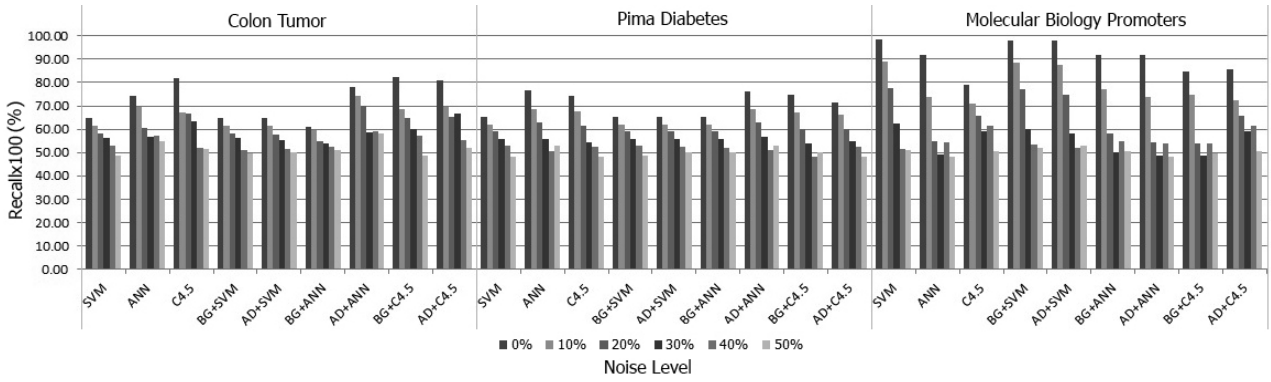

Figure 3 Recall results

และ BG+C4.5 ส่วนข้อมูล Molecular Biology Promoters ที่ข้อมูลรบกวนระดับ 0% แบบจำลองที่มีค่าการเรียกคืนสูงสุดคือ SVM, BG+SVM และ AD+SVM เมื่อเพิ่มข้อมูลรบกวนที่ 50% แบบจำลองที่มีค่าการเรียกคืนสูงสุดคือ AD+SVM, BG+SVM และ SVM ตามลำดับ Figure 3

การเปรียบเทียบค่า F-Measure

ค่า F-Measure[23]เป็นค่าที่แสดงถึงประสิทธิภาพโดยรวมของแบบจำลองที่พยากรณ์ได้ถูกต้องของชุดข้อมูล Pima Diabetes, Molecular Biology Promoters และ Colon Tumor ที่มีข้อมูลรบกวนสามารถแสดงค่า F-Measure ของการพยากรณ์ได้ดัง Table 4

จากตารางที่ 4 แสดงการเปรียบเทียบค่า F-Measure ที่แบบจำลองสามารถพยากรณ์ได้ จากการทดลองพบว่าข้อมูล Colon Tumor ที่ข้อมูลรบกวนระดับ 0% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ BG+C4.5, C4.5 และ AD+C4.5 เมื่อเพิ่มข้อมูลรบกวนที่ระดับ 50% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ AD+ANN, AD+C4.5 และ C4.5 ในข้อมูล Pima Diabetes ที่ข้อมูลรบกวนระดับ 0% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ ANN, AD+ANN และ AD+C4.5 เมื่อเพิ่มข้อมูลรบกวนที่ระดับ 50% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ AD+ANN, BG+C4.5 และ ANN

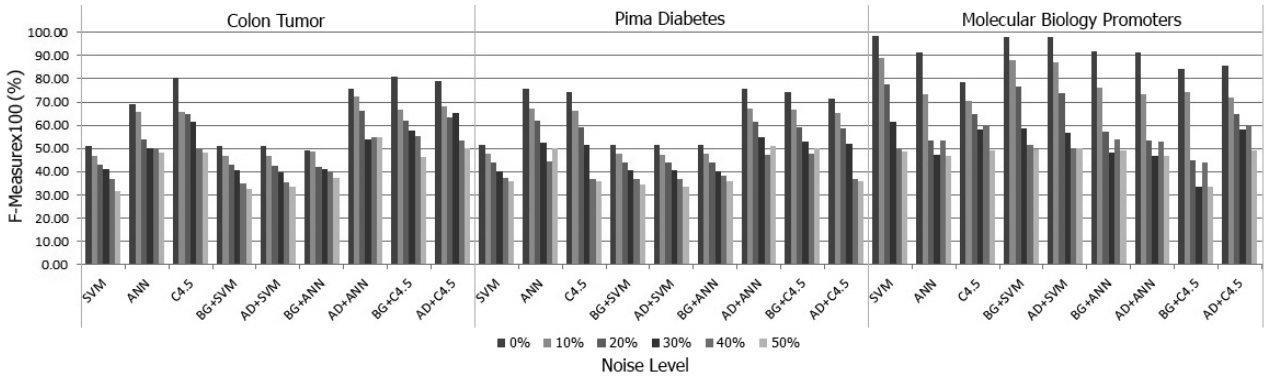

Figure 4 F-Measure results

Table 4 F-measure score of the data sets

F-Measurex 100 (%)									
Noise	SVM	ANN	C4.5	BG+SVM	AD+SVM	BG+ANN	AD+ANN	BG+C4.5	AD+C4.5
Colon Tumor									
0%	50.98	69.29	80.50	50.98	50.98	49.24	75.61	**80.76**	78.85
10%	46.98	65.66	65.55	46.98	46.98	48.75	**72.29**	66.52	68.09
20%	42.98	54.08	65.04	42.98	42.61	42.15	**66.01**	62.06	63.22
30%	40.98	49.99	61.66	40.61	39.88	41.27	53.85	57.87	**65.14**
40%	36.98	49.83	49.97	34.88	35.19	40.10	54.78	**55.46**	53.35
50%	31.81	48.01	48.22	32.76	33.71	37.54	**54.68**	46.17	49.98
Pima Diabetes									
0%	51.35	**75.90**	74.14	51.35	51.46	51.35	75.73	74.30	71.46
10%	47.59	**67.36**	66.28	47.56	47.47	47.59	67.33	66.53	65.16
20%	43.87	**61.77**	59.32	43.88	43.83	43.93	61.69	59.17	58.83
30%	40.14	52.55	51.45	40.42	40.53	40.28	**54.84**	52.95	52.14
40%	37.12	44.30	36.89	36.93	36.81	38.29	47.33	**47.78**	36.89
50%	36.08	49.89	35.87	34.31	33.65	35.86	**51.18**	49.99	35.87
Molecular Biology Promoters									
0%	**98.38**	91.53	78.68	98.19	98.10	91.89	91.53	84.24	85.56
10%	**88.82**	73.14	70.41	88.15	87.13	76.38	73.14	74.15	71.91
20%	**77.42**	53.62	64.84	76.79	74.03	57.35	53.22	44.77	64.84
30%	**61.65**	47.20	58.19	58.53	56.60	48.27	46.76	33.28	58.01
40%	49.81	53.23	**60.03**	51.57	49.88	53.81	52.98	43.75	**60.03**
50%	48.83	46.67	49.13	49.86	**50.21**	49.01	46.60	33.46	49.13

ส่วนข้อมูล Molecular Biology Promoters ที่ข้อมูลรบกวนระดับ 0% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ SVM, BG+SVM และ AD+SVM เมื่อเพิ่มข้อมูลรบกวนที่ 50% แบบจำลองที่มีประสิทธิภาพสูงสุดคือ AD+SVM, BG+SVM และ AD+C4.5 ตามลำดับ Figure 4

วิจารณ์และสรุปผลการทดลอง

ในงานวิจัยนี้มีจุดประสงค์เพื่อเปรียบเทียบความทนทานต่อข้อมูลรบกวนของแบบจำลองSVM, ANN, C4.5, BG+SVM, AD+SVM, BG+ANN, AD+ANN, BG+C4.5 และ AD+C4.5 กับชุดข้อมูลมาตรฐาน 3 ชุด จากการทดลองคณะผู้วิจัยพบว่า

แบบจำลอง SVM มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Molecular Biology Promoters และมีความทนต่อการพยากรณ์ข้อมูลMolecular Biology Promoters ที่ข้อมูลรบกวนไม่เกินระดับ30% แต่จะมีประสิทธิภาพที่ต่ำมากเมื่อนำไปพยากรณ์ข้อมูล Pima Diabetes และ Colon Tumor

แบบจำลอง ANN มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Pima Diabetes และ ข้อมูล Molecular Biology Promoters และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Pima Diabetes ที่ข้อมูลรบกวนไม่เกินระดับ 20% แต่จะมีประสิทธิภาพที่ต่ำเมื่อนำไปพยากรณ์ข้อมูล Colon Tumor

แบบจำลอง C4.5มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Colon Tumor แต่มีความทนที่น้อยต่อการพยากรณ์ข้อมูลทั้งสาม

แบบจำลอง BG+SVM มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Molecular Biology Promoters และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Molecular Biology Promoters แต่จะมีประสิทธิภาพที่ต่ำมากเมื่อนำไปพยากรณ์ข้อมูล Pima Diabetes และ Colon Tumor

แบบจำลอง AD+SVM มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Molecular Biology Promoters และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Molecular Biology Promoters แต่จะมีประสิทธิภาพที่ต่ำมากเมื่อนำไปพยากรณ์ข้อมูล Pima Diabetes และ Colon Tumor

แบบจำลอง BG+ANN มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Molecular Biology Promoters แต่จะมีประสิทธิภาพที่ต่ำเมื่อนำไปพยากรณ์ข้อมูล Colon Tumor และ Pima Diabetes

แบบจำลอง AD+ANN มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูลทั้งสาม และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Colon Tumor และ Pima Diabetes

แบบจำลอง BG+C4.5 มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Colon Tumor และ Pima Diabetes และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Colon Tumor และ Pima Diabetesแต่จะมีประสิทธิภาพที่ต่ำเมื่อนำไปพยากรณ์ข้อมูล Molecular Biology Promoters

แบบจำลอง AD+C4.5 มีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูล Colon Tumor และ Pima Diabetes และมีความทนที่ดีต่อการพยากรณ์ข้อมูล Colon Tumor แต่จะมีประสิทธิภาพที่ต่ำเมื่อนำไปพยากรณ์ข้อมูล Molecular Biology Promoters

ซึ่งจะพบว่าแบบจำลองที่เป็นเทคนิคแบบเดี่ยวจะมีประสิทธิภาพที่ดีในการพยากรณ์ข้อมูลที่มีข้อมูลรบกวน ไม่เกิน 20% แต่เมื่อเพิ่มข้อมูลรบกวนถึง 50% แบบจำลองที่เป็นเทคนิคแบบรวมกลุ่มจะสามารถทำการพยากรณ์ข้อมูลได้ดีกว่า จึงสามารถสรุปได้ว่าแบบจำลองที่เป็นเทคนิคแบบรวมกลุ่มมีความทนทานกับข้อมูลที่มีข้อมูลรบกวนได้ดีกว่าแบบจำลองที่เป็นเทคนิคแบบเดี่ยว

จากงานวิจัยนี้สามารถนำผลการวิจัยในเทคนิคที่นำเสนอกับชุดข้อมูล ไปประกอบการเลือกใช้เทคนิคที่เหมาะสมกับข้อมูลที่มีอยู่ได้ และสามารถนำเทคนิคนี้ไปพัฒนาต่อยอดเพื่อให้แบบจำลองมีประสิทธิภาพที่ดียิ่งขึ้นได้โดยการกำหนดค่าพารามิเตอร์ที่เหมาะสมกับชุดข้อมูลในอนาคต

เอกสารอ้างอิง

1. เดช ธรรมศิริ, พยุง มีสัจ. “การเรียนรู้แบบรวมกลุ่มด้วยโครงข่ายประสาทเทียมเอดาบูท สำหรับการจำแนกข้อมูล”. National Conference on Computing and Technology. 545-50, 2554.
2. Yu Wang, Cheng De Lin. “Learning by Bagging and Adaboost based on Support Vector Machine “. IEEE International Conference on Industrial Informatics. Vienna 2, 663 - 8 2007.
3. David F. Nettleton, Albert Orriols-Puig, Albert Fornells. “A Study of The Effect of Different Type of Noise on The Precision of Supervised Learning Techniques”. Artificial Intelligence Review, 33:275-306 2010.
4. José A., Sáez a., Galar M, Luengo J, Herrera F. “Tackling The Problem of Classification With Noisy Data Using Multiple Classifier Systems: Analysis of The Performance and Robustness”. Information Sciences, 247:1-20 2013.
5. Xing Zhu, Xindong Wu. “Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts”. Artificial Intelligence Review, 22:177-210 2004.
6. Xingquan Zhu, Xindong Wu, Qijun Chen. “Eliminating Class Noise in Large Datasets”. Twentieth International Conference on Machine Learning; Washington DC; 920-6, 2003.
7. Hu H, Li J-Y, Wang H, Daggard G, Wang L-Z. “Robustness Analysis of Diversifiled Ensemble Decision Tree Algorithms for Microarray Data Classification”. The Seven International Conference on Machine Learning and Cybernetics; Kunming; 115-20, 2008.
8. Dietterich T. “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization”. Machine Learning, 40(2):139-57 2000.
9. Nittaya Kerdprasop, Kittisak Kerdprasop. “Software development for noise-tolerant decision-tree induction”. 2009.
10. Hyun-Chul Kim, Shaoning Pang, Hong-Mo Je, Daijin Kim, Sung-Yang Bang. “Support Vector Machine Ensemble with Bagging”. Lecture Notes in Computer Science, 2388(2002):397-408 2002.

11. Thongkam J, Sukmak V. “Bagging Random Tree for Analyzing Breast Cancer Survival”. KKU Research Journal, 17(1):1-13 2012.
12. Xuchun Li, Lei Wang, Eric Sung. “Adaboost with SVM-based Component Classifiers”. Engineering Applications of Artificial Intelligence, 21(2008):785-95 2007.
13. Cortes C, Vapnik V. “Support-Vector Networks”. Machine Leaming, 20:273-97 1995.
14. Frank Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain.”. Psychological Review, 65(6):386-408 1958.
15. Yufei Yuan, Michael J. Shaw. “Induction of fuzzy decision trees”. Fuzzy Sets and Systems, 69(2):125-39 1995.
16. Breiman L. “Bagging Predictor”. Machine Learning, 24(2):123-40 1996.
17. Freund Y, Robert E.Schapire. “A Decision-Theoretic Generalization of Online Learning and An Application To Boosting”. Journal of Computer and System Sciences, 55:119-39 1997.
18. Guo-Zheng Li. “Colon Tumor”. http://levis.tongji.edu.cn/gzli/data/ColonTumor.zip2002.
19. National Institute of Diabetes and Digestive and Kidney Diseases. “Diabetes”. http://repository.seasr.org/Datasets/UCI/arff/diabetes.arff1990.
20. Harley C. , Reynolds R. . “Molecular-biology_promoters”. http://repository.seasr.org/Datasets/UCI/arff/molecular-biology_promoters.arff1990.
21. University of Waikato. “Weka Machine Learning”. http://www.cs.waikato.ac.nz/ml/weka/2013.
22. Ron Kohavi. “A study of cross-validation and bootstrap for accuracy estimation and model selection”. 14th international joint conference on Artificial intelligence. San Francisco; 2, 1137-43 1995.
23. Powers D.M.W. “ Evaluation : From Precision, Recall and F-measure To ROC, Informedness & Correlation “. Journal of Machine Learning Technologies, 2(1):37-63 2011.