การเปรียบเทียบประสิทธิภาพเทคนิคการลดมิติของข้อมูลสำหรับค้นหาปัจจัย และสร้างโมเดลการจำแนกกลุ่มการระบายน้ำของประตูระบายน้ำ

An Efficiency Comparison of Dimensionality Reduction Techniques for Study of Factors and Model Building for Drainage Classification of Floodgate

นิธินันท์ มาตา[1], พนิดา หล่อวงศ์ตระกูล[2] และพยุง มีสัจ[3]

[1]สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์
ถ.จิระ ต.ในเมือง อ.เมือง จ.บุรีรัมย์ 31000
[2]สาขาวิชาเทคโนโลยีสารสนเทศและการสื่อสาร คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัย เทคโนโลยีราชมงคลอีสาน
ถ.สุรนารายณ์ ต.ในเมือง อ.เมือง จ.นครราชสีมา 30000
[3]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
1518 ถ. ประชาราษฎร์ 1 แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800
[1]mata0711@gmail.com, [2] panidajlo@gmail.com, [3]phayung.m@it.kmutnb.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการลดมิติของข้อมูลสำหรับสร้างโมเดลการจำแนกกลุ่มการระบายน้ำของประตูระบายน้ำผักไห่ และหาปัจจัยที่สำคัญสำหรับเป็นข้อมูลในการจำแนกกลุ่มการระบายน้ำ โดยใช้ชุดข้อมูลปัจจัยที่เกี่ยวข้องกับการระบายน้ำของประตูระบายน้ำผักไห่ พบว่าโมเดลการจำแนกกลุ่มที่มีประสิทธิภาพเหมาะสมกับข้อมูลการระบายน้ำมากที่สุดคือ Decision Tree แบบ J48 (Accuracy = 95.048%, Precision = 0.950, Recall = 0.949) และนำโมเดล J48 มาเพิ่มประสิทธิภาพในการจำแนกกลุ่มและหาปัจจัยที่จำเป็นสำหรับการพยากรณ์การระบายน้ำนั้น พบว่าโมเดล J48 ใช้ร่วมกับเทคนิคการลดคุณลักษณะของข้อมูลแบบ CFS (Correlation-based Feature Selection) ทำให้ประสิทธิภาพในการจำแนกกลุ่มการระบายน้ำเพิ่มขึ้น จาก 95.048% เป็น 96.286% นอกจากนี้ทำให้ทราบถึงปัจจัยที่สำคัญซึ่งจำเป็นต้องใช้ในการพยากรณ์การระบายน้ำของประตูระบายน้ำฝักไห่ มี 4 ปัจจัย ได้แก่ เดือน ปริมาณน้ำท้ายประตูระบายน้ำ ปริมาณน้ำที่ปล่อยจากเขื่อนเจ้าพระยา และปริมาณน้ำที่ระบายสู่แม่น้ำน้อย

คำสำคัญ: การลดมิติของข้อมูล วิธีเลือกคุณลักษณะข้อมูล การระบายน้ำของประตูระบายน้ำ

Abstract

The aim of this research was to compare the performance of classification algorithms. Feature selection techniques are used to reduce the dimensions of large data for modeling of drainage classifiers. Pakhai floodgate were sued in experiments to find major factors for the prediction of drainage by applying the factors associated with the drainage of the floodgate. The results from the experiments showed that the best method was Decision Tree (J48) (Accuracy = 95.048%, Precision = 0.950, Recall = 0.949). To improve the accuracy of the classification of drainage and find a major factor for the prediction of drainage found J48 with CFS (Correlation-based Feature Selection) increased performance in predicting drainage from 95.048% to 96.286%. In addition, found the key factors which need to be used to predict the drainage of Pakhai floodgates 4 factors include Month, Water

above floodgate, Water drainage from Chao Praya Dam, and Water inflow to Noi River.

Keywords: Dimensionality Reduction, Feature Selection Methods, Drainage of Floodgate

1. บทนำ

ประตูระบายน้ำผักไห่เป็นหนึ่งในประตูระบายน้ำที่อยู่บนเส้นทางการระบายน้ำของแม่น้ำน้อย [1] โดยระบายน้ำจากต้นเขื่อนเจ้าพระยา (สถานีวัดระดับน้ำ C2) ไปบรรจบที่คลองโผงเผง คลองบางบาล และไหลลงไปรวมกับแม่น้ำเจ้าพระยาส่วนหนึ่ง และอีกส่วนจะถูกระบายลงสู่คลองผักไห่ – เจ้าเจ็ด การบริหารจัดการน้ำของเขื่อนเจ้าพระยาผ่านแม่น้ำน้อยนั้นหากเป็นช่วงเวลาที่ปริมาณน้ำในแม่น้ำเจ้าพระยามีปริมาณน้อยและไม่เกิดสภาพน้ำอัดเอ่อไหลย้อน ประตูระบายน้ำผักไห่จะสามารถระบายน้ำได้เต็มศักยภาพ ในทางตรงกันข้ามหากเกิดกรณีน้ำไหลย้อนจากแม่น้ำเจ้าพระยาเข้าสู่คลองโผงเผง และคลองบางบาล จะทำให้ปริมาณน้ำที่เหนือประตูและท้ายประตูระบายน้ำผักไห่สูงซึ่งส่งผลกระทบทำให้เกิดน้ำท่วมบ้านเรือนประชาชนบริเวณดังกล่าว ดังนั้นจำเป็นต้องมีการศึกษาปัจจัยต่างๆ ที่มีผลต่อการระบายน้ำ เพื่อลดการเกิดผลกระทบจากน้ำท่วม และน้ำแล้งแก่ประชาชนที่อาศัยอยู่บริเวณปลายแม่น้ำน้อยให้ได้มากที่สุด

ในงานวิจัยนี้มีวัตุประสงค์เพื่อเปรียบเทียบประสิทธิภาพเทคนิคการลดมิติของข้อมูลในการหาปัจจัยที่สำคัญสำหรับเป็นข้อมูลในการพยากรณ์การระบายน้ำของประตูระบายน้ำผักไห่ โดยแบ่งขั้นตอนออกเป็น 2 ส่วนคือการหาโมเดลการพยากรณ์ที่เหมาะสมกับข้อมูลที่ใช้ประกอบการตัดสินใจในการระบายน้ำแต่ละกรณี และการหาปัจจัยที่สำคัญต่อการระบายน้ำที่ทำให้โมเดลมีความแม่นยำมากยิ่งขึ้น โดยใช้การเปรียบเทียบเทคนิคการลดมิติของข้อมูล เพื่อจำแนกประเภทกรณีการระบายน้ำที่ประตูระบายน้ำผักไห่ออกเป็น 6 กรณี ได้แก่ 1) ลดการระบายเพื่อการเกษตร 2) ปิดการระบายน้ำเพื่อเก็บน้ำไว้ใช้ทางการเกษตร 3) ระบายน้ำตามความเหมาะสม 4) ระบายน้ำเพื่อรองรับน้ำในหน้าฝน 5) ปิดการระบายน้ำเนื่องจากเกิดภาวะน้ำไหลย้อนกลับ และ 6) เปิดประตูระบายน้ำเพื่อให้น้ำไหลย้อนขึ้นด้านเหนือประตูระบายน้ำลงทุ่งผักไห่

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การจำแนกประเภทข้อมูล (Data Classification)

เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นความแตกต่างระหว่างกลุ่มของข้อมูลได้ และเพื่อทำนายว่าข้อมูลนี้ควรจัดอยู่ในกลุ่มใด ซึ่งโมเดลที่ใช้จำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้จะขึ้นอยู่กับการวิเคราะห์กลุ่มของข้อมูลทดลอง (Training data) โดยนำ Training data มาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในกลุ่มเดียวกันบ้าง ผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลจัดประเภทข้อมูล (Classifier model) [2] มีเทคนิคที่นิยมใช้งานได้แก่ ต้นไม้ตัดสินใจ (Decision Tree) [3], โครงข่ายประสาทเทียม (Neural Network) [4], การตัดสินใจโดยใช้กฎ (Rule-Base) และ นาอีพเบย์ (Baysian Naivebayes) [5]

2.2 การคัดเลือกคุณลักษณะของข้อมูล (Feature Selection)

เป็นการเลือกคุณลักษณะของข้อมูลมีความสำคัญน้อยออก เพื่อดูประสิทธิภาพการทำนายหลังจากที่ได้คุณลักษณะของข้อมูลบางตัวออก ซึ่งส่วนใหญ่จะให้ค่าความถูกต้องสูงขึ้น [6]

2.3 Info Gain Attribute Evaluation

ใช้การประเมินค่าของลักษณะเฉพาะโดยวัด Information Gain ซึ่งเป็นตัววัดความสัมพันธ์ของลักษณะเฉพาะให้กับกลุ่มนั้น ๆ [7] การหาค่า Information Gain (IG) สามารถคำนวณได้

2.4 Correlation based Feature Selection (CFS)

ใช้หลักการเพื่อหาคุณลักษณะที่สามารถรวมกันแล้วให้ค่าที่สามารถให้ผลการทำนายเป็นกลุ่ม กลุ่มเดียวที่แข็งแกร่งที่สุดโดยต้องการกลุ่มของคุณลักษณะของข้อมูลที่มีขนาดเล็กที่สุด [8]

2.5 Wrapper Subset Evaluation

เป็นอัลกอริทึมในการเรียนรู้เพื่อหาเป้าหมายเพื่อคาดคะเน

มูลค่าของกลุ่มคุณลักษณะข้อมูล [9] โดยใช้ Cross-validation สำหรับการคำนวณหาค่าความถูกต้องโดยหลักการของ Wrapper ประกอบด้วย Search Algorithms เพื่อหากลุ่มของคุณลักษณะข้อมูลที่เหมาะสม แล้วนำกลุ่มของ คุณลักษณะข้อมูลที่ได้มาคำนวณหาค่าความถูกต้อง

2.6 งานวิจัยที่เกี่ยวข้อง

การเปรียบเทียบหาเทคนิคการคัดเลือกคุณลักษณะที่ให้ประสิทธิภาพการจำแนกข้อมูลที่ดีที่สุดเพื่อนำมาทำการจำแนกและหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา โดยเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะ 3 วิธี จากการทดสอบพบว่าเทคนิคการคัดเลือกคุณลักษณะโดยใช้ Wrapper Subset Evaluation ร่วมกับ Bayesian belief network ให้ค่าความถูกต้องสูงที่สุด [10]

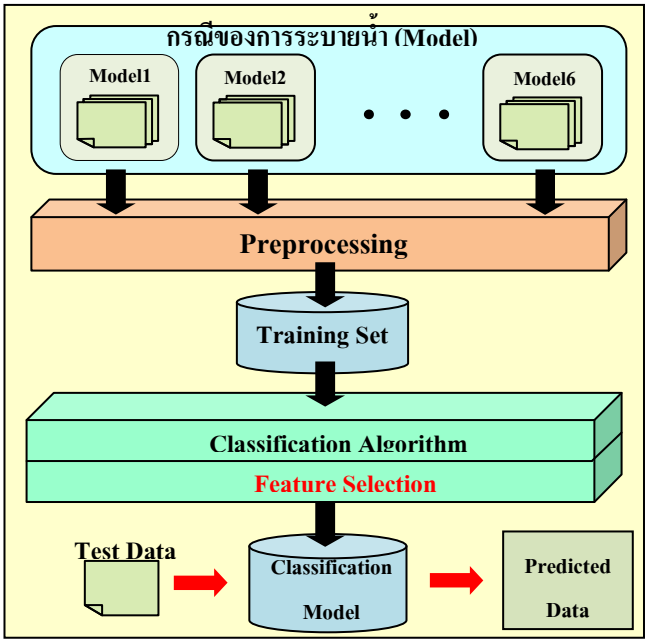
3. วิธีดำเนินงานวิจัย

3.1 ข้อมูลที่ใช้ในการวิจัย

เพื่อการสร้างโมเดลการพยากรณ์ และวัดประสิทธิภาพการลดมิติของข้อมูล ได้ใช้ข้อมูลที่เป็นปัจจัยที่เกี่ยวข้องกับการระบายน้ำของประตูระบายน้ำผักไห่ และปริมาณการระบายน้ำในแต่ละวัน ปี 2554 – 2557 ซึ่งมีข้อมูลจำนวน 731 ระเบียน โดยมีรายละเอียดดังตารางที่ 1

3.2 โมเดลและวิธีเลือกคุณลักษณะข้อมูลที่ใช้ในการเปรียบเพื่อหาประสิทธิภาพ

ในงานวิจัยนี้ได้ใช้การวัดค่าประสิทธิภาพซึ่งพิจารณาจากค่าความถูกต้อง (Accuracy) ความแม่นยำ (Precision) และค่าความระลึก (Recall) จากเทคนิคในการจำแนกประเภทข้อมูล 4 เทคนิค ได้แก่ Bayesian Naivebayes, Neural Network (Multilayer Perceptron), Rules Based (JRip) และ Decision Tree (J48) ทุกเทคนิคใช้วิธีการ 10-fold cross validation เพื่อหาประสิทธิภาพการพยากรณ์เหมาะสม หลังจากนั้นนำโมเดลที่ได้ทำการลดมิติของข้อมูล คือ Info Gain Attribute Evaluation, CFs และ Wrapper Subset Evaluation ให้เหลือปัจจัยที่สำคัญต่อการพยากรณ์เท่านั้น เพื่อให้ได้โมเดลการพยากรณ์ที่มีประสิทธิภาพมากยิ่งขึ้น



ภาพที่ 1 : ขั้นตอนการดำเนินงาน

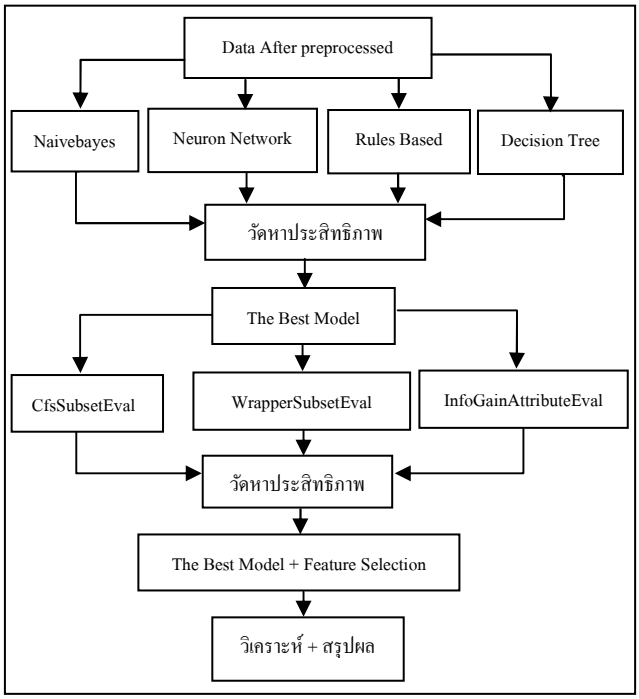
ตารางที่ 1 : รายละเอียดของข้อมูล (input)

No	Name	Detail
1	Month	เดือน
2	Season	ฤดู
3	OverPakhai	ปริมาณน้ำเหนือประตูระบายน้ำ
4	UnderPakhai	ปริมาณน้ำท้ายประตูระบายน้ำ
5	RainfallC13	ปริมาณน้ำฝนที่เขื่อนเจ้าพระยา
6	RainfallC2	ปริมาณน้ำฝนที่สถานีวัดน้ำ C2
7	InflowNoiRiver	ปริมาณน้ำที่ระบายสู่แม่น้ำน้อย
8	C2	ปริมาณน้ำจากสถานีวัดน้ำ C2
9	ChaoPraya	ปริมาณน้ำจากเขื่อนเจ้าพระยา

3.3 ขั้นตอนการดำเนินการทดลอง

งานวิจัยนี้แบ่งขั้นตอนการทำงานออกเป็น 4 ขั้นตอน ดังภาพที่ 1 ประกอบด้วย ขั้นตอนการเตรียมข้อมูล (Data Collection) ขั้นตอนกรองข้อมูล (Data Preprocessing) ขั้นตอนการหาโมเดลการพยากรณ์ (Classification Algorithm) และขั้นตอนการหาปัจจัยที่สำคัญที่ทำให้โมเดลพยากรณ์ได้มี

ประสิทธิภาพมากขึ้น โดยที่ขั้นตอนของการหาโมเดลพยากรณ์คำตอบ และการหาปัจจัยที่สำคัญ แสดงดังภาพที่ 2



ภาพที่ 2 : ขั้นตอนของการหาโมเดลพยากรณ์คำตอบ และการหาปัจจัยที่สำคัญในการพยากรณ์

3.4 การเปรียบเทียบโมเดลและวิธีเลือกคุณลักษณะข้อมูล

ขั้นตอนการเปรียบเทียบโมเดลและวิธีเลือกคุณลักษณะข้อมูลพิจารณาค่าประสิทธิภาพจากค่าความถูกต้อง (Accuracy) ความแม่นยำ (Precision) และค่าความระลึก (Recall)

4. ผลการทดลอง

4.1 ผลการวิเคราะห์การเปรียบเทียบโมเดลการพยากรณ์การระบายน้ำที่ประตูระบายน้ำผักไห่

ตารางที่ 2 : ผลการเปรียบเทียบค่าความถูกต้องของกลุ่มข้อมูลเมื่อถูกจำแนกประเภทด้วยอัลกอริทึมต่างกัน

Model	10- fold cross validation		
	Accuracy	Recall	Precision
Naivebayes	78.9546	0.856	0.790
Multilayer Perceptron	92.7098	0.911	0.927
JRip	92.022	0.921	0.920
J48	95.0481	0.949	0.950

จากตารางที่ 2 พบว่าการเปรียบประสิทธิภาพในการจำแนกข้อมูลของโมเดล Naivebayes, Multilayer Perceptron, JRip และ J48 นั้นโมเดลที่มีประสิทธิภาพความแม่นยำในการจำแนกประเภทข้อมูลมากที่สุดคือ Decision Tree (J48) หลังจากนั้นนำโมเดล J48 เพิ่มประสิทธิภาพในการจำแนกมากยิ่งขึ้น โดยการทำการคัดเลือกคุณลักษณะของข้อมูลเพื่อให้ได้ปัจจัยที่สำคัญในการสร้างแบบจำลองการพยากรณ์การระบายน้ำ

4.2 ผลการวิเคราะห์การเปรียบเทียบประสิทธิภาพการลดมิติของข้อมูลโดยใช้การเลือกคุณลักษณะของข้อมูล

ตารางที่3 : ผลการเปรียบเทียบประสิทธิภาพของเทคนิคการลดคุณลักษณะของข้อมูลเมื่อประยุกต์ใช้โมเดล J48

Model	Accuracy	Recall	Precision
J48	95.048	0.949	0.950
J48+CFS	96.286	0.963	0.963
J48+Wrapper	95.185	0.951	0.952
J48+InfoGain	95.185	0.951	0.952

จากตารางที่ 3 พบว่าหลังจากทดลองนำโมเดล J48 มาเพิ่มประสิทธิภาพในการพยากรณ์การระบายน้ำและหาปัจจัยที่จำเป็นสำหรับการพยากรณ์การระบายน้ำนั้น โมเดล J48 ใช้ร่วมกับเทคนิคการลดคุณลักษณะของข้อมูลแบบ CFS ทำให้ประสิทธิภาพในการพยากรณ์การระบายน้ำเพิ่มขึ้นและพบว่าปัจจัยที่สำคัญซึ่งจำเป็นต้องใช้ในการพยากรณ์การระบายน้ำของประตูระบายน้ำฝักไห่ มี 4 ปัจจัย ได้แก่ เดือน ปริมาณน้ำท้ายประตูระบายน้ำ ปริมาณน้ำที่ปล่อยจากเขื่อนเจ้าพระยา และปริมาณน้ำที่ระบายสู่แม่น้ำน้อย

5. บทสรุป

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการลดมิติของข้อมูลสำหรับสร้างโมเดลการจำแนกการระบายน้ำของประตูระบายน้ำผักไห่ และหาปัจจัยที่สำคัญที่ใช้สำหรับเป็นข้อมูลในการพยากรณ์การระบายน้ำ

โดยใช้ชุดข้อมูลปัจจัยที่เกี่ยวข้องกับการระบายน้ำของประตูระบายน้ำผักไห่ และปริมาณการระบายน้ำในแต่ละวัน พบว่าโมเดลการจำแนกข้อมูลที่มีประสิทธิภาพเหมาะสมกับข้อมูลการระบายน้ำมากที่สุดคือ Decision Tree แบบ J48 (Accuracy = 95.048, Precision = 0.950, Recall = 0.949) และนำโมเดล J48 มาเพิ่มประสิทธิภาพในการการจำแนกและหาปัจจัยที่จำเป็นสำหรับการพยากรณ์การระบายน้ำนั้น พบว่าโมเดล J48 ใช้ร่วมกับเทคนิคการลดคุณลักษณะของข้อมูลแบบ CFs ทำให้ประสิทธิภาพในการจำแนกการระบายน้ำเพิ่มขึ้น จาก 95.048% เป็น 96.286% นอกจากนี้ทำให้ทราบถึงปัจจัยที่สำคัญซึ่งจำเป็นต้องใช้ในการจำแนกการระบายน้ำของประตูระบายน้ำฝักไห่ มี 4 ปัจจัย ได้แก่ เดือน ปริมาณน้ำท้ายประตูระบายน้ำ ปริมาณน้ำที่ปล่อยจากเขื่อนเจ้าพระยา และปริมาณน้ำที่ระบายสู่แม่น้ำน้อย สำหรับพยากรณ์การระบายน้ำที่ประตูระบายน้ำผักไห่ออกเป็น 6 กรณี ได้แก่ ลดการระบายเพื่อการเกษตร ปิดการระบายน้ำเพื่อเก็บน้ำไว้ใช้ทางการเกษตร ระบายน้ำตามความเหมาะสม ระบายน้ำเพื่อรองรับน้ำในหน้าฝน ปิดการการระบายน้ำเนื่องจากเกิดภาวะน้ำไหลย้อนกลับ เปิดประตูระบายน้ำเพื่อให้น้ำไหลย้อนขึ้นด้านเหนือประตูระบายน้ำลงทุ่งผักไห่

ทั้งนี้ผลการวิจัยในครั้งนี้สามารถนำไปสร้างเป็นระบบพยากรณ์การระบายน้ำเพื่อการเปิดปิดระบบประตูระบายน้ำแบบอัตโนมัติได้ และในการพยากรณ์ควรหาปัจจัยอื่นที่เกี่ยวข้องกับการระบายน้ำที่คาดว่าจะส่งผลกระทบต่อกรณีของการระบายน้ำเพื่อให้สามารถสร้างโมเดลในการพยากรณ์ได้มีประสิทธิภาพเพิ่มมากขึ้น

เอกสารอ้างอิง

[1] การบริหารจัดการน้ำในพื้นที่ลุ่มต่ำ (ทุ่งผักไห่) เข้าถึงได้จาก www://www.kmcenter.rid.go.th/kmc12 สืบค้นเมื่อ 20 ต.ค. 2557

[2] ภรัณยา อำมฤครัตน์ และ พยุง มีสัจ ,“การเปรียบเทียบประสิทธิภาพการลดมิติข้อมูลและจำแนกข้อมูลโดยวิธีการทางเครือข่ายประสาทเทียม”, The 11th Graduate Research Conference Khon Kean University 2010 ,2553.

[3] มลธิดา ฤทธิ์สมบูรณ์ และสุชา สมานชาติ, “การพัฒนาระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อสินค้าโดยใช้เทคนิคต้นไม้ตัดสินใจ”, เทคโนโลยีสารสนเทศ ปีที่ 4 ฉบับที่ 7 มกราคม – มิถุนายน 2551.

[4] พยุง มีสัจ, “ระบบฟัซซี่และโครงข่ายประสาทเทียม”. พิมพ์ครั้งที่ 1. กรุงเทพฯ : ศูนย์ผลิตตำราเรียน มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2555.

[5] Arpad Kelemen and Others, “Naive Bayesian Classifier for Microarrayb Data,” IEEE Transactions on Knowledge and Data Engineering, 2003.

[6] ธรรมศักดิ์ เธียรนิเวศน์, “การลดขนาดของข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่”. วิทยานิพนธ์ สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี, 2548.

[7] T. Genchev, P. Zervas, N. Fakotakis, and G.Kokkinakis. “Benchmarking feature selection techniques on the speaker verification task”, 5th International. Symposium on Communication systems, networks and digital processing, 314 – 318, 2006.

[8] M.A. Hall and G. Holmes, “Benchmarking Attribute Selection Techniques for Discrete Class Data Mining”, IEEE, 2003.

[9] H.B. Borges, and J.C. Nievola, “Attribute Selection methods comparison for classification of diffuse large B-Cell lymphoma”, Proceedings of the Fourth International Conference on Machine Learning and Applications, 201 – 206, 2005.

[10] วสุคนธ์ทิพย์ วงพันธ์ และอนงค์นาฏ ศรีวิหก, “เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะเพื่อหาปัจจัยที่มีผลต่อพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา”, มหาวิทยาลัยขอนแก่น, 2551.