การคัดเลือกคุณลักษณะเพื่อสร้างโมเดลสำหรับการพยากรณ์ผลสัมฤทธิ์ทางการเรียนด้วยเทคนิคเหมืองข้อมูล

The Feature Selection to Creating Models for Predicting Learning Achievement using Data Mining Techniques

ทิพย์หทัย ทองธรรมชาติ[1]
Tiphathai Thongthammachart[1]

[1]อาจารย์ประจำโปรแกรมวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อทำการคัดเลือกคุณลักษณะโดยใช้เทคนิค Information Gain และเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลของเทคนิคเหมืองข้อมูล 2 เทคนิค คือ เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับและเทคนิคต้นไม้ตัดสินใจแบบ C4.5 เทคนิคที่ให้ค่าประสิทธิภาพสูงที่สุดจะถูกนำมาใช้ในการสร้างโมเดลสำหรับพยากรณ์ผลการเรียนของนักศึกษา ข้อมูลที่ใช้ในการวิจัยเป็นข้อมูลนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร และสาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด ระหว่างปีการศึกษา 2554–2559 จำนวน 358 ชุดข้อมูล จำแนกเป็นข้อมูลชุดฝึกสอน จำนวน 231 คน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2554–2557 ข้อมูลชุดทดสอบ จำนวน 70 คน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2558 และข้อมูลชุดตรวจสอบ จำนวน 57 คน ซึ่งเป็นนักศึกษาที่สำเร็จการศึกษาประจำปีการศึกษา 2559 หลังจากใช้เทคนิคการคัดเลือกคุณลักษณะแล้ว ลดเหลือเพียง 16 ตัวแปร หลังจากนั้นได้ทำการรวมกลุ่มวิชาเข้าด้วยกัน ปรากฏว่าเหลือตัวแปรต้น 5 ตัวแปร และตัวแปรตาม 1 ตัวแปร ผลการเปรียบเทียบพบว่า เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ ให้ค่าความถูกต้อง ร้อยละ 85.71 ซึ่งมากกว่าเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ซึ่งมีค่าร้อยละ 76.62

คำสำคัญ: เทคนิคเหมืองข้อมูล / การพยากรณ์ / การคัดเลือกคุณลักษณะ / ผลสัมฤทธิ์ทางการเรียน

Abstract

The objective of this research is to use Information Gain for feature selection and compared performance of two types of Data Mining Techniques which were Back Propagation Neural Network and Decision Tree (C4.5). The most significant techniques will be applied in creating a predicting learning achievement model for students. The data obtained from Kampheang Phet Rajabhat Universty and Kampheang Phet Rajabhat Universty Maesot in Faculty of Management, Business Computer Program, during the academic year 2011 - 2014 is 358 records. There are training data (231 records) who graduated in 2011-2016, testing data (70 records) who graduated in 2015 and validation data (57 records) that graduated in 2016. After using Information Gain, the researcher decreased the amount of research variables to be 16, and, combined the research variables with subject groups. This revealed that only five independent variables and one dependent variables appeared. The results of the study found that Back Propagation Neural Network (BPNN) provided accuracy rates (85.71 %) more than Decision Tree (C4.5) (76.62 %).

Keyword: Data Mining Techniques / Predicting / Feature Selection / Learning Achievement

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการให้คำแนะนำของอาจารย์ที่ปรึกษาจะพิจารณาจากผลการเรียนที่ผ่านมา โดยดูจากรายวิชาในแต่ละสาขาวิชาของนักศึกษาเป็นหลัก ดังนั้น หากมีเครื่องมือที่ช่วยในการวิเคราะห์ความถนัด และจุดอ่อนทางการศึกษาของนักศึกษาเป็นรายบุคคล ส่งผลให้กระบวนการตัดสินใจ ให้คำแนะนำของอาจารย์ที่ปรึกษา รวมทั้งกระบวนการพัฒนาคุณภาพการศึกษามีประสิทธิภาพมากยิ่งขึ้น การทำเหมืองข้อมูล (Data Mining) เป็นกระบวนการที่สกัดข้อมูลที่มีประโยชน์ เพื่อให้ได้สารสนเทศที่มีเหตุผลและสามารถนำไปใช้ช่วยในการตัดสินใจ สามารถอธิบายรูปแบบที่น่าสนใจจากข้อมูล รวมถึงใช้เพื่อการทำนาย หรือคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคต เป็นการนำข้อมูลที่เก็บรวบรวมในอดีตมาสร้างเป็นตัวแบบ (Model) แล้วนำตัวแบบนั้นไปใช้ในการทำนายหรือพยากรณ์การเกิดเหตุการณ์ที่จะเกิดขึ้นในอนาคต (Han and Kamber, 2006; Tan, Steinbach and Kumar, 2006) ซึ่งเทคนิคที่ใช้ในการทำเหมืองข้อมูลแบ่งออกเป็น 3 ส่วน ได้แก่ การหากฎความสัมพันธ์ (association rule) การจำแนกประเภทข้อมูล (data classification) และการจัดกลุ่มข้อมูล (data clustering)

ปัจจุบันการทำเหมืองข้อมูล ได้เข้ามามีบทบาทในการวิเคราะห์ข้อมูลด้านการศึกษา เพื่อการวางแผนดูแลนักศึกษา โดยนำข้อมูลประวัตินักศึกษา ข้อมูลประวัติการเรียน มาใช้ในการวิเคราะห์หาสาเหตุหรือปัจจัยที่ส่งผลต่อผลสัมฤทธิ์ทางการเรียน ระดับผลการเรียนของนักศึกษา ดังเช่น การใช้เทคนิคการทาเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียน โรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกาแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา ใช้กระบวนการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งใช้วิธี Correlation-based Feature Selection (CFS) และวิธี Information Gain (IG) แล้วใช้เทคนิคเหมืองข้อมูลแบบโครงข่ายประสาทเทียมแบบมัลติเลเยอร์เพอร์เซ็ปตรอน (MLP) ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และต้นไม้ตัดสินใจ (Decision Tree) มาสร้างตัวแบบพยากรณ์และเปรียบเทียบตัวแบบ ด้วยการทดสอบประสิทธิภาพแบบ 10-Fold Cross Validation (เสกสรรค์ วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์ และดวงดาว วิชาดากุล, 2558) การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล ซึ่งมีการเปรียบเทียบประสิทธิภาพระหว่างตัวแบบพยากรณ์ที่ถูกพัฒนาขึ้นด้วยเทคนิคต้นไม้ช่วยตัดสินใจ (Decision Tree) กับเทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network: ANN) (ธีรพงษ์ สังข์ศรี, 2557) การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์ มีการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิต โดยใช้เทคนิคการคัดเลือกคุณสมบัติที่สำคัญ แล้วสร้างตัวแบบการพยากรณ์ด้วยเทคนิค BPNN และเทคนิค SVMs (นิภาพร ชนะมาร และพรรณี สิทธิเดช, 2557) ตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล เป็นการเปรียบเทียบตัวแบบการจำแนก 4 เทคนิค คือ Decision Tree, Naïve Bayes, k-NN และ Rule Induction (ธาดา จันตะคุณ, 2559) ดังนั้นจึงสรุปไม่ได้ว่าอัลกอริทึมใดที่ให้ค่าประสิทธิภาพสูงที่สุด ทั้งนี้ผลจากการวิจัยจะเป็นแนวทางให้มหาวิทยาลัยและผู้ที่เกี่ยวข้องสามารถหาทางช่วยเหลือและให้คำแนะนำในเรื่องการเรียนแก่นักศึกษากลุ่มดังกล่าวได้รวดเร็วขึ้น

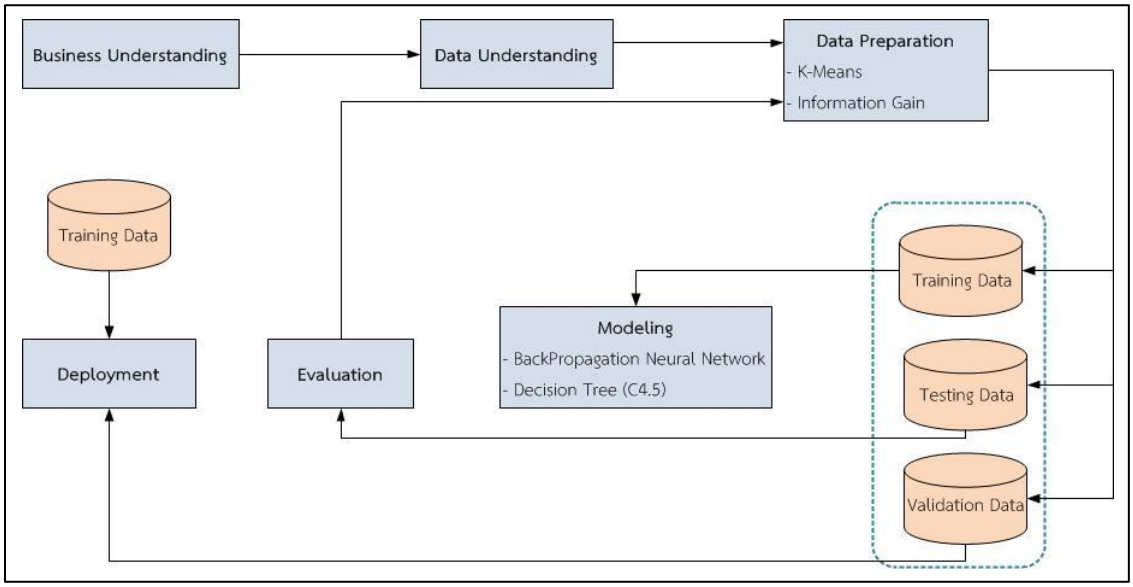
จากเหตุผลดังกล่าว จึงทำให้งานวิจัยนี้มุ่งที่จะทำการคัดเลือกโดยใช้เทคนิค Information Gain และเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลของเทคนิคเหมืองข้อมูล 2 เทคนิค คือ เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ผลการศึกษาจะสามารถนำไปใช้สำหรับการวางแผนดูแลให้คำแนะนำนักศึกษาที่คาดว่าจะอยู่ในกลุ่มที่มีระดับผลการเรียนต่ำ ซึ่งหากนักศึกษามีผลการเรียนอยู่ในระดับดีจะสามารถลดอัตราการพ้นสภาพของนักศึกษาลงได้ และเมื่อหาแบบจำลองที่มีความน่าเชื่อถือได้ก็สามารถที่จะนำข้อมูลนักศึกษาที่เรียนในระดับชั้นปีที่ 1 มาทำการทดสอบโดยผ่านแบบจำลองนี้ จะสามารถพยากรณ์ได้ว่านักศึกษามีโอกาสสำเร็จการศึกษาในสาขาวิชาที่ตนเองได้เลือกเรียนหรือไม่ และถ้านักศึกษาคนดังกล่าวอยู่ในกลุ่มที่มีผลการเรียนต่ำ ทางสาขาวิชาจะได้ให้คำแนะนำหรือเข้าไปแก้ปัญหาให้ได้อย่างถูกต้องและเหมาะสม

วัตถุประสงค์ของการวิจัย

1. เพื่อทำการคัดเลือกคุณลักษณะโดยใช้เทคนิค Information Gain ที่ส่งผลต่อระดับผลการเรียนของนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร และสาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด
2. เพื่อเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลของเทคนิคเหมืองข้อมูล 2 เทคนิค คือ เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5

วิธีดำเนินการวิจัย

การวิจัยนี้ได้ดำเนินงานโดยประยุกต์ตามแนวทางในการทำเหมืองข้อมูล ที่เรียกว่า กระบวนการมาตรฐานอุตสาหกรรม หรือ CRISP-DM (Chapman *et al.*, 2000) ที่ได้รับความนิยมมากในปัจจุบัน ซึ่งมีกระบวนการดำเนินการวิจัย ประกอบด้วย 6 ขั้นตอนหลัก ดังภาพที่ 1 รายละเอียดการทำงานแต่ละขั้นตอน มีดังนี้



ภาพที่ 1 กระบวนการดำเนินการวิจัย

1. ศึกษาโจทย์ที่ต้องการทำ (Business Understanding)

ผู้วิจัยได้ทำการศึกษาโครงสร้างหลักสูตรบริหารธุรกิจบัณฑิต สาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร จากโครงสร้างของหลักสูตรได้มีการกำหนดเป็นหมวดวิชาศึกษาทั่วไป (กลุ่มวิชาภาษาและการสื่อสาร กลุ่มวิชามนุษย์ศาสตร์ กลุ่มวิชาสังคมศาสตร์ กลุ่มวิชาคณิตศาสตร์ วิทยาศาสตร์ และเทคโนโลยี) หมวดวิชาเฉพาะ (จำแนกเป็นกลุ่มวิชาแกน กลุ่มวิชาบังคับ และกลุ่มวิชาเลือก) และหมวดวิชาเลือกเสรี

2. ศึกษาข้อมูลที่ใช้ในงานวิจัย (Data Understanding)

ผู้วิจัยได้เลือกใช้ข้อมูลภูมิหลังและข้อมูลผลการเรียนรายวิชาในชั้นปีที่ 1 และ 2 ของนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ ที่สำเร็จการศึกษาประจำปีการศึกษา 2554 – 2559 ได้ตัวแปรทั้งหมดจำนวน 31 ตัวแปร สามารถจำแนกเป็นตัวแปรต้น 30 ตัวแปร และตัวแปรตาม 1 ตัวแปร

3. การเตรียมข้อมูล (Data Preparation)

3.1 ทำความสะอาดข้อมูล หลังจากสำรวจข้อมูลแล้ว พบว่าข้อมูลยังไม่สมบูรณ์ เช่น ค่าว่าง (Missing Value) และมีสิ่งรบกวน (Noisy Data) แก้ไขโดยการแทนค่าข้อมูลที่ถูกต้องไปแทนที่ข้อมูลเดิม

3.2 แปลงข้อมูล เนื่องจากข้อมูลมีทั้งที่เป็นตัวเลข และข้อมูลที่เป็นตัวอักษร ไม่อยู่ในรูปแบบที่สามารถวิเคราะห์ได้ จึงต้องทำการแทนค่าข้อมูลให้อยู่ในรูปแบบที่สามารถวิเคราะห์ได้

3.3 ผู้วิจัยได้ทำการคัดเลือกข้อมูล โดยทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรต้น กับตัวแปรตาม

3.4 ศึกษาข้อมูลผลสัมฤทธิ์ทางการเรียนของนักศึกษา พบว่าอยู่ระหว่าง 2.00-3.64 ดังนั้นผู้วิจัยจึงใช้เทคนิคเคมีน (k=3) แบ่งกลุ่มผลสัมฤทธิ์ทางการเรียนออกเป็น 3 กลุ่ม คือ กลุ่มเรียนเก่ง (H) กลุ่มเรียนปานกลาง (M) และกลุ่มเรียนอ่อน (L) ดังตารางที่ 1

ตารางที่ 1 ผลสัมฤทธิ์ทางการเรียนในแต่ละกลุ่ม

กลุ่ม	ผลสัมฤทธิ์ทางการเรียน
High	3.00 ขึ้นไป
Medium	2.60-2.99
Low	2.00-2.59

3.5 ใช้เทคนิคการคัดเลือกคุณลักษณะ Information Gain ร่วมกับวิธีการค้นหาแบบจัดลำดับช่วยลดจำนวนตัวแปร ผลของการคัดเลือกตัวแปรต้นจากทั้งหมด 30 ตัวแปร ลดเหลือเพียง 16 ตัวแปรต้น ดังตารางที่ 2 ข้อมูลที่ใช้ในงานวิจัยมีจำนวนทั้งสิ้น 358 คน จำแนกเป็นข้อมูลชุดฝึกสอน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2554–2557 จำนวน 231 คน ข้อมูลชุดทดสอบ จำนวน 70 คน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2558 และข้อมูลชุดตรวจสอบ จำนวน 57 คน ซึ่งเป็นนักศึกษาที่สำเร็จการศึกษาประจำปีการศึกษา 2559

ตารางที่ 2 รายละเอียดของตัวแปรหลังจากใช้เทคนิคการคัดเลือกคุณลักษณะ Information Gain

ลำดับ	ชื่อตัวแปร	รายละเอียด	ค่าข้อมูล
1	Language	กลุ่มวิชาภาษาและการสื่อสาร	Interval
2	Human	กลุ่มวิชามนุษย์ศาสตร์	Interval
3	Social	กลุ่มวิชาสังคมศาสตร์	Interval
4	Math	กลุ่มวิชาคณิตศาสตร์ วิทยาศาสตร์ และเทคโนโลยี	Interval
5	Net	การจัดระบบเครือข่ายและการสื่อสารข้อมูลธุรกิจด้วยคอมพิวเตอร์	Interval
6	Acc1	การบัญชี 1	Interval
7	Bus	ความรู้เบื้องต้นเกี่ยวกับการประกอบธุรกิจ	Interval
8	IT	เทคโนโลยีสารสนเทศเบื้องต้น	Interval
9	Pro	การเขียนโปรแกรมคอมพิวเตอร์และอัลกอริทึม	Interval
10	Eco1	เศรษฐศาสตร์จุลภาค	Interval
11	Str	โครงสร้างข้อมูล	Interval
12	Eco2	เศรษฐศาสตร์มหภาค	Interval
13	OS	การติดตั้งและการจัดการระบบปฏิบัติการคอมพิวเตอร์	Interval
14	Acc2	การบัญชี 2	Interval
15	Mark	หลักการตลาด	Interval
16	Org	องค์การและการจัดการ	Interval
17	GPA	ผลสัมฤทธิ์ทางการเรียน	Ordinal

4. การสร้างตัวแบบ (Modeling)

ในขั้นตอนนี้ เป็นการนำข้อมูลหลังการคัดเลือกตัวแปรที่สำคัญด้วยเทคนิค Information Gain ที่เหลือจำนวนตัวแปรต้น 16 ตัวแปร เพื่อทดลองการจำแนกประเภทข้อมูลด้วยและเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ (BackPropagation Neuron Network: BPNN) และเทคนิคต้นไม้ตัดสินใจ (Decision Tree) แบบ C4.5 หลังจากนั้นทำการเปรียบเทียบการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษา พบว่าข้อมูลมีค่าความถูกต้องต่ำ ดังนั้นผู้วิจัยจึงได้ทำการทดลองเพิ่มเติม โดยการรวมกลุ่มวิชาเข้าด้วยกัน ดวยวิธีการคำนวณคาเฉลี่ย ปรากฏว่าเหลือตัวแปรต้น 5 ตัวแปร และตัวแปรตาม 1 ตัวแปร ดังตารางที่ 3

ตารางที่ 3 รายละเอียดของตัวแปร

ลำดับ	ชื่อตัวแปร	รายละเอียด	ค่าข้อมูล
1	Language	กลุ่มวิชาภาษาและการสื่อสาร	Interval
2	Human	กลุ่มวิชามนุษย์ศาสตร์	Interval
3	Social	กลุ่มวิชาสังคมศาสตร์	Interval
4	Math	กลุ่มวิชาคณิตศาสตร์ วิทยาศาสตร์ และเทคโนโลยี	Interval
5	Force	กลุ่มวิชาบังคับ	Interval
6	GPA	ผลสัมฤทธิ์ทางการเรียน	Ordinal

จากผลการทดลองพบว่า ข้อมูลภูมิหลังที่ถูกตัดออกไปทำให้ค่าความถูกต้องสูงขึ้น ดังนั้นหากต้องการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษา ก็สามารถใช้ข้อมูลที่จำเป็น นั่นคือตัวแปรต้น 5 ตัวแปร และแสดงค่าความถูกต้องดังตารางที่ 4

ตารางที่ 4 ค่าความถูกต้องของข้อมูลชุดฝึกสอน

ข้อมูล	BPNN	C4.5
ชุดฝึกสอน	85.71%	76.62%

5. การประเมินผล (Evaluation)

ในงานวิจัยนี้ใช้การทดสอบแบบไขว้ทบ (k - Fold Cross Validation) แบบ 10 ส่วน ดังภาพที่ 2 แล้วทำการทดสอบเพื่อประเมินประสิทธิภาพ

การวัดค่าประสิทธิภาพของเทคนิควิธีต่างๆ จะต้องทำการเลือกข้อมูลสำหรับเรียนรู้ (Training Set) และข้อมูลสำหรับทดสอบ (Testing Set) ในงานวิจัยนี้เลือกใช้วิธีสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง k กลุ่ม (k-Fold Cross Validation) โดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วนๆ ให้เท่าๆ กัน ต่อจากนั้นให้นำข้อมูลบางส่วนมาทำการเรียนรู้ และนำข้อมูลบางส่วนมาทำการทดสอบแบบจำลองที่ได้จากการเรียนรู้ โดยในการทำงานจะทำการเลือกสุ่มข้อมูลออกเป็น k ชุดที่เท่าๆ กัน ในการทดลองครั้งแรก ข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ในการทดลองครั้งที่ 2 ข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดเรียนรู้ ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบและข้อมูลชุดเรียนรู้ ซึ่งจะมีการทดลองทั้งหมด k ครั้ง ในงานวิจัยนี้ได้เลือกใช้ค่า k = 10 ดังอธิบายในภาพที่ 2

	ข้อมูลชุดเรียนรู้									ข้อมูลชุดทดสอบ
รอบที่ 1	2	3	4	5	6	7	8	9	10	1
รอบที่ 2	1	3	4	5	6	7	8	9	10	2
รอบที่ 3	1	2	4	5	6	7	8	9	10	3
รอบที่ 4	1	2	3	5	6	7	8	9	10	4
รอบที่ 5	1	2	3	4	6	7	8	9	10	5
รอบที่ 6	1	2	3	4	5	7	8	9	10	6
รอบที่ 7	1	2	3	4	5	6	8	9	10	7
รอบที่ 8	1	2	3	4	5	6	7	9	10	8
รอบที่ 9	1	2	3	4	5	6	7	8	10	9
รอบที่ 10	1	2	3	4	5	6	7	8	9	10

ภาพที่ 2 10-Fold Cross Validation

การคำนวณประสิทธิภาพของแบบจำลอง สามารถคำนวณได้จากตาราง Confusion Matrix ซึ่งเป็นตารางสรุปจำนวนข้อมูลที่ตัวแบบมีการจำแนกได้อย่างถูกต้องและไม่ถูกต้อง ผลการทดลองแสดงดังตารางที่ 3

ตารางที่ 5 The Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

แล้วทำการวัดประสิทธิภาพตัวแบบการพยากรณ์โดยใช้เกณฑ์การวัดประสิทธิภาพของตัวแบบรู้จำด้วยวิธี Predictive Modeling (Bramer, 2007) ซึ่งประกอบด้วยค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าประสิทธิภาพโดยรวม (F–Measure) และค่าความถูกต้อง (Accurary) ซึ่งมีค่าอยู่ระหว่าง 0 - 1 ซึ่ง 1 หมายถึงประสิทธิภาพดี ดังสมการ (1) (2) (3) และ (4) (ภาสพิชญ์ ชูใจ, 2557) ตามลำดับ

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

โดยที่

TP คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าจริง)
TN คือ ค่าที่พยากรณ์ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าไม่จริง)
FP คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าจริง พยากรณ์ว่าไม่จริง)
FN คือ ค่าที่พยากรณ์ไม่ถูกต้อง (ข้อมูลบอกว่าไม่จริง พยากรณ์ว่าจริง)

ค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ย (Root Mean Squared Error: RMSE) ดังสมการ (5)

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad (5)$$

โดยที่

y_j คือ ค่าข้อมูลจริง
$\hat{y}_j$ คือ ค่าที่ได้จากการพยากรณ์

ในงานวิจัยนี้ได้เลือกใช้วิธีประเมินประสิทธิภาพของการพยากรณ์ด้วยค่าความถูกต้อง โดยใช้ข้อมูลฝึกสอนทดลองปรับค่าพารามิเตอร์ที่เหมาะสมของทั้งเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5 และได้นำมาทดลองกับข้อมูลชุดทดสอบเปรียบเทียบประสิทธิภาพการทำงานของข้อมูลทั้งสองชุดเพื่อป้องกันการเกิด Over-fitting นอกจากนั้นยังได้ทำการทวนสอบผลการทดลองกับข้อมูลอีกชุดหนึ่ง คือ ชุดตรวจสอบ เพื่อเพิ่มความเชื่อมั่นของตัวจำแนกประเภท ดังแสดงในตารางที่ 6

ตารางที่ 6 ค่าความถูกต้องของข้อมูลชุดทดสอบ และชุดตรวจสอบ

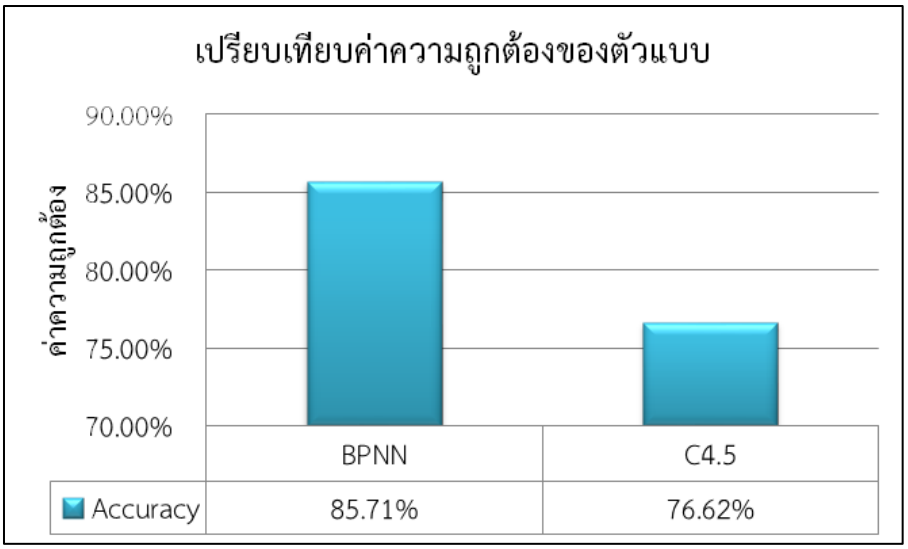
ข้อมูล	BPNN	C4.5
ชุดทดสอบ	85.71%	84.29%
ชุดตรวจสอบ	82.46%	77.19%

6. นำตัวแบบมาใช้งาน (Deployment)

หลังจากทำการประเมินผลตัวจำแนกของข้อมูลชุดฝึกสอนข้อมูลชุดทดสอบ และข้อมูลชุดตรวจสอบเรียบร้อยแล้วได้ผล สามารถนำตัวแบบที่ได้สร้างขึ้นมาใช้ประโยชน์จริงในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษาที่สำเร็จการศึกษาในปี พ.ศ. 2559 และในรุ่นต่อๆ ไปได้

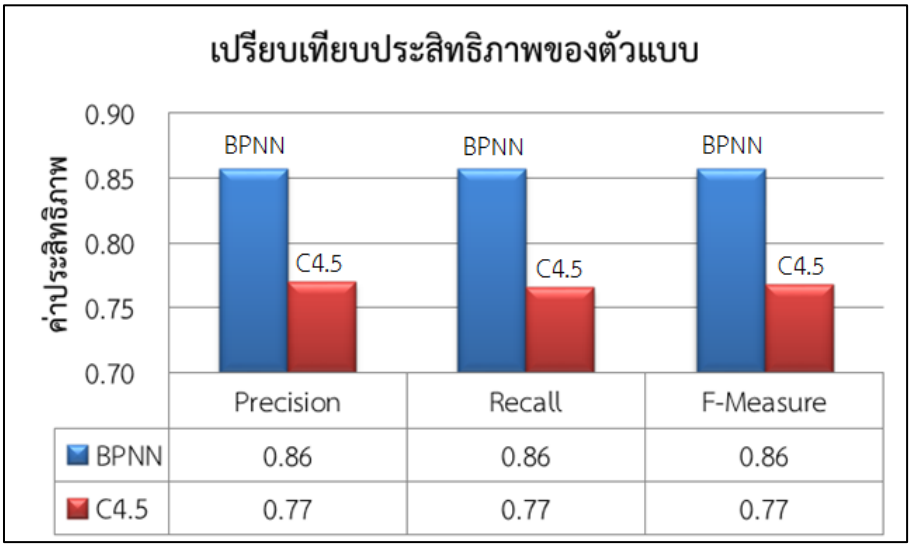
ผลการวิจัย

จากการนำข้อมูลภูมิหลังและข้อมูลผลการเรียนรายวิชาในชั้นปีที่ 1 และ 2 ของนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร ที่สำเร็จการศึกษาประจำปีการศึกษา 2554 – 2559 จำนวนทั้งสิ้น 358 คน จำแนกเป็นข้อมูลชุดฝึกสอน จำนวน 231 คน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2554–2557 ข้อมูลชุดทดสอบ จำนวน 70 คน ซึ่งเป็นผู้สำเร็จการศึกษาประจำปีการศึกษา 2558 และข้อมูลชุดตรวจสอบ จำนวน 57 คน ซึ่งเป็นนักศึกษาที่กำลังจะสำเร็จการศึกษาประจำปีการศึกษา 2559 ไปพยากรณ์ด้วยเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ด้วยวิธีไขว้ทบ 10 ส่วน (10-Fold Cross Validation) พบว่าตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับให้ค่าความถูกต้องในการพยากรณ์สูงกว่าตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจแบบ C4.5 ดังภาพที่ 4



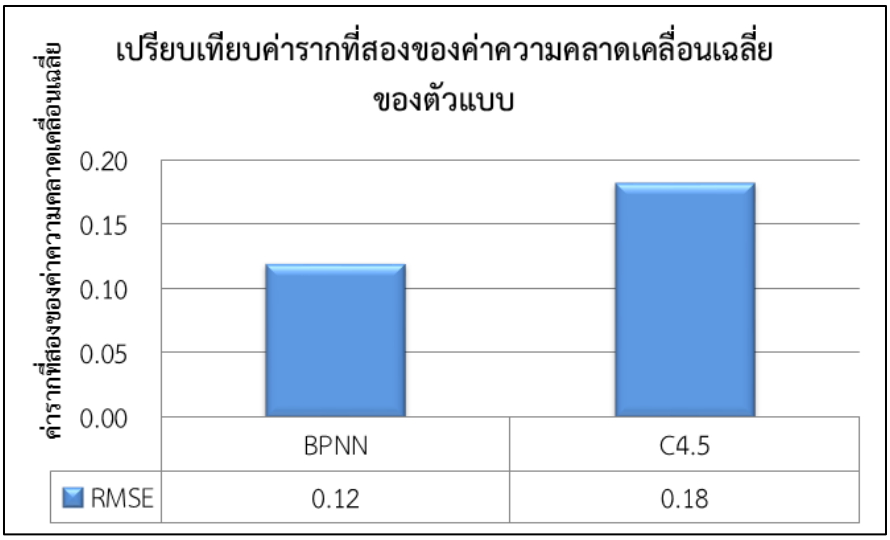
ภาพที่ 4 แผนภูมิเปรียบเทียบค่าความถูกต้องของตัวแบบ

ผลการทดสอบประสิทธิภาพของตัวแบบโดยใช้ค่าความแม่นตรง ค่าความระลึก และค่าประสิทธิภาพโดยรวม พบว่าตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับให้ประสิทธิภาพในการพยากรณ์สูงกว่าตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจแบบ C4.5 ดังภาพที่ 5



ภาพที่ 5 แผนภูมิเปรียบเทียบประสิทธิภาพของตัวแบบ

นอกจากนี้ เมื่อทำการเปรียบเทียบค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ย พบว่า ตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ ให้ค่าความคลาดเคลื่อนน้อยที่สุดเท่ากับ 0.12 และตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจแบบ C4.5 ให้ค่าความคลาดเคลื่อนสูงที่สุดเท่ากับ 0.18 ดังภาพที่ 6



ภาพที่ 6 แผนภูมิเปรียบเทียบค่ารากที่สองของค่าความคลาดเคลื่อนเฉลี่ยของตัวแบบ

ดังนั้นตัวแบบที่เหมาะสมที่สุดในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษา คือ ตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ เนื่องจากเทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับนั้นให้ประสิทธิภาพค่าความถูกต้องในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของของนักศึกษาดีที่สุด และให้ค่าความคลาดเคลื่อนน้อยที่สุด

สรุปและอภิปรายผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อทำการคัดเลือกคุณลักษณะโดยใช้เทคนิค Information Gain ที่ส่งผลต่อระดับผลการเรียนของนักศึกษา และเปรียบเทียบประสิทธิภาพในการพยากรณ์ข้อมูลของเทคนิคเหมืองข้อมูล 2 เทคนิค คือ เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ และเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ของนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร และสาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด เป็นการวิเคราะห์หาองค์ความรู้ที่เป็นประโยชน์เพื่อพยากรณ์ผลสัมฤทธิ์ของนักศึกษา และเพื่อหาแนวทางการแนะแนวการศึกษาที่เหมาะสมแก่นักศึกษาได้อย่างมีประสิทธิภาพมากขึ้น จากการคัดเลือกคุณลักษณะพบว่าเหลือตัวแปรต้น 5 ตัวแปร และตัวแปรตาม 1 ตัวแปร ผลการวิจัยพบว่า เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ ให้ค่าความถูกต้อง ร้อยละ 85.71 ซึ่งมากกว่าเทคนิคต้นไม้ตัดสินใจแบบ C4.5 ซึ่งมีค่าร้อยละ 76.62

การคัดเลือกคุณลักษณะด้วยวิธี Information Gain ร่วมกับการใช้เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับเหมาะสมที่สุดกับข้อมูลที่ใช้ในการสร้างตัวแบบพยากรณ์ในงานวิจัยนี้มากที่สุด และเมื่อดูคุณลักษณะที่ส่งผลในการพยากรณ์ผลสัมฤทธิ์การเรียนของนักศึกษาสาขาวิชาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏกำแพงเพชร และสาขาวิชาคอมพิวเตอร์ธุรกิจ มหาวิทยาลัยราชภัฏกำแพงเพชร แม่สอด ซึ่งผลในการพยากรณ์สามารถนำไปใช้เป็นแรงจูงใจในการศึกษาเพื่อทำให้นักศึกษามีผลการเรียนที่สูงขึ้น และเป็นการวางแผนการเรียน อีกทั้งเป็นเครื่องมือให้อาจารย์ที่ปรึกษาให้คำแนะนำในการศึกษาแก่นักศึกษาได้

ข้อเสนอแนะ

ข้อเสนอแนะในการนำผลการวิจัยไปใช้

สามารถนำโมเดลดังกล่าวไปพัฒนาเป็นซอฟต์แวร์ในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนแก่นักศึกษาได้

ข้อเสนอแนะในการทำวิจัยครั้งต่อไป

1. การใช้เทคนิคเหมืองข้อมูลเป็นการค้นหาความรู้จากฐานข้อมูลที่มีอยู่มาพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษา ซึ่งในการวิจัยครั้งต่อไปควรมีการเพิ่มคุณลักษณะอื่นๆ ที่น่าจะมีผลต่อการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนักศึกษามาพิจารณาเป็นคุณลักษณะในการพยากรณ์

2. ควรมีข้อมูลมากพอในการทำเหมืองข้อมูล อาจใช้เทคนิคเหมืองข้อมูลเทคนิคอื่นๆ มาเปรียบเทียบเพื่อให้ได้ผลการพยากรณ์ที่มีประสิทธิภาพมากขึ้น

เอกสารอ้างอิง

ธาดา จันตะคุณ. (2559). ตัวแบบการจำแนกการเลือกหลักสูตรการศึกษา คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม โดยใช้เทคนิคเหมืองข้อมูล. **การประชุมวิชาการครุศาสตร์อุตสาหกรรม ระดับชาติ ครั้งที่ 9** (น. 336-343). กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ธีรพงษ์ สังข์ศรี. (2557). การวิเคราะห์พฤติกรรมสำหรับการเลือกสมัครสาขาวิชาเรียนและการเปรียบเทียบตัวแบบพยากรณ์จำนวนนักศึกษาใหม่โดยใช้เทคนิคการทำเหมืองข้อมูล. **The 10th National Conference on Computing and Information Technology** (น. 963-968). กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

นิภาพร ชนะมาร และพรรณี สิทธิเดช. (2557). การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์. **วารสารมหาวิทยาลัยราชภัฏสกลนคร, 6**(12), 31-45.

ภาสพิชญ์ ชูใจ. (2557). **การเรียนรู้ร่วมกันสำหรับปัญหาการจำแนกข้อมูลไม่สมดุล.** ปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.

เสกสรรค์ วิลัยลักษณ์, วิภา เจริญภัณฑารักษ์ และดวงดาว วิชาดากุล. (2558). การใช้เทคนิคการทำเหมืองข้อมูลเพื่อพยากรณ์ผลการเรียนของนักเรียนโรงเรียนสาธิตแห่งมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ศูนย์วิจัยและพัฒนาการศึกษา. **Veridian E-Journal, Science and Technology Silpakorn University, 2**(2), 1-17.

Chapman, P.. (2000). **CRISP-DM 1.0 - Step-by-step data mining guide.** Technical report. The CRISP-DM

Han, J., and Kamber, M. (2006). **Data Mining: Concepts and Techniques.** (2nd ed). Morgan Kaufmann.

Bramer, M. (2007). **Principles of Data Mining.** Springer-Verlag London Limited, pp 173-176.

Tan, P. N., Steinbach, M. and Kumar, V. (2006). **Introduction to Data Mining.** Addison-Wesley.