วารสารมหาวิทยาลัยราชภัฏสกลนคร ; 6 (12) : กรกฎาคม – ธันวาคม 2557
Sakon Nakhon Rajabhat University Journal ; 6 (12) : July – December 2014

การวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติและการพยากรณ์

Learning Attributes Analysis by Feature Selection and Prediction

นิภาพร ชนะมาร [1]
พรรณี สิทธิเดช [2]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลในการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิต โดยใช้เทคนิคการคัดเลือกคุณสมบัติที่สำคัญ แล้วสร้างตัวแบบการพยากรณ์ด้วยเทคนิค BPNN และเทคนิค SVMs จากข้อมูลที่คัดเลือกซึ่งเป็นปัจจัยการเรียนรู้ที่สำคัญ ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลของนิสิตที่ศึกษาและสำเร็จการศึกษาแล้วในหลักสูตรเดียวกัน จำนวน 180 ระเบียน ผลการวิเคราะห์ปัจจัยการเรียนรู้ พบว่า ข้อมูลภูมิหลังไม่ใช่ข้อมูลสำคัญในการทำนายผลสัมฤทธิ์ทางการเรียนของนิสิต ตัวแปรที่สำคัญ 10 ตัวแปร เป็นรายวิชาที่มีความสอดคล้องกับกรอบมาตรฐานคุณวุฒิระดับปริญญาตรีสาขาคอมพิวเตอร์ พ.ศ. 2552 ผลการทดลองสร้างตัวแบบการพยากรณ์ด้วยเทคนิค BPNN และเทคนิค SVMs จากตัวแปรที่สำคัญ 10 ตัวแปรมีค่าความผิดพลาดอยู่ในระดับต่ำกว่า ตัวแบบการพยากรณ์ที่ใช้ตัวแปรตั้งต้น 22 ตัวแปร นอกจากนั้น เมื่อทดลองใช้เทคนิคการรวมกลุ่ม ด้วยวิธี Baggingร่วมกับ BPNN และ SVMs พบว่าผลการพยากรณ์ของ Baggingร่วมกับ BPNN (Bagging BPNN) มีค่าความผิดพลาดอยู่ในระดับต่ำสุด (RMSE=0.1051) ใช้ในการพยากรณ์ได้อย่างมีประสิทธิภาพ ผลของงานวิจัยนี้ให้ประโยชน์ในการวิเคราะห์ปัจจัยการเรียนรู้และการพยากรณ์ผลสัมฤทธิ์ทางการเรียนของนิสิตซึ่งจะช่วยให้นิสิตสามารถพยากรณ์ผลการเรียนของตนเองและปรับพฤติกรรมการเรียน ได้เช่น การเพิ่มถอนรายวิชาให้เหมาะสมกับศักยภาพตนเอง

คำสำคัญ : การวิเคราะห์ปัจจัยการเรียนรู้, การทำเหมืองข้อมูล, การคัดเลือกคุณสมบัติ, เทคนิคการพยากรณ์

[1] นิสิตปริญญาเอก ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร
[2] อาจารย์ ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

วารสารมหาวิทยาลัยราชภัฏสกลนคร ; 6 (12) : กรกฎาคม – ธันวาคม 2557
Sakon Nakhon Rajabhat University Journal ; 6 (12) : July – December 2014

Abstract

This research aims to apply data mining techniques in predicting academic achievement of students using feature selection techniques to select important features. Then, build prediction models using BPNN and SVMs techniques from key factors for learning. The 180 data records used in this work were collected from students who studied and graduated in the same program. The learning attributes analysis result shows that students' background information is not important in predicting the academic achievement of students but only 10 courses that they studied in the first and second years. These courses are compliant with the qualifications framework of undergraduate program in Computer Science. The BPNN and SVM spreadiction models were trained by using 10 important parameters and got less error than using all parameters. Furthermore, the ensemble technique; Bagging, was combined to BPNN and SVMs models. It is found that Bagging with BPNN (Bagging BPNN) gave the lowest errors (RMSE = 0.1051). The model can be used to predict learning achievement of the students efficiently. The experiment results of this research will be helpful for learning achievement attributes analysis and for predicting learning achievement in order to help students predict their own learning achievement and adjust their learning behavior such as add or withdraw courses according to their potential.

Keywords : Learning Attributes Analysis, Data Mining, Feature Selection, Prediction Techniques

บทนำ

คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวรได้เริ่มเปิดหลักสูตรปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ มาตั้งแต่ พ.ศ. 2537 มีการปรับปรุงหลักสูตรครั้งสำคัญใน พ.ศ. 2548 ซึ่งหลักสูตรปรับปรุง พ.ศ. 2548 ได้ใช้ต่อเนื่องยาวนานกับนิสิตรหัส 2548 จนถึง รหัส 2554 หลังจากนั้นมีการปรับปรุงเพียงเล็กน้อย จนกระทั่งมีการปรับปรุงครั้งใหญ่เพื่อให้สอดรับกับกรอบมาตรฐานคุณวุฒิระดับอุดมศึกษาแห่งชาติ (Thai Qualifications Framework for Higher Education, TQF:HEd) (NQF. 2006; สำนักงานคณะกรรมการการอุดมศึกษา. 2552ก) ซึ่งเป็นกรอบที่มาตรฐานการเรียนรู้ของแต่ละระดับคุณวุฒิมีการกำหนดปริมาณการเรียนรู้ในแต่ละองค์ความรู้ของแต่ละหลักสูตร ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร ได้ปรับปรุงหลักสูตรปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2555 ตามกรอบมาตรฐานคุณวุฒิระดับปริญญาตรีสาขาคอมพิวเตอร์ พ.ศ. 2552 (สำนักงานคณะกรรมการการอุดมศึกษา. 2552ข)

จากการบริหารจัดการหลักสูตรปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2548มาเป็นเวลาต่อเนื่องยาวนานกับนิสิตรหัส 2548 จนถึง รหัส 2554 พบว่า นิสิตจำนวนมากประสบปัญหาการมีเกรดเฉลี่ยต่ำ บางคนต้องย้ายสาขาหรือต้องพ้นสภาพการเป็นนิสิต ซึ่งเป็นเรื่องที่น่าเสียดาย กรรมการบริหารหลักสูตรและอาจารย์ที่ปรึกษาพยายามช่วยกันแก้ปัญหาโดยการจัดกิจกรรมพบที่ปรึกษา

บ่อยครั้ง จากการสัมภาษณ์นิสิตพบว่า นิสิตมีความรู้พื้นฐานไม่เพียงพอต่อการเรียนในบางวิชา อันส่งผลต่อการเรียนรายวิชานั้นและเกรดเฉลี่ยโดยรวม

ปัจจุบันมีการพัฒนาและประยุกต์ใช้ศาสตร์ทางด้านคอมพิวเตอร์และเทคโนโลยีในการวิเคราะห์ข้อมูล โดยเฉพาะอย่างยิ่งการนำวิธีการทำเหมืองข้อมูลมาช่วยในการวิเคราะห์ข้อมูลให้สามารถสกัดองค์ความรู้ที่มีประโยชน์และนำมาแก้ปัญหาต่าง ๆ ได้มากมาย ในทางการศึกษาสามารถใช้ประโยชน์ของการวิเคราะห์ข้อมูลด้วยการทำเหมืองข้อมูล เป็นเครื่องมืออีกอย่างหนึ่งที่ช่วยให้สถาบันการศึกษาสามารถพัฒนาองค์กรให้มีคุณภาพ ช่วยหาแนวทางพัฒนาระบบการศึกษา ซึ่งมีนักวิจัยทำการวิเคราะห์ข้อมูลและได้ตัวแบบที่มีประสิทธิภาพในการแก้ปัญหาแตกต่างกันไป เช่น จิราพร ยิ่งกว่าชาติ (2549) ใช้เทคนิคการเรียนรู้แบบข่ายงานเบย์มาทำนายผลสำเร็จการศึกษาของนิสิต พบปัจจัยที่มีผลต่อการสำเร็จการศึกษาของนิสิต 3 ปัจจัยได้แก่ เกรดเฉลี่ยในชั้นปีแรก อาชีพของมารดา และรายได้ของครอบครัว สอดคล้องกับงานวิจัยของภัทรพงศ์ พงศ์ภัทรกานต์ (2553) ได้วิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรี พบปัจจัยที่สำคัญ 3 ปัจจัย คือ ขนาดโรงเรียนเดิม อาชีพของบิดา และอาชีพของมารดา

ดังนั้นงานวิจัยนี้จึงมุ่งที่จะวิเคราะห์ปัจจัยการเรียนรู้ที่สำคัญในสาขาวิชาวิทยาการคอมพิวเตอร์ด้วยหลักการทำเหมืองข้อมูล โดยการคัดเลือกคุณสมบัติที่สำคัญและทำการพยากรณ์ผลสัมฤทธิ์การเรียนรู้ของนิสิต โดยใช้ข้อมูลผลการเรียนของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ตามหลักสูตร ฉบับปรับปรุง พ.ศ. 2548 ซึ่งเป็นหลักสูตรเดียวกัน และสำเร็จการศึกษาแล้วเพื่อศึกษาความเหมาะสมของรายวิชาในหลักสูตรและเพื่อประโยชน์ในการปรับปรุงหลักสูตรต่อไป

บทความนี้จะได้นำเสนอแนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง วิธีดำเนินการวิจัย สรุปผลการวิจัย และอภิปรายผล ตามลำดับ

แนวคิดทฤษฎีที่เกี่ยวข้อง

1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูลเป็นกระบวนการที่สกัดข้อมูลที่มีประโยชน์ เพื่อให้ได้สารสนเทศที่มีเหตุผลและสามารถนำไปใช้ช่วยในการตัดสินใจ สามารถอธิบายรูปแบบที่น่าสนใจจากข้อมูล รวมถึงใช้เพื่อการทำนาย หรือคาดการณ์สิ่งที่น่าจะเกิดในอนาคต เป็นการนำข้อมูลที่เก็บรวบรวมมาในอดีต มาสร้างตัวแบบ (Model) แล้วนำตัวแบบนั้นไปใช้ทำนายหรือพยากรณ์การเกิดเหตุการณ์ในอนาคต (Han and Kamber. 2001; Tan, Steinbach and Kumar. 2006) กระบวนการในการวิเคราะห์ข้อมูลและสร้างตัวแบบที่ได้รับความนิยมมากในปัจจุบัน เรียกว่า แนวคิดกระบวนการมาตรฐานอุตสาหกรรม (CRIPS-DM: CRoss-Industry Standard Process for Data Mining) (Chapman et al. 2000) ประกอบด้วย 6 ขั้นตอนหลัก คือ 1) การทำความเข้าใจโจทย์ (Business Understanding) มุ่งเน้นทำความเข้าใจวัตถุประสงค์ของโครงการ วางแผน และออกแบบเบื้องต้นเพื่อให้บรรลุวัตถุประสงค์ 2) การทำความเข้าใจข้อมูล (Data Understanding) โดยเริ่มต้นด้วยการเก็บรวบรวมข้อมูล ทำความคุ้นเคยกับข้อมูล ระบุปัญหาคุณภาพของข้อมูลเชิงลึก รวมถึงการตั้งสมมติฐานสำหรับการค้นหาความรู้ที่ซ่อนอยู่ 3) การเตรียมข้อมูล (Data Preparation) เป็นขั้นตอนที่ครอบคลุมกิจกรรมทั้งหมดใน

การสร้างชุดข้อมูล รวมถึงการบันทึกข้อมูล การเลือกตัวแปรที่สำคัญ การเปลี่ยนแปลงข้อมูล การทำความสะอาดข้อมูล 4) การสร้างตัวแบบ (Modeling) เป็นการสร้างตัวแบบด้วยเทคนิคเหมืองข้อมูล การเลือกข้อมูลที่ใช้และการปรับพารามิเตอร์ของเทคนิคเหมืองข้อมูล เพื่อเปรียบเทียบประสิทธิภาพของหลายเทคนิคเหมืองข้อมูลที่ใช้แก้ปัญหา 5) การประเมินผล (Evaluation) ก่อนที่จะนำตัวแบบไปใช้งานขั้นสุดท้ายจะต้องประเมินผลตัวแบบอย่างละเอียด ตรวจสอบความถูกต้องของการบรรลุวัตถุประสงค์ ในตอนท้ายของขั้นตอนนี้จะต้องมีการตัดสินใจเลือกรูปแบบที่เหมาะสมที่สุด 6) การใช้งาน (Deployment) เป็นการประยุกต์ใช้ตัวแบบ โดยส่วนใหญ่จะถูกนำไปใช้ประโยชน์ในการตัดสินใจขององค์กรตามสภาพปัญหาและวัตถุประสงค์ของโครงการ

2. การคัดเลือกคุณสมบัติ (Feature Selection)

การคัดเลือกคุณสมบัติเป็นเทคนิคที่ช่วยลดจำนวนตัวแปรที่จะใช้ในตัวแบบพยากรณ์ อาจกระทำเพื่อเลือกตัวแปรที่ดีที่สุดเพียงตัวเดียว หรือเลือกกลุ่มของตัวแปรที่มีความสำคัญต่อการพยากรณ์ กระบวนการคัดเลือกคุณสมบัติเป็นกระบวนการที่สำคัญในการเตรียมข้อมูลของการทำเหมืองข้อมูล เพื่อทำให้การสร้างตัวแบบพยากรณ์มีประสิทธิภาพ เพราะจะช่วยลดมิติของข้อมูลและอาจช่วยให้การเรียนรู้วิธีการพยากรณ์ดำเนินการได้เร็วขึ้นและมีประสิทธิภาพมากขึ้น ในงานวิจัยนี้ ทดลองใช้การคัดเลือกคุณสมบัติ 3 วิธี ดังต่อไปนี้

2.1 การคัดเลือกคุณสมบัติแบบ Correlation-based Feature Selection เป็นการคัดเลือกกลุ่มคุณสมบัติอย่างง่าย ใช้หลักการคำนวณค่าความสัมพันธ์ระหว่างคุณสมบัติย่อยต่อค่าพยากรณ์ ซึ่งอาจใช้คำนวณด้วยค่าสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson's correlation) และมีการจัดอันดับตามค่าความสัมพันธ์เพื่อประเมินค่าความสามารถในการพยากรณ์ของแต่ละคุณสมบัติ นอกจากนั้นยังพิจารณาคัดเลือกกลุ่มของคุณสมบัติที่มีความสัมพันธ์ภายในระหว่างคุณสมบัติย่อยกันเองต่ำเพื่อลดความซ้ำซ้อนของอิทธิพลการพยากรณ์ (Hall and Smith. 1998)

2.2 การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection เป็นการคัดเลือกคุณสมบัติที่ต้องกำหนดตัวชี้วัดความสอดคล้องมั่นคงไว้ก่อนเป็นอันดับแรก จากนั้นใช้ตัวชี้วัดนี้ เพื่อประเมินความมั่นคงของกลุ่มคุณสมบัติที่เกี่ยวข้องกับค่าพยากรณ์เดียวกัน (Liu and Setiono. 1996; Liu et al. 1998) ตัวชี้วัดความมั่นคงที่ถูกกำหนดให้เป็นตัวชี้วัดนี้ อาจกำหนดมาตรวัดเช่นเดียวกับการวัดระยะทางของคุณสมบัติย่อยและสามารถแปลผลคุณสมบัติที่สอดคล้องกันมากด้วยค่าเข้าใกล้ศูนย์ กระบวนการคัดเลือกคุณสมบัติจะกระทำซ้ำและเลือกกลุ่มคุณสมบัติที่มีค่าตัวชี้วัดน้อยและจำนวนคุณสมบัติเท่าเดิมหรือลดลงเท่านั้น วิธีการค้นหากลุ่มคุณสมบัติที่สอดคล้องมั่นคงนี้เป็นวิธีที่รวดเร็วและสามารถทราบความสัมพันธ์ระหว่างคุณสมบัติได้ (Hall and Holmes. 2003)

2.3 การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection เป็นวิธีคัดเลือกตัวแปรโดยมีหลักการเช่นเดียวกับการเลือกตัวแปรของการสร้างต้นไม้ตัดสินใจ เพื่อให้ได้ตัวแปรที่เป็นตัวแบ่งข้อมูลออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด (Homogeneous) ด้วยมาตรวัดการได้ประโยชน์จากการแบ่งกลุ่มย่อยเรียกว่า อัตราส่วนเกน (Gain Ratio) ซึ่งเป็นอัตราส่วนของค่าเกน (Gain หรือ Information Gain) กับ ค่าสารสนเทศการแบ่งกลุ่ม (Split Info) อันเป็นการลดอิทธิพลของตัวแปรที่มีค่าหลาย

ค่า ผลที่ได้รับจากการใช้เทคนิคนี้จะได้ลำดับของตัวแปรซึ่งตัวแปรที่อยู่ลำดับแรกๆ จะถือว่ามีอิทธิพลในการพยากรณ์ตัวแปรเป้าหมายมากกว่าตัวแปรในลำดับถัดไป ทำให้เราสามารถพิจารณาเลือกจำนวนตัวแปรที่เหมาะสมได้อย่างมีประสิทธิภาพ (Tan, Steinbach and Kumar. 2006; Asha, Manjunath and Jayaram. 2010)

3. เทคนิคการพยากรณ์ (Prediction)

เทคนิคการพยากรณ์เป็นกระบวนการสร้างรูปแบบทางคณิตศาสตร์หนึ่งอย่างหรือมากกว่าโดยอาศัยข้อมูลจากอดีตมาสร้างเป็นตัวแบบ (Model) ในการพยากรณ์ (Sang. 2012) ในงานวิจัยนี้ได้ทดลองสร้างตัวแบบในการพยากรณ์ด้วยเทคนิคต่อไปนี้

3.1 เทคนิคโครงข่ายประสาทเทียมแบบแพร่กลับ (Backpropagation Neural Network: BPNN) เป็นขั้นตอนวิธีโครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer Perceptron: MLP) เรียนรู้โดยการปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนด (Nodes) ให้เหมาะสมโดยการปรับค่าจะขึ้นกับความแตกต่างของค่าเอาท์พุตที่คำนวณได้กับค่าจริงในข้อมูลฝึกสอนพื้นฐาน การเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่กลับจะใช้ข้อมูลที่ผ่านเข้าไปในลำดับชั้นที่แตกต่างกันของโครงข่ายโดยอินพุตเวกเตอร์ (Input Vector) จะถูกนำเข้าและจะส่งผลกระทบต่อการแพร่กระจายเข้าไปในชั้นของระบบโครงข่ายชั้นต่อชั้นแบบตัวเดินหน้าผ่านค่าน้ำหนัก (Weight) จนได้ผลลัพท์ของโครงข่าย ซึ่งในทางกลับกันของกระบวนการแพร่กลับ ค่าน้ำหนักของโครงข่ายจะถูกปรับเปลี่ยนไปโดยขึ้นอยู่กับค่าความผิดพลาดของผลที่คำนวณจากโครงข่าย โดยลำดับชั้นการทำงานของ BPNN ที่นิยมใช้ มี 3 ลำดับชั้นได้แก่ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นนำออก (Output Layer) (Baha, Emine and Dursun. 2012; Rojas. 1996)

3.2 เทคนิคซัพพอร์ตเวกเตอร์แมทชีน (Support Vector Machines: SVMs) เป็นเทคนิคการจำแนกประเภทข้อมูลที่มีพื้นฐานมาจากทฤษฎีการเรียนรู้ทางสถิติคล้ายเทคนิคโครงข่ายประสาทเทียม (Luts et al. 2010; Cristianini and Shawe-Taylor. 2000) หลักการของSVMs คือการใช้ข้อมูลนำเข้าที่ใช้ฝึกสอนเป็นเวคเตอร์ในสเปซ N มิติ เช่น ถ้าในกรณีของ 2 มิติและ 3 มิติจะเป็นจุดที่อยู่ในระนาบ (x, y) และสเปซ (x, y, z) ตามลำดับ จากนั้นทำการสร้างไฮเปอร์เพลน (Hyperplane) ที่จะแยกกลุ่มของเวคเตอร์อินพุตออกเป็นกลุ่มต่าง ๆในกรณีที่เป็น 2 มิติไฮเปอร์เพลนก็คือเส้นตรงและกรณีที่เป็น 3 มิติไฮเปอร์เพลนก็คือระนาบ กระบวนการสอนให้ระบบเรียนรู้ของเทคนิค SVMs ทำโดยเลือกเส้นหรือระนาบจำแนกประเภทข้อมูลที่เหมาะสมที่สุดโดยการคำนวณค่าสัมประสิทธิ์ของสมการที่เหมาะสมซึ่งจะอยู่ในรูปแบบของฟังก์ชันที่คำนวณได้จากข้อมูลที่นำมาฝึกสอนโดยใช้ทฤษฎีการหาค่าเหมาะสมที่สุด (Optimization Theory) ในบางกรณีที่ข้อมูลสองกลุ่มอาจจะจับกลุ่มในตำแหน่งต่างๆแบบไม่เชิงเส้น (Non-Linear) ทำให้ไม่สามารถใช้วิธีแบบเชิงเส้นได้ ข้อเด่นของเทคนิค SVMs จะใช้ฟังก์ชันเคอร์เนล (Kernel Function) ทำการแปลงค่าเวคเตอร์ในสเปซข้อมูลนำเข้าไปสู่สเปซข้อมูลแบบอื่นได้ ทำให้สามารถรองรับข้อมูลที่มีมิติเพิ่มมากขึ้นได้

3.3 เทคนิคการรวมกลุ่มตัวจำแนกประเภท(Ensemble Classify)เป็นเทคนิคในการรวมเอากลุ่มของตัวจำแนกข้อมูลอย่างง่ายหลายตัวจำแนกมาช่วยแก้ปัญหาเดียวกันด้วยการนำเอาตัวจำแนกข้อมูลเหล่านั้นมาร่วมกันตัดสินใจ ซึ่งเทคนิคที่นิยม คือ วิธีการ Bagging และวิธีการ Boosting (Shixin. 2003;

Freund and Schapire. 1995) ในงานวิจัยนี้เลือกใช้วิธีการ Bagging ซึ่งเป็นวิธีการสุ่มข้อมูลย่อยจากชุดข้อมูลฝึกสอนแบบสุ่มซ้ำได้ (Bootstraps Replacement) จำนวนหลายเซตย่อย นำมาสร้างตัวจำแนกที่แตกต่างกัน แล้วจึงนำผลการทำนายของแต่ละตัวจำแนกย่อยมาพิจารณาร่วมกัน ที่นิยมใช้ คือ การลงคะแนนเสียงข้างมาก (Breiman.1996)

วิธีดำเนินการวิจัย

การวิจัยนี้ได้ดำเนินงานโดยประยุกต์ตามแนวทางในการทำเหมืองข้อมูลที่เรียกว่า กระบวนการมาตรฐานอุตสาหกรรม หรือ CRIPS-DM (Cross Reference Industry Standard for Data Mining)(Chapman et al. 2000) ที่ได้รับความนิยมมากในปัจจุบัน ซึ่งมีขั้นตอนการดำเนินการวิจัย ดังภาพที่ 1 รายละเอียดการทำงานแต่ละขั้นตอน มีดังนี้

1. ขั้นการจัดเตรียมข้อมูล

ข้อมูลที่ใช้ในงานวิจัยนี้ คือ ข้อมูลของนิสิตที่ศึกษาหลักสูตรปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2548 จำนวน 180 ระเบียน ประกอบด้วย คุณสมบัติ 23 ตัวแปร แบ่งเป็นตัวแปรอิสระ 22 ตัวแปร ได้แก่ ข้อมูลภูมิหลังต่างๆ และข้อมูลผลการเรียนของรายวิชาที่ศึกษาในแผนการเรียนชั้นปีที่หนึ่งและชั้นปีที่สอง ตัวแปรตามหรือตัวแปรพยากรณ์ คือ เกรดเฉลี่ยของนิสิตเมื่อสำเร็จการศึกษา โดยข้อมูลที่ไม่เกี่ยวข้องกับการทำนาย เช่น สถานะการศึกษา ที่อยู่ หมายเลขโทรศัพท์และข้อมูลที่ไม่มีค่า (Missing Value) ซึ่งได้ถูกคัดออก ดังแสดงในตาราง 1

2. ขั้นการคัดเลือกคุณสมบัติ

ผู้วิจัยได้ทดลองสร้างตัวแบบการพยากรณ์จากข้อมูลทั้งหมดที่มีตัวแปรอิสระ 22 ตัวแปร ด้วยเทคนิคBPNN และเทคนิคSVMs ได้ผลการพยากรณ์ที่มีค่ารากที่สองของกำลังสองของข้อผิดพลาด (Root Mean Square Error: RMSE) เท่ากับ 0.2444 และ 0.1246 ตามลำดับ หลังจากนั้น จึงทำการวิเคราะห์ปัจจัยการเรียนรู้ด้วยการคัดเลือกคุณสมบัติที่สำคัญ โดยใช้เทคนิคการคัดเลือกคุณสมบัติ 3 วิธี ได้แก่ การคัดเลือกคุณสมบัติแบบ Correlation-based Feature Selection การคัดเลือกคุณสมบัติแบบ Consistency-based Feature Selection และ การคัดเลือกคุณสมบัติแบบ Gain Ratio Feature Selection ผลการทดลองทั้งสามเทคนิคสามารถลดจำนวนของคุณสมบัติจาก 22 ตัวแปร เหลือ 9ตัวแปร 10 ตัวแปร และ 11 ตัวแปร ตามลำดับ ดังแสดงในตาราง 2

3. ขั้นการสร้างตัวแบบในการพยากรณ์

หลังจากคัดเลือกคุณสมบัติที่สำคัญด้วยเทคนิคการคัดเลือกคุณสมบัติ 3 วิธี แล้ว จึงนำข้อมูลตามผลการคัดเลือกมาสร้างตัวแบบในการพยากรณ์ผลสัมฤทธิ์ทางการเรียน ด้วยเทคนิคBPNNและเทคนิค SVMs ซึ่งในงานวิจัยนี้ได้ใช้เทคนิค 10-fold Cross Validation ซึ่งเป็นการแบ่งข้อมูลออกเป็น 10 ส่วนเท่าๆ กัน ทำการสร้างตัวแบบ 10 ครั้ง โดยใช้ข้อมูลครั้งละเพียง 9 ส่วนเป็นส่วนของการฝึกสอน และข้อมูลอีก 1 ส่วนที่เหลือเป็นส่วนของการทดสอบ พบว่า ชุดข้อมูลที่มีจำนวนคุณสมบัติ 10 ตัว ให้ค่าความถูกต้องในการพยากรณ์ของทั้งสองเทคนิคสูงกว่าชุดข้อมูลอื่นๆ จากนั้นจึงนำชุดข้อมูลนี้มาทำการทดลองอีกครั้ง โดยมีการปรับค่าพารามิเตอร์ที่สำคัญเพื่อเพิ่มประสิทธิภาพให้กับตัวแบบพยากรณ์ ดังนี้

1) การปรับค่าพารามิเตอร์ของเทคนิค BPNN คือ ปรับค่าจำนวนโหนดในชั้นซ่อน (Hidden Layer) เป็นจำนวน 5 โหนด 6 โหนด 7 โหนด 8 โหนด 9 และ 10 โหนด ตามลำดับ กำหนดค่าอัตราการเรียนรู้ (Learning Rate) เป็น 0.2 กำหนดค่าโมเมนตัม (Momentum) เป็น 0.2 และกำหนดจำนวนรอบในการทดสอบ (Training Time) เป็น 100 รอบ จากนั้นทำการเปรียบเทียบประสิทธิภาพ ผลการทดลองพบว่า ตัวแบบโครงสร้าง BPNN ที่ดีที่สุดมีจำนวนชั้นซ่อนเป็น 10 โหนด(10:10:1)

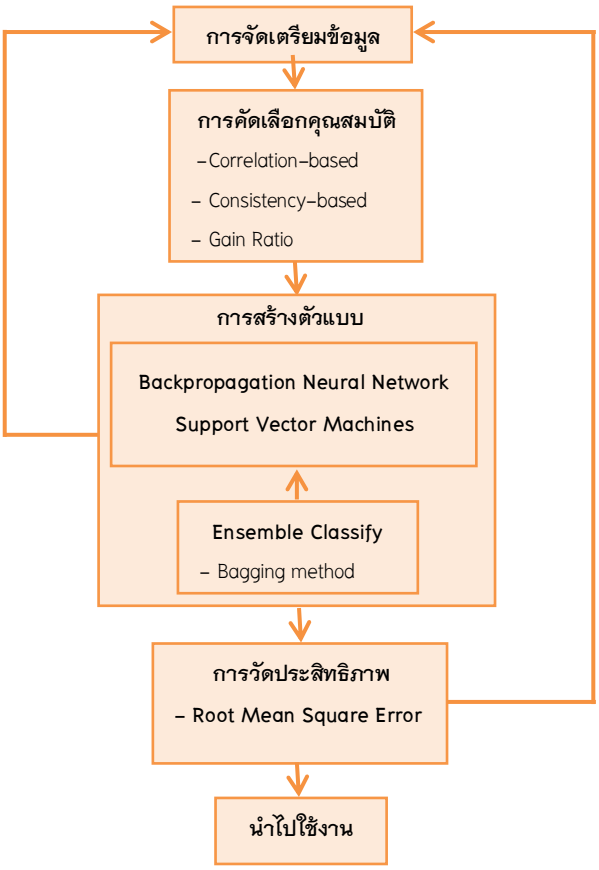
2) การปรับค่าพารามิเตอร์ของเทคนิค SVMs คือกำหนดค่าคอมเพลกซิตี้ (Complexity) หรือค่า C เป็น 1 และเลือกเคอร์เนล เป็นแบบโพลินอเมียลฟังก์ชัน (Polynomial Function) และ แบบเรเดียลเบสิสฟังก์ชัน (Radial Basis Function) จากนั้นทำการเปรียบเทียบประสิทธิภาพ ผลการทดลองพบว่า ได้ตัวแบบพยากรณ์ด้วยเทคนิค SVMs ด้วยเคอร์เนลแบบโพลินอเมียลฟังก์ชัน

3) เมื่อได้ตัวแบบพยากรณ์ที่ดีที่สุดของเทคนิคการพยากรณ์ทั้ง 2 เทคนิค ผู้วิจัยได้ทดลองเพิ่มประสิทธิภาพการพยากรณ์ด้วยเทคนิคการรวมกลุ่มตัวจำแนกประเภท โดยเลือกใช้วิธีการ Baggingให้กับตัวแบบพยากรณ์ทั้งสองเทคนิค ผลการทดลองพบว่า ตัวแบบพยากรณ์ทั้งสองเทคนิคนั้นมีประสิทธิภาพในการพยากรณ์เพิ่มขึ้น

4. ขั้นการวัดประสิทธิภาพตัวแบบในการพยากรณ์

การวัดประสิทธิภาพของตัวแบบการพยากรณ์ที่ทำการทดลองในงานวิจัยนี้ เลือกใช้การวัดประสิทธิภาพของตัวแบบด้วยค่ารากที่สองของกำลังสองของข้อผิดพลาด (Root Mean Square Error: RMSE) ซึ่งเป็นมาตรวัดที่นิยมใช้เพื่อวัดความแตกต่างระหว่างค่าที่พยากรณ์และค่าจริง สรุปผลการประเมินประสิทธิภาพ แสดงในตาราง 3 และภาพที่ 2

วารสารมหาวิทยาลัยราชภัฏสกลนคร ; 6 (12) : กรกฎาคม – ธันวาคม 2557
Sakon Nakhon Rajabhat University Journal ; 6 (12) : July – December 2014



ภาพที่ 1 ขั้นตอนการดำเนินงานวิจัย

ตาราง 1 ตัวแปรที่ใช้ในการวิเคราะห์

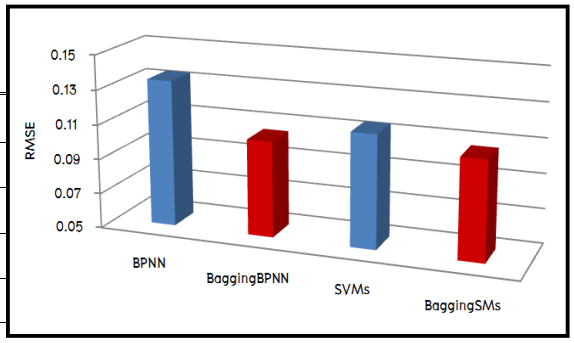
ลำดับ	ชื่อตัวแปร	รายละเอียด	ชนิดข้อมูล
1	Entrance	Entrance Method	Nominal
2	Sex	Gender	Nominal
3	In_Fa	Father's income	Ordinal
4	Occu_Fa	Father's career	Nominal
5	In_Mo	Mother's income	Ordinal
6	Occu_Mo	Mother career	Nominal
7	Acad_Fa	Father's degree	Ordinal
8	Acad_Mo	Mother's degree	Ordinal
9	Old_GPA	Grade 12 GPA	Interval
10	GENERAL	General subjects	Interval
11	251100	Philosophy of Sciences	Interval
12	252111	Introductory Mathematics	Interval
13	256103	Introductory Chemistry	Interval
14	258101	Introductory Biology	Interval
15	261103	Introductory Physics	Interval
16	252112	Calculus	Interval
17	254271	Introduction to Programming	Interval
18	252351	Discrete Mathematics	Interval
19	254251	Data Structure	Interval
20	254261	Computer Architecture	Interval
21	254275	Object Oriented Programming	Interval
22	254351	Database System	Interval
23	GPA	Learning Achievement	Interval

ตาราง 2 แสดงผลการคัดเลือกคุณสมบัติ

วิธีการคัดเลือกคุณสมบัติ	ผลการคัดเลือกคุณสมบัติ										
	General	251100	252111	258101	252112	252351	254271	254275	254251	254261	254351
Consistency-based Feature Selection	√	√	√	√	√		√	√		√	√
Correlation-based Feature Selection	√		√	√	√	√	√	√	√	√	√
Gain Ratio Feature Selection	√	√	√	√	√	√	√	√	√	√	√

ตาราง 3 เปรียบเทียบการวัดประสิทธิภาพของตัวแบบพยากรณ์

ตัวแบบพยากรณ์	ค่า RMSE
BPNN	0.1344
BaggingBPNN	**0.1051**
SVMs	0.1150
BaggingSVMs	0.1071



ภาพที่ 2 ผลการประเมินประสิทธิภาพของตัวแบบพยากรณ์ด้วยวิธีต่าง ๆ

จากตาราง 3 และภาพที่ 2 จะเห็นได้ว่าระหว่างตัวแบบพยากรณ์ BPNN และ SVMs ที่ยังไม่มีการเพิ่มประสิทธิภาพด้วยวิธีการ Bagging นั้น พบว่า ตัวแบบพยากรณ์ SVMs มีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบพยากรณ์ BPNN เพราะค่า RMSE น้อยกว่า คือ 0.1150 และ 0.1344 ตามลำดับ แต่เมื่อมีการเพิ่มประสิทธิภาพในการพยากรณ์ของตัวแบบพยากรณ์ (BaggingBPNN และ BaggingSVMs) พบว่า ตัวแบบพยากรณ์ทั้งสองแบบมีประสิทธิภาพในการพยากรณ์ดีขึ้นจากตัวแบบพยากรณ์เดิม และพบว่าตัวแบบพยากรณ์ BaggingBPNN กลับมีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบพยากรณ์BaggingSVMs เพราะค่า RMSE น้อยกว่า คือ 0.1051 และ 0.1071 ตามลำดับ

สรุปผลการวิจัยและข้อเสนอแนะ

งานวิจัยนี้ทำการวิเคราะห์ปัจจัยการเรียนรู้ที่สำคัญในสาขาวิชาวิทยาการคอมพิวเตอร์ และสร้างตัวแบบพยากรณ์ผลสัมฤทธิ์ทางการเรียน (เกรดเฉลี่ย) เมื่อสำเร็จการศึกษา โดยใช้ข้อมูลภูมิหลังต่าง ๆ ได้แก่ วิธีการคัดเลือกเข้าศึกษาระดับอุดมศึกษา เพศ รายได้ของบิดามารดา อาชีพของบิดามารดา วุฒิการศึกษาของบิดามารดา เกรดเฉลี่ยของนักเรียนเมื่อสำเร็จมัธยมศึกษา นำมารวมกับข้อมูลผลการเรียนของรายวิชาที่ศึกษาในแผนการเรียนชั้นปีที่หนึ่งและชั้นปีที่สอง รวมทั้งสิ้นเป็นจำนวนตัวแปรอิสระ 22 ตัวแปร สร้างตัวแบบ

การพยากรณ์ตัวแปรตาม ได้แก่ เกรดเฉลี่ยเมื่อสำเร็จการศึกษา ข้อมูลที่ใช้ในงานวิจัยนี้ได้มาจากข้อมูลของนิสิตที่ศึกษาหลักสูตรเดียวกัน ในระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2548 และสำเร็จการศึกษาแล้ว จำนวน 180 ระเบียน

การดำเนินการวิจัย ได้นำข้อมูลมาคัดเลือกคุณสมบัติที่สำคัญด้วยเทคนิคการคัดเลือกคุณสมบัติ 3 วิธี สามารถลดจำนวนของคุณสมบัติจาก 22 ตัวแปรเหลือ 10 ตัวแปร แล้วนำข้อมูลมาใช้สร้างตัวแบบการพยากรณ์ด้วยเทคนิคBPNNและเทคนิคSVMs ทำการปรับค่าพารามิเตอร์ของตัวแบบให้มีประสิทธิภาพสูงขึ้น รวมทั้งเพิ่มเทคนิคการรวมกลุ่มตัวจำแนกประเภทด้วยวิธี Bagging ได้ตัวแบบการพยากรณ์ที่มีประสิทธิภาพ มีค่า RMSE อยู่ในระดับต่ำมาก โดย BaggingBPNN มีประสิทธิภาพสูงสุดที่ค่า RMSE เท่ากับ 0.1051

จากผลการวิเคราะห์คัดเลือกคุณสมบัติที่สำคัญที่ได้จำนวน 10 ตัวแปร พบว่า ตัวแปรดังกล่าว เป็นรายวิชาที่มีความสอดคล้องกับกรอบมาตรฐานคุณวุฒิระดับปริญญาตรีสาขาคอมพิวเตอร์ พ.ศ. 2552 (สำนักงานคณะกรรมการการอุดมศึกษา.2552ข) รวมทั้งยังคงเป็นรายวิชาที่บรรจุไว้ในหลักสูตรปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ ฉบับปรับปรุง พ.ศ. 2555 การศึกษาความเหมาะสมของรายวิชาในหลักสูตรนั้นจะเป็นประโยชน์ในการบริหารหลักสูตรและการปรับปรุงหลักสูตรในอนาคต ซึ่งคุณสมบัติอันพึงประสงค์ของบัณฑิตที่เป็นที่ต้องการของสังคมจะช่วยรักษามาตรฐานของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ของประเทศไทยให้เทียบเคียงกับประเทศใกล้เคียง อันจะนำไปสู่การแลกเปลี่ยนนิสิตระหว่างประเทศในประชาคมอาเซียน นอกจากประโยชน์ในการบริหารหลักสูตรของสถาบันการศึกษาแล้ว ตัวแบบการพยากรณ์เกรดเฉลี่ยของนิสิตยังจะช่วยให้นิสิตสามารถประเมินตนเองและปรับพฤติกรรมการเรียน เช่น การเพิ่มถอนรายวิชาให้เหมาะสมกับศักยภาพของตนเอง กล่าวคือ หากเป็นรายวิชาบังคับ นิสิตอาจจะชะลอการเรียนในรายวิชาที่อาจส่งผลต่อการพ้นสภาพ และหากเป็นรายวิชาเลือก นิสิตก็จะมีข้อมูลช่วยการตัดสินใจเลือกรายวิชาที่เหมาะสมกับตนเอง

อย่างไรก็ตามจากข้อจำกัดทางด้านข้อมูลของจำนวนนิสิตที่สำเร็จการศึกษาด้วยหลักสูตรเดียวกันจึงไม่สามารถใช้ข้อมูลของนิสิตที่ศึกษาหลักสูตรเก่าได้ ทำให้ข้อมูลที่ใช้ในการศึกษามีน้อย การวิเคราะห์ข้อมูลจึงยังไม่สามารถอ้างอิงและทดสอบได้มากนัก ดังนั้นจึงควรใช้ข้อมูลของนิสิตที่กำลังจะสำเร็จการศึกษาในการวิเคราะห์และทดสอบเพิ่มเติมรวมทั้งวิเคราะห์ด้วยตัวแบบการพยากรณ์ขั้นสูงอื่น ๆ ต่อไป

เอกสารอ้างอิง

จิราพร ยิ่งกว่าชาติ. (2549). **การประยุกต์ใช้การเรียนรู้แบบเบย์กับการสร้างแบบจำลองสำหรับทำนายผลสำเร็จการศึกษาของนักศึกษา.**วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม. (อัดสำเนา).

ภัทรพงศ์ พงศ์ภัทรกานต์. (2553). **การวิเคราะห์ปัจจัยที่ส่งผลต่อการพ้นสภาพของนักศึกษาระดับปริญญาตรีโดยใช้คอมมิตตีแมชชีน**. The 6th National Conference on Computing and Information Technology. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือกรุงเทพฯ. หน้า 491–496.

สำนักงานคณะกรรมการการอุดมศึกษา. (2552ก). **กรอบมาตรฐานคุณวุฒิระดับอุดมศึกษา (TQF: HEd)**. สืบค้นเมื่อวันที่ 8 มีนาคม 2557, จากเว็บไซต์ http://www.mua.go.th/users/tqf-hed/news/news6.php hed/news/news6.php

_____. (2552ข). **มาตรฐานคุณวุฒิระดับปริญญาตรีสาขาคอมพิวเตอร์ พ.ศ. 2552**. สืบค้นเมื่อวันที่ 8 มีนาคม 2557, จากเว็บไซต์ http://www.mua.go.th/users/tqf-hed/news/FilesNews/FilesNews6/computer_m1.pdf

Asha, G. K., Manjunath, A. S. and Jayaram, M. A. (2010). Comparative Study Of Attribute Selection using gain ratio and correlation based feature selection. **International Journal of Information Technology and Knowledge Management**, Vol. 2(2), pp. 271–277.

Baha, S., Emine, U. and Dursun, D. (2012). Predicting and analyzing secondary education placement-test scores: A data mining approach. **Expert Systems with Applications**. Vol.39, pp. 9468–9476.

Breiman, L. (1996). Bagging predictors. **Machine Learning**. Vol. 24(2), pp. 123–140.

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

Cristianini, N. and Shawe–Taylor, J. (2000). **An Introduction to Support Vector Machines and Other Kernel–Based Learning Methods**. Cambridge : Cambridge University Press.

Freund, Y. and Schapire, R.E. (1995). A decision–theoretic generalization of on–line learning and an application to boosting. **The 2nd European Conference on Computational Learning Theory**, London. pp. 23–37.

Hall, M. and Smith, L. (1998). Practical feature subset selection for machine learning. McDonald, C. (Ed.), **In Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98**, Perth, 4–6 February, pp. 181–191, Berlin: Springer.

Hall, M. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. **IEEE Transactions on Knowledge and Data Engineering**. Vol. 15, pp. 1437–1447.

Han, J. and Kamber, M. (2001). **Data Mining Concepts and Techniques**. Morgan Kaufmann Publishers.

Liu, H. and Setiono, R. (1996). **A probabilistic approach to feature selection: a filter solution**. 13th International Conference on Machine Learning (ICML), Italy. pp.319.

Liu, H. et al. (1998). **A monotonic measure for optimal feature selection**. Proceedings in 10th European Conference on Machine Learning. Chemnitz, Germany. April 21–23, pp. 101–106.

Luts, J. et al. (2010). A tutorial on support vector machine–based methods for classification problems in chemometrics. **AnalyticaChimicaActa**. 665, pp. 129–145.

NQF, (2006). **National Qualifications Framework for Higher Education in Thailand.** Implementation Handbook, Thailand.

Rojas, R. (1996). **Neural Networks: A Systematic Introduction**. Berlin: Springer–Verlag.

Sang, C. S. (2012). **Practical Applications of Data Mining.** U.S.A. :Jones & Bartlett Publishers,

Shixin, Y. (2003). **Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data**. Ph.D. Dissertation, University of Antwerp. (Copy).

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

Reference

Asha, G. K., Manjunath, A. S. and Jayaram, M. A. (2010). Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection. **International Journal of Information Technology and Knowledge Management**, Vol 2(2), pp. 271–277.

Baha, S., Emine, U. and Dursun, D. (2012). Predicting and analyzing secondary education placement–test scores: A data mining approach. **Expert Systems with Applications**, Vol.39, pp.9468–9476.

Breiman, L. (1996). Bagging predictors.**Machine Learning**. Vol. 24(2),pp. 123–140.

Chapman, P. et al. (2000). CRISP–DM 1.0 – Step–by–step data mining guide. **Technical Report**, The CRISP–DM consortium.

Cristianini, N. and Shawe–Taylor, J. (2000).**An Introduction to Support Vector Machines and Other Kernel–Based Learning Methods**. Cambridge University Press, Cambridge.

Freund, Y. and Schapire, R.E. (1995). A decision–theoretic generalization of on–line learning and an application to boosting. **The 2nd European Conference on Computational Learning Theory**, London. pp. 23–37.

Hall, M. and Smith, L. (1998).Practical feature subset selection for machine learning.**In C. McDonald (Ed.), Computer Science '98 Proceedings of the 21st Australasian Computer Science Conference ACSC'98**, Perth, 4–6 February, pp. 181–191, Berlin: Springer.

Hall, M. and Holmes, G. (2003).Benchmarking Attribute Selection Techniques for Discrete Class Data Mining.**IEEE Transactions on Knowledge and Data Engineering**. Vol. 15, 1437–1447.

Han, J. and Kamber, M. (2001).Data Mining Concepts and Techniques. ***Morgan Kaufmann Publishers.***

Yingkuachat, J. (2006).**An Application of Bayesian Learning to Construction of Models for Prediction of Student's Graduation**. M.Sc. Thesis, Sripatum University. (Copy)

Liu, H., Setiono, R. (1996). A Probabilistic Approach to Feature Selection: a Filter Solution. In: 13th Int. Conference on Machine Learning (ICML), pp. 319–327.

Liu, et al., (1998).A monotonic measure for optimal feature selection.**Proceedings 10th European Conference on Machine Learning Chemnitz**, Germany, April 21–23, pp. 101–106.

Luts, J. et al. (2010) A tutorial on support vector machine–based methods for classification problems in chemometrics.**AnalyticaChimicaActa**, 665, pp. 129–145.

NQF, (2006).**National Qualifications Framework for Higher Education in Thailand.**Implementation Handbook, Thailand.

Office of the Higher Education Commission.(2009 a).**Thai Qualifications Framework for Higher Education (TQF: HEd)**. Last viewed at 8th March 2014. http://www.mua.go.th/users/tqf-hed/news/news6.php.

_____. (2009 b).**Qualification Standard for Bachelor's Degree in Computer Science B.E. 2552.** Last viewed at 8th March 2014, http://www.mua.go.th/users/tqf-hed/news/FilesNews/FilesNews6/computer_m1.pdf

Pongpatrarakan, P. (2010). **Factor Analysis Influencing Retirement in Undergraduate Degree Students By The Use of Committee Machine.** The 6 th National Conference on Computing and Information Technology. pp. 491 – 496.

Rojas, R. (1996). Neural Networks: A Systematic Introduction. **Springer-Verlag**, Berlin.

Sang, C. S. (2012).Practical Applications of Data Mining.**Jones& Bartlett Publishers**, U.S.A.

Shixin, Y. (2003). **Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data**. Ph.D. dissertation, University of Antwerp.

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