สำนักหอสมุด มหาวิทยาลัยบูรพา
ต.แสนสุข อ.เมือง จ.ชลบุรี 20131

การจำแนกกลุ่มนิสิตชั้นปี 1 เพื่อการให้ความช่วยเหลือด้านวิชาการ โดยการใช้แบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาที่มีการคัดเลือกปัจจัย

กมลวรรณ แสงระวี

23 ส.ค. 2559
365231 TH ∞ 24467

งานนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารเทศ
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
สิงหาคม 2558
ลิขสิทธิ์เป็นของมหาวิทยาลัยบูรพา

CLASSIFYING FIRST YEAR COLLEGE STUDENT FOR SUPPORTING IN ACADEMIC ASPECT BY USING PREDICTION MODEL FOR GRADUATION WITH FACTOR SELECTION

KAMONWAN SANGRAWEE

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT
FOR THE MASTER DEGREE OF SCEINCE IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATICS BURAPHA UNIVERSITY
AUGUST 2015.

COPYRIGHT 2015 BURAPHA UNIVERSITY

คณะกรรมการควบคุมงานนิพนธ์และคณะกรรมการสอบงานนิพนธ์ได้พิจารณางานนิพนธ์ของ กมลวรรณ แสงระวี ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมงานนิพนธ์

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

คณะกรรมการสอบงานนิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.อัณณ์นุพันธ์ รอดทุกข์)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)

คณะวิทยาการสารสนเทศ อนุมัติให้รับงานนิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพา

..... คณบดีคณะวิทยาการสารสนเทศ
(ผู้ช่วยศาสตราจารย์ ดร.สุวรรณา รัศมีขวัญ)
วันที่.....เดือน สิงหาคม พ.ศ. 2558

กิตติกรรมประกาศ

งานนิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณา ช่วยเหลือ แนะนำ และให้คำปรึกษาอย่างดียิ่งจากผู้ช่วยศาสตราจารย์ ดร. สุวรรณา รัศมีขวัญ อาจารย์ผู้ควบคุมงานนิพนธ์หลัก และผู้ช่วยศาสตราจารย์ ดร. กฤษณะ ชินสาร อาจารย์ผู้ควบคุมงานนิพนธ์รอง ที่ได้กรุณาถ่ายทอดความรู้ แนวคิด วิธีการ คำแนะนำ และตรวจสอบแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่ยิ่ง ผู้วิจัยกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณผู้ทรงคุณวุฒิ ผู้ช่วยศาสตราจารย์ ดร.อัณณ์นุพันธ์ รอดทุกข์ คณาจารย์ และผู้เชี่ยวชาญทุกท่าน ที่ให้คำแนะนำในความรู้ด้านวิชาการ แนวทางในการแก้ไขปัญหาต่าง ๆ ที่ทำให้งานนิพนธ์สำเร็จได้ด้วยดี

ขอขอบพระคณะวิทยาการสารสนเทศ และกองทะเบียนประมวลผลการศึกษามหาวิทยาลัยบูรพา ที่ให้ความอนุเคราะห์และความสะดวกในการเก็บข้อมูลเพื่อการวิจัยและให้ความร่วมมือในการเก็บข้อมูลเป็นอย่างดี

ขอขอบพระคุณ คุณพ่อจำเนียร แสงระวี คุณแม่วิไล แสงระวี และญาติพี่น้องทุกท่านที่คอยช่วยเหลือสนับสนุนทั้งด้านกำลังใจด้วยดีตลอดมา

ขอขอบพระคุณ นายจักรวาล คุณะดิลก คู่ชีวิตที่คอยช่วยเหลือสนับสนุนทั้งด้านกำลังกาย กำลังใจและกำลังทรัพย์ด้วยดีตลอดมา

ขอขอบคุณเพื่อนนิสิตสาขาเทคโนโลยีสารสนเทศทุกท่านที่ได้ให้คำแนะนำและส่งเสริมกำลังใจตลอดมา นอกจากนี้ยังมีผู้ที่ให้ความร่วมมือช่วยเหลืออีกหลายท่าน ซึ่งผู้วิจัยไม่สามารถกล่าวนามในที่นี้ได้หมด จึงขอขอบคุณทุกท่านเหล่านั้นไว้ ณ โอกาสนี้ด้วย

คุณค่าทั้งหลายที่ได้รับจากงานนิพนธ์ฉบับนี้ ผู้วิจัยขอมอบเป็นกตัญญูกตเวทีแด่บิดา มารดา และบูรพาจารย์ที่เคยอบรมสั่งสอน ตลอดจนผู้มีพระคุณทุกท่าน

กมลวรรณ แสงระวี

52920431: สาขาวิชา: เทคโนโลยีสารสนเทศ; วท.ม. (เทคโนโลยีสารสนเทศ)
คำสำคัญ: การคัดเลือกคุณลักษณะของข้อมูล / การจำแนกข้อมูล / โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ

กมลวรรณ แสงระวี: การจำแนกกลุ่มนิสิตชั้นปี 1 เพื่อการให้ความช่วยเหลือด้านวิชาการโดยการใช้แบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาที่มีการคัดเลือกปัจจัย (CLASSIFYING FIRST YEAR COLLEGE STUDENT FOR SUPPORTING IN ACADEMIC ASPECT BY USING PREDICTION MODEL FOR GRADUATION WITH FACTOR SELECTION) อาจารย์ผู้ควบคุมงานนิพนธ์: ผู้ช่วยศาสตราจารย์สุวรรณา รัศมีขวัญ, Ph.D. 83 หน้า. ปี พ.ศ. 2558.

งานนิพนธ์นี้นำเสนอวิธีการเลือกคุณลักษณะของข้อมูลนิสิตชั้นปีที่ 1 ที่มีผลต่อการสำเร็จการศึกษาของนิสิต เพื่อใช้ในการจำแนกกลุ่มนิสิตชั้นปีที่ 1 เพื่อการให้ความช่วยเหลือด้านวิชาการ กลุ่มตัวอย่างที่ใช้ในงานวิจัยนี้คือนิสิตระดับปริญญาตรีชั้นปีที่ 1 จำนวน 699 คน สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ที่ศึกษาระหว่างปีการศึกษา พ.ศ. 2550 - 2554 โดยมีปัจจัยตั้งต้นจำนวน 25 ปัจจัยถูกนำมาใช้ในการทำนายการสำเร็จการศึกษาประกอบด้วยประวัติส่วนตัวต่าง ๆ ของนิสิต ประวัติการศึกษาของการเรียนในชั้นปีที่ 1 งานนิพนธ์นี้ได้นำเสนอเทคนิคการคัดเลือกปัจจัยที่เหมาะสมจำนวน 3 เทคนิค ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation-based Features Selection) และ 3) เทคนิค Wrapper Approach มาใช้ในการคัดเลือกปัจจัยที่เหมาะสมของข้อมูลก่อนการนำไปทำการจำแนกประเภทด้วยแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back - Propagation Neural Network: BPNN) ผลจากการทดลองพบว่า การคัดเลือกคุณลักษณะข้อมูลด้วยเทคนิค CFS ร่วมกับ BPNN ให้ค่าความถูกต้องของการจำแนกข้อมูลเฉลี่ยสูงสุดให้ค่าความถูกต้องเฉลี่ย 88.06% และสามารถลดคุณลักษณะที่ต้องนำมาใช้ในการจำแนกจาก 25 เหลือ 7 คุณลักษณะ เมื่อพิจารณาผลการทดลองใช้เทคนิค CFS ร่วมกับ BPNN ในการทดสอบแยกชุดข้อมูลนิสิตตามรายปีการศึกษา 2553 และ 2554 พบว่าค่าความถูกต้องของการจำแนกกลุ่มนิสิตเท่ากับ 81.76% และ 80.45% ตามลำดับ ทำให้สรุปได้ว่าตัวแบบ (Model) สามารถคัดเลือกคุณลักษณะที่เหมาะสมถึงแม้ว่ามีความผันแปรของพฤติกรรมของข้อมูล ซึ่งทำให้ตัวแบบ (Model) นี้สามารถนำไปใช้เป็นแนวทางการคัดเลือกปัจจัยที่เหมาะสมสำหรับข้อมูลของนิสิตคณะต่าง ๆ ที่มีรูปแบบของพฤติกรรมข้อมูลที่แตกต่างจากชุดข้อมูลที่ใช้ในงานนิพนธ์นี้

52920431: MAJOR: INFORMATION TECHNOLOGY; M.Sc. (INFORMATION TECHNOLOGY)
KEYWORD: FEATURE SELETION/ CLASSIFICATION/ ARTIFICIAL NEURAL NETWORKS

KAMONWAN SANGRAWEE: CLASSIFYING FIRST YEAR COLLEGE STUDENT FOR SUPPORTING IN ACADEMIC ASPECT BY USING PREDICTION MODEL FOR GRADUATION WITH FACTOR SELECTION. ADVISOR: ASSISTANT PROFESSOR SUWANNA RASMEQUAN, Ph.D., 83 P. 2015.

This independent study presents a methodology for selecting attribute of the first year student data affecting their graduation to use for classifying first year students into proper group for academic assistant. The sample used in this research was 699 first year students in computer science program at Burapha University during the academic year 2007 to 2011. Twenty five Input factors, containing personal profiles and educational records of their first academic year, were trained for prediction of the student graduation. This research implemented three methods for feature selection to determine suitable factors of the input data. The methods included Consistency-based Feature Selection technique (CS), Correlation-based Features Selection technique (CFS), and Wrapper approach. The suitable factors for each method were applied to categorize the data by using Back-Propagation Neural Network (BPNN). The results showed that CFS was the best feature selection method since the average accuracy of data classification calculated from BPNN was the highest value at 88.06%. Feature selection by CFS with data classification by BPNN can reduce the attributes for prediction of the student graduation to 7 attributes from 25 input attributes. The CFS with BPNN method also performed well when categorizing the students separately in the year 2010 and 2011. The accuracy of data classification were equal to 81.76% and 80.45%, respectively. In conclusion, the CFS with BPNN method can determine the suitable factors and classify the students accurately under the variation of data character. Therefore, the method can be used in selecting attributes of the first year student data affecting their graduation in other academic programs with different data character from this research.

สารบัญ

หน้า

บทคัดย่อภาษาไทย ..... ข

บทคัดย่อภาษาอังกฤษ ..... ค

สารบัญ ..... ง

สารบัญตาราง ..... ฉ

สารบัญภาพ ..... ช

บทที่

1 บทนำ 1

ความเป็นมาและความสำคัญของปัญหา 1

วัตถุประสงค์ของการศึกษา 3

ขอบเขตของการศึกษา 3

แนวทางในการดำเนินงานนิพนธ์ 3

ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา 4

เครื่องมือที่ใช้ในการพัฒนาระบบงาน 5

นิยามศัพท์เฉพาะ 5

2 เอกสารและงานวิจัยที่เกี่ยวข้อง 6

การคัดเลือกคุณลักษณะ (Feature selection) 6

การจำแนกประเภท (Classification) 17

การวัดประสิทธิภาพ (Evaluation) 24

งานวิจัยที่เกี่ยวข้อง 26

3 วิธีการดำเนินการวิจัย 29

การรวบรวมข้อมูลที่เกี่ยวข้องและจัดเตรียมข้อมูลเพื่อการประมวลผล 29

การรวบรวมข้อมูลนิสิต 29

การตรวจสอบความถูกต้องของข้อมูล 30

การแปลงรูปแบบข้อมูล 30

การเตรียมชุดข้อมูลสำหรับใช้ในการวิเคราะห์ 37

การวิเคราะห์ค่าสถิติเพื่อเปรียบเทียบคุณลักษณะของข้อมูล 39

การวิเคราะห์และการสร้างแบบจำลองเพื่อการพยากรณ์โอกาสสำเร็จการศึกษา 51

สารบัญ (ต่อ)

บทที่		หน้า
	การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยไม่มีการคัดเลือกปัจจัย.....	51
	การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาแบบมีการคัดเลือกปัจจัย.....	52
4	ผลการวิเคราะห์ข้อมูล.....	55
	การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยไม่มีการคัดเลือกปัจจัย.....	57
	การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาแบบมีการคัดเลือกปัจจัย.....	
	การทดลองที่ 1: การทดลองเพื่อศึกษาว่าจำนวนคุณลักษณะที่เพิ่มขึ้นของปัจจัยนำเข้าของนิสิต	57
	การทดลองที่ 2: การทดลองเพื่อค้นหาชุดคุณลักษณะของปัจจัยนำเข้าที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา.....	71
5	สรุปผลการวิจัย.....	78
	สรุปผลการทดลอง.....	78
	ข้อดี.....	79
	ข้อจำกัด.....	79
	ข้อเสนอแนะ.....	79
บรรณานุกรม.....		81
ประวัติย่อของผู้วิจัย.....		83

สารบัญตาราง

ตารางที่ หน้า

2-1 ตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจในการออกไปเล่นกอล์ฟ..... 11

2-2 ความถี่ที่สังเกตได้ของลักษณะ Outlook 11

2-3 ความน่าจะเป็นและความน่าจะเป็นแบบมีเงื่อนไขของลักษณะ Outlook..... 12

2-4 ความถี่ที่สังเกตได้ของลักษณะ Outlook และ Temperature 13

2-5 ความน่าจะเป็นและความน่าจะเป็นแบบมีเงื่อนไขของลักษณะ Outlook และ Temperature..... 13

2-6 สรุปค่าความสัมพันธ์..... 14

2-7 การหากลุ่มแอตตริบิวต์ที่เหมาะสม..... 14

2-8 แสดงการคำนวณค่า Consistency ของคุณลักษณะ Outlook และ Humidity..... 16

2-9 แสดงการคำนวณค่า Consistency ของคุณลักษณะ Outlook, Humidityและ Windy..... 16

2-10แสดงการกำหนดค่าในการวัดประสิทธิภาพ..... 24

3-1 ปัจจัยที่ใช้ในการสร้างแบบจำลองการจำแนกประเภทนิสิต จำนวน 25 ปัจจัย..... 29

3-2 คำอธิบายตัวแปรชื่อตัวแปรและค่าตัวแปรตาม..... 32

3-3 คำอธิบายตัวแปร ชื่อตัวแปรและค่าตัวแปรอิสระ..... 32

3-4 จำนวนชุดข้อมูลนิสิตเป็นรายปีการศึกษา..... 37

3-5 ตารางแสดงปัจจัยร่วมของจำนวนปัจจัยที่ใช้ในการทดสอบ..... 38

4-1 ตารางแสดงค่าความถูกต้องของการจำแนกข้อมูลนิสิต เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษาที่แตกต่างกัน..... 57

4-2 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2550..... 58

4-3 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2551..... 60

4-4 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2552..... 61

4-5 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา2550–2552. 62

4-6 ตารางแสดงค่าความถูกต้องของการจำแนกข้อมูลนิสิตแต่ละปีการศึกษา เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา และการใช้เทคนิคการคัดเลือกปัจจัยที่แตกต่างกัน..... 64

4-7 สรุปปัจจัยที่เหมาะสมที่สุดสำหรับข้อมูลนิสิตในแต่ละปีการศึกษาและเทคนิคการคัดเลือกปัจจัยที่ให้ค่าการจำแนกประเภทนิสิตถูกต้องสูงสุด..... 70

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4-8 แสดงปัจจัยที่ถูกคัดเลือกจากปัจจัยตั้งต้น 25 ปัจจัยด้วยเทคนิค CS CFS และ Wrapperและค่าความถูกต้องของการจำแนกในข้อมูลนิสิตในแต่ละปีการศึกษา..... 72

4-9 แสดงการเปรียบเทียบค่าความถูกต้องจากคุณลักษณะที่ถูกคัดเลือกที่ได้ผลลัพธ์ที่ดีที่สุดกับการใช้คุณลักษณะทั้งหมด..... 74

4-10ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือก..... 76

4-11ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือกด้วยเทคนิค CS, CFS และ Wrapper ในข้อมูลนิสิตปีการศึกษา 2554..... 76

สารบัญภาพ

ภาพที่		หน้า
2-1	ขั้นตอนการคัดเลือกคุณลักษณะโดยทั่วไป	7
2-2	แสดงขั้นตอนของการค้นหาแบบดีที่สุดก่อน (Best - first search)	8
2-3	วิธีการแบบกรองและแบบแวพเปอร์สำหรับการเลือกลักษณะเด่น	17
2-4	กระบวนการในการจำแนกข้อมูลในขั้นตอนการเรียนรู้	18
2-5	กระบวนการในการจำแนกข้อมูลในขั้นตอนการทดสอบ	19
2-6	สถาปัตยกรรมของโครงข่ายประสาทเทียม	20
2-7	สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ	22
2-8	คำนิยามตัวแปร	23
2-9	การเลือกสุ่มข้อมูลแบบความเที่ยงตรง 10 กลุ่ม (10 - Fold Cross Validation)	23
3–1	แสดงจำนวนปัจจัยโดยมาตราวัดแบบมาตราแบบนามบัญญัติ, อันดับ และอันตรภาค	31
3-2	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของสถานะสำเร็จการศึกษา	37
3-3	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเพศ	39
3-4	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพของบิดา	39
3-5	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพของมารดา	40
3-6	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านกลุ่มโรงเรียน (รัฐ/เอกชน)	40
3-7	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านประเภทนิสิต (ปกติ/พิเศษ)	40
3-8	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพสมรสของบิดา มารดา	41
3-9	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านอาชีพของบิดา	41
3-10	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านอาชีพของมารดา	42
3-11	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านรายได้บิดา	42
3-12	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านรายได้มารดา	43
3-13	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านวิธีเข้าศึกษาของนิสิต	43
3-14	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านภูมิภาคที่อยู่อาศัย (ที่อยู่นิสิต)	44
3-15	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านจำนวนพี่น้องทั้งหมด (รวมตัวเอง)	44

สารบัญภาพ (ต่อ)

ภาพที่		หน้า
3-16	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเป็นบุตรคนที่.....	45
3-17	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านจำนวนพี่น้องที่กำลังศึกษาอยู่(รวมตัวเอง)	45
3-18	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านผลการเรียนเฉลี่ยสถานศึกษาเดิม.....	46
3-19	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดภาคเรียนที่ 1 ชั้นปีที่1.....	46
3-20	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดภาคเรียนที่ 2 ชั้นปีที่1.....	47
3-21	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่1.....	47
3-22	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่1.....	48
3-23	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาแคลคูลัส 1...	48
3-24	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาแคลคูลัส 2...	49
3-25	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาโปรแกรมมิ่ง 1	49
3-26	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาโปรแกรมมิ่ง 2	50
3-27	กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาภาษาอังกฤษ	50
3-28	ตัวแบบ (Model) การคัดเลือกปัจจัยที่เหมาะสมสำหรับการจำแนกประเภทนิสิต.....	51
4-1	ปัจจัยที่ใช้ในการทดลอง.....	57
4-2	ค่าความถูกต้องของการจำแนกข้อมูลนิสิตเมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตเพิ่มขึ้น.....	58
4-3	ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค CS	64

สารบัญภาพ (ต่อ)

ภาพที่ หน้า

4-4 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค CFS..... 64

4-5 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค Wrapper..... 64

4-6 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2550 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน..... 65

4-7 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2551เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน..... 66

4-8 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2552 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน..... 67

4-9 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2550 - 2552 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน..... 68

4-10 เปรียบเทียบค่าความถูกต้องการจำแนกประเภทนิสิตจากปัจจัยที่ดีที่สุดถูกคัดเลือกโดยเทคนิค CS,CFS และ Wrapper กับค่าความถูกต้องจากปัจจัยการใช้ทั้งหมด 25 ปัจจัย..... 75

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

การศึกษาเป็นรากฐานที่สำคัญในการพัฒนาทรัพยากรมนุษย์ จึงเป็นสิ่งที่ทุกประเทศควรให้ความสำคัญเป็นอันดับต้นๆ และประเทศไทยเองก็ให้ความสำคัญเป็นอย่างมาก มีการเร่งรัดพัฒนาการศึกษาอย่างต่อเนื่อง ดังจะเห็นได้จากการสนับสนุนทางการศึกษาจากรัฐบาลที่ให้ความสำคัญของการศึกษามากยิ่งขึ้น โดยการวางนโยบายกระตุ้นให้ประชาชนได้รับการศึกษาที่สูงขึ้น กล่าวคือ ให้ประชาชนต้องมีการศึกษาขั้นต่ำเก้าปี (ระดับมัธยมศึกษาปีที่ 3) และมีการกำหนดการศึกษาขั้นพื้นฐานสิบสองปี (ระดับมัธยมศึกษาปีที่ 6) ไว้ในรัฐธรรมนูญแห่งราชอาณาจักรไป พุทธศักราช 2550 ที่กำหนดให้รัฐจะต้องจัดการศึกษาให้กับประชาชนอย่างทั่วถึงเท่าเทียมเป็นธรรมเป็นเวลาไม่น้อยกว่า 12 ปี โดยไม่เก็บค่าใช้จ่าย ดังนั้นประชาชนจึงหันมาให้ความสำคัญกับการศึกษาเพิ่มขึ้นและส่งบุตรหลานเข้าเรียนในระดับที่สูงขึ้นอย่างต่อเนื่อง โดยเฉพาะอย่างยิ่งสถาบันการศึกษาในระดับอุดมศึกษาทั้งภาครัฐและภาคเอกชนที่มีการเปิดเพิ่มขึ้นกว่า 100 แห่งทั่วประเทศ เพื่อตอบรับต่อความต้องการของนิสิตที่ศึกษาต่อในระดับอุดมศึกษาที่เพิ่มสูงขึ้น

อย่างไรก็ตาม ปัจจุบันการศึกษาในระดับอุดมศึกษาต้องประสบปัญหาสำคัญอย่างหนึ่ง คือ มีจำนวนผู้สนใจเรียนในระดับอุดมศึกษาเพิ่มขึ้น แต่มีนิสิตบางส่วนที่ไม่สามารถสำเร็จการศึกษาตามระยะเวลาที่หลักสูตรกำหนด และบางส่วนต้องออกจากการศึกษากลางคันหรือไม่สามารถสำเร็จการศึกษา ซึ่งนิสิตทั้งสองกลุ่มนี้ทำให้เกิดผลเสียหายตามมา กล่าวคือ ส่วนผู้เรียนเสียเงินและเวลาที่ใช้ในการศึกษาไปโดยเปล่าประโยชน์ สถานศึกษาต้องเสียเวลาในการบริหารจัดการ เสียทรัพยากรในการลงทุนและเสียโอกาสในการสร้างบัณฑิต ดังนั้นจากปัญหาดังกล่าวซึ่งถือว่าเป็นปัญหาที่สำคัญมากที่สถานศึกษาต้องให้ความสำคัญในการป้องกันและหาทางแก้ไข หากสถานศึกษาสามารถทราบถึงปัจจัยหรือสาเหตุที่ส่งผลต่อโอกาสการสำเร็จการศึกษาตามระยะเวลาที่หลักสูตรกำหนดวางไว้สามารถจำแนกสถานภาพของนิสิตในแต่ละกลุ่มได้ และคาดการณ์โอกาสการสำเร็จการศึกษาของนิสิตที่เข้ามาศึกษาในสถาบันศึกษาว่ามีโอกาสสำเร็จการศึกษาตามระยะเวลากำหนดมากหรือน้อยเพียงใด ก็จะสามารถใช้เป็นแนวทางในการแก้ไขปัญหาให้กับนิสิตในแต่ละกลุ่มได้อย่างเหมาะสม ย่อมจะเป็นผลดีอย่างยิ่งในการจัดการเรียนการสอนของสถานศึกษานั้นๆ แต่ทั้งนี้ ข้อมูลนิสิตส่วนใหญ่ที่ถูกรวบรวมไว้ในฐานข้อมูลของสถาบันนั้น มีจำนวนชุดข้อมูลและจำนวนปัจจัยของข้อมูลอยู่เป็นจำนวนมาก การนำข้อมูลเหล่านั้นมาสร้างเป็นแบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาจึงทำให้ต้องใช้เวลามากในการประมวลผล และปัจจัยเหล่านั้นอาจจะมีบางปัจจัยที่ไม่เป็นประโยชน์ต่อ

การสร้างแบบจำลอง ดังนั้น ขั้นตอนการคัดเลือกปัจจัย จึงเป็นขั้นตอนที่สำคัญขั้นตอนหนึ่งในการเตรียมข้อมูลสำหรับการสร้างแบบจำลองโดยการคัดเลือกปัจจัย คือ การทำให้ขนาดมิติของข้อมูลเดิมลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด และได้ความถูกต้องของผลลัพธ์มากที่สุด ซึ่งเทคนิควิธีการคัดเลือกปัจจัยมีอยู่หลากหลายวิธีที่จะนำมาใช้ก่อนนำข้อมูลดังกล่าวไปประมวลผลต่อ โดยแต่ละเทคนิคของการคัดเลือกปัจจัยจะมีประสิทธิภาพที่แตกต่างกันขึ้นอยู่กับลักษณะของข้อมูล

ดังนั้นในงานนิพนธ์นี้จึงเสนอการวิเคราะห์เปรียบเทียบเทคนิคการคัดเลือกปัจจัยที่เหมาะสมโดยมีปัจจัยตั้งต้นจำนวน 25 ปัจจัย ซึ่งประกอบไปด้วยประวัติส่วนตัวของนิสิต ประวัติการศึกษา และสถานภาพของการสำเร็จการศึกษาของนิสิตในชั้นปีที่ 1 ระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพาที่ใช้เป็นกรณีศึกษา โดยในงานนิพนธ์นี้ผู้จัดทำได้นำเทคนิคการคัดเลือกปัจจัยแบบการกรอง (Filter Approach) ซึ่งเป็นเทคนิคการคัดเลือกปัจจัยโดยพิจารณาค่าความสัมพันธ์ของตัวแปรอิสระกับค่าตัวแปรเป้าหมาย ทำให้ใช้เวลาในการประมวลผลน้อยและลดโอกาสในการเกิดโอเวอร์ฟิตติง (Overfitting) เปรียบเทียบการคัดเลือกปัจจัยกับเทคนิคการคัดเลือกปัจจัยแบบการควบรวม (Wrapper Approach) ซึ่งมีวิธีการคัดเลือกปัจจัยโดยพิจารณาตัวแปรทุกตัวแปร จึงทำให้ใช้เวลาในการประมวลผลค่อนข้างสูงในกรณีที่มีจำนวนมิติข้อมูล (Attribute) จำนวนมาก และมีความเสี่ยงในการเกิดโอเวอร์ฟิตติง (Overfitting) สูงในกรณีที่มีจำนวนชุดข้อมูลน้อย (Observations) มาใช้ในการคัดเลือกปัจจัยที่เหมาะสมร่วมกับวิธีการจำแนกประเภทนิสิตแบบจำลองโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back - Propagation Neural Network: BPNN) เนื่องจากการเรียนรู้แบบแพร่ย้อนกลับนั้นจะมีความสามารถในการจดจำรูปแบบ และสามารถที่จะเรียนรู้ความสัมพันธ์ของรูปแบบได้มากมาย จึงเป็นที่นิยมใช้สำหรับการประยุกต์ด้านต่าง ๆ ที่หลากหลาย ในงานนิพนธ์นี้จึงใช้โครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับในการพย่ากรณ์โอกาสการสำเร็จศึกษา เพื่อสร้างแบบจำลองในการพยากรณ์โอกาสสำเร็จการศึกษาสำหรับการจำแนกประภทนิสิต เพื่อใช้ประกอบการจำแนกกลุ่มนิสิตเพื่อให้ความช่วยเหลือเฉพาะกลุ่มได้อย่างเหมาะสม ซึ่งสามารถใช้เป็นแนวทางประกอบการแก้ปัญหาออกกลางคันของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศในชั้นปีที่ 1 ซึ่งมีอัตรานิสิตออกกลางคันสูงถึงร้อยละ 30 ผลจากการดำเนินงานนิพนธ์นี้จะทำให้ได้เทคนิคการคัดเลือกปัจจัยที่ให้ประสิทธิภาพการจำแนกข้อมูลที่เหมาะสมที่สุด สามารถบอกถึงปัจจัยที่สัมพันธ์กับการสำเร็จการศึกษาของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ และปัจจัยที่ได้รับการคัดเลือกสามารถนำไปสร้างแบบจำลองการพยากรณ์การสำเร็จการศึกษาของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ สามารถนำผลพยากรณ์ที่ได้ไปประกอบการจำแนกกลุ่มนิสิตสำหรับการจัดการเรียนการสอนของแต่ละกลุ่มอย่างเหมาะสม และนำไปใช้เป็นแนวทางในการสร้างระบบและกลไกที่จะช่วยให้สามารถติดตามกำกับดูแล ให้คำปรึกษาแก่นิสิตในแต่ละกลุ่มได้อย่างเหมาะสม

วัตถุประสงค์ของการศึกษา

1. เพื่อหาแบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาที่มีการคัดเลือกปัจจัยสำหรับการจำแนกกลุ่มนิสิตชั้นปีที่ 1 เพื่อการให้ควาช่วยเหลือด้านวิชาการ

2. เพื่อสร้างตัวแบบ (Model) สำหรับการคัดเลือกปัจจัยที่เหมาะสมสำหรับการจำแนกประเภทนิสิตในกรณีที่คุณลักษณะของนิสิตแตกต่างไปจากชุดข้อมูลที่ใช้ในการทดลองนี้

ขอบเขตของการศึกษา

1. ชุดข้อมูลที่ใช้ในงานนิพนธ์นี้ ใช้ชุดข้อมูลนิสิตจากฐานข้อมูลของกองทะเบียนและประมวลผลนิสิต มหาวิทยาลัยบูรพา โดยประกอบไปด้วยข้อมูลประวัติส่วนตัวและประวัติการศึกษาของนิสิตระดับปริญญาตรี ชั้นปีที่ 1 และข้อมูลผลการสำเร็จการศึกษาของนิสิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ตั้งแต่ปีการศึกษา 2550 – 2554 จำนวน 699 ชุดข้อมูล

2. เทคนิคที่นำมาใช้ในการวิเคราะห์ปัจจัยที่มีผลต่อการสำเร็จการศึกษาในงานนิพนธ์นี้ ประกอบไปด้วยเทคนิค 2 ประเภท คือ เทคนิคการคัดเลือกปัจจัยแบบการกรอง (Filter Approach) ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และเทคนิคการคัดเลือกปัจจัยแบบการควบรวม (Wrapper Approach) ซึ่งเทคนิคเหล่านี้ถูกนำมาทดลองใช้ในการคัดเลือกปัจจัยที่เหมาะสม โดยใช้วิธีโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back-Propagation Neural Network: BPNN) เพื่อการจำแนกกลุ่มของนิสิตจากสถานภาพของนิสิตในปีสุดท้าย ซึ่งแบ่งออกเป็น 3 กลุ่มได้แก่ กลุ่มที่ 1 สำเร็จการศึกษาตามที่หลักสูตรกำหนด 4 ปี กลุ่มที่ 2 สำเร็จการศึกษาไม่เป็นไปตามที่หลักสูตรกำหนด และกลุ่มที่ 3 ไม่สำเร็จการศึกษา

3. การทดสอบความแม่นยำและการเปรียบเทียบผล ในงานนิพนธ์นี้ได้ใช้เครื่องมือการวัดประสิทธิภาพ ดังนี้ อัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

แนวทางในการดำเนินงานนิพนธ์

งานนิพนธ์นี้ใช้ข้อมูลประวัติส่วนตัวของนิสิต ประวัติการศึกษาของนิสิตในชั้นปีที่ 1 และสถานภาพของการสำเร็จการศึกษาในปีสุดท้ายตามหลักสูตรที่มหาวิทยาลัยได้กำหนดไว้ ของนิสิตระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ตั้งแต่ปีการศึกษา 2550 – 2554 มาทำการคัดเลือกปัจจัยที่เหมาะสมและสร้างตัวแบบการคัดเลือกปัจจัยที่เหมาะสมและนำไปสู่การจำแนกกลุ่มข้อมูลนิสิต ซึ่งมีแนวทางดังต่อไปนี้

1. การศึกษาและรวบรวมข้อมูล

1.1 ศึกษางานวิจัยและทฤษฎีที่เกี่ยวข้อง

1.2 ศึกษาเทคนิคที่เกี่ยวข้อง ได้แก่ เทคนิคการคัดเลือกปัจจัยโดยการใช้วิธีการ ดังนี้ เทคนิคการวิธีการคัดเลือกคุณลักษณะแบบการกรอง (Filter Approach) และแบบการควบรวม (Wrapper Approach) จากนั้นจึงทำการศึกษาตัวแบบโครงสร้างของโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back-Propagation Neural Network: BPNN) สำหรับใช้ในการจำแนกข้อมูล

1.3 ศึกษาข้อมูลทะเบียนประวัตินิสิต และพิจารณาลักษณะข้อมูลที่จะใช้ในการดำเนินงานวิจัย

1.4 ศึกษาเทคโนโลยีและเครื่องมือสนับสนุนในการวิเคราะห์ข้อมูล เช่น โปรแกรม WEKA (Waikato Environment for Knowledge Analysis)

2. การจัดเตรียมข้อมูล

ในงานวิจัยนี้ได้ทำการรวบรวมข้อมูลประวัติส่วนตัว และประวัติการศึกษา ของนิสิตระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา ตั้งแต่ปีการศึกษา 2550 – 2554 จำนวน 699 คน โดยมีปัจจัยตั้งต้นจำนวน 25 ปัจจัย จากนั้นทำการแปลงค่าข้อมูลบางส่วนให้เป็นตัวเลขเพื่อสะดวกต่อการวิเคราะห์ข้อมูล

3. การดำเนินการวิเคราะห์ข้อมูลเพื่อหาปัจจัยและแบบจำลองในการพยากรณ์โอกาสการสำเร็จการศึกษา

ทำการทดลองโดยใช้เทคนิคที่นำเสนอในข้อที่ 2 เพื่อหาปัจจัยและแบบจำลองในการพยากรณ์โอกาสการสำเร็จการศึกษา

4. การทดสอบ

การดำเนินการทดสอบในงานนิพนธ์นี้ได้แยกชุดข้อมูลเป็น 2 กลุ่ม คือ กลุ่มที่ 1 สำหรับใช้ในการทดลอง ได้แก่ ชุดข้อมูลของนิสิตในปีการศึกษา 2550 , 2551, 2552 และ2550 - 2552 และ กลุ่มที่ 2 คือข้อมูล Unseen Data ได้แก่ ชุดข้อมูลของนิสิตในปีการศึกษา 2553 และ 2554 โดยทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบด้วยวิธีการ 10 - Fold Cross Validation

ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา

1. ทำให้ทราบถึงปัจจัยที่มีผลต่อการสำเร็จการศึกษา ของนิสิตชั้นปีที่ 1 ระดับปริญญาตรี สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ

2. ได้แบบจำลองสำหรับการพยากรณ์โอกาสสำเร็จการศึกษาของนิสิต

3. ได้ตัวแบบ (Model) สำหรับใช้ในการคัดเลือกปัจจัยที่เหมาะสมสำหรับการจำแนกประเภทนิสิต

เครื่องมือที่ใช้ในการพัฒนาระบบงาน

ในการพัฒนาระบบสารสนสำหรับการพยากรณ์โอกาสสำเร็จการศึกษาของนิสิตที่เริ่มเข้าศึกษาในสถาบันการศึกษาในปีการศึกษาแรก (ชั้นปีที่ 1) ประกอบไปด้วยเครื่องคอมพิวเตอร์ อุปกรณ์ต่อพ่วงต่างๆ อุปกรณ์เครือข่าย โปรแกรมที่ใช้ในการพัฒนาและภาษาที่ใช้ในการพัฒนาระบบ แบ่งออกเป็นดังนี้

1. ด้านฮาร์ดแวร์ (Hardware) ที่ใช้ในการพัฒนาทั้งส่วนรับ และส่วนให้บริการมีดังนี้
    - หน่วยความจำสำรอง (Ram) อย่างน้อย 2GB
    - หน่วยความจำสำรอง (Hard disk) อย่างน้อย 3GB
2. ด้านโปรแกรมคอมพิวเตอร์ที่ใช้ในการพัฒนา
    - ระบบโปรแกรมปฏิบัติการ Windows 7 home basic
    - โปรแกรมที่ช่วยในการวิเคราะห์ WEKA version 3.6.1

นิยามศัพท์เฉพาะ

1. อุดมศึกษา หมายถึง การศึกษาที่สูงขึ้นจากระดับมัธยมศึกษาโดยคำว่าอุดมศึกษา มีรากศัพท์มาจากศัพท์ภาษาบาลีอุตม หมายถึง สูงสุด และศัพท์ภาษาสันสกฤต ศึกษา หมายถึง การเล่าเรียนดังนั้น คำว่า "อุดมศึกษา" จึงหมายถึง การเรียนขั้นสูงสุด
2. การศึกษาขั้นพื้นฐาน หมายถึง การศึกษาระดับชั้นมัธยมศึกษาปี ที่ 6
3. การศึกษาขั้นต่ำ หมายถึง การศึกษาระดับชั้นมัธยมศึกษาปี ที่ 3

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในงานนิพนธ์นี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้อง ได้แก่ 1) การคัดเลือกคุณลักษณะ 2) การจำแนกประเภท (Classification) 3) การวัดประสิทธิภาพ (Evaluation) 4) งานวิจัยที่เกี่ยวข้อง รวมทั้งอภิปรายงานวิจัยที่เกี่ยวข้อง ดังรายละเอียดดังนี้

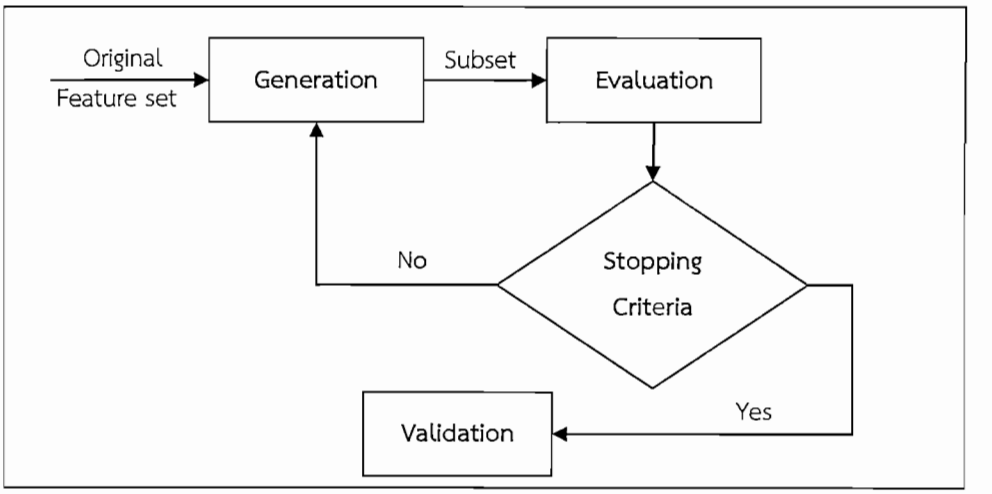
การคัดเลือกคุณลักษณะ

คุณลักษณะของข้อมูล คือ ลักษณะหรือคุณสมบัติที่ใช้ระบุองค์ประกอบหรือรายละเอียดของชุดข้อมูล คุณลักษณะของข้อมูลที่ดีต้องมีความถูกต้องและเชื่อถือได้ เก็บเฉพาะข้อมูลที่จำเป็นต้องใช้และครบถ้วนสมบูรณ์ อย่างไรก็ตามหากชุดข้อมูลมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรและเวลา การนำข้อมูลไปใช้งานอาจทำให้ประสิทธิภาพลดลงได้ (จิราภรณ์ ถมแก้ว, 2554)

การคัดเลือกคุณลักษณะ (Feature selection: FS) คือ การคัดเลือกคุณลักษณะของชุดข้อมูล (Feature Subset) ที่มีจำนวนมิติสูงให้มีมิติน้อยลงกว่าข้อมูลต้นฉบับ หรือเป็นการคัดเลือกคุณลักษณะ (ปัจจัย) สำคัญที่อยู่ในชุดข้อมูลนั้นออกมาจากคุณลักษณะทั้งหมดที่มีอยู่ โดยการคัดเลือกคุณลักษณะที่เหมาะสมนั้นจะช่วยลดคุณลักษณะที่ไม่เกี่ยวข้องกับการพยากรณ์รวมไปถึงการลดความซ้ำซ้อนของการเก็บข้อมูล (วีระยุทธ พิมพาภรณ์, 2557)

1. ขั้นตอนการเลือกคุณลักษณะเด่นที่เหมาะสม

กระบวนการของการคัดเลือกลักษณะเด่นประกอบด้วย 4 ขั้นตอนพื้นฐานแสดงดังในภาพที่ 2-1 โดยในขั้นตอนแรกจะเป็นการสร้างซับเซตของลักษณะเด่น (Subset generation) ที่ได้รับการคัดเลือกจากการประเมินบนหลักการของกลยุทธ์ในการค้นหา (Search strategy) โดยแต่ละซับเซตที่รับการคัดเลือกก็จะถูกประเมินผล (Evaluation) และเปรียบเทียบกับซับเซตที่ดีที่สุดก่อนหน้านี้ ถ้าซับเซตใหม่ดีกว่า ซับเซตใหม่จะแทนที่ซับเซตที่ดีที่สุดก่อนหน้านี้ ทั้งนี้กระบวนการสร้างซับเซตและการประเมินซับเซตจะทำซ้ำไปเรื่อย ๆ จนกว่ากฎเกณฑ์การหยุด (Stopping criterion) จะเป็นจริง แล้วจึงนำซับเซตที่ได้มาตรวจสอบโดยอาศัยขั้นตอนวิธีจำแนกประเภท (Classification) เพื่อเป็นการประเมินคุณภาพของซับเซตที่ถูกเลือกมา



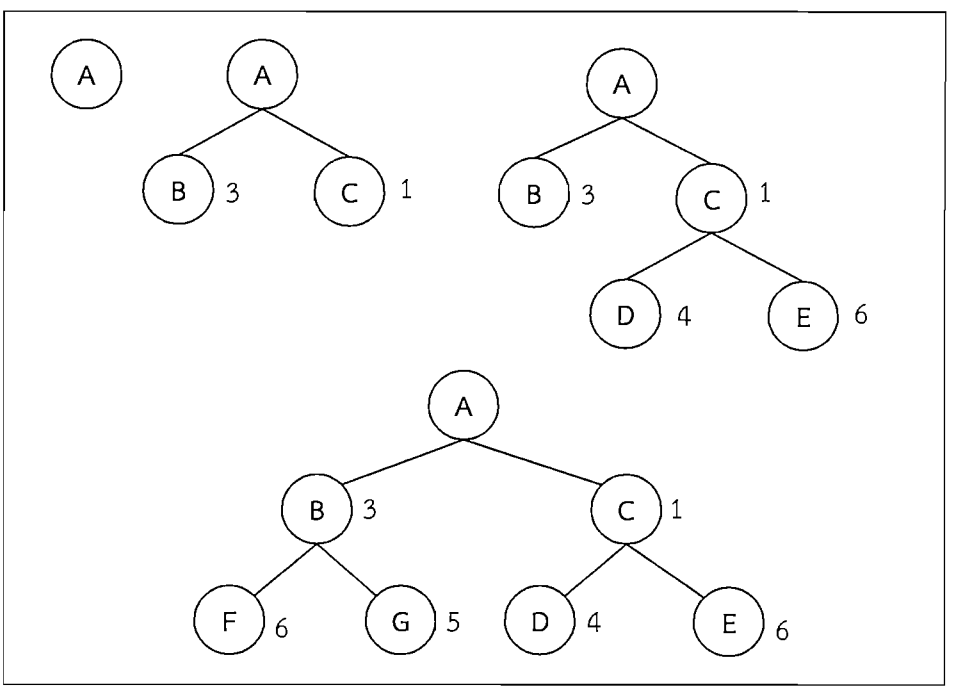
ภาพที่ 2-1 ขั้นตอนการคัดเลือกคุณลักษณะโดยทั่วไป (Novakovic, Strbac, & Bulatovic, 2011)

2. การค้นหาชุดของคุณลักษณะที่เหมาะสมโดยใช้อัลกอริทึมในการค้นหา (Search Algorithm)

อัลกอริทึมในการค้นหา (Search Algorithm) เป็นการค้นหาชุดคุณลักษณะของการค้นหาสำหรับเซตข้อมูลที่มีขนาด N ลักษณะเด่น โดยจะมีซับเซตที่ต้องถูกคัดเลือกเป็น 2^n ซับเซต ซึ่งทำให้สเปซของการค้นหาเป็นแบบเอ็กโปแน็นเชียล (Exponential) แน่นอนว่าย่อมเป็นไปไม่ได้สำหรับการค้นหาแบบละเอียดถี่ถ้วน (Exhaustive Search) ถึงแม้ว่า N จะมีขนาดปานกลางก็ตามทำให้การใช้เทคนิคการค้นหาแบบนี้มีข้อเสียในเรื่องการคำนวณทำให้เวลาที่ใช้การค้นหาชุดของคุณลักษณะจะใช้เวลาที่นานมากถ้าข้อมูลมีขนาดใหญ่ เพราะฉะนั้นวิธีนี้จึงเหมาะกับข้อมูลที่มีขนาดเล็ก

โดยในงานนิพนธ์นี้ได้ใช้วิธีการค้นหาชุดของคุณลักษณะแบบดีที่สุดก่อน (Best - first search) เป็นกระบวนการค้นหาข้อมูลที่ได้นำเอาข้อดีของทั้งการค้นหาแบบลึกก่อน (Depth first search) และการค้นหาแบบกว้างก่อน (Breadth first search) มารวมกันเป็นวิธีการเดียวโดยที่แต่ละขั้นของการค้นหาในโหนดลูกนั้นการค้นหาแบบดีที่ดีก่อนจะเลือกเอาโหนดที่ดีที่สุด (Most promising) และการที่จะทราบว่ าโหนดใดดีที่สุดนี้สามารถทำได้โดยอาศัยฮิวริสติกฟังกชัน ซึ่งฮิวริสติกฟังก์ชันนี้จะทำหน้าที่เหมือนตัววัดผล และให้ผลของการวัดนี้ออกมาเป็นคะแนน ภาพที่ 2-2 เป็นตัวอย่างของการค้นหาแบบดีที่สุดก่อน ขั้นตอนนี้เริ่มจากตอน 1 สร้างโหนดราก (Root node) ในขั้นตอน 2 สร้างโหนดลูก B และ C แล้วตรวจสอบโหนด B และ C ด้วยฮิวริสติกฟังก์ชันได้ผลออกมาเป็นคะแนน คือ 3 และ 1 ตามลำดับจากนั้นให้เลือกโหนด C เป็นโหนดต่อไปที่เราสนใจเพราะมีค่าน้อยกว่า (หมายเหตุ ในการเลือกนี้จะเลือกค่ามากสุดหรือน้อยสุดก็ได้ขึ้นอยู่ กับลักษณะของปัญหา)

แล้วสร้างโหนดลูกให้กับโหนด C ในขั้นตอน 3 ได้โหนด D และ E แล้วตรวจสอบคะแนนได้ 4 และ 6 ตามลำดับ จากนั้นทำการเปรียบเทียบค่าของโหนดท้ายสุดหรือเทอร์มินอลโหนด (Terminal node) ทุกโหนดว่าโหนดใดมีค่าดีที่สุดในที่นี่จะต้องเลือกโหนด B เพราะมีคะแนนเพียง 3 (เลือกคะแนนต่ำสุด) แล้วสร้างโหนดลูกตามขั้นตอน 4 ได้ F และ G แล้วตรวจสอบคะแนนได้ 6 และ 5 คะแนน ตามลำดับทำเช่นนี้เรื่อยๆ จนพบคำตอบหรือจนไม้สามารถสร้างโหนดต่อไปได้อีก (บุญเจริญ ศิริเนาวกุล, 2550)



ภาพที่ 2-2 แสดงขั้นตอนของการค้นหาแบบดีที่สุดก่อน (Best - first search)

3. การประเมินค่าของกลุ่มคุณลักษณะที่ผ่านการค้นหาได้มาโดยใช้อัลกอริทึมในการประเมินค่า (Evaluator Algorithm)

ขั้นตอนการประเมินค่าของกลุ่มคุณลักษณะที่ถูกคัดเลือก (Evaluator Algorithm) เป็นวิธีการประเมินค่ากลุ่มคุณลักษณะที่ถูกคัดเลือกมาจาก Search Method ซึ่งเป็นการประเมินค่าชุดของคุณะลักษณะที่ได้ค้นหามาว่าชุดคุณะลักษณะใดเป็นชุดของคุณะลักษณะที่เหมาะสมจะถูกนำไปประมวลผลในการจำแนกประเภทข้อมูลต่อไป แล้วให้ประสิทธิภาพที่ดีที่สุด โดยฟังก์ชันการประเมินค่าจึงถูกนำมาใช้วัดความสามารถของคุณลักษณะที่ผ่านการค้นหาเป็นขั้นตอนที่ทำต่อจากอัลกอริทึมในการค้นหาชุดของคุณลักษณะที่เหมาะสมเพื่อประเมินค่าชุดคุณะลักษณะที่ได้มานั้นเหมาะสม

หรือไม่ หากพบว่าชุดคุณลักษณะที่เหมาะสมแล้วจึงหยุดการคัดเลือกคุณลักษณะ โดยที่อัลกอริทึมที่ใช้ในการประเมินค่าชุดคุณลักษณะแบ่งได้เป็น 2 ประเภทคือ

3.1 การคัดเลือกคุณลักษณะแบบการกรอง (Filter Approach) เป็นวิธีการคัดเลือกคุณลักษณะที่ง่ายต่อการตีความ โดยจะกำจัดคุณลักษณะที่ไม่เกี่ยวข้อง (Irrelevant feature) ต่อการจำแนกประเภทซึ่งวัดเป็นคะแนน และทำการเลียงลำดับคุณลักษณะตามคะแนนที่ได้ โดยส่วนใหญ่อาศัยคุณสมบัติความสัมพันธ์ระหว่างข้อมูลกับกลุ่มที่กำหนดไว้ล่วงหน้าคุณสมบัตินี้ คือ ฟังก์ชันการประเมินค่า สำหรับวิธีการกรองนั้นสามารถใช้ฟังก์ชันในการประเมินค่าด้วยมาตรวัดระยะทาง มาตรวัดสารสนเทศ มาตรวัดความไม่เป็นอิสระ หรือมาตรวัดความคงเส้นคงวา วิธีการคัดเลือกคุณลักษณะแบบคัดกรองที่นิยมใช้ส่วนใหญ่ ได้แก่ Correlation-based Feature Selection (CFS) Consistency-based Subset Evaluation (CS) และ Chi - Square ดังนั้นในงานนิพนธ์นี้ได้นำวิธีการทั้ง 2 ดังกล่าว มาใช้เป็นวิธีการคัดเลือกคุณลักษณะ ดังมีรายละเอียดต่อไปนี้

1. วิธีการ Correlation - based Feature Selection (CFS)

วิธีการ Correlation - based Feature Selection (CFS) (Mark A Hall ,2000) เป็นวิธีการที่ประเมินผลซับเซตของแอตทิบวต์ด้วยวิธีการแบบฮิวริสติก (Heuristic) เพื่อใช้ในการพิจารณาความดีของแต่ละคุณลักษณะในการทำนายคลาสด้วยระดับความสัมพันธ์กันภายในระหว่างคุณลักษระเด่นกับคลาส สมการ 2.1 จะกำหนดแต้ม (Score) ที่สูงให้กับซับเซตที่บรรจุด้วยแอตทิบิวต์ที่มีความสัมพันธ์ที่สูงกับคลาสและมีค่าความสัมพันธ์ที่ต่ำกับแอตทิบิวต์อื่น ๆ

$$Merit_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k+1)\,\overline{r_{ff}}}} \qquad (2.1)$$

จากสมการ 2.1 ตัวแปร $Merit_s$ คือ ตัวบ่งความดีแบบฮิวริสติกของซับเซต S ที่บรรจุด้วย k แอตทิบิวต์ ในขณะที่ $\overline{r_{cf}}$ เป็นค่าเฉลี่ยของความสัมพันธ์กันระหว่างแอตทิบิวต์กับคลาส และ $\overline{r_{ff}}$ คือ ค่าเฉลี่ยของความสัมพันธ์กันระหว่างแอตทิบิวต์ โดยจากสมการ 2.1 จะเห็นได้ว่าตัวตั้งในสมการเป็นการแสดงถึงกลุ่มของแอตทิบิวต์ที่มีผลต่อการทำนายคลาส ในขณะที่ตัวหารในสมการแสดงถึงการซ้ำซ้อนกันของลักษณะเด่น โดยฮิวริสติกที่นำเสนอนี่จะมองว่าแอตทิบิวต์ที่ไม่มีความสัมพันธ์กับคลาส (Irrelevant feature) จะให้ผลการทำนายคลาสที่ไม่ดี ขณะเดียวกันลักษณะเด่นที่ซ้ำซ้อนกันจะถูกคัดแยกออกไป ซึ่งจะมีค่าความสัมพันธ์กันที่ค่อนข้างสูงกับลักษณะเด่นตัวอื่น ๆ เนื่องจากวิธีการ CFS มองว่าลักษณะเด่นต่าง ๆ จะมีความอิสระกัน ดังนั้น CFS จึงขาดประสิทธิภาพในการหาซับซ้อนของลักษณะเด่นที่มีการกระทำกันภายในสูง ๆ ได้ อย่างเช่นปัญหา Parity (ความ

คล้ายคลึงกัน) แต่ถ้าเป็นซับเซตที่มีการกระทำกันภายในที่ไม่มากแล้ว วิธีการ CFS ก็สามารถที่จะหาคำตอบได้

โดยที่

$\overline{r_{cf}}$ คือ ค่าเฉลี่ยของความสัมพันธ์กันระหว่างแอตทิบิวต์กับคลาส

$\overline{r_{ff}}$ คือ ค่าเฉลี่ยของความสัมพันธ์กันระหว่างแอตทิบิวต์

$$r_{cf} = \frac{Gini' \times \sum_{x \in X} p(x)^2}{(1 - \sum_{c \in C} p(c)^2) \sum_{c \in C}' p(c)^2} \qquad (2.2)$$

เมื่อ c คือ สมาชิกของคลาส C

x คือ สมาชิกของแอตทิบิวต์ f

$$Gini' = \left[\sum_{c \in C} p(c)(1 - p(c)) \right] - \sum_{x \in X} \left(\frac{p(x)^2}{\sum_{x \in X} p(x)^2} \sum_{c \in C} p\langle c|x \rangle (1 - p\langle c|x \rangle) \right) \quad (2.3)$$

$p(*)$ คือ ความน่าจะเป็นที่เหตุการณ์ * จะเกิดขึ้น

$p(c \mid x)$ คือ ความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ c เมื่อเหตุการณ์ x ได้เกิดขึ้นแล้ว

จากตารางที่ 2-1 เป็นตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจในการออกไปเล่นกอล์ฟโดยพิจารณาจากสภาพอากาศต่าง ๆ ซึ่งมีลักษณะ 4 ลักษณะประกอบการตัดสินใจ ได้แก่ สภาพอากาศ (Outlook), อุณหภูมิ (Temperature), ความชื้น (Humidity) และ ความแรงของลม (Windy) และคลาสที่เป็นคำตอบในการตัดสินมีค่าที่เป็นไปได้คือ Yes หรือ No

ตารางที่ 2-1 ตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจในการออกไปเล่นกอล์ฟ

NO.	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rain	mild	high	true	No

การคำนวณค่าความสัมพันธ์กันระหว่างแอตทิบิวต์กับคลาส (r_{cf}) ได้ดังนี้

1. สรุปข้อมูลจำนวนความถี่และความน่าจะเป็นที่เกี่ยวข้องของ Outlook กับคลาสได้ผลตามตารางที่ 2-2 และ 2-3

ตารางที่ 2-2 ความถี่ที่สังเกตได้ของลักษณะ Outlook

	Yes (y)	No (n)	รวม
Sunny (s)	2	3	5
Overcast (o)	4	0	4
Rain (r)	3	2	5
รวม	9	5	14

ตารางที่ 2-3 ความน่าจะเป็นและความน่าจะเป็นแบบมีเงื่อนไขของลักษณะ Outlook

	Yes (y)	No (n)	p( )
Sunny (s)	p(y|s) =2/5	p(n|s) =3/5	p(s) = 5/14
Overcast (o)	p(y|o) =1.00	p(n|o) =0.00	p(o) = 4/14
Rain (r)	p(y|r) =3/5	p(n|r) =2/5	p(r) = 5/14
p( )	p(y) = 9/14	p(n) = 5/14	

2 คำนวณ Ginni สำหรับค่าความสัมพันธ์กันระหว่างแอตทิบิวต์กับคลาส จากสมการที่ 2.3

$$Gini' = [p(y)(1-p(y)+p(n)(1-p(n)] - \left[\begin{array}{l} \left[\frac{p(s)^2}{p(s)^2+p(o)^2+p(r)^2}\right]\left[p\langle y|s\rangle(1-p\langle y|s\rangle)+p\langle n|s\rangle(1-p\langle n|s\rangle)\right]+ \\ \left[\frac{p(o)^2}{p(s)^2+p(o)^2+p(r)^2}\right]\left[p\langle y|o\rangle(1-p\langle y|o\rangle)+p\langle n|o\rangle(1-p\langle n|o\rangle)\right]+ \\ \left[\frac{p(r)^2}{p(s)^2+p(o)^2+p(r)^2}\right]\left[p\langle y|r\rangle(1-p\langle y|r\rangle)+p\langle n|r\rangle(1-p\langle n|r\rangle)\right]+ \end{array}\right]$$

= 0.0955

3 คำนวณค่าความสัมพันธ์กันระหว่างแอตทิบิวต์กับคลาส จากสมการที่ 2.2

$$r_{cf} = \frac{0.0955\times[p(s)^2+p(o)^2+p(r)^2]}{[1-(p(y)^2+p(n)^2)](p(y)^2+p(n)^2)}$$

= 0.130

การคำนวณค่าความสัมพันธ์กันระหว่างแอตทิบิวต์ (r_{ff}) มีวิธีการคำนวนเหมือนค่าความสัมพันธ์ระหว่างแอตทิบิวต์กับคลาส รายละเอียดที่แตกต่างคือการนำแอตทิบิวต์ตัวหนึ่งทำหน้าที่เปรียบเสมือนคลาสในสมการที่ *2.2* และ *2.3* ตัวอย่างการคำนวณแสดงได้จากความสัมพันธ์ของ Outlook กับ Temperature ได้ดังนี้

1.สรุปข้อมูลจำนวนความถี่และความน่าจะเป็นที่เกี่ยวข้องของ Outlook กับ Temperature ได้ผลตามตารางที่ 2-4 และ 2-5

ตารางที่ 2-4 ความถี่ที่สังเกตได้ของลักษณะ Outlook และ Temperature

	Hot (h)	Mild (m)	Cool (c)	รวม
Sunny (s)	2	2	1	5
Overcast (o)	2	1	1	4
Rain (r)	0	3	2	5
รวม	4	6	4	14

ตารางที่ 2-5 ความน่าจะเป็นและความน่าจะเป็นแบบมีเงื่อนไขของลักษณะ Outlook และ Temperature

	Hot (h)	Mild (m)	Cool (c)	p( )
Sunny (s)	p(h|s) =2/5	p(m|s) =2/5	p(c|s) =1/5	p(s) = 5/14
Overcast (o)	p(h|o) =2/4	p(m|o) =1/4	p(c|o) =1/4	p(o) = 4/14
Rain (r)	p(h|r) =0.00	p(m|r) =3/5	p(c|r) =2/5	p(r) = 5/14
p( )	p(h) = 4/14	p(m) = 6/14	p(c) = 4/14	

2 คำนวณ Ginni สำหรับค่าความสัมพันธ์กันระหว่างแอตทิบิวต์ จากสมการที่ 2.3

$$Gini' = [p(h)(1-p(h)+p(m)(1-p(m)+p(c)(1-p(c)] - \left[\left[\frac{p(s)^2}{p(s)^2+p(o)^2+p(r)^2}\right][p\langle h|s\rangle(1-p\langle h|s\rangle)+p\langle m|s\rangle(1-p\langle m|s\rangle)+p\langle c|s\rangle(1-p\langle c|s\rangle)]+ \left[\frac{p(o)^2}{p(s)^2+p(o)^2+p(r)^2}\right][p\langle h|o\rangle(1-p\langle h|o\rangle)+p\langle m|o\rangle(1-p\langle m|o\rangle)+p\langle c|o\rangle(1-p\langle c|o\rangle)]+ \left[\frac{p(r)^2}{p(s)^2+p(o)^2+p(r)^2}\right][p\langle h|r\rangle(1-p\langle h|r\rangle)+p\langle m|r\rangle(1-p\langle m|r\rangle)+p\langle c|r\rangle(1-p\langle c|r\rangle)]\right]$$

= 0.0773

3 คำนวณค่าความสัมพันธ์กันระหว่างแอตทิบิวต์ จากสมการที่ 2.2

$$r_{ff} = \frac{0.0773 \times [p(s)^2 + p(o)^2 + p(r)^2]}{[1-(p(h)^2 + p(o)^2 + p(c)^2)](p(h)^2 + p(o)^2 + p(c)^2)}$$

$$= 0.116$$

ตารางที่ 2-6 สรุปค่าความสัมพันธ์ (Mark A. Halll , 1999)

	Outlook	Temperature	Humidity	Windy	Class
Outlook	1.000	0.116	0.022	0.007	0.130
Temperature		1.000	0.248	0.028	0.025
Humidity			1.000	0.000	0.185
Windy				1.000	0.081

ตารางที่ 2-7 การหากลุ่มแอตตริบิวต์ที่เหมาะสม

รอบที่	Feature set	k	$\overline{r_{cf}}$	$\overline{r_{ff}}$	$Merit_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k+k(k+1)\,\overline{r_{ff}}}}$
0	[ ]	0	N/A	N/A	0.0
1	[Outlook]	1	0.130	1.000	0.130
	[Temperature]	1	0.025	1.000	0.025
	[Humidity]	1	0.185	1.000	**0.185**
	[Windy]	1	0.081	1.000	0.081
2	**[Outlook, Humidity]**	2	0.158	0.022	**0.220**
	[Temperature, Humidity]	2	0.105	0.258	0.133
	[Humidity, Windy]	2	0.133	0.0	0.188
3	[Outlook,Temperature, Humidity]	3	0.113	0.132	0.175
	[Outlook, Humidity, Windy]	3	0.132	0.0096	**0.226**
4	[Outlook,Temperature, Humidity, Windy]	4	0.105	0.0718	0.191

จากตารางที่ 2-7 แสดงการหา $Merit_s$ โดยใช้ค่าความสัมพันธ์ดังแสดงในตารางตารางที่ 2-6

2. วิธีการ Consistency - based Subset Evaluation (CS)

วิธีการ Consistency - based Subset Evaluation (CS) เป็นวิธีการเลือกคุณลักษณะเด่นบนหลักการของความสอดคล้องกันของคลาสข้อมูล (Consistency) วิธีการนี้จะพิจารณาจากการรวมกันของ Attribute โดยที่ค่าของAttribute เหล่านั้นจะทำการแบ่งเซตของข้อมูลออกเป็นซับเซตย่อย ๆ ที่ประกอบด้วยหนึ่งคลาสที่เป็นส่วนใหญ่ (Majority class) ตามปกติแล้วในการค้นหานั้น ซับเซตที่มีขนาดเล็ก ๆ จะให้ความสอดคล้องกันของคลาส (Consistency) ที่สูง โดยอัลริทึมของ CS ที่กล่าวนี้จะใช้การวัดความสอดคล้องกันของคลาส (Consistency merit) ดังสมการ 2.2 ซึ่งสามารถคำนวณได้ดังนี้

$$Consistency_s = 1 - \frac{\sum_{i=1}^{j} |D_i| - |M_i|}{N} \qquad (2.4)$$

โดยที่ s เป็นซับเซตของแอตทิบิวต์และ J คือจำนวนของค่าที่เป็นไปได้ทั้งหมดจากการรวมกันของแอตทิบิวต์ใน s ขณะที่ $|D_i|$ คือ จำนวนของแถวข้อมูลทั้งหมดที่เกิดขึ้นจากการรวมกันของแอตทิบิวต์ตัวที่ i และในขณะที่ $|M_i|$ คือ ขนาดของคลาสที่เป็นคลาสส่วนใหญ่จากการรวมกันแอตทิบิวต์ตัวที่ i และ N คือ จำนวนแถวของข้อมูลในเซตข้อมูล

จากสมการ 2.4 นั้นเทอมที่ลบออกจาก 1 เราจะเรียกว่าอัตราความไม่สอดคล้องกัน (Inconsistency rate) การคำนวณค่า Consistency สามารถพิจารณาได้จากตัวอย่างชุดข้อมูลฝึกที่ใช้ประกอบการตัดสินใจในการออกไปเล่นกอล์ฟโดยพิจารณาจากสภาพอากาศต่าง ๆ ในตารางที่ 2-1 โดยจะแสดงการคำนวณค่า Consistency ของคุณลักษณะ Outlook และHumidity ได้ผลตามตารางที่ 2-8 และคุณลักษณะ Outlook, Humidity และ Windy ได้ผลตามตารางที่ 2-9

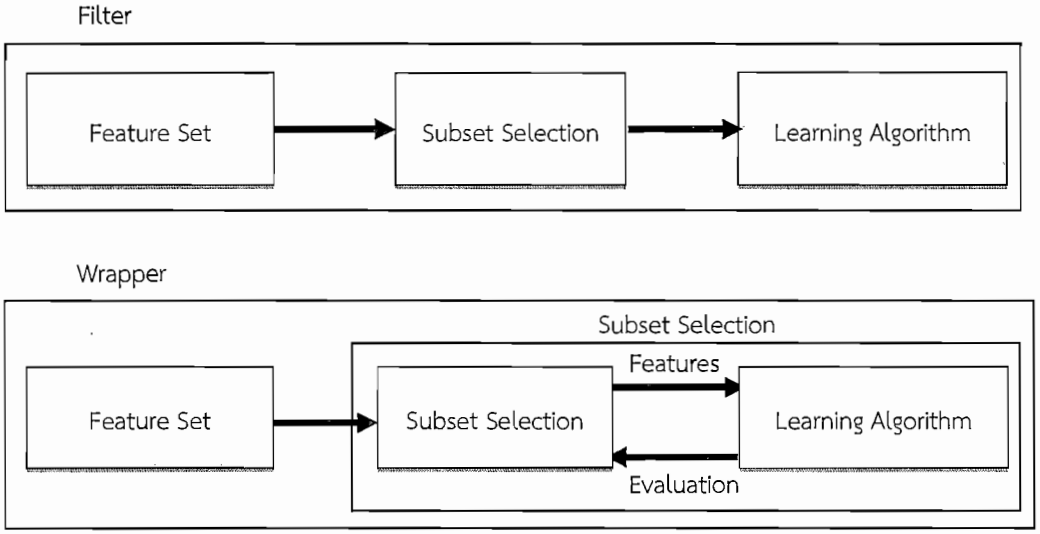
ตารางที่ 2-8 แสดงการคำนวณค่า Consistency ของคุณลักษณะ Outlook และ Humidity

i	outlook	humidity	Class	$\|D_i\|$	$\|M_i\|$	$\|D_i\|-\|Mi\|$
1	sunny	high	No, No, No	3	3	0
2	overcast	high	Yes, Yes	2	2	0
3	rain	high	Yes, No	2	1	1
4	rain	normal	Yes, No, Yes	3	2	1
5	overcast	normal	Yes, Yes	2	2	0
6	sunny	normal	Yes, Yes	2	2	0
					รวม	2
					Consistency = 1-(2/14) = 0.857	

ตารางที่ 2-9 แสดงการคำนวณค่า Consistency ของคุณลักษณะ Outlook, Humidity และ Windy

i	outlook	humidity	Windy	Class	$\|D_i\|$	$\|M_i\|$	$\|D_i\|-\|Mi\|$
1	sunny	high	FALSE	No, No	2	2	0
2	sunny	high	TRUE	No	1	1	0
3	overcast	high	FALSE	Yes	1	1	0
4	rain	high	FALSE	Yes	1	1	0
5	rain	high	FALSE	Yes, Yes	2	2	0
6	rain	normal	TRUE	No	1	1	0
7	overcast	normal	TRUE	Yes	1	1	0
8	sunny	normal	FALSE	Yes	1	1	0
9	normal	Yes	TRUE	Yes	1	1	0
10	normal	Yes	TRUE	Yes	1	1	0
11	normal	Yes	FALSE	Yes	1	1	0
12	high	Yes	TRUE	No	1	1	0
						รวม	0
						Consistency = 1-(0/14) = 1.00	

3.2 การคัดเลือกคุณลักษณะแบบการควบรวม (Wrapper Approach) การคัดเลือกคุณลักษณะโดยวิธีการแบบควบรวมต้องมีการนำอัลริทึมในการทำเมืองข้อมูลมาทำงานร่วมกับขั้นตอนการประเมินค่าซับเซต ซึ่งจะทำให้ได้ซับเซตที่มีความแม่นยำในการจำแนกประเภทมากกว่าแบบการกรองที่ใช้เงื่อนไขทางสถิติหรือเงื่อนไขทางทฤษฎีสารสนเทศ (Information Theory) ในการประเมินค่า ซึ่งการคัดเลือกแบบการควบรวมนี้ทำให้การคัดเลือกคุณลักษณะใช้เวลามากกว่าการคัดเลือกแบบการกรอง เนื่องจากมีการเรียนรู้หลายครั้งเท่ากับจำนวนครั้งที่สร้างซับเซตจึงทำให้ใช้เวลามาก และเป็นวิธีที่มีความซับซ้อนในการคำนวณสูงมาก

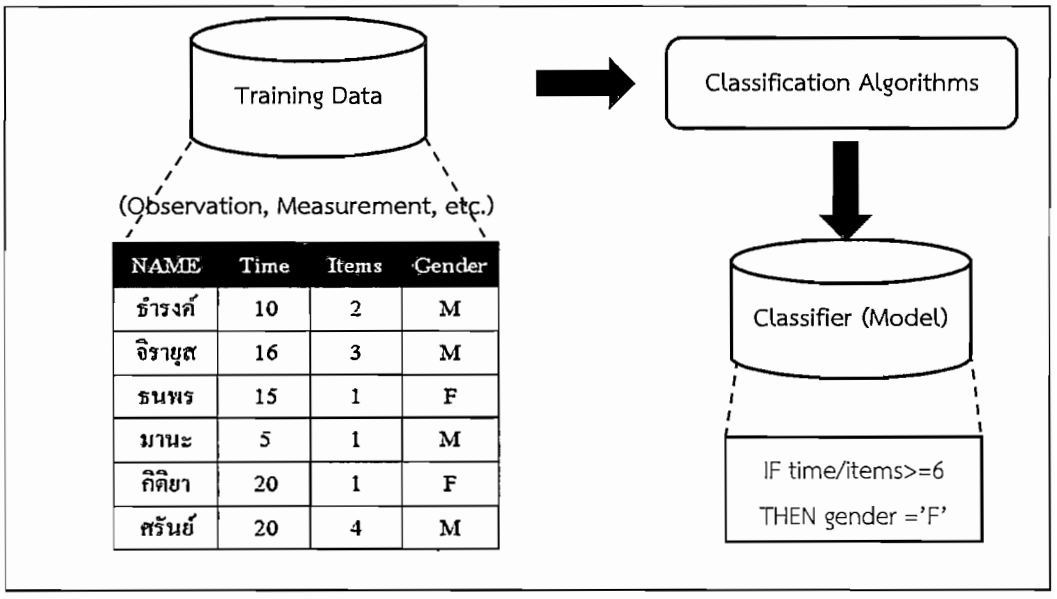


ภาพที่ 2-3 วิธีการแบบกรองและแบบแวพเปอร์สำหรับการเลือกลักษณะเด่น (สมบัติ ฝอยทอง, 2555)

การจำแนกประเภท (Classification)

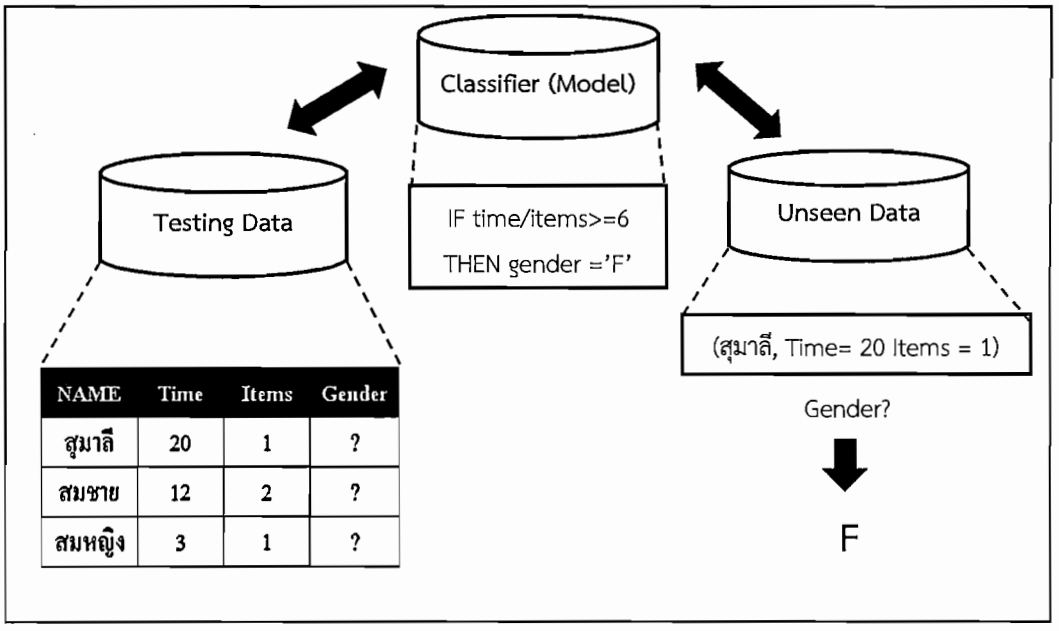
การจำแนกประเภท (Classification) เป็นการวิเคราะห์ข้อมูลสำหรับลักษณะของงานที่กำหนดกลุ่มให้กับหน่วยตัวอย่าง โดยอาศัยคุณลักษณะของหน่วยตัวอย่างซึ่งในที่นี้หมายถึงตัวแปรทำนาย เพื่อที่จะสามารถจำแนกกลุ่มได้ต้องทราบล่วงหน้าว่ามีกลุ่มอะไรบ้าง (Predefined Categories) และมีข้อมูลของหน่วยตัวอย่างในแต่ละกลุ่มก่อน การจำแนกประเภทข้อมูลนี้ได้นำไปประยุกต์ใช้ในหลายด้าน เช่น การจัดกลุ่มลูกค้าทางการตลาดการตรวจสอบความผิดปกติ การคัดแยกอีเมลขยะจากอีเมลปกติโดยพิจารณาจากหัวข้อเรื่องหรือเนื้อหาของอีเมล การวิเคราะห์ทางการแพทย์ และ การทำนายการประสบความสำเร็จในการศึกษาระดับปริญญาตรีโดยอาศัยผลการทดสอบต่าง ๆ และผลการศึกษาในระดับชั้นมัธยมปลาย เป็นต้น

ในการดำเนินการจำแนกประเภท เทคนิคการจำแนกประเภทจะสร้างตัวแบบการจำแนกประเภท (Classification Model) อย่างเป็นระบบจากข้อมูลนำเข้า (Input Data) ตัวอย่างเทคนิคการจำแนกประเภท ได้แก่ การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back - Propagation Neural Network) ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) และการจำแนกข้อมูลด้วยการสร้างต้นไม้ตัดสินใจ (Decision Tree Classifier) เป็นต้น แต่ละเทคนิคใช้อัลกอลิทึมการเรียนรู้เพื่อสร้างตัวแบบที่มีความสอดคล้องกับข้อมูลมากที่สุด โดยแสดงความสัมพันธ์ระหว่างตัวแปรทำนาย และกลุ่มของข้อมูลนำเข้า ตัวแบบที่สร้างขึ้นนี้นอกจากที่จะมีความสอดคล้องกับข้อมูลนำเข้าแล้ว ยังควรมีความสามารถในการจำแนกประเภทข้อมูลใหม่ที่ไม่ได้ใช้ในการสร้างตัวแบบได้อย่างถูกต้องอีกด้วย กล่าวคือ ข้อมูลสอน (Training Data) จะถูกนำไปใช้ในการสร้างตัวแบบการจำแนกประเภท โดยตัวแบบที่ได้รับการทดสอบว่ามีความถูกต้องสูงสุดก็จะถูกนำไปใช้การทำนายข้อมูลใหม่ หรือข้อมูลทดสอบ (Test Data)



NAME	Time	Items	Gender
ธำรงค์	10	2	M
จิรายุส	16	3	M
ธนพร	15	1	F
มานะ	5	1	M
กิติยา	20	1	F
ศรันย์	20	4	M

ภาพที่ 2-4 กระบวนการในการจำแนกข้อมูลในขั้นตอนการเรียนรู้



NAME	Time	Items	Gender
สุมาลี	20	1	?
สมชาย	12	2	?
สมหญิง	3	1	?

ภาพที่ 2-5 กระบวนการในการจำแนกข้อมูลในขั้นตอนการทดสอบ

ทฤษฎีของโครงข่ายประสาทเทียม

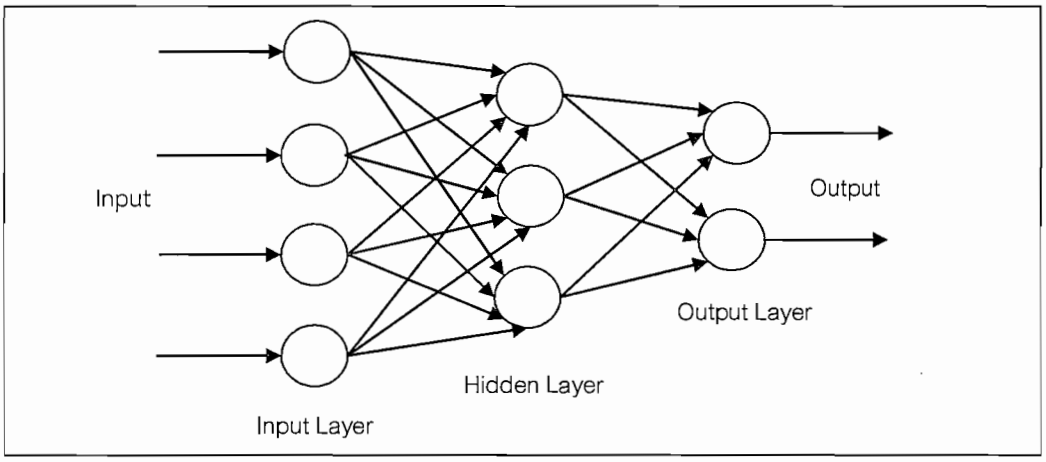
โครงข่ายประสาทเทียม (Artificial Neural Networks) ถูกสร้างขึ้นเพื่อเรียนแบบการทำงานของสมองมนุษย์มีการทำงานแบบขนานจำนวนมาก ในที่นี้เลือกใช้โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Neural Network: MLP) หรือโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (Back - Propagation Neural Network) ซึ่งเป็นโครงข่ายประสาทเทียมแบบต้องมีผู้สอน (Supervised Learning)

1. สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นมีลักษณะหลักๆ ดังนี้คือ

1.1 จำนวนชั้นต่าง ๆ โครงข่ายประสาทเทียมจะประกอบด้วยชั้นต่าง ๆ คือ ชั้นอินพุต (Input layer) ชั้นเอาต์พุต (Output layer) และชั้นซ่อน (Hidden layer) ซึ่งจะอยู่ระหว่างชั้นอินพุตและชั้นเอาต์พุต

1.2 การเชื่อมต่อระหว่างชั้นต่าง ๆ การเชื่อมต่อระหว่างชั้นต่าง ๆ นั้น ทุก ๆ โหนดในชั้นอินพุตจะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นซ่อนชั้นแรกและทุก ๆ โหนดในชั้นซ่อนชั้นแรกจะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นถัดไปจนในที่สุดทุก ๆ โหนดในชั้นซ่อนชั้นสุดท้าย จะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นเอาต์พุต

1.3 การทำงานของชั้นต่าง ๆ ชั้นอินพุตไม่มีการประมวลผลทำหน้าที่รับสัญญาณเข้าแล้วกระจายออกไปยังแต่ละโหนด ในชั้นถัดไปเท่านั้น ส่วนชั้นซ่อนและชั้นเอาต์พุต นั้นมีการประมวลผล ภาพที่ 2-6 แสดงสถาปัตยกรรมของโครงข่ายประสาทเทียมแบบเพอร์เซพตรอนหลายชั้น ซึ่งจะประกอบไปด้วย ชั้นของอินพุต, ชั้นซ่อนจำนวน 1 ชั้น และชั้นของเอาต์พุต โดยแต่ละโหนดจะถูกเชื่อมต่อกันเป็นโครงข่าย



ภาพที่ 2-6 สถาปัตยกรรมของโครงข่ายประสาทเทียม

2. วิธีการเรียนรู้แบบแพร่ย้อนกลับ (Back - Propagation Neural Network)

2.1 ขั้นตอนการเรียนรู้ของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ แบ่งเป็นขั้นตอนดังต่อไปนี้

2.1.1 กำหนดจำนวนโหนดอินพุต (N) จำนวนโหนดเอาต์พุต (J) จำนวนโหนดของชั้นซ่อน (M) จำนวน 1 ชั้นซ่อน และกำหนดจำนวนข้อมูลอินพุตและข้อมูลเอาต์พุตต่อจากนั้นทำการรับจำนวนรอบสูงสุดที่จะทำการเรียนรู้ (R) และค่าผิดพลาดที่ยอมรับได้

2.1.2 กำหนดค่าตั้งค่าพารามิเตอร์ของอัตราการเรียนรู้ (η) ให้อยู่ในช่วง [0,1]

2.1.3 การสุ่มน้ำหนักเริ่มต้นให้กับทุก ๆ เส้นเชื่อมโยงภายในโครงข่ายประสาทเทียมในทั้ง 2 ชั้น โดยให้มีค่าอยู่ระหว่าง [-1,1]

2.1.4 รับค่าอินพุตของข้อมูลชุดแรกเพื่อใช้ในการคำนวณหาค่าเอาต์พุตของโครงข่ายประสาทเทียม

2.1.5 คำนวณค่าผลรวมของโหนดในชั้นซ่อน (S_m) ก่อนทำการปรับค่ารวมด้วยฟังก์ชันกระตุ้น (Activation Function) ฟังก์ชันซิกมอยด์ ซึ่งจะได้ค่าโหนดในชั้นซ่อนที่อยู่ในช่วง [0,

1] สำหรับแต่ละโหนดของชั้นซ่อนค่าเอาต์พุตของชั้นซ่อนก่อนทำการปรับค่า โดยมีรายละเอียดสมการดังนี้

ค่าเอาต์พุตของชั้นซ่อนก่อนทำการปรับค่า มีสมการดังนี้

$$S_m = \sum_{n=1}^{N} x_n * w_{nm} \tag{2.5}$$

ค่าเอาต์พุตของชั้นซ่อนหลังทำการปรับค่า มีสมการดังนี้

$$y_m = f(s_m) \tag{2.6}$$

ฟังก์ชั่นที่ใช้ปรับค่า f(x) มีสมการดังนี้

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.7}$$

2.1.6 คำนวณค่าเอาต์พุตของโหนดในชั้นเอาต์พุตด้วยสมการที่ 2.6 ก่อนทำการปรับค่าผลรวมด้วยฟังก์ชันกระตุ้นฟังชันซิกมอยด์สมการที่ 2.7 ซึ่งจะได้ของโหนดในชั้นเอาต์พุตที่อยู่ในช่วง [0,1] สำหรับค่าของผลลัพธ์ในโหนดของชั้นเอาต์พุต หลังทำการปรับค่า แสดงดังสมการที่ 2.8

$$v_j = \sum_{m=1}^{M} y_m * w_{mj} \tag{2.8}$$

ค่าเอาต์พุตของชั้นเอาต์พุตหลังทำการปรับค่ามีสมการที่ 2.9

$$z_j = f(v_j) \tag{2.9}$$

2.1.7 หาค่าความผิดพลาดและปรับน้ำหนัก นำเอาต์พุตที่ได้กับเอาต์พุตที่ได้กำหนดไว้มาหาค่าความผิดพลาดของข้อมูล ถ้าค่าผิดพลาดของข้อมูลน้อยกว่าค่าผิดพลาดที่ยอมรับได้ทำการรับข้อมูลชุดต่อไปถ้าไม่ใช่ปรับน้ำหนักแล้วทำการรับข้อมูลของชุดถัดไปแล้วจึงกลับไปทำข้อ 3.2.2.5 แต่ถ้าเป็นข้อมูลชุดสุดท้ายทำข้อ 3.2.2.8

ค่าความผิดพลาดในแต่ละชุดของข้อมูลตัวอย่าง มีสมการที่ 2.10

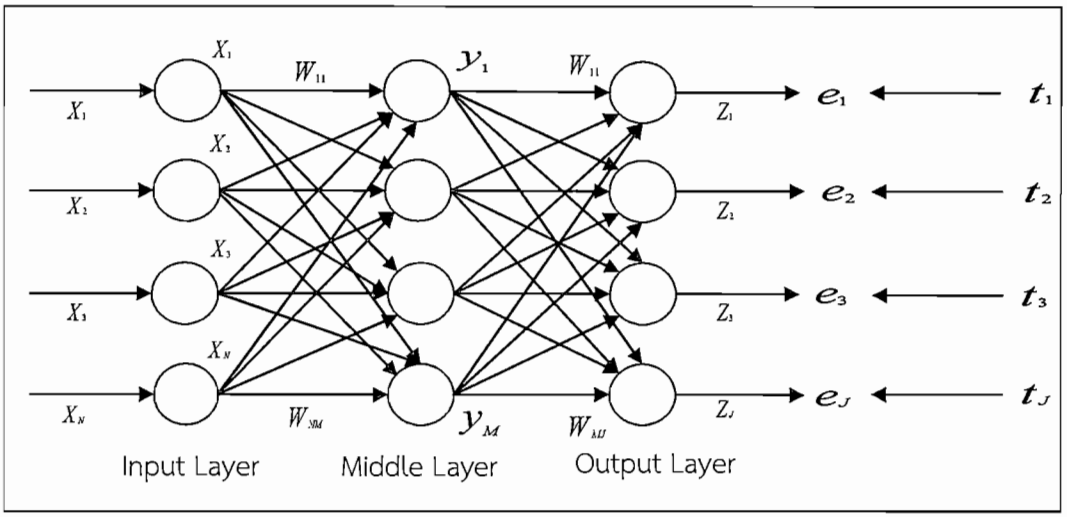
$$e^{(q)} = \frac{1}{2}\sum(t_j - z_j)^2 \qquad (2.10)$$

การปรับน้ำหนัก ระหว่างโหนดในชั้นซ่อนและชั้นเอาต์พุต ดังสมการที่ 2.11

$$w_{mj}^{(\gamma+1)} = w_{mj}^{(\gamma)} + \eta\left\{(t_j^{(q)} - z_j^{(q)}) * \left[z_j^{(q)}(1 - z_j^{(q)})\right] * y_m^{(q)}\right\} \qquad (2.11)$$

การปรับน้ำหนัก ระหว่างโหนดในชั้นอินพุตและชั้นซ่อน ดังสมการที่ 2.12

$$w_{nm}^{(\gamma+1)} = w_{nm}^{(\gamma)} + \eta\left\{\sum_{j=1}^{J}(t_j^{(q)} - z_j^{(q)})\left[z_j^{(q)}(1 - z_j^{(q)})\right]w_{mj}^{(\gamma)}\right\} * \left[y_m^{(q)}(1 - y_m^{(q)})\right]\left[x_n^{(q)}\right] \qquad (2.12)$$



ภาพที่ 2-7 สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ

2.1.8 คำนวณหาค่าผิดพลาดรวมเฉลี่ย (Mean Squared Error: MSE) ในทุก ๆ แถวข้อมูล โดยนำค่าผิดพลาดของชุดข้อมูลแต่ละแถวของชุดข้อมูลมารวมกันแล้วทำการหาค่าเฉลี่ย เพื่อใช้ในการตรวจสอบว่าผลลัพธ์ของทุกๆ ข้อมูลในแต่ละรอบนั้นมีค่าน้อยกว่าค่าผิดพลาดที่ยอมรับได้ในทุก ๆ ข้อมูลหรือไม่ ถ้าความผิดพลาดยอมรับได้ให้จบการเรียนรู้แต่ถ้าค่าความผิดพลาดเฉลี่ยมากกว่าค่าที่ยอมรับได้ให้ตรวจสอบว่าได้ทำการเรียนรู้ครบตามจำนวนรวบที่กำหนดไว้หรือไม่ ถ้าครบแล้วให้จบการเรียนรู้แต่ถ้ายังไม่ครบให้กลับไปทำข้อ 2.1.4 ใหม่ซึ่งก็คือเริ่มต้นการเรียนรู้รอบใหม่

การหาค่าผิดพลาดรวมเฉลี่ย ดังสมการที่ 2.13

สำนักหอสมุด มหาวิทยาลัยบูรพา
ต.แสนสุข อ.เมือง จ.ชลบุรี 20131

$$E = \frac{1}{Q}\sum_{q=1}^{Q} e^{(q)} \qquad (2.13)$$

2.2 ความหมายของตัวแปรต่างๆที่ใช้ แสดงตัวแปลต่างๆ พร้อมความหมายดังนี้

x_n = อินพุตโหนดที่ n มีทั้งหมด N โหนด

s_m = เอาต์พุตของชั้นซ่อน ก่อนทำการปรับค่า (activation) เป็น y_m

y_m = เอาต์พุตของชั้นซ่อน หลังทำการปรับค่าของโหนดที่ m มีทั้งหมด M โหนด

v_j = เอาต์พุตของชั้นเอาต์พุต ก่อนทำการปรับค่า (activation) เป็น z_j

z_j = ค่าเอาต์พุตที่ได้ทำการปรับค่าแล้วของชั้นเอาต์พุตโหนดที่ j มีทั้งหมด J โหนด

t_j = ค่าเอาต์พุตที่ต้องการที่ชั้นเอาต์พุตโหนดที่ j มีทั้งหมด J โหนด

w_{nm} = น้ำหนักของเส้นเชื่อมระหว่างชั้นอินพุต กับชั้นซ่อน

w_{mj} = น้ำหนักของเส้นเชื่อมระหว่างชั้นซ่อน กับชั้นเอาต์พุต

η = อัตราการเรียนรู้มีค่าอยู่ระหว่าง 0 ถึง 1

r = จำนวนรอบที่จะทำการเรียนรู้ มี R เป็นจำนวนรอบที่กำหนด

q = จำนวนชุดของข้อมูลตัวอย่าง มี Q เป็นตัวกำหนด

$e^{(q)}$ = ค่าผิดพลาดของข้อมูลตัวอย่าง

E = ค่าผิดพลาดรวมเฉลี่ยของข้อมูลตัวอย่าง

ภาพที่ 2-8 คำนิยามตัวแปร

2.3 ข้อดีและข้อเสียของการเรียนรู้แบบแพร่ย้อนกลับ

ข้อดี คือ การเรียนรู้แบบแพร่ย้อนกลับนั้นจะมีความสามารถในการจดจำรูปแบบ และสามารถที่จะเรียนรู้ความสัมพันธ์ของรูปแบบได้มากมาย ประกอบกับมีความยืดหยุ่นของการเรียนรู้ที่ความหลากหลาย โดยสามารถออกแบบทางเลือกจาก จำนวนชั้น เส้นเชื่อมโยง จำนวนโหนดในแต่ละชั้นที่ได้ทำการกำหนดขึ้น เป็นต้น ด้วยความยืดหยุ่นนี้เองทำให้การเรียนรู้แบบแพร่กลับสามารถแก้ปัญหางานประยุกต์ได้อย่างมากมาย อย่างไรก็ตามการเรียนรู้แบบแพร่กลับต้องการตัวอย่างรูปแบบที่จะเรียนรู้

365231

ข้อจำกัด คือ การใช้เวลามากในการฝึกสอนโครงข่ายประสาทเทียมให้เกิดการเรียนรู้สำหรับการเรียนรู้แก้ปัญหา

การวัดประสิทธิภาพ (Evaluation)

การประเมินประสิทธิภาพของตัวแบบการจำแนกประเภท นั้นจะอาศัยผลการทำนายการเป็นสมาชิกกลุ่มข้อมูล เมื่อใช้ตัวแบบดังกล่าวกับข้อมูลทดสอบ โดยนับจากจำนวนตัวอย่างที่ได้รับการทำนายที่ถูกต้อง และไม่ถูกต้อง ค่าจำนวนตัวอย่างเหล่านี้ถูกนำมาสร้างเป็นตารางแสดงผลลัพธ์การจำแนกประเภทที่เรียกว่า Confusion Matrix แสดงดังตารางที่ 2-10 ซึ่งใช้สำหรับกรณีปัญหาที่ต้องการจำแนกประเภทข้อมูลเป็นสองกลุ่ม ค่าที่แสดงในตารางเป็นจำนวนตัวอย่างในแต่ละสถานการณ์ ดังนี้

True positive (TP)	หมายถึง จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งกำลังสนใจอยู่
True negative (TN)	หมายถึง จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาสซึ่งไม่ได้สนใจอยู่
False positive (FP)	หมายถึง จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งกำลังสนใจอยู่
False negative (FN)	หมายถึง จำนวนข้อมูลที่ทำนายผิดมาเป็นคลาสซึ่งไม่ได้สนใจอยู่

จำนวนหน่วยตัวอย่างที่ได้รับการทำนายได้อย่างถูกต้องมีจำนวน TN+TP ตัวอย่าง และจำนวนหน่วยตัวอย่างที่ได้รับการทำนายไม่ถูกต้องมีจำนวน FN+FP ตัวอย่าง

ตารางที่ 2-10 แสดงการกำหนดค่าในการวัดประสิทธิภาพ

		ค่าความจริง (Actual)	
		ยอมรับ (Positive)	ปฏิเสธ (Negative)
ค่าทำนาย (Predicted)	ยอมรับ (Positive)	True positive (TP)	False positive (FP)
	ปฏิเสธ (Negative)	False negative (FN)	True negative (TN)

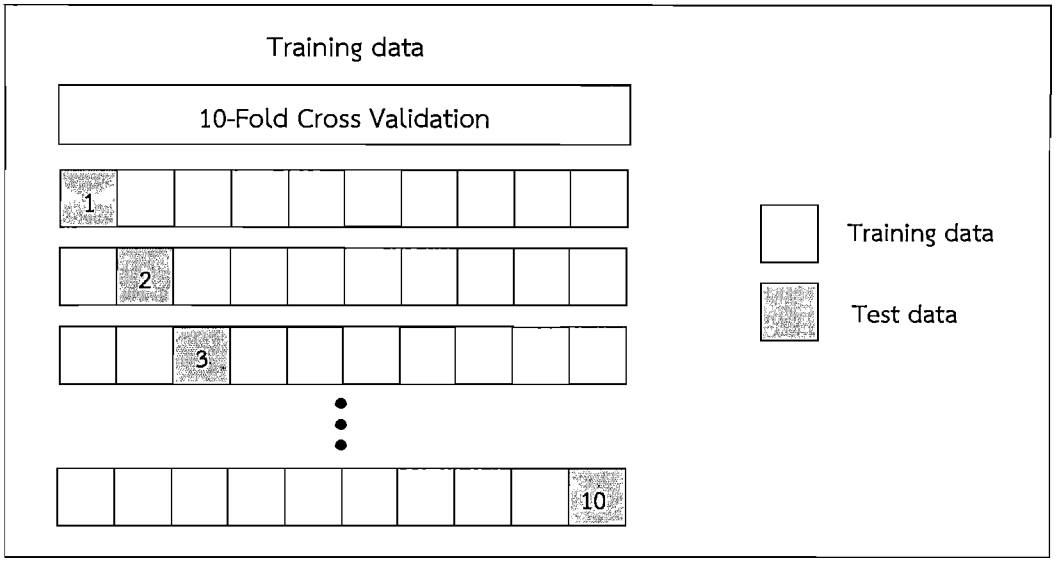
โดยการวัดประสิทธิภาพในการจำแนกประเภทกลุ่มข้อมูลวิธีที่นิยมมากวิธีหนึ่ง คือ อัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

1. อัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR) คือ ค่าที่แสดงว่าตัวแบบสามารถทำนายได้ถูกต้องมากน้อยเพียงใดเมื่อเทียบอัตราส่วนกับข้อมูลทั้งหมด ได้จากสมการที่ 2.12

$$\mathrm{CCR} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \qquad (2.12)$$

ในการตัดสินประสิทธิภาพของตัวแบบจำแนกกลุ่มตัวแบบที่ดีที่สุด คือ ตัวแบบที่มีค่าอัตราความถูกต้องของการจำแนกประเภทสูงที่สุด

การวัดประสิทธิภาพจะต้องทำการเลือกข้อมูลชุดสอน (Training Set) ข้อมูลชุดทดสอบ (Test Set) และ ข้อมูลสำหรับทดสอบโมเดล (Unseen Test Data) โดยใช้วิธีการ Cross Validation ซึ่งเป็นวิธีการคาดการค่าความผิดพลาดของโมเดลหรือวิธีการที่นำเสนอ โดยพื้นฐานของวิธีการ Cross Validation คือการสุ่มตัวอย่างโดยเริ่มจากการแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบในการเลือกสุ่มข้อมูลแบบความเที่ยงตรงจำนวน K กลุ่ม (K - Fold Cross Validation) นั้น จะเลือกสุ่มข้อมูลออกเป็น K ชุดเท่าๆกัน ในการทดลองครั้งแรกข้อมูลชุดที่ 1 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ในการทดลองครั้งที่สองข้อมูลชุดที่ 2 เป็นข้อมูลชุดทดสอบและข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั่งข้อมูลทุกชุดได้ถูกนำมาเป็นข้อมูลชุดทดสอบซึ่งจะมีการทดสอบทั้งหมด K ครั้ง ตัวอย่างการเลือกสุ่มข้อมูลแบบ ความเที่ยงตรง K กลุ่ม เมื่อ K = 10 แสดงดังภาพที่ 2-9 ข้อดีของการเลือกสุ่มข้อมูลแบบความเที่ยงตรง K กลุ่ม คือ ข้อมูลทุกตัวจะถูกนำมาเป็นข้อมูลชุดสอนและข้อมูลชุดทดสอบ ข้อเสียคือใช้เวลานานในการทดลอง เนื่องจากต้องทดลองข้อมูลทั้งหมด K ครั้ง ทั้งนี้นิยมกำหนดให้ค่า K มีค่าเท่ากับ 10



ภาพที่ 2-9 การเลือกสุ่มข้อมูลแบบความเที่ยงตรง 10 กลุ่ม (10 - Fold Cross Validation)

งานนิพนธ์นี้ได้นำทฤษฎีต่าง ๆ มาช่วยในการคัดเลือกปัจจัยที่ผลต่อการสำเร็จการศึกษา ประกอบด้วยเทคนิคการคัดเลือกปัจจัยของข้อมูลก่อนการนำไปการจำแนกประเภท ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach ซึ่งภายหลังการคัดเลือกปัจจัยที่มีผลต่อการสำเร็จการศึกษาแล้วเสร็จจะถูกนำไปใช้ในตัวแบบการจำแนกประเภทนิสิตด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) และโดยทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบประกอบด้วยข้อมูลชุดสอน (Training Set) และข้อมูลชุดทดสอบ (Test Set) โดยใช้วิธีการ 10-Fold Cross Validation และการวัดประสิทธิภาพ (Evaluation) ด้วยอัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

งานวิจัยที่เกี่ยวข้อง

การคัดเลือกคุณลักษณะที่เหมาะสมจะช่วยลดความซ้ำซ้อนของข้อมูลและสามารถกำจัดคุณลักษณะที่ไม่เกี่ยวข้องออกไป มีส่วนช่วยเพิ่มประสิทธิภาพการทำงานของอัลกอริทึมได้เป็นอย่างดี โดยการคัดเลือกคุณลักษณะประกอบไปด้วยอัลกอริทึมในการค้นหาและการประเมินค่าของชุดคุณลักษณะที่ผ่านการค้นหา ซึ่งมีมากมากหลายเทคนิค ดังนั้นผู้จัดทำงานนิพนธ์จึงศึกษาเอกสารเพื่อหาเทคนิคใดที่เหมาะสมและนำมาใช้ในการทดลองครั้งนี้

สุคนธ์ทิพย์ วงศ์พันธ์ (2008) ได้นำเสนองานวิจัยเรื่อง การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อจำแนกพฤติกรรมการกระทำผิดของนักเรียนระดับ

อาชีวศึกษา เป็นงานวิจัยที่นำเสนอ การเปรียบเทียบการทำเหมืองข้อมูล 2 วิธี ได้แก่ 1) วิธีการ Simple Classificationเป็นการจำแนกข้อมูลโดยไม่มีการคัดเลือกคุณลักษณะที่เหมาะสมและใช้อัลกอริทึมในการจำแนกประเภท 4 อัลกอริทึม ได้แก่ 1. ตัวจำแนกประเภทเบย์อย่างง่าย (Naive Bayes) 2. ข่ายงานความเชื่อแบบเบย์ (Bayesian Network) 3. Decision Tree: C4.5 และ Neural Network 2) วิธีการ Hybrid Classification เป็นการนำวิธีการคัดเลือกคุณลักษณะที่เหมาะสม โดยทำการเปรียบเทียบอัลกอริทึมการค้นหาซับเซตคุณลักษณะระหว่าง Genetic Search กับ Forward Selection ร่วมกับวิธีการประเมินคุณค่าของซับเซตคุณลักษณะจำนวน 3 วิธีการได้แก่ 1) CFS 2) Consistency และ 3) Wrapper จากนั้นจึงนำไปประมวลผลต่อในอัลกอริทึมเพื่อจำแนกประเภทข้อมูลซึ่งเป็นวิธีเดียวกับวิธี Simple Classification โดยผลการทดลองพบว่าวิธีการ Hybrid Classification ที่ใช้ Genetic Search ร่วมกับ Wrapper โดยใช้อัลกอริทึม C4.5 ให้ค่าความถูกต้องสูงสุดเท่ากับ 83.01% และสามารถลดคุณลักษณะที่ต้องนำมาใช้ในการจำแนกจาก 43 เหลือเพียง 15 คุณลักษณะ สำหรับคุณลักษณะ 15 คุณลักษณะของข้อมูลชุดนี้ก็เพียงพอที่จะนำมาใช้เพื่อให้ได้ประสิทธิภาพที่เหมาะสมที่สุด โดยปัจจัยที่มีอิทธิพลต่อการกระทำความผิดของนักเรียน ดังนี้ เพศ, เกรดเฉลี่ยเดิมก่อนเข้าศึกษา, กรุ๊ปเลือด, ผู้ปกครอง, ส่วนสูง, จังหวัดที่อยู่ปัจจุบันของนักเรียน, จังหวัดที่อยู่ของผู้ปกครอง, อายุ, รายได้บิดา, อาชีพของมารดา, เกรดวิชาพิมพ์ดีดไทย1, เกรดวิชาพิมพ์ดีดไทย1, การได้รับทุนการศึกษา, เชื้อชาติ และเกรดวิชาเอกสารธุรกิจ

Karamouzis.ST (2009) ได้นำเสนองานวิจัยเรื่อง Sensitivity Analysis of Neural Network Parameters for Identifying the Factors for College Student Success เป็นงานวิจัยที่ศึกษาเกี่ยวกับการพยากรณ์อัตราการสำเร็จการศึกษาของนักศึกษาในสถาบันการศึกษาโดยได้ประยุกต์ใช้โครงข่ายประสาท เทียม (ANN) ในการพยากรณ์การสำเร็จการศึกษาจากข้อมูลประวัติส่วนตัวของนักศึกษา จำนวน 1,407 ตัวอย่าง พบว่าความแม่นยำในการพยากรณ์โอกาสการสำเร็จการศึกษามีค่าถึง 77%

ทิพย์สุดา วงศ์คำดี (2009) ได้นำเสนองานวิจัยเรื่อง การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้โครงข่ายประสาทเทียม เป็นการเปรียบเทียบแบบจำลองการพยากรณ์ผลสำเร็จการศึกษาของนักศึกษา โดยใช้วิธีการวิเคราะห์จำแนกประเภท เทคนิคโครงข่ายประสาทเทียม และเทคนิคโครงข่ายประสาทเทียมที่มีการคัดเลือกปัจจัยโดยการใช้วิธีการทางสถิติ โดยใช้ข้อมูลประวัติส่วนตัวของนักศึกษา ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น ผลการวิจัยพบว่า การพยากรณ์ผลสำเร็จการศึกษาของนักศึกษาด้วยโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (Back propagation Neural Network: BPNN) ที่มีการคัดเลือกปัจจัยโดยการใช้วิธีการทางสถิติ มีประสิทธิภาพการพยากรณ์ได้ดีกว่าโครงข่ายประสาทเทียมที่นำเข้าปัจจัยทั้งหมดและการวิเคราะห์จำแนกประเภท (Discriminate Analysis) โดยสามารถลดคุณลักษณะที่ต้อง

นำมาใช้ในการจำแนกจาก 10 เหลือเพียง 4 คุณลักษณะ ได้แก่ เกรดเฉลี่ย ม.6 เกรดเฉลี่ย ณ ขณะนั้น การศึกษาของบิดา และอาชีพของบิดา ด้วยค่าความถูกต้องของการจำแนกกลุ่ม 94.54 %, 93.54% และ 81.5% ตามลำดับ

ธนาวุฒิ นิลมณี (2013) ได้นำเสนองานวิธีการเลือกคุณลักษณะของข้อมูลที่มีผลกระทบต่อความสำเร็จการศึกษาของนักศึกษาและการจำแนกข้อมูลด้วยวิธีการโครงข่ายประสาทเทียม โดยใช้ข้อมูลนักศึกษาปริญญาตรี ระหว่างปี พ.ศ. 2550 - 2554 จำนวน 10,966 ชุดข้อมูล จำนวน 33 ปัจจัย และนำมาคัดเลือกคุณลักษณะของข้อมูลที่เหมาะสม โดยใช้ประเภท Filter และ Wrapper เป็นอัลกอริทึมในการคัดเลือกคุณลักษณะข้อมูลที่เหมาะสม จากนั้นทำการเปรียบเทียบการจำแนกข้อมูลโดยใช้โครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (BPNN) และโครงข่ายประสาทเทียมแบบฟังก์ชันรัศมีฐาน (RBFNN) ผลการทดลองพบว่า การเลือกคุณลักษณะข้อมูลแบบ Greedy Stepwise Method กับ ClaSE (Classifier SubsetEval) คัดเลือกจาก 33 ปัจจัยเหลือเพียง 5 ปัจจัย คือ ปีการศึกษาที่เข้า, คณะที่ศึกษา, สถานภาพของนักศึกษา, เกรดเฉลี่ยสะสมและหน่วยกิตสะสม โดยใช้ Classifiers แบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (BPNN) เป็นเทคนิคที่ดีที่สุดให้ค่าความถูกต้องของการจำแนกเท่ากับ 99.83%

บทที่ 3
วิธีดำเนินงานนิพนธ์

งานนิพนธ์นี้ได้นำเสนอการใช้แบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาที่มีการคัดเลือกปัจจัยในการจำแนกกลุ่มนิสิตชั้นปีที่ 1 เพื่อให้การช่วยเหลือด้านวิชาการโดยการนำเทคนิคการคัดเลือกปัจจัยแบบการกรอง (Filter Approach) และเทคนิคการคัดเลือกปัจจัยแบบการควบรวม (Wrapper Approach) ในการคัดเลือกปัจจัยที่เหมาะสมและมีอิทธิพลต่อการสำเร็จการศึกษา เพื่อใช้เป็นปัจจัยนำเข้าในขั้นตอนการสร้างแบบจำลองการจำแนกประเภทนิสิตโดยโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) ที่เหมาะสมสำหรับใช้ในการพยากรณ์โอกาสการสำเร็จการศึกษาของนิสิตซึ่งจะถูกใช้ในการจำแนกกลุ่มนิสิตเพื่อให้การช่วยเหลือด้านวิชาการ ดังมีวิธีรายละเอียดต่อไปนี้

การรวบรวมข้อมูลที่เกี่ยวข้องและจัดเตรียมข้อมูลเพื่อการประมวลผล (Pre-processing)

1. การรวบรวมข้อมูลนิสิต (Student Data)

การรวบรวมข้อมูลนิสิตสำหรับการสร้างกลุ่มประชากรตัวอย่างนำมาจากฐานข้อมูลของกองทะเบียนและประมวลผลนิสิต มหาวิทยาลัยบูรพา โดยมีชุดข้อมูลประกอบด้วย 1) ข้อมูลส่วนตัวของนิสิต 2) ข้อมูลประวัติการศึกษาของนิสิตระดับปริญญาตรีในชั้นปีที่ 1 และ 3) ข้อมูลผลการสำเร็จการศึกษา ของนิสิตในหลักสูตร 4 ปี สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา จำนวน 5 ปี ย้อนหลังตั้งแต่ปีการศึกษา 2550 - 2554 จำนวนทั้งสิ้น 699 ชุด ข้อมูลโดยปัจจัยหรือคุณลักษณะนำเข้าที่คาดว่ามีความสัมพันธ์ที่ส่งผลต่อการสำเร็จการศึกษาซึ่งข้อมูลทั้งหมดประกอบไปด้วย ประกอบด้วย 25 ปัจจัย โดยแต่ละค่าปัจจัยมีรายละเอียดดังตารางที่ 3-1

ตารางที่ 3–1 ปัจจัยที่ใช้ในการสร้างแบบจำลองการจำแนกประเภทนิสิต จำนวน 25 ปัจจัย

ลำดับที่	ปัจจัย
1	เพศ
2	สถานภาพของบิดา
3	สถานภาพของมารดา
4	สถานภาพสมรสของบิดามารดา
5	อาชีพของบิดา
6	อาชีพของมารดา
7	รายได้ของบิดา

ตารางที่ 3–1 (ต่อ)

ลำดับที่	ปัจจัย
8	รายได้ของมารดา
9	ภูมิภาคของสถานศึกษาเดิม (ที่อยู่นิสิต)
10	จำนวนพี่น้องทั้งหมด (รวมตัวเอง)
11	บุตรคนที่
12	บุตรคนที่ที่กำลังศึกษา
13	ประเภทของสถานบันศึกษาเดิม(รัฐ/เอกชน)
14	เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา
15	วิธีเข้าศึกษา
16	ประเภทนิสิต(ปกติ/พิเศษ)
17	เกรดภาคเรียนที่ 1 ชั้นปีที่ 1
18	เกรดภาคเรียนที่ 2 ชั้นปีที่ 1
19	เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1
20	เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1
21	เกรดรายวิชาแคลคูลัส 1
22	เกรดรายวิชาแคลคูลัส 2
23	เกรดรายวิชาโปรแกรมมิ่ง 1
24	เกรดรายวิชาโปรแกรมมิ่ง 2
25	เกรดรายวิชาภาษาอังกฤษ 1

2. การตรวจสอบความถูกต้องของข้อมูล (Data Cleaning)

การตรวจสอบความถูกต้องของข้อมูลในงานนิพนธ์นี้เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่า มีความถูกต้อง เหมาะสม และครบถ้วน เพื่อลดความผิดพลาดที่อาจจะเกิดขึ้นเนื่องจากข้อมูลของนิสิตในระบบมีความไม่สมบูรณ์ อาทิเช่น ข้อมูลอาชีพของบิดามารดา ข้อมูลสถานการมีชีวิตอยู่ของบิดามารดา ข้อมูลรายได้ของบิดามารดา ผลที่ได้จากการตรวจสอบความถูกต้องของข้อมูล พบว่า มีค่าของข้อมูลสูญหาย (Missing Values) ซึ่งจะทำให้เกิดความคาดเคลื่อนในการพยากรณ์ ผู้จัดทำงานนิพนธ์จึงได้คัดเลือกชุดข้อมูลที่ไม่สมบูรณ์ออกจากกลุ่มข้อมูลตัวอย่าง

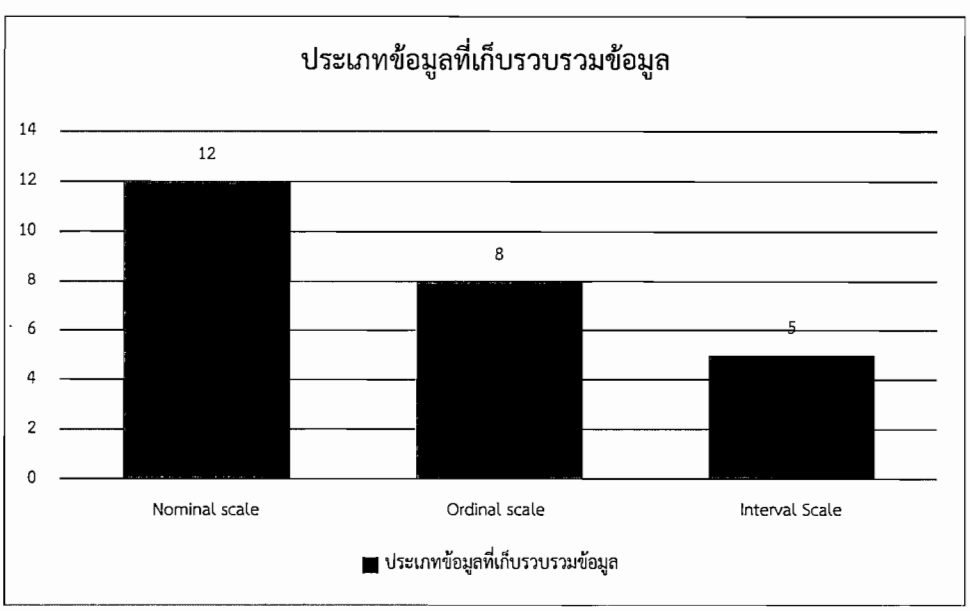
3. การแปลงรูปแบบข้อมูล (Data Transformation)

การแปลงรูปแบบข้อมูล (Data Transformation) เป็นการเปลี่ยนแปลงรูปแบบข้อมูลที่ได้รวบรวมมาให้ตรงกับรูปแบบข้อมูลของโปรแกรมคอมพิวเตอร์สำเร็จรูปที่จะนำมาวิเคราะห์และใช้งาน จึงจำเป็นต้องจัดให้อยู่ในรูปแบบเดียวกันก่อนนำไปใช้ในการประมวลผล โดยแบ่งข้อมูลที่เก็บรวบรวมข้อมูลเป็น 3 แบบ คือ

3.1 การเก็บรวบรวมข้อมูลของปัจจัยที่มีมาตราวัดแบบมาตรานามบัญญัติ (Nominal scale) ได้แก่ เพศ, สถานภาพของบิดา, สถานภาพของมารดา, สถานภาพสมรสของบิดามารดา, อาชีพของบิดา, อาชีพของมารดา, รายได้ของบิดา, รายได้ของมารดา, ภาคของสถานศึกษาเดิม (ที่อยู่นิสิต), ประเภทของสถานบันศึกษาเดิม(รัฐ/เอกชน), วิธีเข้าศึกษา และประเภทนิสิต(ปกติ/พิเศษ)

3.2 การเก็บรวบรวมข้อมูลของปัจจัยที่มีมาตราวัดแบบมาตราอันดับ (Ordinal scale) ได้แก่ จำนวนพี่น้องทั้งหมด (รวมตัวเอง), บุตรคนที่, บุตรคนที่ที่กำลังศึกษา, เกรดรายวิชาแคลคูลัส 1 , เกรดรายวิชาแคลคูลัส 2, เกรดรายวิชาโปรแกรมมิ่ง 1, เกรดรายวิชาโปรแกรมมิ่ง 2 และเกรดรายวิชาภาษาอังกฤษ 1

3.3 การเก็บรวบรวมข้อมูลของปัจจัยที่มีมาตราวัดแบบมาตราอันตรภาค (Interval Scale) เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา, เกรดภาคเรียนที่ 1 ชั้นปีที่ 1, เกรดภาคเรียนที่ 2 ชั้นปีที่ 1, เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1และเกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1



ภาพที่ 3–1 แสดงจำนวนปัจจัยโดยมาตราวัดแบบมาตราแบบนามบัญญัติ, อันดับ และอันตรภาค

โดยข้อมูลแต่ละปัจจัยจะถูกแปลงให้เป็นตัวเลขก่อนนำเข้าประมวลผล รายละเอียดการการเปลี่ยนแปลงรูปแบบข้อมูล ดังต่อไปนี้

รายละเอียดของสัญลักษณ์ที่ใช้แทนตัวแปรสำหรับใช้ในงานวิจัย

1) สัญลักษณ์ที่ใช้แทนตัวแปรตามหรือข้อมูลผลลัพธ์ (Output) แสดงรายละเอียดของสัญลักษณ์และคำอธิบายความหมาย ดังตารางที่ 3-2

ตารางที่ 3-2 คำอธิบายตัวแปรชื่อตัวแปรและค่าตัวแปรตาม

ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร
ผลการสำเร็จการศึกษา	STATUS	Yes = สำเร็จการศึกษาภายในระยะเวลา 4 ปี Yes++ = สำเร็จการศึกษามากกว่า 4 ปี NO = ไม่สำเร็จการศึกษา

2) สัญลักษณ์ที่ใช้แทนตัวแปรอิสระหรือข้อมูลนำเข้า (Input) แสดงรายละเอียดของสัญลักษณ์และคำอธิบายความหมาย ดังตารางที่ 3-3

ตารางที่ 3-3 คำอธิบายตัวแปร ชื่อตัวแปรและค่าตัวแปรอิสระ

ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร
เพศ	SEX	ชาย =1 หญิง =2
สถานภาพของบิดา	STATUSFA	ถึงแก่กรรม =1 มีชีวิต =2
สถานภาพของมารดา	STATUSMO	ถึงแก่กรรม =1 มีชีวิต =2
สถานภาพสมรสของบิดามารดา	STATUSFM	อยู่ด้วยกัน =1 บิดาถึงแก่กรรม =2 มารดาถึงแก่กรรม =3 บิดาและมารดาถึงแก่กรรม =4 บิดาและมารดาแต่งงานใหม่ =5 แยกกันอยู่ =6 หย่าร้าง =7 ไม่ระบุ =8

ตารางที่ 3-3 (ต่อ)

ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร	
อาชีพของบิดา	OCCFA	รับราชการ	=1
		พนักงานราชการ/ลูกจ้างหน่วยงานราชการ	=2
		พนักงานรัฐวิสาหกิจ	=3
		พนักงานลูกจ้างเอกชน	=4
		ค้าขาย,ธุรกิจส่วนตัวและอาชีพอิสระ/	=5
		เกษตรกร/ประมง	=6
		ไม่ประกอบอาชีพ	=7
		ไม่ทราบ	=8
อาชีพของมารดา	OCCMO	รับราชการ	=1
		พนักงานราชการ/ลูกจ้างหน่วยงานราชการ	=2
		พนักงานรัฐวิสาหกิจ	=3
		พนักงานลูกจ้างเอกชน	=4
		ค้าขาย,ธุรกิจส่วนตัวและอาชีพอิสระ/	=5
		เกษตรกร/ประมง	=6
		ไม่ประกอบอาชีพ	=7
		ไม่ทราบ	=8
รายได้ของบิดา	INCFA	ไม่มีรายได้	=1
		< 150,000 บาทต่อปี (< 12,500 บาทต่อเดือน)	=2
		150,000 - 300,000 บาทต่อปี (12,500 - 25,000 บาทต่อเดือน)	=3
		> 300,000 บาทต่อปี (> 25,000 บาทต่อเดือน)	=4
		ไม่ระบุ	=5
รายได้ของมารดา	INCMO	ไม่มีรายได้	=1
		< 150,000 บาทต่อปี (< 12,500 บาทต่อเดือน)	=2
		150,000 - 300,000 บาทต่อปี (12,500 - 25,000 บาทต่อเดือน)	=3
		> 300,000 บาทต่อปี (> 25,000 บาทต่อเดือน)	=4
		ไม่ระบุ	=5
ภาคของสถานศึกษาเดิม (ที่อยู่นิสิต)	PROVINCE	ภาคเหนือ	=1
		ภาคใต้	=2
		ภาคกลาง	=3
		ภาคตะวันออก	=4
		ภาคตะวันตก	=5
		ภาคตะวันออกเฉียงเหนือ	=6
จำนวนพี่น้องทั้งหมด (รวมตัวเอง)	NUMSON	ตามจำนวน	

ตารางที่ 3-3 (ต่อ)

ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร	
บุตรคนที่	SON	ตามจำนวน	
บุตรคนที่ที่กำลังศึกษา	SONED	ตามจำนวน	
ประเภทของสถานบันศึกษาเดิม	TYPESCHOOL	รัฐบาล	=1
		เอกชน	=2
เกรดเฉลี่ยสถานศึกษาเดิม	ENTRYGPA	0.00 -1.99	=1
		2.00 - 2.50	=2
		2.51 – 3.00	=3
		3.01 - 4.00	=4
ประเภทนิสิต (ปกติ/พิเศษ)	PROGRAM	ภาคปกติ	=1
		ภาคพิเศษ	=2
วิธีเข้าศึกษา	BYTEDES	Admission กลาง	=1
		โครงการความสามารถพิเศษ	=2
		สอบที่มหาวิทยาลัยโดยตรง	=3
เกรดภาคเรียนที่ 1 ชั้นปีที่ 1	GPA-1	0.00-1.74	=1
		1.75-1.99	=2
		2.00-2.50	=3
		2.51-3.00	=4
		3.01-4.00	=5
เกรดภาคเรียนที่ 2 ชั้นปีที่ 1	GPA-2	0.00-1.74	=1
		1.75-1.99	=2
		2.00-2.50	=3
		2.51-3.00	=4
		3.01-4.00	=5
เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1	ACC-GPA-1	0.00-1.74	=1
		1.75-1.99	=2
		2.00-2.50	=3
		2.51-3.00	=4
		3.01-4.00	=5
เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1	ACC-GPA-2	0.00-1.74	=1
		1.75-1.99	=2
		2.00-2.50	=3
		2.51-3.00	=4
		3.01-4.00	=5

ตารางที่ 3-3 (ต่อ)

ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร	
เกรดรายวิชาแคลคูลัส 1	Calculus1	A	=9
		B+	=8
		B	=7
		C+	=6
		C	=5
		D+	=4
		D	=3
		F	=2
		W/ไม่ลงเรียน	=1
เกรดรายวิชาแคลคูลัส 2	Calculus2	A	=9
		B+	=8
		B	=7
		C+	=6
		C	=5
		D+	=4
		D	=3
		F	=2
		W/ไม่ลงเรียน	=1
เกรดรายวิชาโปรแกรมมิ่ง 1	Pro1	A	=9
		B+	=8
		B	=7
		C+	=6
		C	=5
		D+	=4
		D	=3
		F	=2
		W/ไม่ลงเรียน	=1
เกรดรายวิชาโปรแกรมมิ่ง 2	Pro2	A	=9
		B+	=8
		B	=7
		C+	=6
		C	=5
		D+	=4
		D	=3
		F	=2
		W/ไม่ลงเรียน	=1

ตารางที่ 3-3 (ต่อ)

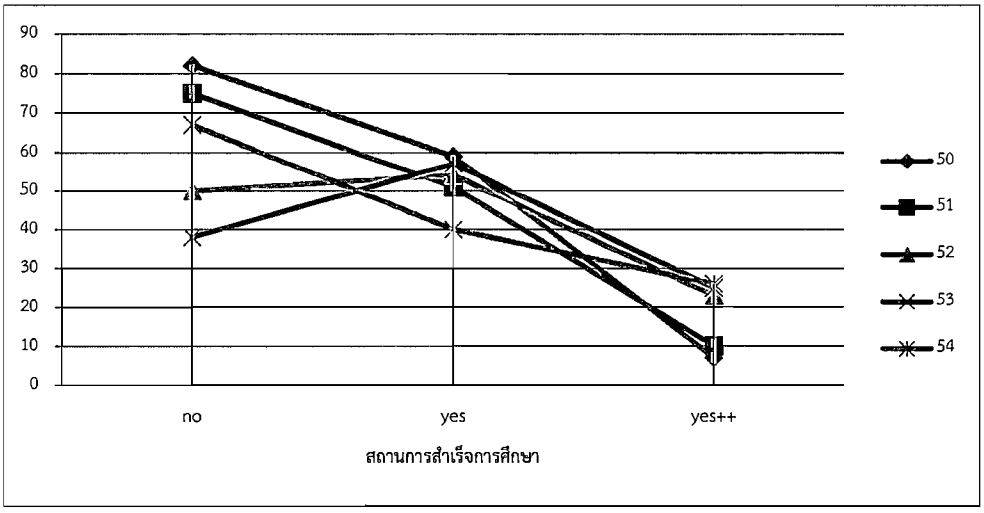
ตัวแปร	ชื่อตัวแปร	ค่าของตัวแปร	
เกรดรายวิชา ภาษาอังกฤษ1	Eng1	A	=9
		B+	=8
		B	=7
		C+	=6
		C	=5
		D+	=4
		D	=3
		F	=2
		W/ไม่ลงเรียน	=1

4. การเตรียมชุดข้อมูลสำหรับใช้ในการวิเคราะห์

การเตรียมชุดข้อมูลสำหรับใช้ในการวิเคราะห์ เป็นการเตรียมชุดข้อมูลนิสิตสำหรับใช้ในการทดลองวิเคราะห์หาปัจจัยโดยมีจำนวน 6 ชุด ซึ่งเป็นข้อมูลนิสิตตั้งแต่ปีการศึกษา 2550 – 2554 โดยแบ่งชุดข้อมูลในการทดลองออกเป็นรายปีการศึกษาตามรหัสนิสิต ได้แก่ 1) ข้อมูลนิสิตปีการศึกษา 2550 2) ข้อมูลนิสิตปีการศึกษา 2551 3) ข้อมูลนิสิตปีการศึกษา 2552 4) ข้อมูลนิสิตปีการศึกษา 2550 – 2552 และ ข้อมูลทดสอบ ได้แก่ 1) ข้อมูลนิสิตปีการศึกษา 2553 2) ข้อมูลนิสิตปีการศึกษา 2554 ดังแสดงในตารางที่ 3-4 และในแต่ละชุดข้อมูลของนิสิตตามปีการศึกษาจะมีจำนวนและคุณลักษณะของข้อมูลนิสิตที่ใช้ในการทดสอบรายละเอียดดังตารางที่ 3-5

ตารางที่ 3-4 จำนวนชุดข้อมูลนิสิตเป็นรายปีการศึกษา

ปีการศึกษาที่รับเข้า	จำนวนข้อมูล	ไม่สำเร็จการศึกษา (no)		สำเร็จการศึกษาภายในระยะเวลา 4 ปี (yes)		สำเร็จการศึกษามากกว่า 4 ปี (yes++)	
		จำนวน	ร้อยละ	จำนวน	ร้อยละ	จำนวน	ร้อยละ
2550	148	82	55.40	59	39.87	7	4.73
2551	140	75	53.57	51	37.50	10	7.4
2552	125	49	39.20	53	42.40	23	18.40
2553	153	40	26.14	69	45.10	44	28.76
2554	133	64	50.38	40	30.08	26	19.55



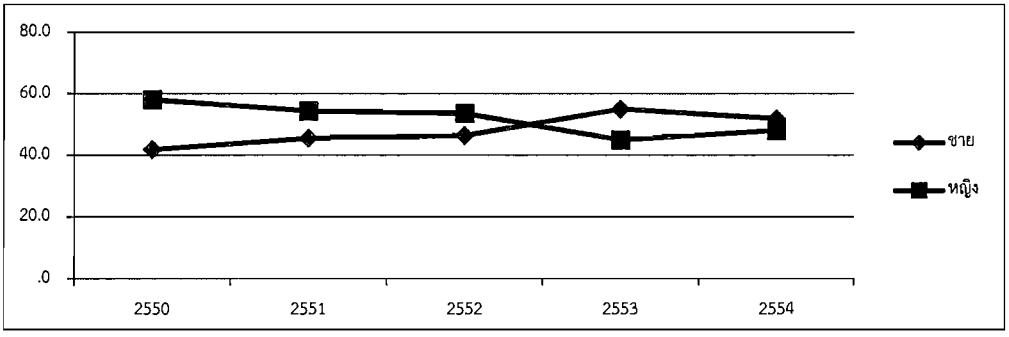
ภาพที่ 3-2 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของสถานะสำเร็จการศึกษา

ตารางที่ 3-5 ตารางแสดงปัจจัยร่วมของจำนวนปัจจัยที่ใช้ในการทดสอบ

ลำดับ	ปัจจัย	ปัจจัยที่ใช้ในการทดสอบ			
		14 ปัจจัย	16 ปัจจัย	20ปัจจัย	25 ปัจจัย
1	เพศ (SEX)	✓	✓	✓	✓
2	สถานภาพสมรสของบิดา มารด (STATUSFM)	✓	✓	✓	✓
3	อาชีพของบิดา (OCCFA)	✓	✓	✓	✓
4	อาชีพของมารดา (OCCMO)	✓	✓	✓	✓
5	รายได้ของบิดา (INCFA)	✓	✓	✓	✓
6	รายได้ของมารดา (INCMO)	✓	✓	✓	✓
7	เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา (ENTRYGPA)	✓	✓	✓	✓
8	สถานภาพของบิดา (STATUSFA)	✓	✓	✓	✓
9	สถานภาพของมารดา (STATUSMO)	✓	✓	✓	✓
10	ภูมิภาคของสถานศึกษาเดิม (ที่อยู่นิสิต) (PROVINCE)	✓	✓	✓	✓
11	จำนวนพี่น้องทั้งหมด (รวมตัวเอง) (NUMSON)	✓	✓	✓	✓
12	บุตรคนที่ (SON)	✓	✓	✓	✓
13	บุตรคนที่ที่กำลังศึกษา (SONED)	✓	✓	✓	✓
14	ประเภทของสถานบันศึกษาเดิม (รัฐ/เอกชน) (TYPESCHOOL)	✓	✓	✓	✓
15	วิธีเข้าศึกษา (BYTEDES)	-	✓	✓	✓
16	ประเภทนิสิต (ปกติ/พิเศษ) (PROGRAM)	-	✓	✓	✓
17	เกรดภาคเรียนที่ 1 ชั้นปีที่ 1 (GPA-1)	-	-	✓	✓
18	เกรดภาคเรียนที่ 2 ชั้นปีที่ 1 (GPA-2)	-	-	✓	✓
19	เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1 (ACC-GPA-1)	-	-	✓	✓
20	เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 (ACC-GPA-2)	-	-	✓	✓
21	เกรดรายวิชาแคลคูลัส 1	-	-	-	✓
22	เกรดรายวิชาแคลคูลัส 2	-	-	-	✓
23	เกรดรายวิชาโปรแกรมมิ่ง 1	-	-	-	✓
24	เกรดรายวิชาโปรแกรมมิ่ง 2	-	-	-	✓
25	เกรดรายวิชาภาษาอังกฤษ 1	-	-	-	✓

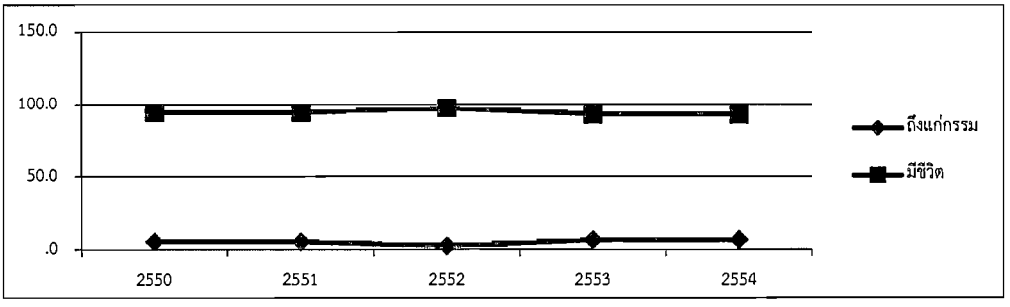
5 การวิเคราะห์ค่าสถิติเพื่อเปรียบเทียบคุณลักษณะของข้อมูล

การวิเคราะห์ค่าสถิติเพื่อเปรียบเทียบคุณลักษณะและปัจจัยของข้อมูลที่คาดว่าจะมีผลต่อการสำเร็จการศึกษาของนิสิต ดังแสดงในตารางที่ 3-1 การวิเคราะห์ค่าสถิติของข้อมูลนั้นจะใช้สถิติค่าความถี่ (Frequency) และค่าร้อยละ (Percentage) เพื่อเปรียบเทียบแนวโน้มของคุณลักษณะข้อมูลของนิสิตในแต่ละปีการศึกษา ดังแสดงด้วยในตารางที่ ก- 1, ก-2, ก-3, ก-4, ก-5 และกราฟในภาพที่ 3-3, 3-4, 3-5, 3-6, 3-7, 3-8, 3-9 ,3-10, 3-11, 3-12, 3-13, 3-14, 3-15, 3-16, 3-17, 3-18, 3-19, 3-20, 3–21, 3–22, 3–23, 3–24, 3–25, 3–26 ถึง ภาพที่ 3-27



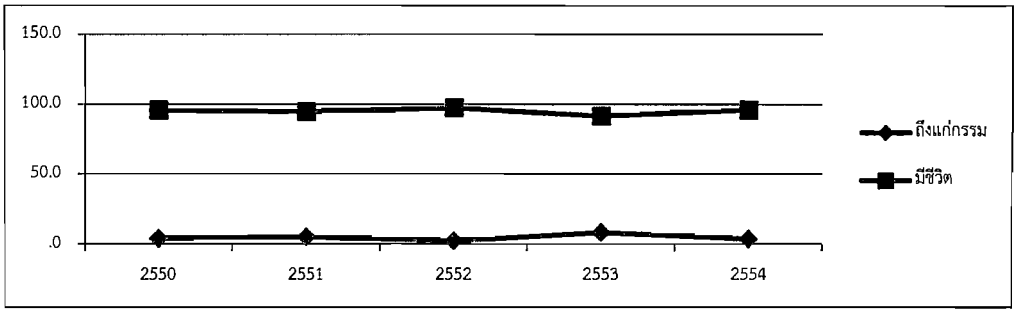
ภาพที่ 3-3 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเพศ

จากภาพที่ 3-3 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2552 ส่วนใหญ่เป็นข้อมูลนิสิตเพศหญิง และข้อมูลนิสิตปีการศึกษา 2553 ถึง 2554 ส่วนใหญ่เป็นข้อมูลนิสิตเพศชาย



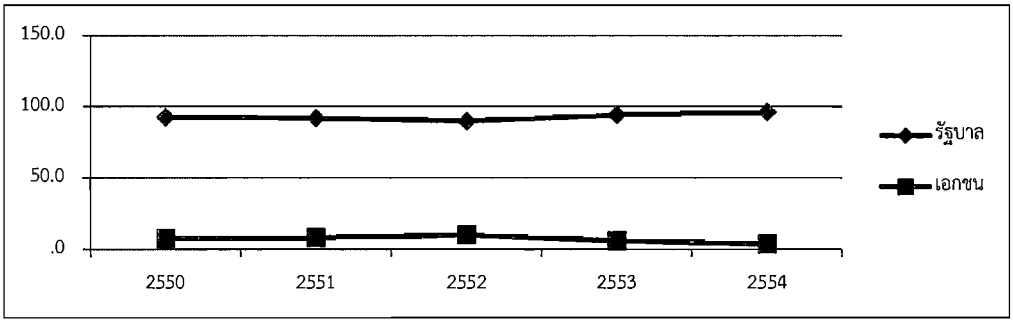
ภาพที่ 3-4 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพของบิดา

จากภาพที่ 3-4 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่มีสถานภาพของบิดาของนิสิตสถานภาพมีชีวิต



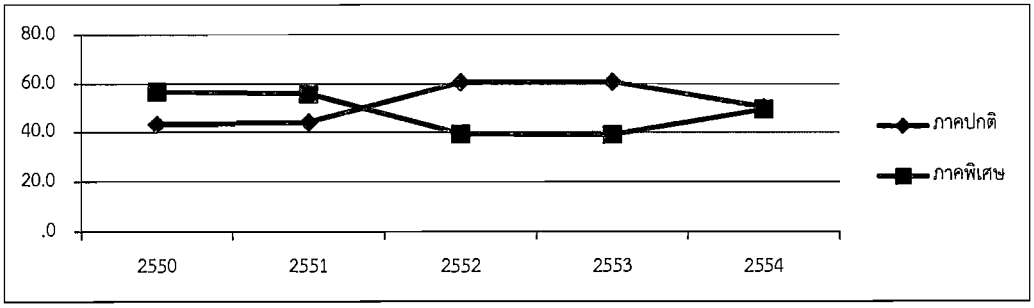
ภาพที่ 3-5 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพของมารดา

จากภาพที่ 3-5 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่สถานภาพของมารดาของนิสิตสถานภาพมีชีวิต



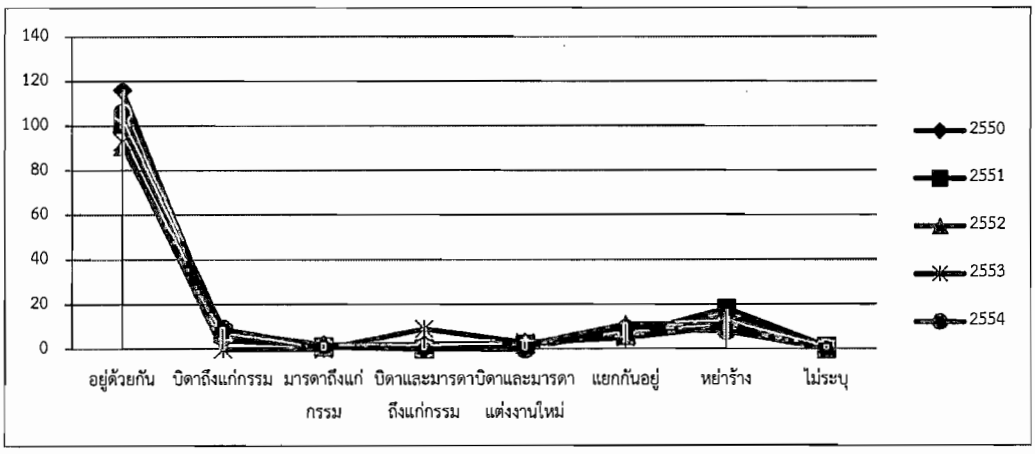
ภาพที่ 3-6 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านกลุ่มโรงเรียน (รัฐ/เอกชน)

จากภาพที่ 3-6 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่ได้รับการศึกษาอยู่กลุ่มโรงเรียนรัฐบาล



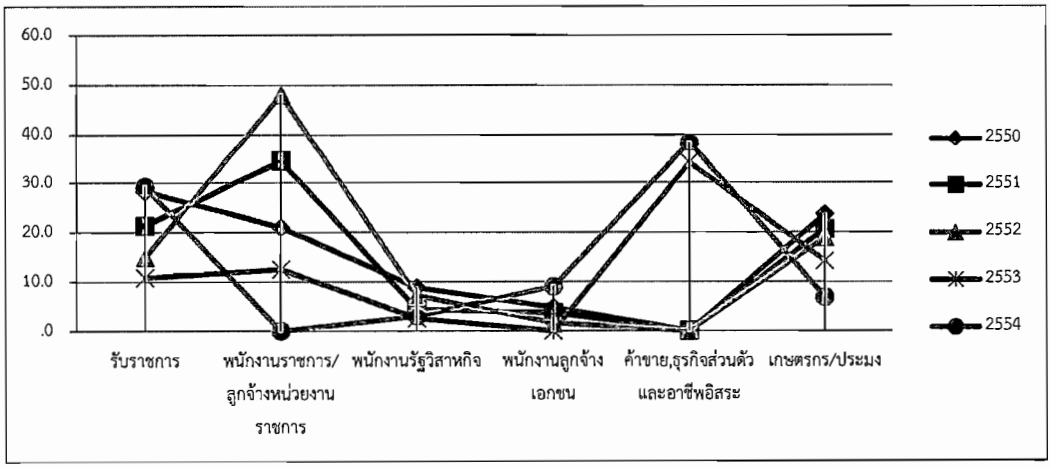
ภาพที่ 3-7 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านประเภทนิสิต (ปกติ/พิเศษ)

จากภาพที่ 3-7 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2551 ส่วนใหญ่เป็นนิสิตหลักสูตร 4 ปี ภาคพิเศษ และข้อมูลนิสิตปีการศึกษา 2552 ถึง 2553 ส่วนใหญ่เป็นนิสิตหลักสูตร 4 ปี ภาคปกติ และข้อมูลนิสิตปีการศึกษา 2554 ส่วนใหญ่เป็นนิสิตหลักสูตร 4 ปี ภาคปกติกับพิเศษเท่ากัน



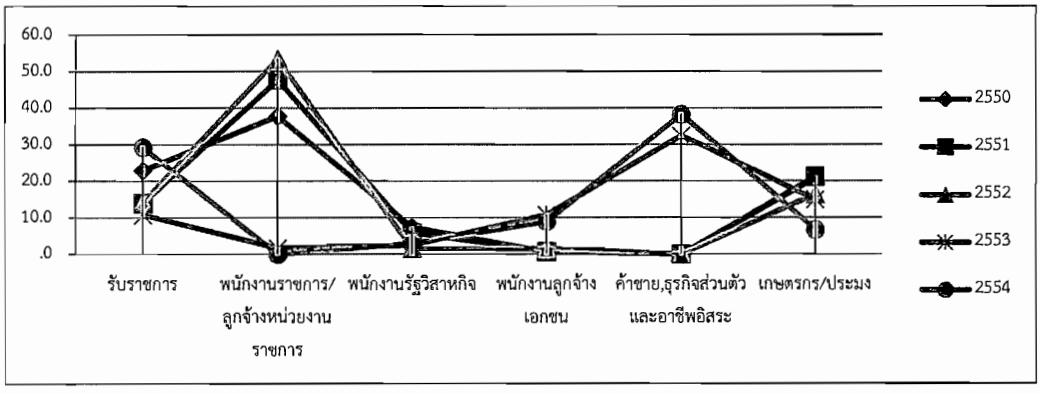
ภาพที่ 3-8 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านสถานภาพสมรสของบิดา มารดา

จากภาพที่ 3-8 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่สถานภาพสมรสของบิดามารดาของนิสิตมีสถานภาพอยู่ด้วยกัน



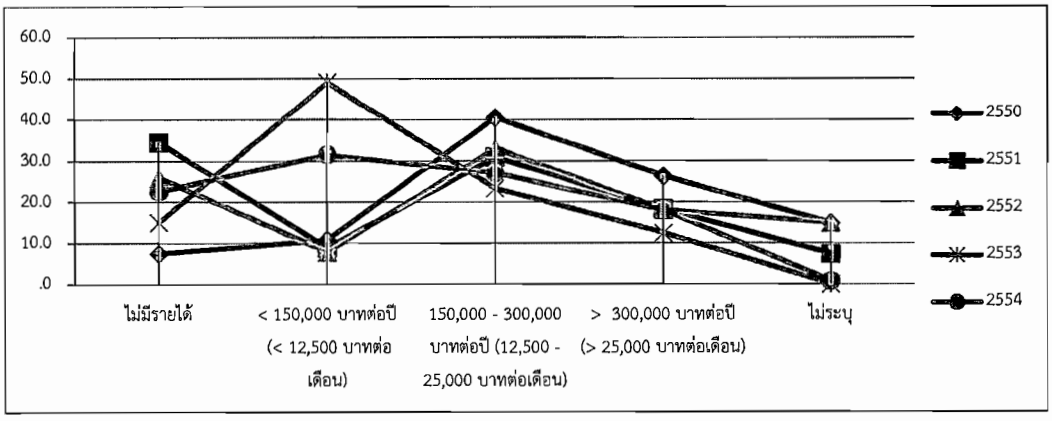
ภาพที่ 3-9 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านอาชีพของบิดา

จากภาพที่ 3-9 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ส่วนใหญ่มีบิดาประกอบอาชีพรับราชการ และข้อมูลนิสิตปีการศึกษา 2551 และ 2552 ส่วนใหญ่มีบิดาประกอบอาชีพพนักงานราชการ/ลูกจ้างหน่วยงานราชการ และข้อมูลนิสิตปีการศึกษา 2553 และ 2554 ส่วนใหญ่มีบิดาประกอบอาชีพค้าขาย, ธุรกิจส่วนตัวและอาชีพอิสระ ตามลำดับ



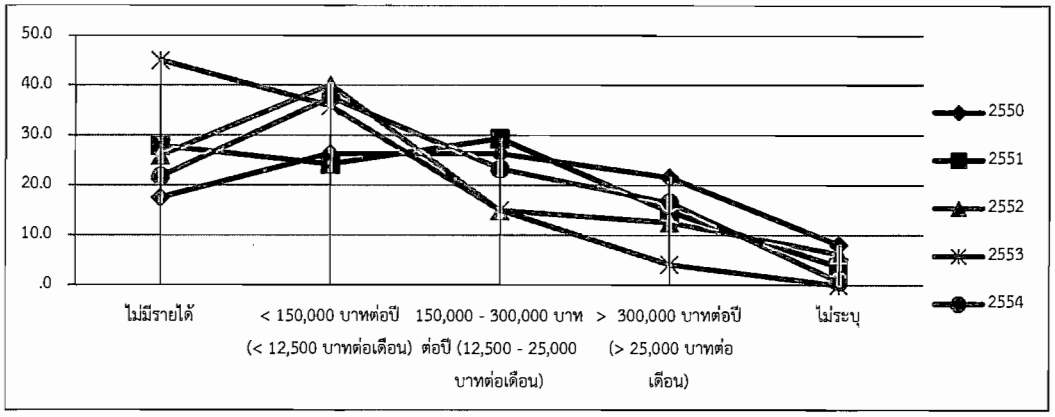
ภาพที่ 3-10 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านอาชีพของมารดา

จากภาพที่ 3-10 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2552 ส่วนใหญ่มีมารดาประกอบอาชีพพนักงานราชการ/ลูกจ้างหน่วยงานราชการ และข้อมูลนิสิตปีการศึกษา 2553 ส่วนใหญ่มีมารดาประกอบอาชีพค้าขาย,ธุรกิจส่วนตัวและอาชีพอิสระ และข้อมูลนิสิตปีการศึกษา 2554 ส่วนใหญ่มีมารดาประกอบอาชีพค้าขาย,ธุรกิจส่วนตัวและอาชีพอิสระ ตามลำดับ



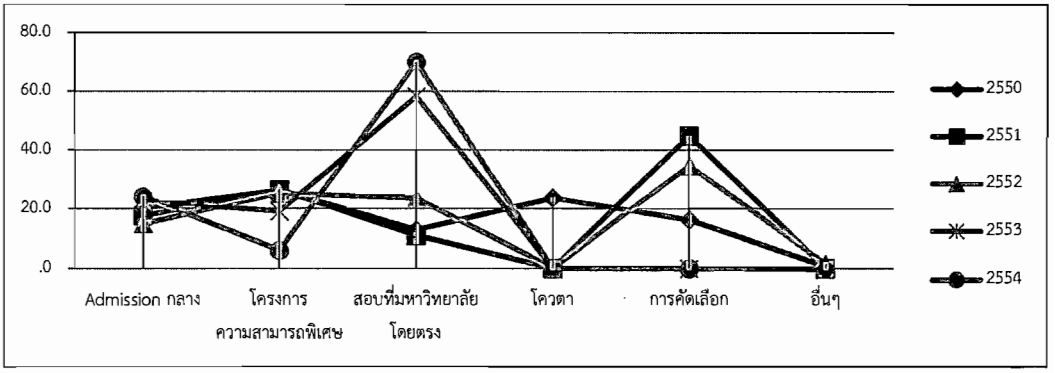
ภาพที่ 3-11 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านรายได้บิดา

จากภาพที่ 3-11 พบว่าข้อมูลนิสิตปีการศึกษา 2550 และ 2552 รายได้บิดาส่วนใหญ่มีรายได้ 150,000 - 300,000 บาทต่อปี (12,500 - 25,000 บาทต่อเดือน) และข้อมูลนิสิตปีการศึกษา 2551 รายได้บิดาส่วนใหญ่ไม่มีราย และข้อมูลนิสิตปีการศึกษา 2553 ถึง 2554 รายได้บิดาส่วนใหญ่มีรายได้น้อยกว่า 150,000 บาทต่อปี (< 12,500 บาทต่อเดือน) ตามลำดับ



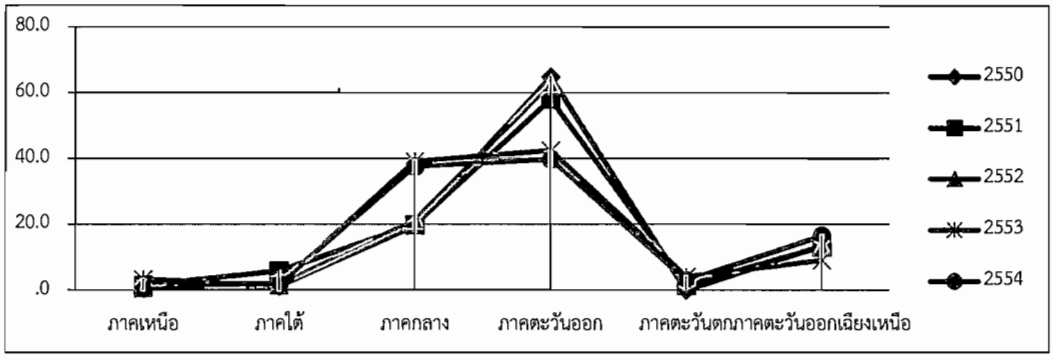
ภาพที่ 3-12 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านรายได้มารดา

จากภาพที่ 3-12 พบว่าข้อมูลนิสิตปีการศึกษา 2550 รายได้มารดาส่วนใหญ่มีรายได้ 150,000 - 300,000 บาทต่อปี (12,500 - 25,000 บาทต่อเดือน) และ น้อยกว่า 150,000 บาทต่อปี (< 12,500 บาทต่อเดือน) พบว่าข้อมูลนิสิตปีการศึกษา 2551 รายได้มารดาส่วนใหญ่มีรายได้ 150,000 - 300,000 บาทต่อปี (12,500 - 25,000 บาทต่อเดือน) พบว่าข้อมูลนิสิตปีการศึกษา 2552, 2553 และ2554 รายได้มารดาส่วนใหญ่มีรายได้ น้อยกว่า 150,000 บาทต่อปี (< 12,500 บาทต่อเดือน) ตามลำดับ



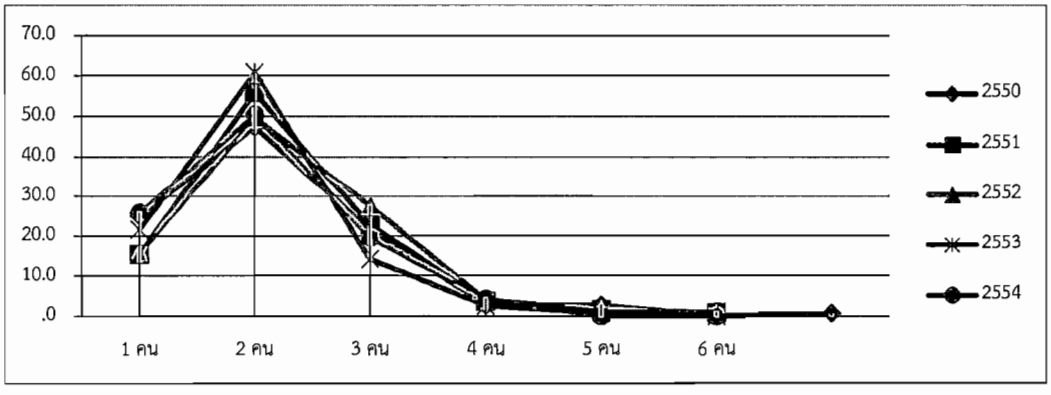
ภาพที่ 3-13 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านวิธีเข้าศึกษาของนิสิต

จากภาพที่ 3-13 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ส่วนใหญ่เข้ารับศึกษาโดยวิธีการคัดเลือกแบบโครงการความสามารถพิเศษ ข้อมูลนิสิตปีการศึกษา 2551และ 2552 ส่วนใหญ่เข้ารับศึกษาโดยวิธีการคัดเลือกแบบการสอบคัดเลือก ข้อมูลนิสิตปีการศึกษา 2553 และ2554 ส่วนใหญ่เข้ารับศึกษาโดยวิธีการสอบที่มหาวิทยาลัยโดยตรง ตามลำดับ



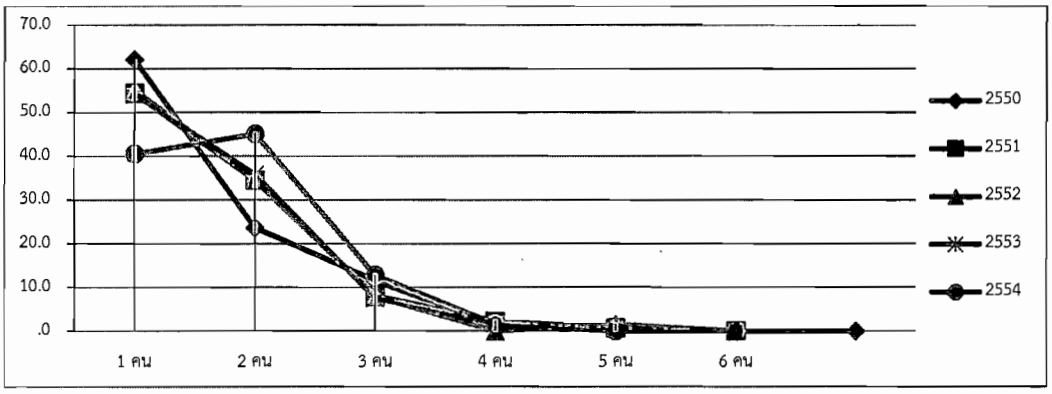
ภาพที่ 3-14 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านภูมิภาคที่อยู่อาศัย (ที่อยู่นิสิต)

จากภาพที่ 3-14 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่อาศัยอยู่ในภูมิภาคภาคตะวันออก รองลงมาคือภาคกลาง



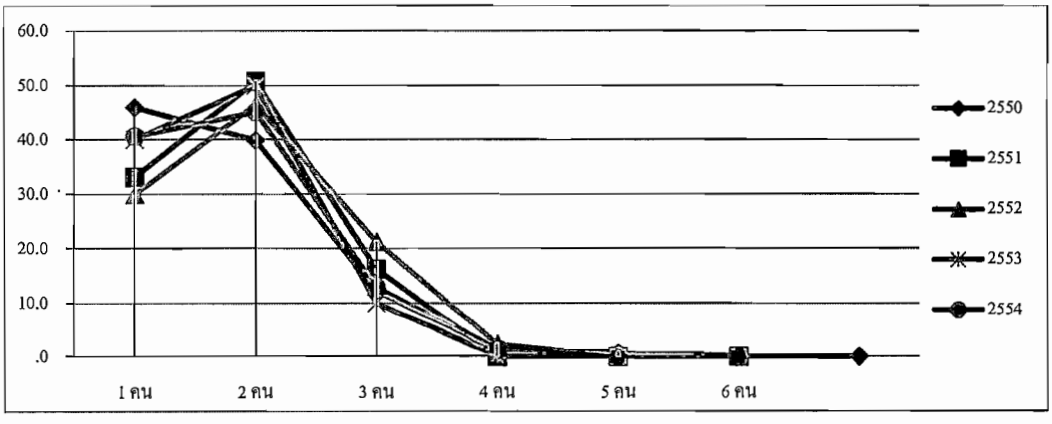
ภาพที่ 3-15 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านจำนวนพี่น้องทั้งหมด (รวมตัวเอง)

จากภาพที่ 3-15 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2554 ส่วนใหญ่มีจำนวนพี่น้องทั้งหมด (รวมตัวเอง) 2 คน



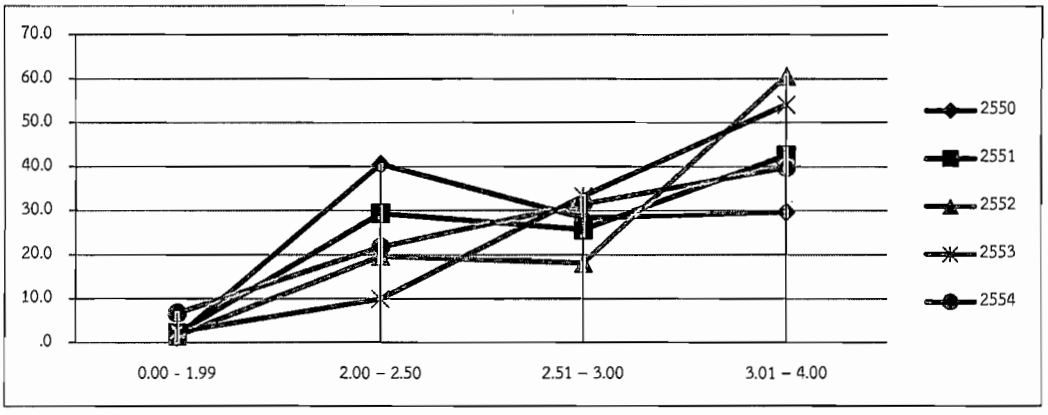
ภาพที่ 3-16 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเป็นบุตรคนที่

จากภาพที่ 3-16 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ถึง 2553 ส่วนใหญ่เป็นบุตรคนที่ 1 คน และข้อมูลนิสิตปีการศึกษา 2554 ส่วนใหญ่เป็นบุตรคนที่ 2 คน



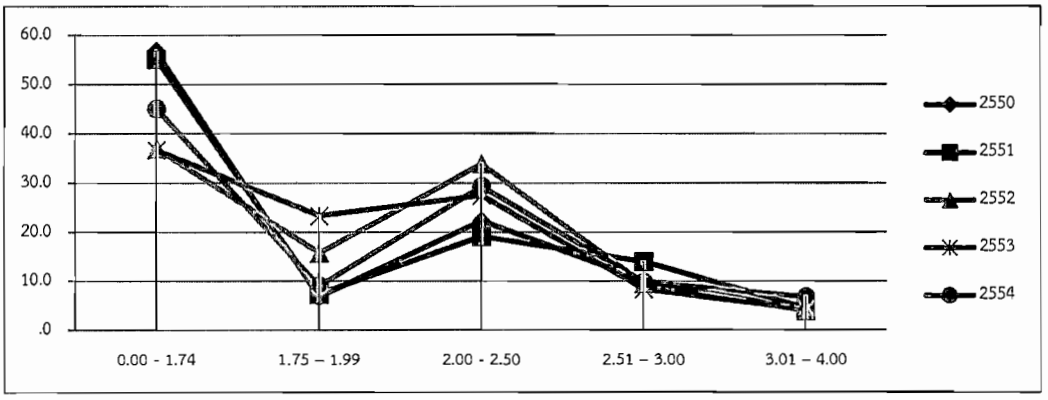
ภาพที่ 3-17 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านจำนวนพี่น้องที่กำลังศึกษาอยู่ (รวมตัวเอง)

จากภาพที่ 3-17 พบว่าข้อมูลนิสิตปีการศึกษา 2550 ส่วนใหญ่มีบุตรที่กำลังศึกษาจำนวน 1 คน และข้อมูลนิสิตปีการศึกษา 2551 ถึง 2554 ส่วนใหญ่มีบุตรที่กำลังศึกษาจำนวน 2 คน



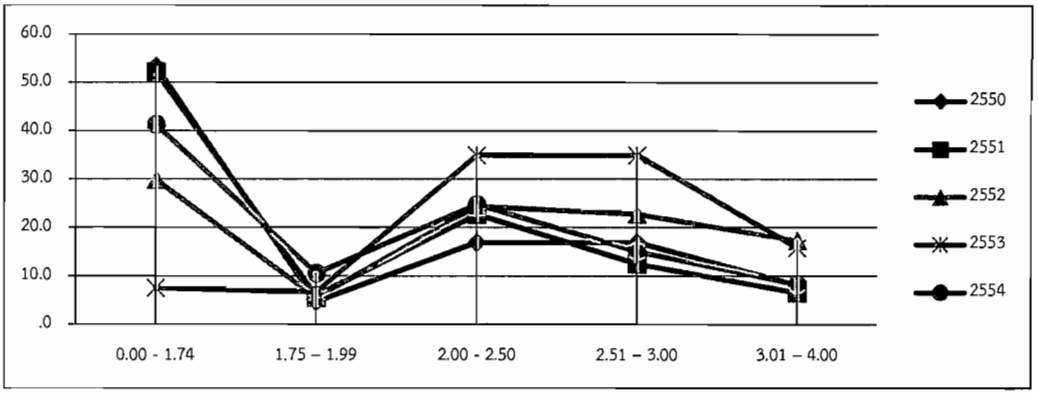
ภาพที่ 3-18 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านผลการเรียนเฉลี่ยสถานศึกษาเดิม

จากภาพที่ 3-18 พบว่าเกรดเฉลี่ยสถานศึกษาเดิมนิสิตส่วนใหญ่อยู่ในช่วงเกรดเกรด 3.01 – 4.00 แต่มีข้อมูลนิสิตปีการศึกษา 2550 เท่านั้นที่มีเกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา อยู่ในช่วงเกรด 2.00 - 2.50



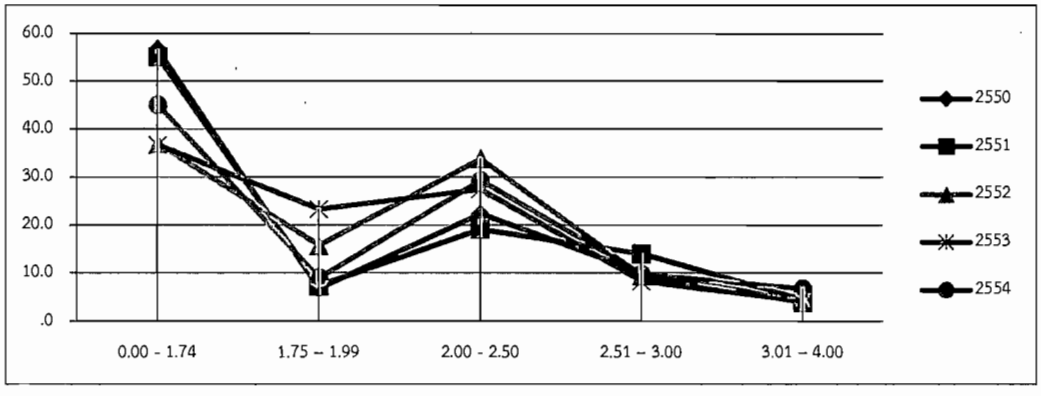
ภาพที่ 3-19 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดภาคเรียนที่ 1 ชั้นปีที่1

จากภาพที่ 3-19 พบว่า เกรดภาคเรียนที่ 1ชั้นปีที่ 1นิสิตส่วนใหญ่มีอยู่ในช่วงเกรด 0.00 - 1.74 รองลงมาคือ เกรด 2.00 - 2.50



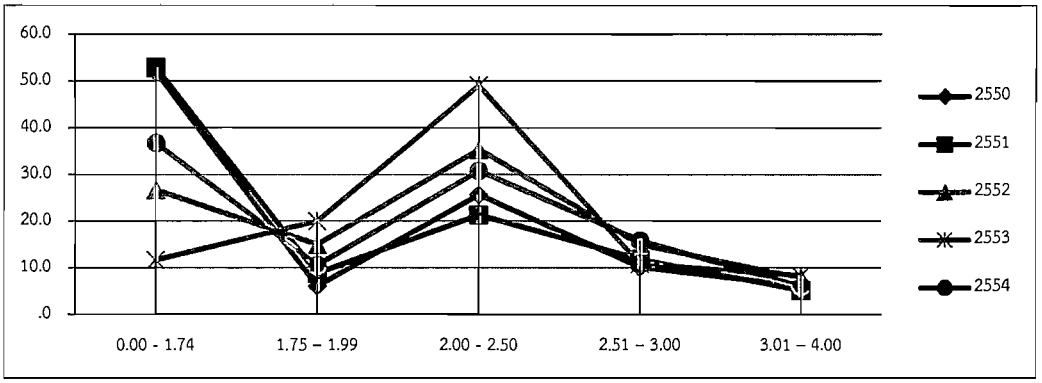
ภาพที่ 3-20 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดภาคเรียนที่ 2 ชั้นปีที่1

จากภาพที่ 3-20 พบว่าข้อมูลนิสิตปีการศึกษา 2550, 2551, 2552 และ 2554 มีเกรดภาคเรียนที่ 2 ชั้นปีที่ 1นิสิตส่วนใหญ่อยู่ในช่วง 0.00 - 1.74 รองลงมา คือ เกรด 2.00 - 2.50 แต่มีข้อมูลนิสิตปีการศึกษา 2553 เท่านั้นที่มีเกรดภาคเรียนที่ 1ชั้นปีที่ 1 อยู่ในช่วงเกรด 2.00 - 2.50 และ 2.51 - 3.00



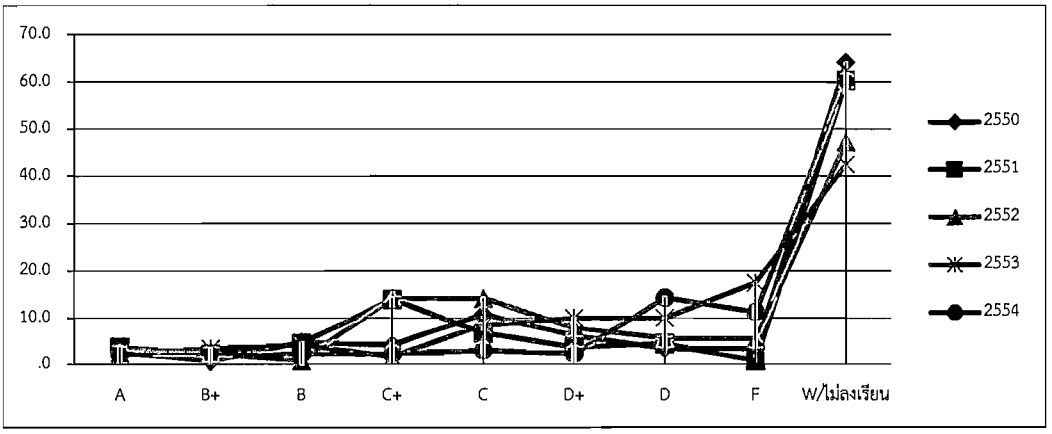
ภาพที่ 3-21 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่1

จากภาพที่ 3-21 พบว่า เกรดเฉลี่ยสะสมภาคเรียนที่ 1ชั้นปีที่ 1นิสิตส่วนใหญ่มีอยู่ในช่วงเกรด 0.00 - 1.74 รองลงมาคือ เกรด 2.00 - 2.50



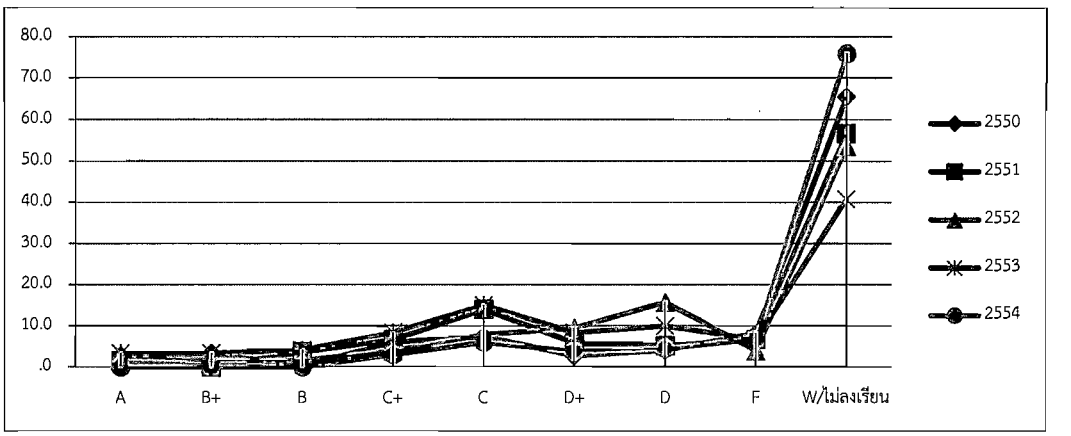
ภาพที่ 3-22 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่1

จากภาพที่ 3-22 พบว่าข้อมูลนิสิตปีการศึกษา 2550, 2551, 2552 และ 2554 มีเกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1นิสิตส่วนใหญ่อยู่ในช่วง 0.00 - 1.74 รองลงมา คือ เกรด 2.00 - 2.50 แต่มีข้อมูลนิสิตปีการศึกษา 2553 เท่านั้นที่มีเกรดภาคเรียนที่ 1ชั้นปีที่ 1 อยู่ในช่วงเกรด 2.00 - 2.50 รองลงมา คือ เกรด 1.75 - 1.99



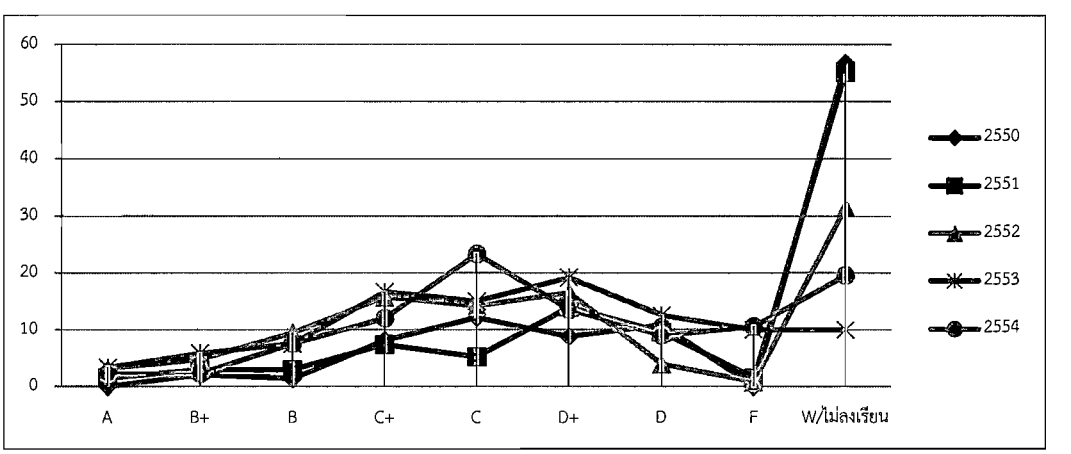
ภาพที่ 3-23 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาแคลคูลัส 1

จากภาพที่ 3-23 พบว่าข้อมูลปีการศึกษา 2550, 2551, 2552, 2553 และ 2554 มีเกรดรายวิชาแคลคูลัส 1เกรดนิสิตส่วนใหญ่ คือ W/ไม่ลงเรียน



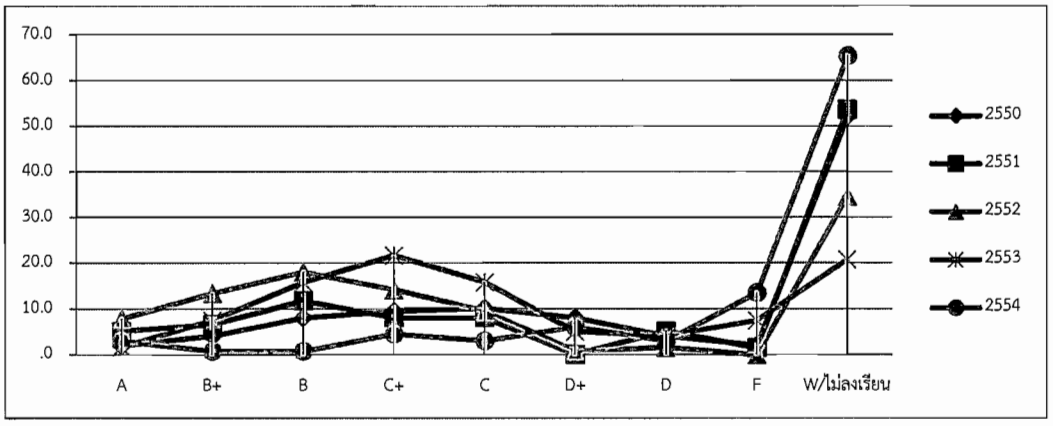
ภาพที่ 3-24 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาแคลคูลัส 2

จากภาพที่ 3-24 พบว่าข้อมูลปีการศึกษา 2550, 2551, 2552, 2553 และ 2554 มีเกรดรายวิชาแคลคูลัส 2 เกรดนิสิตส่วนใหญ่ คือ W/ไม่ลงเรียน



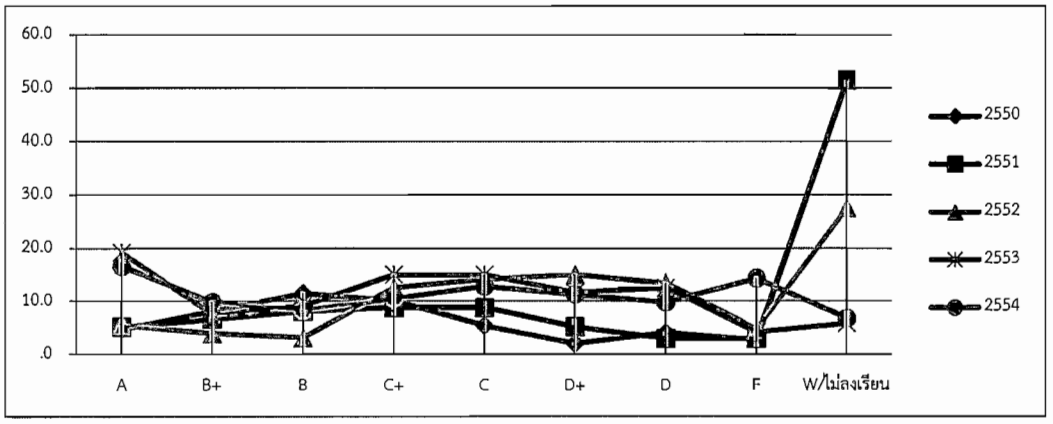
ภาพที่ 3-25 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาโปรแกรมมิ่ง 1

จากภาพที่ 3-25 พบว่าข้อมูลปีการศึกษา 2550, 2551 และ 2552 มีเกรดรายวิชาโปรแกรมมิ่งเกรด 1 นิสิตส่วนใหญ่ คือ W/ไม่ลงเรียน และข้อมูลปีการศึกษา 2553 มีเกรดรายวิชาโปรแกรมมิ่งเกรด 1 นิสิตส่วนใหญ่ คือ เกรด D+ และข้อมูลปีการศึกษา 2554 มีเกรดรายวิชาโปรแกรมมิ่ง 1 เกรดนิสิตส่วนใหญ่ คือ เกรด C



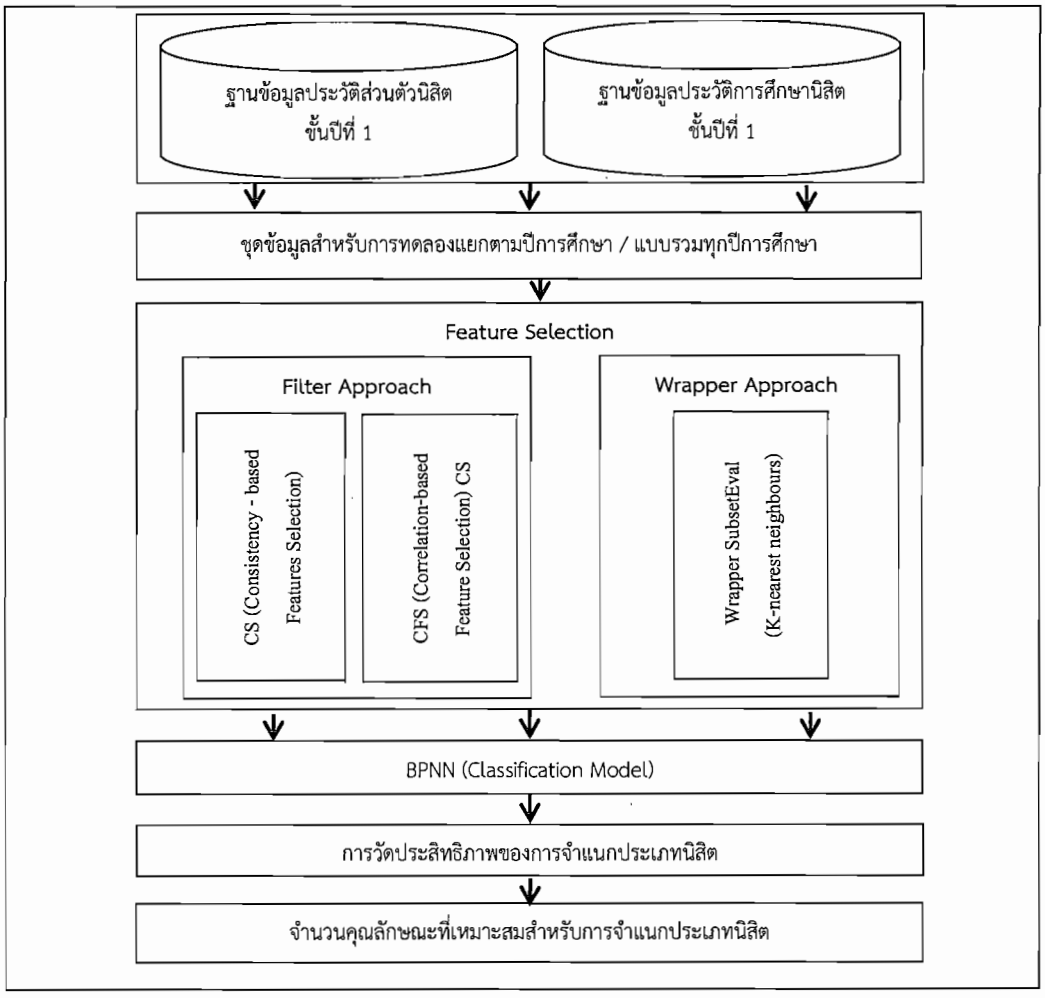
ภาพที่ 3-26 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาโปรแกรมมิ่ง 2

จากภาพที่ 3-26 พบว่าข้อมูลปีการศึกษา 2550, 2551, 2552 และ 2554 มีเกรดรายวิชาโปรแกรมมิ่ง 2 เกรดนิสิตส่วนใหญ่ คือ W/ไม่ลงเรียน และข้อมูลปีการศึกษา 2553 มีเกรดรายวิชาโปรแกรมมิ่ง 2 เกรดนิสิตส่วนใหญ่ คือ เกรด C+



ภาพที่ 3-27 กราฟเปรียบเทียบค่าร้อยละ (Percentage) ของปัจจัยด้านเกรดรายวิชาภาษาอังกฤษ

จากภาพที่ 3-27 พบว่าข้อมูลปีการศึกษา 2550, 2551 และ 2552 มีเกรดรายวิชาภาษาอังกฤษเกรดนิสิตส่วนใหญ่ คือ W/ไม่ลงเรียน และข้อมูลปีการศึกษา 2553 และ 2554 มีเกรดรายวิชาภาษาอังกฤษเกรดนิสิตส่วนใหญ่ คือ เกรด A



ภาพที่ 3-28 ตัวแบบ (Model) การคัดเลือกปัจจัยที่เหมาะสมสำหรับการจำแนกประเภทนิสิต

การวิเคราะห์และการสร้างแบบจำลองเพื่อการพยากรณ์โอกาสสำเร็จการศึกษา

1. การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยไม่มีการคัดเลือกปัจจัย

การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยผู้จัดทำงานนิพนธ์ได้ใช้วิธีการ ซึ่ง เป็นวิธีการที่ไม่มีการคัดเลือกปัจจัยของข้อมูลก่อนการจำแนกประเภทข้อมูล ทำการประมวลผลหาคำตอบเพียงอย่างเดียว อัลกอริทึมที่นำมาใช้ในการจำแนกประเภทข้อมูลนิสิตนี้ คือ ตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) โดยมีขั้นตอนดังนี้

1) นำข้อมูลประวัติส่วนตัวและข้อมูลประวัติการศึกษาของนิสิตที่ถูกเตรียมไว้ในรูปแบบที่สามารถทำการประมวลผลข้อมูลได้ ทำการแปลงไฟล์ข้อมูลให้อยู่ในรูปแบบของ CSV (Comma delimited) ซึ่งข้อมูลในแต่ละฟิลด์จะถูกคั่นด้วนเครื่องหมาย Comma

2) นำข้อมูลจากข้อ 1 มาทำการจำแนกประเภทข้อมูลด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) ด้วยโปรแกรม Weka version 3.6.1 หลังจากนั้นทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบประกอบด้วยข้อมูลสำหรับสอน (Training Data) และข้อมูลสำหรับทดสอบ (Test Data) โดยใช้วิธีการ 10 - Fold Cross Validation และวัดประสิทธิภาพในการจำแนกประเภทกลุ่มข้อมูลด้วยอัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

2. การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาแบบมีการคัดเลือกปัจจัย

การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาแบบมีการคัดเลือกปัจจัย เป็นการทดลองที่จะดำเนินการคัดเลือกปัจจัยของข้อมูลก่อนการนำไปการจำแนกประเภทนิสิตของข้อมูลนิสิตในแต่ละปีการศึกษา สำหรับเทคนิคที่ใช้ในการคัดเลือกปัจจัยของข้อมูลก่อนการนำไปการจำแนกประเภท ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach ซึ่งภายหลังการคัดเลือกปัจจัยที่มีผลต่อการสำเร็จการศึกษาแล้วเสร็จจะถูกนำไปใช้ในตัวแบบการจำแนกประเภทนิสิตซึ่งใช้ในการทดลองนี้ คือ ตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) โดยประเภทนิสิตแบ่งออกเป็น 3 ประเภท คือ 1) สำเร็จการศึกษาภายในระยะเวลา 4 ปี 2) สำเร็จการศึกษามากกว่า 4 ปี และ 3) ไม่สำเร็จการศึกษา เพื่อทำการวิเคราะห์เปรียบเทียบเทคนิคการคัดเลือกปัจจัยที่เหมาะสมและมีอิทธิพลต่อการสำเร็จการศึกษา ส่งผลต่อค่าความถูกต้องในการจำแนกกลุ่มข้อมูลที่ดีที่สุดโดยมีการทดลองดังนี้

การทดลองที่ 1: การทดลองเพื่อศึกษาว่าจำนวนคุณลักษณะที่เพิ่มขึ้นของปัจจัยนำเข้าของนิสิต จะส่งผลต่อการคัดเลือกปัจจัยที่เหมาะสมด้วยเทคนิคการคัดเลือกปัจจัยที่แตกต่างกันอย่างไร

สมมติฐาน1: จำนวนคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา มีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสม ซึ่งมีผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิต

สมมติฐาน2: เทคนิคการคัดเลือกปัจจัยที่แตกต่างกันส่งผลต่อการคัดเลือกปัจจัยและค่าความถูกต้องของการจำแนกประเภทนิสิต

ในการทดลองนี้เป็นการทดลองโดยนำข้อมูลนิสิตในแต่ละปีการศึกษา ได้แก่ 1) 2550 2) 2551 3) 2552 และ 4) 2550 - 2552 มาทำการคัดเลือกปัจจัยที่เหมาะสมก่อนการจำแนกประเภทนิสิตของข้อมูลนิสิตในแต่ละปีการศึกษาโดยเริ่มจากปัจจัยตั้งต้น 14 ปัจจัย และทำการเพิ่มจำนวนปัจจัยจากปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นข้อมูลเดิมก่อนนิสิตเข้าศึกษาในมหาวิทยาลัย แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ25 ปัจจัย ตามลำดับ ดังรายละเอียดในตารางที่ 3.4 และภาพที่ 3–2 และวิเคราะห์

เปรียบเทียบเทคนิคในการเลือกปัจจัยสามารถคัดเลือกปัจจัยที่เหมาะสม ซึ่งภายหลังการคัดเลือกปัจจัยด้วยเทคนิคต่าง ๆ เป็นที่เรียบร้อยแล้วจะถูกนำไปใช้ในการจำแนกประเภทนิสิตด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับเพื่อจำแนกประเภทนิสิต โดยมีขั้นตอนดังนี้

1) การเตรียมข้อมูลให้อยู่ในรูปแบบที่สามารถนำเข้าไปประมวลผลในอัลกอรึทึมโดยนำข้อมูลประวัติส่วนตัวและข้อมูลประวัติการศึกษาของนิสิตที่ถูกเตรียมไว้ในรูปแบบที่สามารถทำการประมวลผลข้อมูลได้ ทำการแปลงไฟล์ข้อมูลให้อยู่ในรูปแบบของ CSV (Comma delimited) ซึ่งข้อมูลในแต่ละฟิลด์จะถูกคั่นด้วนเครื่องหมาย Comma

2) การค้นหาคุณลักษณะที่เหมาะสม (Search Method) เป็นวิธีการค้นหา Attribute ที่เหมาะสมโดยใช้ Best - first search ในการค้นหาชุดของคุณลักษณะ

3) การประเมินค่าของคุณลักษณะที่ถูกคัดเลือก (Evaluator) เป็นวิธีการประเมินค่าของ Attribute ที่ถูกคัดเลือกมาจาก Search Method โดยในเทคนิคการวิเคราะห์ปัจจัย 3 เทคนิค ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach

4) การนำปัจจัยที่ได้รับการคัดเลือกจากข้อ 3 มาจำแนกประเภทนิสิตโดยโปรแกรม Weka version 3.6.1 ด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) โดยทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบประกอบด้วยข้อมูลสำหรับสอน (Training Data) และข้อมูลสำหรับทดสอบ (Test Data) ด้วยวิธีการ 10 - Fold Cross Validation และวัดประสิทธิภาพในการจำแนกประเภทกลุ่มข้อมูลด้วยอัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

5) ทำการทดลองเพิ่มจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา ดังแสดงในตารางที่ 3-5 และดำเนินการตามขั้นตอน 2 ถึงขั้นตอนที่ 5 จากนั้นเปรียบเทียบความถูกต้องของการจำแนกประเภท เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษาที่เพิ่มขึ้น ส่งผลต่อทำให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตเพิ่มขึ้นหรือไม่ และเทคนิคการคัดเลือกปัจจัยเทคนิคใดสามารถการคัดเลือกปัจจัยที่เหมาะสม สามารถจำแนกประเภทนิสิตในแต่ละปีการศึกษาได้ดีที่สุด

การทดลองที่ 2: การทดลองเพื่อค้นหาชุดคุณลักษณะของปัจจัยนำเข้าที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา

สมมติฐาน : มีชุดปัจจัยที่เหมาะสมที่สามารถจำแนกประเภทนิสิตได้ทุกปีการศึกษา ทำให้ค่าความถูกต้องในการจำแนกประเภทนิสิตเฉลี่ยสูงสุด

โดยมีขั้นตอนการดำเนินการดังต่อไปนี้

1) การเตรียมข้อมูลให้อยู่ในรูปแบบที่สามารถนำเข้าไปประมวลผลในอัลกอริทึมโดยนำข้อมูลประวัติส่วนตัวและข้อมูลประวัติการศึกษาของนิสิตที่ถูกเตรียมไว้ในรูปแบบที่สามารถทำการประมวลผลข้อมูลได้ ทำการแปลงไฟล์ข้อมูลให้อยู่ในรูปแบบของ CSV (Comma delimited) ซึ่งข้อมูลในแต่ละฟิลด์จะถูกคั่นด้วนเครื่องหมาย Comma

2) การค้นหาคุณลักษณะที่เหมาะสม (Search Method) เป็นวิธีการค้นหา Attribute ที่เหมาะสมโดยใช้ Best - first search ในการค้นหาชุดของคุณลักษณะที่เหมาะสม

3) การประเมินค่าของคุณลักษณะที่ถูกคัดเลือก (Evaluator) เป็นวิธีการประเมินค่าของ Attribute ที่ถูกคัดเลือกมาจาก Search Method โดยในการวิเคราะห์ปัจจัย 3 เทคนิค ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach

4) การนำชุดปัจจัยที่ได้รับการคัดเลือกในแต่ละปีการศึกษานำมาใช้ในการจำแนกประเภทกลุ่มนิสิตด้วยโปรแกรม Weka version 3.6.1 โดยตัวแบบโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ (BPNN) กับชุดข้อมูลนิสิตในแต่ละปีการศึกษา ได้แก่ 1) 2550 2) 2551 3) 2552 และ 4) 2550 - 2552 โดยทำการแบ่งกลุ่มข้อมูลเพื่อทดสอบประกอบด้วยข้อมูลสำหรับสอน (Training Data) และข้อมูลสำหรับทดสอบ (Test Data) ด้วยวิธีการ 10 - Fold Cross Validation และวัดประสิทธิภาพในการจำแนกประเภทกลุ่มข้อมูลด้วยอัตราความถูกต้องของการจำแนกประเภท (Correct Classification Rate: CCR)

5) หลังจากนั้นทำการทดลองเพิ่มจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษาเป็น 16 ปัจจัย 20 ปัจจัย และ 25 ปัจจัย ดังแสดงในตารางที่ 3-5 และดำเนินการตามขั้นตอน 2 – 4

6) เปรียบเทียบชุดปัจจัยที่ถูกคัดเลือกจากชุดข้อมูลนิสิตในแต่ละปีการศึกษาว่าชุดปัจจัยใดที่สามารถแล้วทำให้ค่าความถูกต้องของการจำแนกประเภทเฉลี่ยสูงสุดของแต่ละเทคนิค

7) เปรียบเทียบปัจจัยได้รับการคัดเลือกทั้ง 3 เทคนิค เพื่อหาเทคนิคการคัดเลือกปัจจัยใดสามารถการคัดเลือกปัจจัยที่เหมาะสมสำหรับใช้ในการจำแนกกลุ่มนิสิตได้ทุกปีการศึกษา

7.1 นำชุดปัจจัยที่ได้รับการคัดเลือกทั้ง 3 เทคนิค มาสร้างตัวแบบการพยากรณ์การสำเร็จการศึกษา โดยมาทำการทดลองกับข้อมูลของนิสิตที่เข้าศึกษาในปีการศึกษา 2550, 2551, 2552 และ 2550 - 2552 ทำการเปรียบเทียบประสิทธิภาพค่าความถูกต้องของแบบจำลองการ

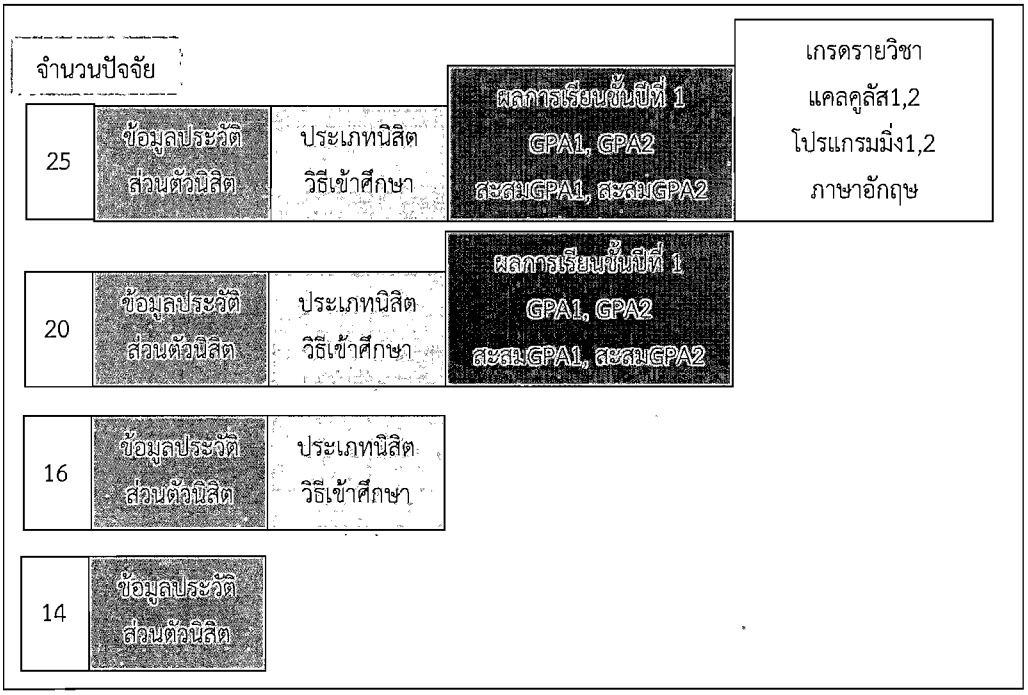
จำแนกประเภทนิสิต เทคนิคใดที่ให้ค่าความถูกต้องของแบบจำลองการจำแนกประเภทนิสิตเฉลี่ยสูงที่สุด

7.2 นำชุดปัจจัยที่ได้รับการคัดเลือกทั้ง 3 เทคนิคมาทดสอบกับข้อมูล Unseen Data ซึ่งเป็นข้อมูลนิสิตที่เข้าศึกษาในปีการศึกษา 2553 และปีการศึกษา 2554 โดยแบ่งชุดข้อมูลสำหรับการทดสอบด้วยวิธี 10 - Fold Cross Validation และวัดประสิทธิภาพด้วยค่าเฉลี่ยความถูกต้องของการจำแนกกลุ่ม (Correct Classification Rate: CCR) ทำการเปรียบเทียบประสิทธิภาพค่าความถูกต้องของแบบจำลองการจำแนกประเภทนิสิต เทคนิคใดที่ให้ค่าความถูกต้องของแบบจำลองการจำแนกประเภทนิสิตสูงที่สุด

จากการดำเนินการทอลองที่ 1 และการทดลองที่ 2 ผู้จัดทำงานนิพนธ์ได้ตัวแบบ (Model) สำหรับการคัดเลือกปัจจัยที่เหมาะสมในการจำแนกประเภทนิสิต ซึ่งผู้สนใจอาจนำไปศึกษาและประยุกต์ใช้กับข้อมูลนิสิตที่มีคุณลักษณะ (Feature) และค่า (Value) ของข้อมูลแตกต่างจากที่ผู้จัดทำงานนิพนธ์ใช้ในการศึกษา ดังรูปภาพ 3-28

บทที่ 4
ผลการวิเคราะห์ข้อมูล

เนื้อหาในบทนี้กล่าวถึงผลการทดลองวิเคราะห์เปรียบเทียบเทคนิคการคัดเลือกปัจจัยที่เหมาะสมสำหรับใช้ในการพยากรณ์โอกาสการสำเร็จการศึกษาของนิสิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยการสารสนเทศ ด้วยเทคนิค 1) เทคนิค CS (Consistency-based Features Selection) 2) เทคนิค CFS (Correlation-based Features Selection) และ 3) เทคนิค Wrapper Approach โดยแสดงจะให้เห็นถึงผลการคัดเลือกปัจจัยเมื่อจำนวนและคุณลักษณะเมื่อมีจำนวนที่เพิ่มขึ้นส่งผลให้ค่าความถูกต้องของการจำแนกประเภทนิสิตอย่างไร ซึ่งจะได้นำปัจจัยที่ได้รับการคัดเลือกไปพัฒนาเป็นแบบจำลองสำหรับการพยากรณ์ผลการสำเร็จการศึกษาของนิสิตที่เหมาะสม ประกอบไปด้วยผลการทดลองดังต่อไปนี้



ภาพที่ 4-1 ปัจจัยที่ใช้ในการทดลอง

ภาพที่ 4–1 แสดงปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา จากปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นข้อมูลประวัติส่วนตัวของนิสิตก่อนนิสิตเข้าศึกษาในมหาวิทยาลัย ได้แก่ เพศ, สถานภาพของบิดา, สถานภาพของมารดา, สถานภาพสมรสของบิดามารดา, อาชีพของบิดา, อาชีพของมารดา, รายได้ของ

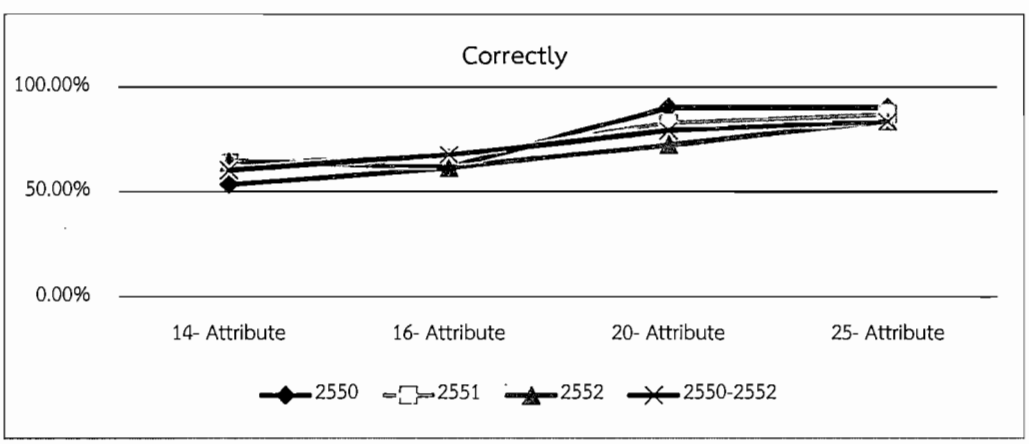
บิดา, รายได้ของมารดา, ภาคของสถานศึกษาเดิม,ประเภทของสถานบันศึกษาเดิม(รัฐ/เอกชน), (ที่อยู่นิสิต), จำนวนพี่น้องทั้งหมด (รวมตัวเอง), บุตรคนที่, บุตรคนที่ที่กำลังศึกษา, เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย, 20 ปัจจัย และ 25 ปัจจัย ดังแสดงในภาพที่ 4-1

การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยไม่มีการคัดเลือกปัจจัย

การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาโดยไม่มีการคัดเลือกปัจจัยเป็นวิธีการที่ไม่มีการคัดเลือกปัจจัยของข้อมูลก่อนการจำแนกประเภทข้อมูล (Classification) แต่มีการเพิ่มจำนวนและคุณลักษณะที่เพิ่มขึ้น โดยเปรียบเทียบถึงผลลัพธ์ค่าความถูกต้องของการจำแนกในแต่ละชุดข้อมูลที่ใช้ในการทดสอบ มีผลการทดลองดังตารางที่ 4–1

ตารางที่ 4-1 ตารางแสดงค่าความถูกต้องของการจำแนกข้อมูลนิสิต เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษาที่แตกต่างกัน

เทคนิค/DATA	Correctly			
	14 - Attribute	16 - Attribute	20 - Attribute	25 - Attribute
2550	53.38%	61.49%	90.54%	85.81%
2551	63.97%	66.18%	83.09%	87.50%
2552	64.57%	61.42%	72.44%	83.46%
2550-2552	60.34%	67.64%	79.08%	83.70%
ค่าเฉลี่ย	62.45%	64.53%	75.76%	83.58%



ภาพที่ 4-2 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตเมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตเพิ่มขึ้น

จากตารางที่ 4-1 จากผลการทดลองโดยการเพิ่มจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา จากปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นข้อมูลประวัติส่วนตัวของนิสิตก่อนนิสิตเข้าศึกษาในมหาวิทยาลัย แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ 25 ปัจจัย ตามลำดับ ดังแสดงภาพที่ 4-1 พบว่าจำนวนและคุณลักษณะปัจจัยนำเข้าของข้อมูลนิสิตของปีการศึกษา 2550 ข้อมูลนิสิตของปีการศึกษา 2551 ข้อมูลนิสิตของปีการศึกษา 2552 และข้อมูลนิสิตของปีการศึกษา 2550 - 2552 ที่เพิ่มขึ้นเมื่อนำมาปัจจัยดังกล่าวไปจำแนกข้อมูลนิสิตมีผลทำให้ค่าความถูกต้องของการจำแนกประเภทนิสิตที่เพิ่มสูงขึ้นตาม

การวิเคราะห์และพยากรณ์โอกาสการสำเร็จการศึกษาแบบมีการคัดเลือกปัจจัย

การทดลองที่ 1: การทดลองเพื่อศึกษาว่าจำนวนคุณลักษณะที่เพิ่มขึ้นของปัจจัยนำเข้าของนิสิต จะส่งผลต่อการคัดเลือกปัจจัยที่เหมาะสมด้วยเทคนิคการคัดเลือกปัจจัยที่แตกต่างกันอย่างไร

สมมติฐาน1: จำนวนคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา มีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสม ซึ่งมีผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิต

สมมติฐาน2: เทคนิคการคัดเลือกปัจจัยที่แตกต่างกันส่งผลต่อการคัดเลือกปัจจัยและค่าความถูกต้องของการจำแนกประเภทนิสิต

ผลการทดลองที่ 1 – 1 จำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา มีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสม ซึ่งมีผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิต

ตารางที่ 4-2 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2550

ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2550-14	INCMO	INCMO	STATUSFA INCFA INCMO STATUSFM TYPESCHOOL ENTRYGPA

ตารางที่ 4-2 (ต่อ)

ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2550-16	INCMO PROGRAM BYTEDES	INCMO PROGRAM	NUMSON SON STATUSFM TYPESCHOOL ENTRYGPA PROGRAM
2550-20	GPA-2 ACC-GPA-2	INCMO PROGRAM GPA-1 GPA-2 ACC-GPA-2	ENTRYGPA ACC-GPA-2
2550-25	PRO2	INCMO BYTEDES ACC-GPA-2 Calculus1 PRO1 PRO2 ENG1	ENTRYGPA ACC-GPA-2

จากตารางที่ 4-2 แสดงถึงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกจากชุดข้อมูลนิสิตปีการศึกษา 2550 ด้วยเทคนิคการคัดเลือกปัจจัย ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach โดยในแต่ละชุดข้อมูลนิสิตมีปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นปัจจัยส่วนตัวของนิสิต แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ25 ปัจจัย ตามลำดับ พบว่า จำนวนปัจจัยตั้งต้น 14 ปัจจัยโดยใช้เทคนิค CS, CFS และWrapper การคัดเลือกปัจจัยนั้น พบว่า มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ รายได้ของมารดา และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 16 ปัจจัย พบว่าทั้ง 3 เทคนิคมีการเลือกปัจจัยที่คล้ายคลึงกัน คือ ประเภทนิสิต(ปกติ/พิเศษ) และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 20 และ 25 ปัจจัย การคัดเลือกปัจจัยโดยเทคนิค Wrapper ในการคัดเลือกปัจจัย นั้น พบว่ามีการคัดเลือกปัจจัยตัวเดียวกัน คือเกรดเฉลี่ย

เดิมก่อนเข้ารับการศึกษา และเกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 และการคัดเลือกปัจจัยด้วยเทคนิค CS และ เทคนิค CFS มีการเลือกปัจจัยที่แตกต่างกันออกไป

ตารางที่ 4-3 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2551

ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2551-14	INCFA ENTRYGPA	INCFA INCMO ENTRYGPA	INCFA TYPESCHOOL ENTRYGPA
2551-16	INCFA INCMO ENTRYGPA BYTEDES	INCFA INCMO ENTRYGPA PROGRAM	INCFA TYPESCHOOL ENTRYGPA
2551-20	INCFA INCMO ENTRYGPA GPA-1 GPA-2	PROGRAM GPA-1 GPA-2	GPA-2
2551-25	PRO2 ENG1	BYTEDES Calculus2 PRO2 ENG1	INCMO PRO2

จากตารางที่ 4-3 แสดงถึงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกจากชุดข้อมูลนิสิตปีการศึกษา 2551 ด้วยเทคนิคการคัดเลือกปัจจัย ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach โดยในแต่ละชุดข้อมูลนิสิตมีปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นปัจจัยส่วนตัวของนิสิต แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ25 ปัจจัย ตามลำดับ พบว่าเมื่อจำนวนปัจจัยตั้งต้น 14 และ 16 ปัจจัย โดยใช้เทคนิค CS, CFS และWrapper ในการคัดเลือกปัจจัยนั้นทำให้เกิดการเลือกปัจจัยที่คล้ายคลึงกัน คือ รายได้ของบิดาและเกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 20 ปัจจัย พบว่าการใช้เทคนิค CS, CFS และWrapper ในการคัดเลือกปัจจัยทำให้เกิดการเลือกปัจจัยที่

คล้ายคลึงกัน คือ เกรดภาคเรียนที่ 2 ชั้นปีที่ 1และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 25 พบว่าการคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ เกรดรายวิชาโปรแกรมมิ่ง 2

ตารางที่ 4 - 4 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2552

ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2552-14	PROVINCE INCMO ENTRYGPA	PROVINCE INCFA INCMO ENTRYGPA	SON INCMO ENTRYGPA
2552-16	PROVINCE INCMO ENTRYGPA	PROVINCE INCFA INCMO ENTRYGPA	SON INCMO ENTRYGPA
2552-20	INCFA INCMO ENTRYGPA PROGRAM GPA-1 GPA-2 ACC-GPA-2	INCFA ENTRYGPA PROGRAM GPA-1 GPA-2 ACC-GPA-2	STATUSFA STATUSFM PROGRAM GPA-1 GPA-2
2552-25	PROVINCE PROGRAM GPA-1 Calculus1 Calculus2 PRO1 PRO2	ENTRYGPA PROGRAM GPA-2 Calculus1 Calculus2 PRO1 PRO2	NUMSON OCCFA PROGRAM Calculus1 PRO2

จากตารางที่ 4-4 แสดงถึงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกจากชุดข้อมูลนิสิตปีการศึกษา 2552 ด้วยเทคนิคการคัดเลือกปัจจัย ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค

Wrapper Approach โดยในแต่ละชุดข้อมูลนิสิตมีปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นปัจจัยส่วนตัวของนิสิต แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ25 ปัจจัย ตามลำดับ พบว่า จำนวนปัจจัยตั้งต้น 14 และ 16 ปัจจัย การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ รายได้ของมารดาและเกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 20 ปัจจัย พบว่า การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ ประเภทนิสิต(ปกติ/พิเศษ) เกรดภาคเรียนที่ 1 ชั้นปีที่ 1 และ เกรดภาคเรียนที่ 2 ชั้นปีที่ 1 และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 25 ปัจจัย พบว่าการคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ ประเภทนิสิต(ปกติ/พิเศษ) เกรดรายวิชาแคลคูลัส 1 และ เกรดรายวิชาโปรแกรมมิ่ง 2

ตารางที่ 4-5 แสดงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกสำหรับข้อมูลนิสิตปีการศึกษา 2550 - 2552

ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2550-2552 14	INCFA INCMO ENTRYGPA	INCFA ENTRYGPA	INCFA INCMO STATUSMO ENTRYGPA
2550-2552 16	INCFA INCMO ENTRYGPA PROGRAM BYTEDES	INCFA ENTRYGPA PROGRAM	INCFA STATUSFM ENTRYGPA PROGRAM BYTEDES
2550-2552 20	INCFA INCMO ENTRYGPA PROGRAM BYTEDES GPA-1 GPA-2 ACC-GPA-2	PROGRAM GPA-1 GPA-2 ACC-GPA-2	PROVINCE GPA-1 ACC-GPA-2

ตารางที่ 4-5 (ต่อ)

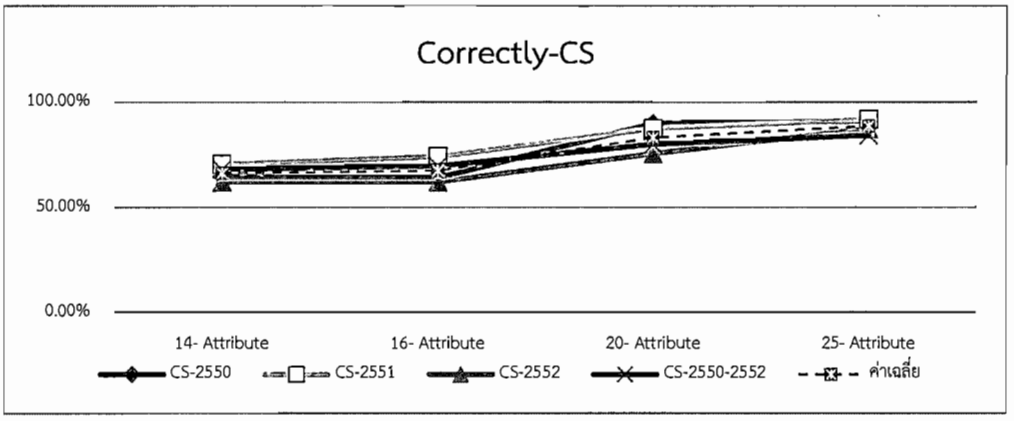
ปีการศึกษา / จำนวนปัจจัยนำเข้า	เทคนิคการคัดเลือกปัจจัย		
	CS - Attribute	CFS - Attribute	Wrapper - Attribute
2550-2552 25	INCFA INCMO ENTRYGPA PROGRAM GPA-1 Calculus1 Calculus2 PRO1 PRO2 ENG1	BYTEDES GPA-1 GPA-2 Calculus1 Calculus2 PRO1 PRO2 ENG1	STATUSFA PRO2

จากตารางที่ 4-2 แสดงถึงปัจจัยและจำนวนปัจจัยที่ถูกคัดเลือกจากชุดข้อมูลนิสิตปีการศึกษา 2550-2552 ด้วยเทคนิคการคัดเลือกปัจจัย ได้แก่ 1) เทคนิค CS (Consistency - based Features Selection) 2) เทคนิค CFS (Correlation - based Features Selection) และ 3) เทคนิค Wrapper Approach โดยในแต่ละชุดข้อมูลนิสิตมีปัจจัยตั้งต้น 14 ปัจจัย ซึ่งเป็นปัจจัยส่วนตัวของนิสิต แล้วทำการเพิ่มจำนวนปัจจัยซึ่งเป็นปัจจัยที่เกิดขึ้นภายหลังจากได้เข้ามาศึกษาแล้ว เป็นจำนวน 16 ปัจจัย 20 ปัจจัย และ25 ปัจจัย ตามลำดับ พบว่า จำนวนปัจจัยตั้งต้น 14 การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ รายได้ของบิดาและเกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 16 การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ รายได้ของบิดา เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา และประเภทนิสิต(ปกติ/พิเศษ) และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 20 การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือ เกรดภาคเรียนที่ 1 ชั้นปีที่ 1 และ เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 และเมื่อเพิ่มจำนวนปัจจัยตั้งต้น 25 การคัดเลือกปัจจัยด้วยเทคนิค CS, CFS และWrapper นั้น มีการเลือกปัจจัยที่คล้ายคลึงกัน คือเกรดรายวิชาโปรแกรมมิ่ง 2

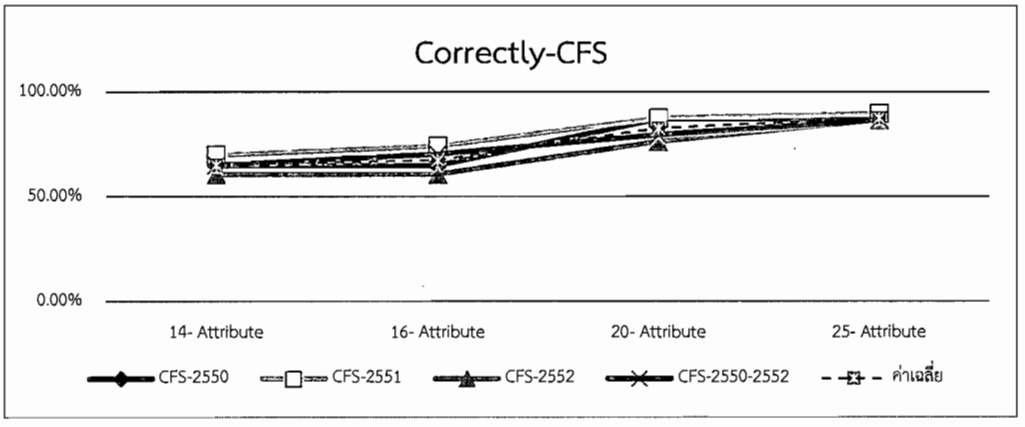
ตารางที่ 4-6 ตารางแสดงค่าความถูกต้องของการจำแนกข้อมูลนิสิตแต่ละปีการศึกษา เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษา และการใช้เทคนิคการคัดเลือกปัจจัยที่แตกต่างกัน

เทคนิค/DATA	BPNN			
	Correct Classification Rate: CCR			
	14-Attribute	16-Attribute	20-Attribute	25-Attribute
CS-2550	64.86%	64.19%	89.86%	**91.22%**
CS-2551	70.59%	74.26%	87.50%	**91.91%**
CS-2552	62.20%	62.20%	75.59%	**87.40%**
CS-2550-2552	68.13%	69.83%	80.05%	84.43%
ค่าเฉลี่ย	66.45%	67.62%	83.25%	88.74%
CFS-2550	64.86%	64.86%	87.16%	87.16%
CFS-2551	69.85%	74.26%	87.50%	89.71%
CFS-2552	60.63%	60.63%	76.38%	86.61%
CFS-2550-2552	65.21%	70.32%	79.56%	**86.86%**
ค่าเฉลี่ย	65.14%	67.52%	82.65%	87.59%
wrapper-2550	65.54%	72.30%	91.22%	91.22%
wrapper-2551	69.12%	69.12%	88.24%	90.44%
wrapper-2552	66.93%	66.93%	77.17%	86.61%
wrapper-2550-2552	67.40%	67.15%	82.00%	86.62%
ค่าเฉลี่ย	67.25%	68.87%	84.65%	88.72%

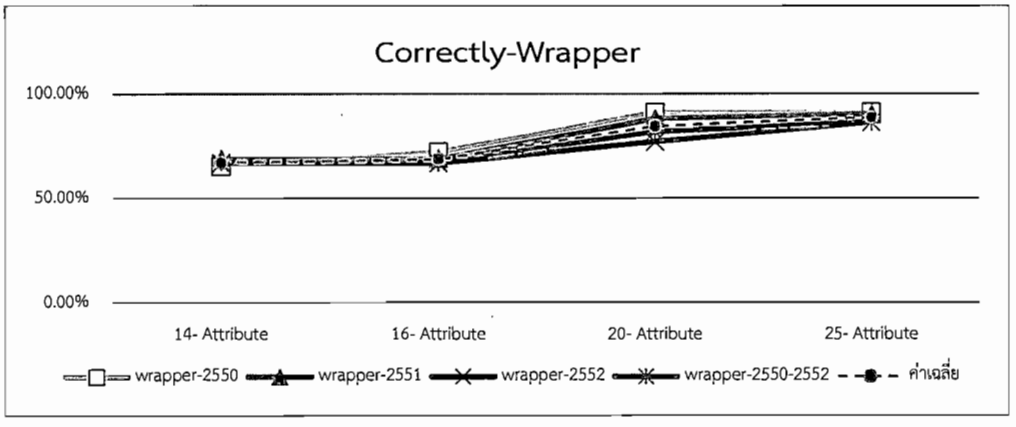
ผลการทดลองการคัดเลือกปัจจัยจากจำนวนคุณลักษณะของปัจจัยนำเข้าของนิสิตในแต่ละปีการศึกษาที่เพิ่มขึ้น และนำเทคนิคการคัดเลือกปัจจัยมาช่วยหาปัจจัยที่เหมาะสมก่อนการจำแนกข้อมูลนิสิต นั้น พบว่า จำนวนคุณลักษณะที่เพิ่มขึ้นมีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสมมากยิ่งขึ้น ซึ่งปัจจัยที่ได้รับการคัดเลือกดังกล่าวส่งผลต่อทำให้ค่าความถูกต้องของการจำแนกประเภทนิสิตที่เพิ่มขึ้นตาม และทั้งนี้ยังพบว่าในข้อมูลนิสิตปีการศึกษา 2550, 2551, 2552 และ 2550 - 2552 ปัจจัยที่ได้รับการคัดเลือกจากปัจจัยตั้งต้น 25 ดังรายละเอียดในตารางที่ 4-6 และแสดงเป็นกราฟดังภาพที่ 4-1, 4-2, 4-3



ภาพที่ 4-3 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค CS

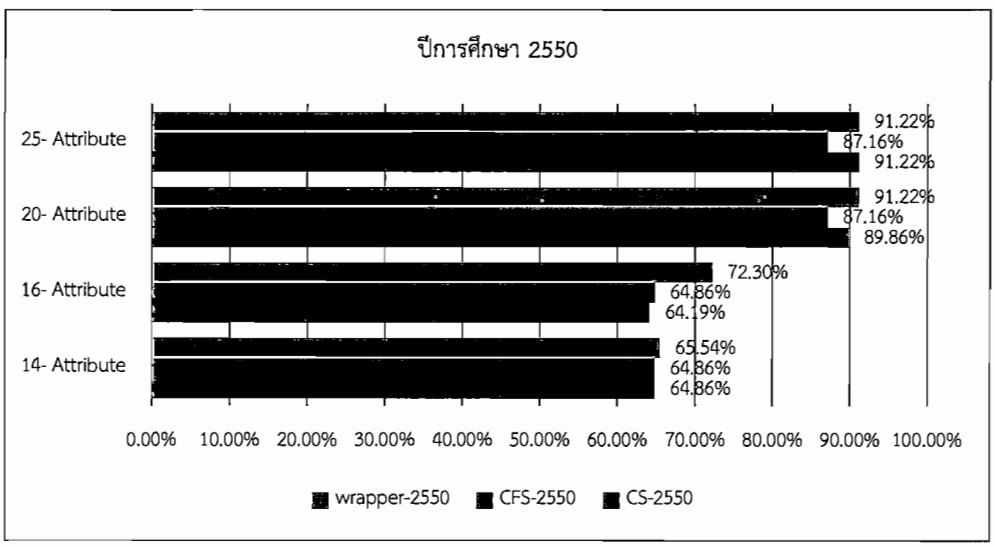


ภาพที่ 4-4 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค CFS



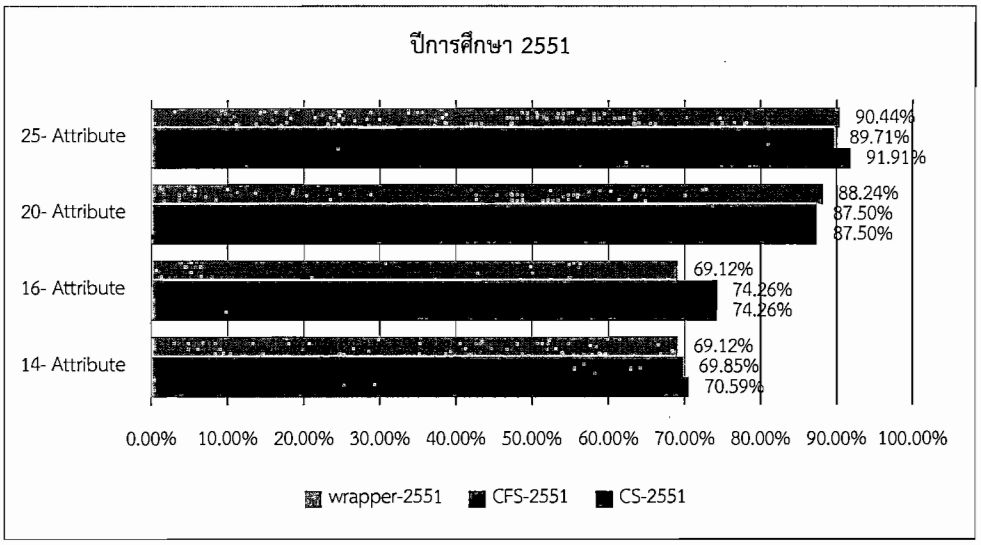
ภาพที่ 4-5 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตโดยเทคนิค Wrapper

ผลการทดลองที่ 1 – 2 เทคนิคการคัดเลือกปัจจัยที่แตกต่างกันส่งผลต่อการคัดเลือกปัจจัยและค่าความถูกต้องของการจำแนกประเภทนิสิต



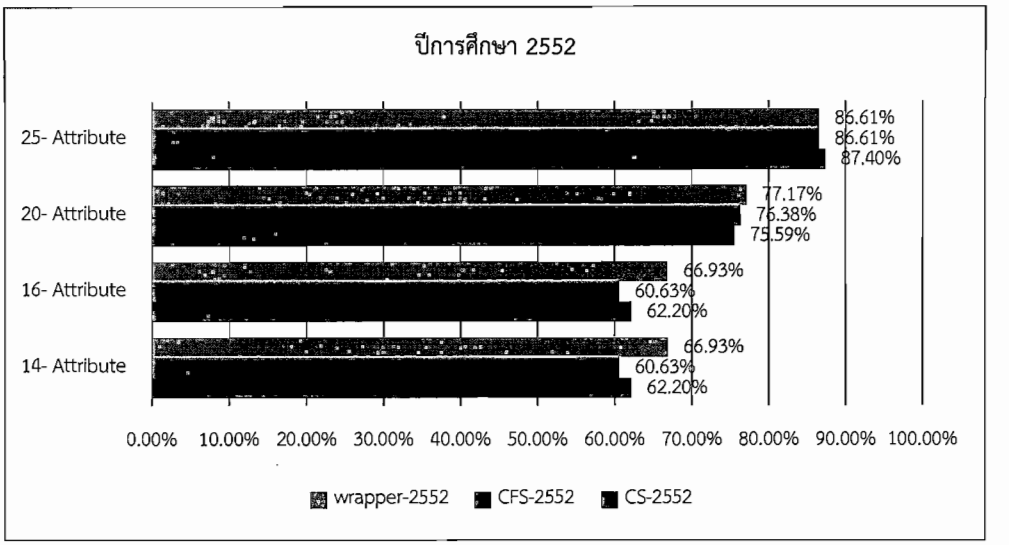
ภาพที่ 4-6 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2550 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน

กลุ่มข้อมูลนิสิตปีการศึกษา 2550 ดังแสดงในตารางที่ 4-2, 4-5 และภาพที่ 4-6 พบว่าเทคนิคการคัดเลือกปัจจัยที่แตกต่างกันมีผลต่อการคัดเลือกชุดปัจจัยที่เหมาะสมแตกต่างกัน และเมื่อเพิ่มจำนวนและคุณลักษณะของปัจจัยตั้งต้นที่เพิ่มขึ้นจะมีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสมส่งผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิตของข้อมูลนิสิตปีการศึกษา 2550 เพิ่มสูงขึ้นตาม โดยจากการทดลองพบว่าการใช้เทคนิค Wrapper ในการคัดเลือกปัจจัยจากจำนวนคุณลักษณะตั้งต้นของข้อมูลนิสิต จำนวน 25 ปัจจัย ปัจจัยที่ได้รับการคัดเลือกมีจำนวน 2 ปัจจัยได้แก่ 1) เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา 2) เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 ซึ่งภายหลังการคัดเลือกปัจจัยได้มีการนำชุดปัจจัยที่ถูกคัดเลือกมาใช้ในการจำแนกนิสิตด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ (BPNN) จะให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิต คือ 91.22% มีค่าความถูกต้องของการจำแนกข้อมูลนิสิตเท่ากับปัจจัยที่ถูกคัดเลือกโดยวิธี CS ปัจจัยที่ได้รับการคัดเลือกมีจำนวน 1 ปัจจัยได้แก่ 1) เกรดรายวิชาโปรแกรมมิ่ง 2



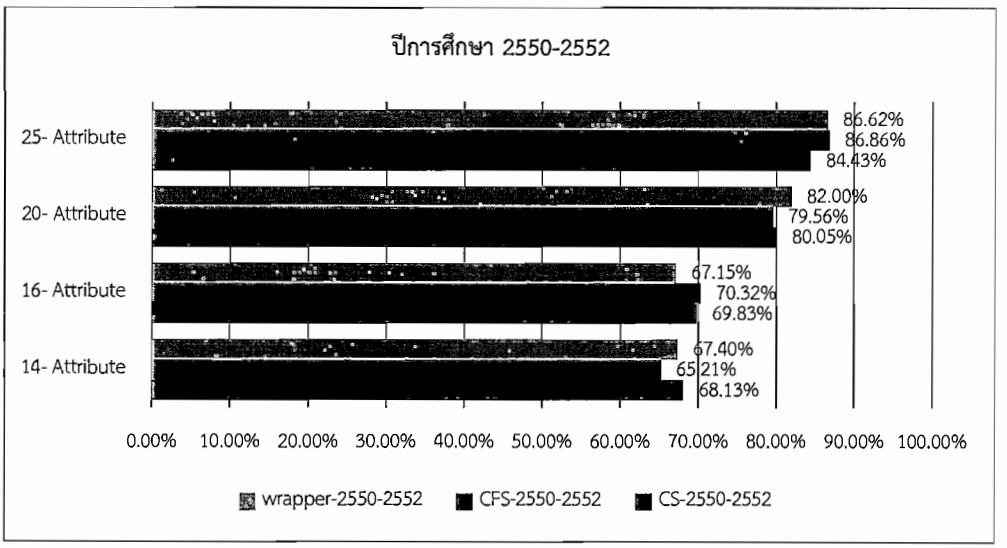
ภาพที่ 4-7 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2551 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน

กลุ่มข้อมูลนิสิตปีการศึกษา 2551 ดังแสดงในตารางที่ 4-3, 4-5 และภาพที่ 4-7 พบว่าเทคนิคการคัดเลือกปัจจัยที่แตกต่างกันมีผลต่อการคัดเลือกชุดปัจจัยที่เหมาะสมแตกต่างกัน และเมื่อเพิ่มจำนวนและคุณลักษณะของปัจจัยตั้งต้นที่เพิ่มขึ้นจะมีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสมส่งผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิตของข้อมูลนิสิตปีการศึกษา 2551 เพิ่มสูงขึ้นตาม โดยจากการทดลองพบว่าการใช้เทคนิค CS ในการคัดเลือกปัจจัยจากจำนวนและคุณลักษณะปัจจัยตั้งต้นของข้อมูลนิสิต จำนวน 25 ปัจจัย ปัจจัยที่ได้รับการคัดเลือกมีจำนวน 2 ปัจจัย ได้แก่ 1) เกรดรายวิชาโปรแกรมมิ่ง 2 2) เกรดรายวิชาภาษาอังกฤษ ซึ่งภายหลังการคัดเลือกปัจจัยได้มีการนำชุดปัจจัยที่ถูกคัดเลือกมาใช้ในการจำแนกนิสิตด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับ จะให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตสูงที่สุด คือ 91.91% เมื่อเปรียบเทียบกับชุดปัจจัยที่ได้รับคัดเลือกจากเทคนิคอื่น



ภาพที่ 4-8 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2552 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน

กลุ่มข้อมูลนิสิตปีการศึกษา 2552 ดังแสดงในตารางที่ 4 - 4, 4 - 5 และภาพที่ 4 - 8 พบว่าเทคนิคการคัดเลือกปัจจัยที่แตกต่างกันมีผลต่อการคัดเลือกชุดปัจจัยที่เหมาะสมแตกต่างกัน และเมื่อเพิ่มจำนวนและคุณลักษณะของปัจจัยตั้งต้นที่เพิ่มขึ้นจะมีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสมส่งผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิตของข้อมูลนิสิตปีการศึกษา 2552 เพิ่มสูงขึ้นตาม โดยจากการทดลองพบว่าการใช้เทคนิค CS ในการคัดเลือกปัจจัยจากจำนวนและคุณลักษณะปัจจัยตั้งต้นของข้อมูลนิสิต จำนวน 25 ปัจจัย ปัจจัยที่ได้รับการคัดเลือกมีจำนวน 7 ปัจจัย ได้แก่ 1) ภูมิภาคของสถานศึกษาเดิม (ที่อยู่นิสิต) 2) ประเภทนิสิต(ปกติ/พิเศษ) 3) เกรดภาคเรียนที่ 1 ชั้นปีที่ 1 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาแคลคูลัส 2 6) เกรดรายวิชาโปรแกรมมิ่ง 1 และ 7) เกรดรายวิชาโปรแกรมมิ่ง 2 ซึ่งภายหลังการคัดเลือกปัจจัยได้มีการนำชุดปัจจัยที่ถูกคัดเลือกมาใช้ในการจำแนกนิสิตด้วยตัวแบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการจำแนกประเภทนิสิต จะให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตสูงที่สุด คือ 87.40% เมื่อเปรียบเทียบกับชุดปัจจัยที่ได้รับคัดเลือกจากเทคนิคอื่น



ภาพที่ 4-9 ค่าความถูกต้องของการจำแนกข้อมูลนิสิตปีการศึกษา 2550 - 2552 เมื่อจำนวนและคุณลักษณะของปัจจัยนำเข้า และใช้เทคนิคการคัดเลือกปัจจัยแตกต่างกัน

กลุ่มข้อมูลนิสิตปีการศึกษา 2550 - 2552 ดังแสดงในตารางที่ 4-5, 4-5 และภาพที่ 4-9 พบว่า เทคนิคการคัดเลือกปัจจัยที่แตกต่างกันมีผลต่อการคัดเลือกชุดปัจจัยที่เหมาะสมแตกต่างกัน และเมื่อเพิ่มจำนวนและคุณลักษณะของปัจจัยตั้งต้นที่เพิ่มขึ้นจะมีอิทธิพลต่อการคัดเลือกปัจจัยที่เหมาะสม ส่งผลต่อค่าความถูกต้องของการจำแนกประเภทนิสิตของข้อมูลนิสิตปีการศึกษา 2550 - 2552 เพิ่มสูงขึ้นตาม โดยจากการใช้เทคนิค CFS ในการคัดเลือกปัจจัยจากจำนวนและคุณลักษณะปัจจัยตั้งต้นของข้อมูลนิสิต จำนวน 25 ปัจจัย ปัจจัยที่ได้รับการคัดเลือกมีจำนวน 8 ปัจจัยได้แก่ 1) วิธีเข้าศึกษา 2) เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1 3) เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาแคลคูลัส 2 6) เกรดรายวิชาโปรแกรมมิ่ง 1 7) เกรดรายวิชาโปรแกรมมิ่ง 2 8) เกรดรายวิชาภาษาอังกฤษ ซึ่งภายหลังการคัดเลือกปัจจัยได้มีการนำชุดปัจจัยที่ถูกคัดเลือกมาใช้ในการจำแนกนิสิตด้วยโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการจำแนกประเภทนิสิต จะให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตสูงที่สุด คือ 86.86% เมื่อเปรียบเทียบกับชุดปัจจัยที่ได้รับคัดเลือกจากเทคนิคอื่น

จากผลการทดลองข้างต้นสามารถสรุปปัจจัยที่เหมาะสมที่สุด และเทคนิควิธีการคัดเลือกปัจจัยของข้อมูลแต่ละปีการศึกษา ดังแสดงในตารางที่ 4-7

ตารางที่ 4-7 สรุปปัจจัยที่เหมาะสมที่สุดสำหรับข้อมูลนิสิตในแต่ละปีการศึกษาและเทคนิคการคัดเลือกปัจจัยที่ให้ค่าการจำแนกประเภทนิสิตถูกต้องสูงสุด

ลำดับ	ปัจจัย	ข้อมูลนิสิตปีการศึกษา				
		2550		2551	2552	2550-2552
1	เพศ (SEX)					
2	สถานภาพสมรสของบิดามารด (STATUSFM)					
3	อาชีพของบิดา (OCCFA)					
4	อาชีพของมารดา (OCCMO)					
5	รายได้ของบิดา (INCFA)					
6	รายได้ของมารดา (INCMO)					
7	เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา (ENTRYGPA)		✓		✓	
8	สถานภาพของบิดา (STATUSF)					
9	สถานภาพของมารดา (STATUSM)					
10	ภูมิภาคของสถานศึกษาเดิม (ที่อยู่นิสิต) (PROVINCE)					
11	จำนวนพี่น้องทั้งหมด (รวมตัวเอง) (NUMSON)					
12	บุตรคนที่ (SON)					
13	บุตรคนที่ที่กำลังศึกษา (SONED)					
14	ประเภทของสถานบันศึกษาเดิม(รัฐ/เอกชน) (TYPE)					
15	วิธีเข้าศึกษา (BYTEDES)					✓
16	ประเภทนิสิต (ปกติ/พิเศษ) (PROGRAM)				✓	
17	เกรดภาคเรียนที่ 1 ชั้นปีที่ 1 (GPA-1)					
18	เกรดภาคเรียนที่ 2 ชั้นปีที่ 1 (GPA-2)		✓		✓	✓
19	เกรดเฉลี่ยสะสมภาคเรียนที่ 1 ชั้นปีที่ 1 (ACC-GPA-1)					
20	เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 (ACC-GPA-2)					
21	เกรดรายวิชาแคลคูลัส 1 (Calculus1)				✓	✓
22	เกรดรายวิชาแคลคูลัส 2 (Calculus2)				✓	✓
23	เกรดรายวิชาโปรแกรมมิ่ง 1 (Pro1)			✓	✓	✓
24	เกรดรายวิชาโปรแกรมมิ่ง 1 (Pro2)	✓			✓	✓
25	เกรดรายวิชาภาษาอังกฤษ			✓		
26	สถานะการสำเร็จการศึกษา					
	เทคนิคการคัดเลือกที่เหมาะสม	CS	Wrapper	CS	CS	CFS

การทดลองที่ 2: การทดลองเพื่อค้นหาชุดคุณลักษณะของปัจจัยนำเข้าที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา

สมมติฐาน : มีชุดปัจจัยที่เหมาะสมที่สามารถจำแนกประเภทนิสิตได้ทุกปีการศึกษา แล้วทำให้ค่าความถูกต้องในการจำแนกประเภทนิสิตเฉลี่ยสูงสุด

เป็นการศึกษาการคัดเลือกปัจจัยที่มีผลต่อการจำแนกประเภทนิสิตที่มีความถูกต้องสูงที่สุด และสามารถนำชุดปัจจัยที่ได้ใช้กับข้อมูลนิสิตทุกปีการศึกษา ทำให้ค่าความถูกต้องเฉลี่ยของความจำแนกข้อมูลสูงสุด โดยทำการทดสอบจากจำนวนและชุดข้อมูลที่ใช้ในการทดลองดังแสดงในตารางที่ 3-4 และ 3-5 สำหรับเทคนิคที่ใช้ในการเลือกปัจจัย ได้แก่ 1) เทคนิค Consistency - based Features Selection (CS) 2) เทคนิค Correlation - based Features Selection (CFS)และ 3) เทคนิค Wrapper Approach ซึ่งภายหลังการคัดเลือกปัจจัยได้มีการใช้อัลกอริทึมตัวแบบโครงข่ายประสาทเทียมแบบแพร่ย้อนกลับในการจำแนกประเภทนิสิต ออกเป็น 3 ประเภท คือ 1) สำเร็จการศึกษาภายในระยะเวลา 4 ปี 2) สำเร็จการศึกษามากกว่า 4 ปี และ 3) ไม่สำเร็จการศึกษา โดยมีผลการทดลองดังนี้

ผลการทดลองที่ 2 การหาชุดปัจจัยที่เหมาะสมที่สามารถจำแนกประเภทนิสิตได้ทุกปีการศึกษาได้เหมาะสมที่สุด

ตารางที่ 4-8 แสดงปัจจัยที่ถูกคัดเลือกจากปัจจัยตั้งต้น 25 ปัจจัยด้วยเทคนิค CS CFS และWrapper และค่าความถูกต้องของการจำแนกในข้อมูลนิสิตในแต่ละปีการศึกษา

วิธี	Filter Approach								Wrapper Approach			
	Consistency - based Features Selection (CS)				Correlation - based Features Selection (CFS)				Wrapper Approach (Wrapper)			
	CS (2550)	CS (2551)	CS (2552)	CS (2550-2552)	CFS (2550)	CFS (2551)	CFS (2552)	CFS (2550-2552)	wrapper (2550)	wrapper (2551)	wrapper (2552)	wrapper (2550-2552)
Attribute	PRO2	PRO2 ENG1	PROVINCE PROGRAM GPA-1 Calculus1 Calculus2 PRO1 PRO2	INCFA INCMO ENTRYGPA PROGRAM GPA-1 Calculus1 Calculus2 PRO1 PRO2 ENG1	INCMO BYTEDES ACC-GPA-2 Calculus1 PRO1 PRO2 ENG1	BYTEDES Calculus2 PRO2 ENG1	ENTRYGPA PROGRAM GPA-2 Calculus1 Calculus2 PRO1 PRO2	BYTEDES GPA-1 GPA-2 Calculus1 Calculus2 PRO1 PRO2 ENG1	ENTRYGPA ACC-GPA-2	INCMO PRO2	NUMSON OCCFA PROGRAM Calculus1 PRO2	STATUSFA PRO2
2550	91.22%	91.22%	85.14%	86.49%	85.14%	89.19%	87.84%	87.16%	91.22%	91.22%	91.22%	91.22%
2551	90.44%	91.91%	86.03%	88.24%	89.71%	89.71%	91.18%	87.50%	86.03%	90.44%	87.24%	90.44%
2552	81.10%	80.32%	87.40%	82.68%	81.89%	79.53%	86.61%	79.53%	67.72%	80.32%	86.61%	81.89%
2550-2552	86.62%	86.37%	87.83%	84.43%	86.62%	84.43%	86.62%	86.86%	83.46%	86.86%	87.10%	86.62%
ค่าเฉลี่ย	87.34%	87.45%	86.60%	85.46%	85.84%	85.71%	88.06%	85.26%	82.10%	87.21%	88.04%	87.54%
ปี 2553	72.50%	80.00%	79.17%	81.67%	80.00%	80.83%	81.67%	79.17%	72.50%	70.00%	76.67%	72.50%
ปี 2554	63.91%	62.41%	77.44%	81.67%	60.90%	77.44%	80.45%	81.20%	63.91%	76.69%	66.92%	63.91%

จากตารางการทดลองตารางที่ 4-8 พบว่าค้นหาชุดคุณลักษณะของปัจจัยที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา จากการคัดเลือกปัจจัยด้วยเทคนิควิธี CS โดยมีตั้งต้นจำนวน 25 ปัจจัย นั้น ชุดปัจจัยที่ได้รับการคัดเลือกจากข้อมูลนิสิตปีการศึกษา 2550 ที่ทำให้มีค่าการจำแนกถูกต้องเฉลี่ยสูงสุดที่ 87.34% โดยชุดปัจจัยที่ถูกคัดเลือกมีจำนวน 2 ปัจจัย คือ 1))เกรดรายวิชาโปรแกรมมิ่ง 2 2) เกรดรายวิชาภาษาอังกฤษ และให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตในปีการศึกษา 2550 เท่ากับ 91.22% ข้อมูลนิสิตปีการศึกษา 2551 เท่ากับ 91.91% ข้อมูลนิสิตปีการศึกษา 2552 เท่ากับ 80.32% และข้อมูลนิสิตปีการศึกษา 2550 - 2552 เท่ากับ 86.37% ตามลำดับ และทดสอบกับข้อมูลสำหรับทดสอบ (Unseen Data) ในข้อมูลนิสิตปีการศึกษา 2553ได้ 80.00% และข้อมูลนิสิตปีการศึกษา 2554 ได้ 62.41%

และพบว่าค้นหาชุดคุณลักษณะของปัจจัยที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา จากการคัดเลือกปัจจัยด้วยเทคนิควิธี CFS โดยมีตั้งต้นจำนวน 25 ปัจจัย นั้น ชุดปัจจัยที่ได้รับการคัดเลือกจากข้อมูลนิสิตปีการศึกษา 2552 ด้วยเทคนิควิธี CFS ที่ทำให้มีค่าการจำแนกถูกต้องเฉลี่ยสูงสุดที่ 88.06% โดยชุดปัจจัยที่ถูกคัดเลือกจำนวน 7 ปัจจัย คือ 1) เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา 2) ประเภทนิสิต(ปกติ/พิเศษ) 3) เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาแคลคูลัส 2 6) เกรดรายวิชาโปรแกรมมิ่ง 1 และ 7) เกรดรายวิชาโปรแกรมมิ่ง 2 และให้ค่าความถูกต้องของการจำแนกข้อมูลนิสิตในปีการศึกษา 2550 เท่ากับ 87.84% ข้อมูลนิสิตปีการศึกษา 2551 เท่ากับ 91.18% ข้อมูลนิสิตปีการศึกษา 2552 เท่ากับ 86.61% และข้อมูลนิสิตปีการศึกษา 2550 - 2552 เท่ากับ 86.62% ตามลำดับ และทดสอบกับข้อมูลสำหรับทดสอบ (Unseen Data) ในข้อมูลนิสิตปีการศึกษา 2553ได้ 81.67% และข้อมูลนิสิตปีการศึกษา 2554 ได้ 80.45%

และพบว่าค้นหาชุดคุณลักษณะของปัจจัยที่สามารถใช้จำแนกประเภทนิสิตได้ทุกปีการศึกษา จากการคัดเลือกปัจจัยด้วยเทคนิควิธี Wrapper Approach โดยมีตั้งต้นจำนวน 25 ปัจจัย นั้น ชุดปัจจัยที่ได้รับการคัดเลือกจากข้อมูลนิสิตปีการศึกษา 2552 ที่ทำให้มีค่าการจำแนกถูกต้องเฉลี่ยสูงสุดที่ 88.04% โดยชุดปัจจัยที่ถูกคัดเลือกมีจำนวน 5 ปัจจัย คือ 1) จำนวนพี่น้องทั้งหมด (รวมตัวเอง) 2) อาชีพของบิดา 3) ประเภทนิสิต (ปกติ/พิเศษ) 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาโปรแกรมมิ่ง 2 และพบว่าค่าความถูกต้องของการจำแนกข้อมูลนิสิตในปีการศึกษา 2550 เท่ากับ 91.22% , ข้อมูลนิสิตปีการศึกษา 2551 เท่ากับ 87.24% , ข้อมูลนิสิตปีการศึกษา 2552 เท่ากับ 86.61% และข้อมูลนิสิตปีการศึกษา 2550 - 2552 เท่ากับ 87.10% ตามลำดับ และทดสอบกับข้อมูลสำหรับทดสอบ (Unseen Data) ในข้อมูลนิสิตปีการศึกษา 2553ได้ 76.67% และข้อมูลนิสิตปีการศึกษา 2554 ได้ 66.92%

ตารางที่ 4-9 แสดงการเปรียบเทียบค่าความถูกต้องจากคุณลักษณะที่ถูกคัดเลือกที่ได้ผลลัพธ์ที่ดีที่สุดกับการใช้คุณลักษณะทั้งหมด

ข้อมูลปีการศึกษา / ปัจจัย (Attributes)	Simple Classification (ใช้ทุก Attributes, 25 Attributes)	ใช้ Attributes ที่ได้จากวิธี CS (2551) (2 Attributes)	ใช้ Attributes ที่ได้จากวิธี CFS (2552) (7 Attributes)	ใช้ Attributes ที่ได้จากวิธี Wrapper (2552) (5 Attributes)
		PRO2 ENG1	ENTRYGPA PROGRAM GPA-2 Calculus1 Calculus2 PRO1 PRO2	NUMSON OCCFA PROGRAM Calculus1 PRO2
2550	85.81%	91.22%	87.84%	91.22%
2551	87.50%	91.91%	91.18%	88.24%
2552	83.46%	80.32%	86.61%	86.61%
2550 - 2552	83.70%	86.37%	86.62%	87.10%
ค่าเฉลี่ย	83.58%	87.45%	**88.06%**	88.04%
Unseen Data 2553	75.83 %	80.00%	81.67%	76.67%
Unseen Data 2554	68.42%	62.41%	80.45%	66.92%

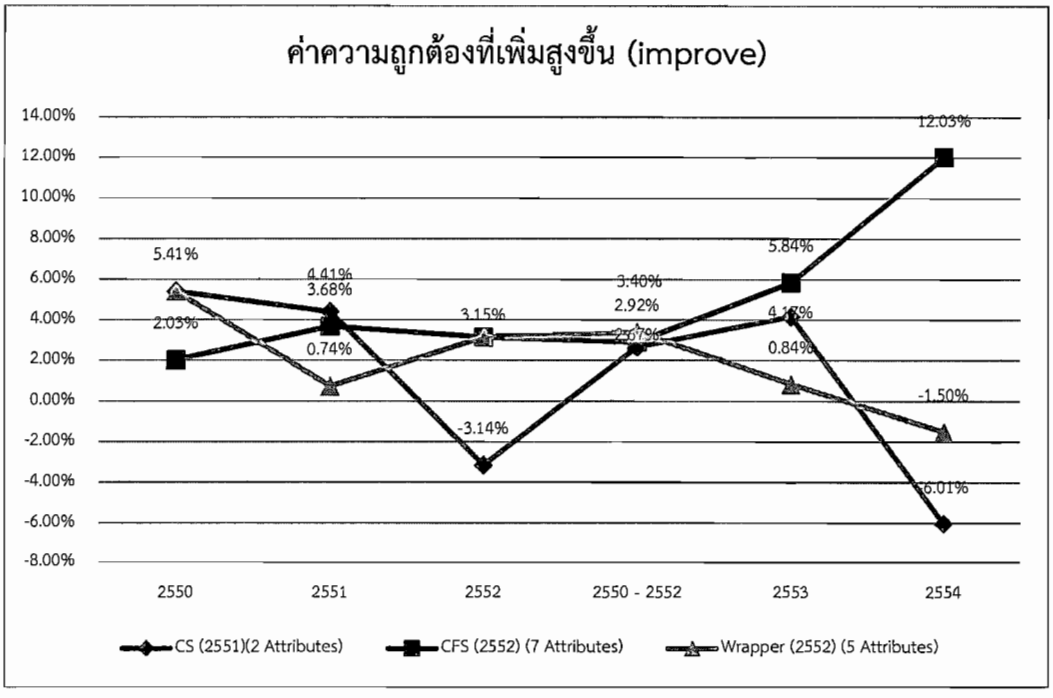
จากตารางที่ 4-9 จากการศึกษาผลการทดลอง พบว่า

1. เทคนิคการคัดเลือกปัจจัยแบบ CFS สามารถคัดเลือกปัจจัยที่เหมาะข้อมูลนิสิตชั้นปีที่ 1 สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ ได้เหมาะสมที่สุดจากชุดปัจจัยที่ถูกคัดเลือกมาจากข้อมูลนิสิตปีการศึกษา 2552 จำนวน 7 ปัจจัย ได้แก่ 1) เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา 2) ประเภทนิสิต(ปกติ/พิเศษ) 3) เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาแคลคูลัส 2 6) เกรดรายวิชาโปรแกรมมิ่ง 1 และ 7) เกรดรายวิชา

โปรแกรมมิ่ง 2 โดยทำให้ค่าความถูกต้องของการจำแนกประเภทนิสิตมีค่าเฉลี่ยมีค่าสูงสุด เท่ากับ 88.06% และเมื่อทำการทดสอบกับชุดข้อมูลปีการศึกษา 2553 และ 2554 (Unseen Data) พบว่าปัจจัยที่ได้รับการคัดเลือกด้วยเทคนิค CFS ให้ผลค่าความถูกต้องของการจำแนกกลุ่มนิสิตในชุดข้อมูลปีการศึกษา 2553 และ 2554 ที่สูงกว่าเมื่อเปรียบเทียบกับการนำปัจจัยที่ได้จากเทคนิค CS และ Wrapper

2. เทคนิคการคัดเลือกปัจจัยแบบ CFS สามารถลดปัจจัยจาก 25 ปัจจัยโดยมีปัจจัยที่เหมาะสมเหลือเพียงจำนวน 7 ปัจจัย(ลดลง 72.00%) เพียงพอสำหรับข้อมูลชุดนี้ที่จะนำมาใช้การจำแนกประเภทนิสิตต่อไป

3. เทคนิคการคัดเลือกปัจจัยแบบ CS, CFS และ Wrapper มีปัจจัยร่วมกัน คือ เกรดรายวิชาโปรแกรมมิ่ง 2 เมื่อนำไปใช้ในการจำแนกประเภทนิสิในชุดข้อมูลปีการศึกษา 2553 และ 2554 สามารถทำให้ค่าความถูกต้องของการจำแนกประเภทนิสิตมีค่า 72.50% และ 63.91%



ภาพที่ 4-10 เปรียบเทียบค่าความถูกต้องการจำแนกประเภทนิสิตจากปัจจัยที่ดีที่สุดถูกคัดเลือกโดยเทคนิค CS,CFS และ Wrapper กับค่าความถูกต้องจากปัจจัยการใช้ทั้งหมด 25 ปัจจัย

จากภาพที่ 4-10 พบว่าคุณลักษณะที่ได้รับการคัดเลือกด้วยเทคนิค CFS ตอบสนองได้ดีต่อการเปลี่ยนแปลงพฤติกรรมของข้อมูล จะเห็นได้ชัดเจนในกรณีผลการทดลองในชุดข้อมูลปีการศึกษา

2553 และ 2554 โดยปัจจัยที่ได้รับการคัดเลือกด้วยเทคนิค CFS นั้น มีผลทำให้ค่าความถูกต้องของการจำแนกกลุ่มนิสิตของชุดข้อมูลปีการศึกษา 2553 และ 2554 มีค่าสูงกว่าเมื่อเปรียบเทียบกับการนำปัจจัยที่ได้จากเทคนิค CS และ Wrapper

ตารางที่ 4-10 ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือกด้วยเทคนิค CS, CFS และ Wrapper เพื่อจำแนกประเภทข้อมูลนิสิตปีการศึกษา 2553

ผลการทำนาย	CS			CFS			Wrapper		
	คำตอบเป้าหมาย			คำตอบเป้าหมาย			คำตอบเป้าหมาย		
	ไม่จบ	จบ	จบเกิน	ไม่จบ	จบ	จบเกิน	ไม่จบ	จบ	จบเกิน
ไม่จบ	30	2	6	**32**	1	5	31	2	5
จบ	0	56	1	2	51	4	1	49	7
จบเกิน	3	9	13	4	6	**15**	3	10	12

จากตารางที่ 4-10 ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือกด้วยเทคนิค CS, CFS และ Wrapper เพื่อจำแนกประเภทข้อมูลนิสิตปีการศึกษา 2553 พบว่าคุณลักษณะที่ถูกคัดเลือกด้วยเทคนิค CFS นั้นสามารถจำแนกกลุ่มนิสิตไม่สำเร็จการศึกษาและกลุ่มสำเร็จการศึกษามากกว่า 4 ปี ได้ดีกว่าคุณลักษณะที่ถูกเลือกด้วยเทคนิค CS และ Wrapper ซึ่งมีผลทำให้ค่าความถูกต้องในการจำแนกประเภทนิสิตสูงกว่าวิธีอื่น

ตารางที่ 4-11 ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือกด้วยเทคนิค CS, CFS และ Wrapper ในข้อมูลนิสิตปีการศึกษา 2554

ผลการทำนาย	CS			CFS			Wrapper		
	คำตอบเป้าหมาย			คำตอบเป้าหมาย			คำตอบเป้าหมาย		
	ไม่จบ	จบ	จบเกิน	ไม่จบ	จบ	จบเกิน	ไม่จบ	จบ	จบเกิน
ไม่จบ	59	8	0	58	1	8	59	6	2
จบ	6	34	0	2	32	6	4	28	8
จบเกิน	12	14	0	4	5	**17**	15	9	2

จากตารางที่ 4-11 ตารางแสดงผลลัพธ์การจำแนกประเภท (Confusion Matrix) โดยใช้ปัจจัยที่ถูกคัดเลือกด้วยเทคนิค CS, CFS และ Wrapper เพื่อจำแนกประเภทข้อมูลนิสิตปีการศึกษา 2554 พบว่า คุณลักษณะที่ถูกคัดเลือกด้วยเทคนิค CFS นั้นสามารถจำแนกกลุ่มสำเร็จการศึกษามากกว่า 4 ปี ได้ดีกว่าคุณลักษณะที่ถูกเลือกด้วยเทคนิค CS และ Wrapper ซึ่งมีผลทำให้ค่าความถูกต้องในการจำแนกประเภทนิสิตสูงกว่าวิธีอื่น

บทที่ 5
สรุปและอภิปรายผล

สรุปผลการทดลอง

งานนิพนธ์นี้มีวัตถุประสงค์ เพื่อหาแบบจำลองการพยากรณ์โอกาสการสำเร็จการศึกษาที่มีการคัดเลือกปัจจัย สำหรับการจำแนกกลุ่มนิสิตชั้นปีที่ 1 เพื่อการให้คาช่วยเหลือด้านวิชาการของนิสิต จากชุดข้อมูลประวัติส่วนตัวนิสิต ชุดข้อมูลประวัติการศึกษาของนิสิตในชั้นปีที่ 1 และข้อมูลผลการสำเร็จการศึกษา ของนิสิตในหลักสูตร 4 ปี สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ ตั้งแต่ปีการศึกษา 2550 – 2554 ที่ได้จากกองทะเบียน จำนวน 699 ระเบียน ซึ่งข้อมูลในแต่ละระเบียนประกอบไปด้วยปัจจัย 25 ปัจจัย จากการดำเนินการทดลองตามวิธีการที่นำเสนอในบทที่ 3 และจากผลลัพธ์ที่ได้แสดงในบทที่ 4 สามารถสรุปผลการดำเนินงานนิพนธ์นี้ได้ดังต่อไปนี้

1. จากการศึกษาและผลการทดลองพบว่า เทคนิคการคัดเลือกปัจจัย Correlation - based Features Selection (CFS) สามารถคัดเลือกปัจจัยจากปัจจัยตั้งต้นทั้งสิ้น 25 ปัจจัย โดยคัดเลือกปัจจัยที่เหมาะสมเหลือเพียงจำนวน 7 ปัจจัย ได้แก่ 1) เกรดเฉลี่ยเดิมก่อนเข้ารับการศึกษา 2) ประเภทนิสิต(ปกติ/พิเศษ) 3) เกรดเฉลี่ยสะสมภาคเรียนที่ 2 ชั้นปีที่ 1 4) เกรดรายวิชาแคลคูลัส 1 5) เกรดรายวิชาแคลคูลัส 2 6) เกรดรายวิชาโปรแกรมมิ่ง 1 และ 7) เกรดรายวิชาโปรแกรมมิ่ง 2 มีค่าเฉลี่ยเท่ากับ 88.06% ซึ่งเป็นค่าสูงสุด โดยวิธีการนี้สามารถพยากรณ์ข้อมูลนิสิตปีการศึกษา 2550 ได้ 87.84% ข้อมูลนิสิตปีการศึกษา 2551 ได้ 91.18% ข้อมูลนิสิตปีการศึกษา 2552 ได้ 86.61% ข้อมูลนิสิตปีการศึกษา 2550 - 2552 ได้ 86.62% ข้อมูลนิสิตปีการศึกษา 2553ได้ 81.67% และข้อมูลนิสิตปีการศึกษา 2554 ได้ 80.45% โดยข้อมูลนิสิตในปีการศึกษา 2550 - 2552 เป็นข้อมูลสำหรับสอน (Training Data) และข้อมูลสำหรับทดสอบ (Test Data) และใช้ข้อมูลนิสิตในปีการศึกษา 2553 - 2554 เป็นข้อมูลสำหรับทดสอบ (Unseen Test Data) ทั้งนี้ในการทดลองผู้นิพนธ์ได้ทำการเปรียบเทียบลักษณะของข้อมูลนำเข้าของนิสิตในแต่ละชั้นปีพบว่า มีรูปแบบพฤติกรรมของข้อมูล (Data Pattern) เป็นไปในทางเดียวกันสำหรับชุดข้อมูลของนิสิตในแต่ละปีการศึกษา สำหรับชุดข้อมูลที่ใช้ในงานนิพนธ์นี้ทั้ง 5 ปีการศึกษา ดังนั้นการนำปัจจัยที่ได้รับจากการคัดเลือกในงานนิพนธ์นี้ไปใช้ในการพยากรณ์โอกาสการสำเร็จการศึกษา จะสามารถใช้ได้อย่างเหมาะสมกับชุดข้อมูลที่มีรูปแบบของพฤติกรรมข้อมูลในทำนองเดียวกัน

2. จากการศึกษาและจากการทดลองในงานนิพนธ์นี้ทำให้ได้ตัวแบบ (Model) สำหรับการคัดเลือกคุณลักษณะที่เหมาะสม ดังแสดงในภาพที่ 3-28 และได้เครื่องมือสำหรับการวิเคราะห์คุณลักษณะ (Feature) ที่ตอบสนอง ทนทาน ต่อการเปลี่ยนแปลงพฤติกรรมของข้อมูล จะเห็นได้

ชัดเจนในกรณีผลการทดลองในชุดข้อมูลปีการศึกษา 2553 และ 2554 ปัจจัยที่ได้รับการคัดเลือกด้วยเทคนิค CFS นั้นเป็นปัจจัยที่น่าเชื่อถือมากกว่าการคัดเลือกปัจจัยด้วยเทคนิค CS และ Wrapper ซึ่งปัจจัยที่ได้จากเทคนิค CFS ให้ผลค่าความถูกต้องของการจำแนกกลุ่มนิสิตในชุดข้อมูลปีการศึกษา 2553 และ 2554 ที่สูงกว่าเมื่อเปรียบเทียบกับการนำปัจจัยที่ได้จากเทคนิค CS และ Wrapper มาวิเคราะห์ และสามารถนำตัวแบบ (Model) ที่ได้ใช้เป็นแนวทางการคัดเลือกปัจจัยที่เหมาะสม สำหรับใช้ในการพยากรณ์โอกาสการสำเร็จการศึกษาในกรณีที่ชุดข้อมูลที่ใช้ในการทดลองมีรูปแบบของพฤติกรรมข้อมูลที่แตกต่างจากชุดข้อมูลที่ใช้ในงานนิพนธ์นี้

ข้อดี

1. สามารถทราบถึงปัจจัยที่มีความสัมพันธ์ต่อการสำเร็จการศึกษา และปัจจัยที่ได้รับการคัดเลือกสามารถนำไปใช้ในการพยากรณ์โอกาสสำเร็จการศึกษาของนิสิตที่ศึกษาอยู่ในสถาบันการศึกษาในปีการศึกษาแรก (ชั้นปีที่ 1) ของนิสิตสาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ ที่จะเกิดขึ้นจริงในอนาคต

2. สามารถนำไปใช้ในการจำแนกกลุ่มนิสิตสำหรับการจัดการเรียนการสอนในแต่ละกลุ่มได้อย่างเหมาะสม และนำใช้เป็นแนวทางในการสร้างระบบและกลไกที่จะช่วยให้สามารถติดตามกำกับดูแล ให้คำปรึกษาแก่นิสิตในแต่ละกลุ่ม เพื่อลดจำนวนของนิสิตที่ไม่สำเร็จการศึกษาหรือออกกลางคัน และสามารถนำกระบวนการค้นหาปัจจัยไปทดลองใช้กับข้อมูลนิสิตสาขาวิชาหรือคณะอื่นได้

ข้อจำกัด

1. ชุดข้อมูลที่นำมาใช้ในการทดลองเป็นชุดข้อมูลที่ไม่สมบูรณ์เนื่องจากมีข้อมูลบ้างรายการมีค่าข้อมูลไม่ครบทุกปัจจัยที่ใช้ในการวิเคราะห์ จึงมีผลต่อการวิเคราะห์ปัจจัยและการสร้างแบบจำลองการจำแนกประเภทที่อาจมีความคลาดเคลื่อน

ข้อเสนอแนะ

1.การวิเคราะห์ปัจจัยในงานนิพนธ์นี้สร้างจากข้อมูลนิสิตในชั้นปีที่ 1 เท่านั้น จะมีประโยชน์มากยิ่งขึ้นอีกหากมีการเพิ่มเติมปัจจัยและคุณลักษณะของประวัติการศึกษานิสิตในปีการศึกษาถัดไปของนิสิตแต่ละรายมาทำการวิเคราะห์หาปัจจัยเหมาะสมเพิ่มเติมในกรณีที่ต้องการพยากรณ์โอกาสการสำเร็จการศึกษาของนิสิตในชั้นปีที่สูงขึ้น

2.ปัจจัยที่มีผลต่อการสำเร็จการศึกษาของนิสิตที่ได้รับการคัดเลือกปัจจัยในงานนิพนธ์นี้และแบบจำลองการจำแนกประเภทนิสิตสามารถนำไปปรับใช้กับคณะหรือมหาวิทยาลัยที่มีลักษณะ

การเก็บข้อมูลนิสิต/นักศึกษาที่ใกล้เคียงกันเท่านั้น เนื่องจากลักษณะข้อมูลนิสิต/นักศึกษาที่ถูกเก็บของแต่ละมหาวิทยาลัยมีความแตกต่างกันออกไป ทำให้ปัจจัยที่ใช้ในการศึกษาก็มีความแตกต่างกันออกไปด้วย

แต่หากพบว่าการเก็บข้อมูลนิสิตมีความแตกต่างจากข้อมูลนิสิตในงานนิพนธ์นี้ ผู้สนใจอาจใช้ตัวแบบการวิเคราะห์คัดเลือกปัจจัยดังแสดงในภาพที่ 3-28 ในการทดลองนี้ไปเป็นขั้นตอนวิธีในการหาปัจจัยที่เหมาะสม

บรรณานุกรม

จิราภรณ์ ถมแก้ว และศรัณย์ อินทโกสุม. (2554). การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ. ใน *การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติ ครั้งที่ 23.* (หน้า 7-12). นครราชสีมา: คณะวิทยาศาสตร์และศิลปศาสตร์ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน.

บุญเจริญ ศิริเนาวกุล. (ม.ป.ป.). *เอกสารการสอน AI.* เข้าถึงได้จาก http://www.cpe.kmutt.ac.th/~boon/AI3heuristic.pdf.

ทิพย์สุดา วงศ์คำดี. (2552) *การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษาโดยใช้โครงข่ายประสาทเทียม*, การศึกษาอิสระปริญญาวิทยาศาสตรมหาบัณฑิต, เทคโนโลยีสารสนเทศ, คณะวิทยาศาสตร์, มหาวิทยาลัยขอนแก่น.

บุษรา ลิ้มพิพัฒนางกูร. (2549). *การพยากรณ์โอกาสสำเร็จการศึกษาของนักศึกษา โดยใช้จีนีติกอัลกอริทึมแบบหลายวัตถุประสงค์ เปรียบเทียบกับวิธีทางระบบโครงข่ายประสาทเทียม: กรณศึกษานักศึกษาภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ.* วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, วิทยาการคอมพิวเตอร์, บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

วีระยุทธ พิมพาภรณ์. (2557). การวิเคราะห์องค์ประกอบบนชุดข้อมูลที่ทับซ้อนเทคนิควิธีการเลือกลักษณะสำคัญแบบพลวัต. *วารสารศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี,* 6, 90-99.

สุคนธ์ทิพย์ วงศ์พันธ์. (2551). *การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะที่เหมาะสมและอัลกอริทึมเพื่อการจำแนกพฤติกรรมการกระทำความผิดของนักเรียนระดับอาชีวศึกษา.* วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, วิทยาการคอมพิวเตอร์, บัณฑิตวิทยาลัย, มหาวิทยาลัยเกษตรศาสตร์.

สมบัติ ฝอยทอง. (2555). *การเลือกลักษณะเด่นโดย ราฟ-มิวชวล บนพื้นฐานของความไม่แน่นอนต่ำสุดและความแน่นอนสูงสุด.* วิทยานิพนธ์ปริญญาการศึกษาดุษฎีบัณฑิต, วิศวกรรไฟฟ้า, บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง.

ธนาวุฒิ นิลมณี. (2557). *การคัดเลือกคุณลักษณะข้อมูลที่มีผลกระทบต่อการสำเร็จการศึกษาของนักศึกษาโดยใช้วิธีการจำแนกข้อมูลโครงข่ายประสาทเทียม.* The 19th National Conference on Computing and Information Technology.

Karamouzis, S.T. (2009). Sensitivity Analysis of Neural Network Parameters for Identifying the Factors for College Student Success. Proceedings of the World Congress on Computer Science and Information Engineering; March 31 - April 2, 2009; Los Angeles. pp.671-675.

M. A. Hall. (2000). Crrelation-based feature selection for discrete and numeric class Machine learning. In Proc. 17th Int. Conf. Mach. Learn., pp.359-366.

H. Liu and R. Setiono. (1996). A probabilistic approach to feature selection: a filter solution, Proc. 13th Int'l Conf. Machine Learning, pp.319-327.

Novakovic, J., Strbac, P., & Bulatovic, D. (2011). Toward optimal feature selection using ranking method and classification algorithms. *Yugoslav Journal of Operations Research,* 21(1), 119-135.