การเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของข้อมูล ระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท

ณรงค์ โพธิ[1] และ สมชาย ปราการเจริญ[2]

[1]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร
[2]ภาควิชาวิทยาการคอมพิวเตอร์และสารสนเทศ คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ กรุงเทพมหานคร
Emails: narong@sueksa.go.th, spk@kmutnb.ac.th

บทคัดย่อ

งานวิจัยนี้ มีวัตถุประสงค์เพื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของข้อมูล ระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท สำหรับเกณฑ์ที่ใช้ในการเปรียบเทียบค่าความแม่นยำ คือ ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ที่ได้จากการประมาณค่าสูญหายของข้อมูล โดยใช้ฐานข้อมูลคุณลักษณะที่พึงประสงค์ของอาจารย์ผู้สอนคอมพิวเตอร์ จากคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ (ปี พ.ศ. 2550) จำนวน 285 ชุด ผลลัพธ์จากการเปรียบเทียบค่าความแม่นยำของข้อมูลที่ได้ใหม่ พบว่า วิธีการวิเคราะห์จำแนกประเภทให้ค่าความแม่นยำ 92.94% ซึ่งสูงกว่าวิธีเซตอย่างหยาบที่ให้ค่าความแม่นยำเพียง 90.75%.

คำสำคัญ-- การประมาณค่าสูญหายของข้อมูล; วิธีเซตอย่างหยาบ; วิธีการวิเคราะห์จำแนกประเภท

1. บทนำ

ผู้ทำการวิจัยข้อมูลทุกคนย่อมต้องการข้อมูลที่มีความสมบูรณ์ครบถ้วนมาใช้ในการวิเคราะห์และพยากรณ์ เพื่อให้ได้ผลลัพธ์ที่มีความแม่นยำและถูกต้องให้มากที่สุด แต่การนำข้อมูลมาใช้งานอาจจะไม่สามารถใช้ข้อมูลได้อย่างเต็มประสิทธิภาพ เนื่องจากอาจจะมีข้อมูลบางส่วนที่สูญหายไปหรือไม่สมบูรณ์เกิดขึ้น เนื่องมาจากขั้นตอนการจัดเก็บข้อมูล ปัญหาจากการป้อนและการตีความหมายข้อมูลผิดพลาด เครื่องมือจัดเก็บหรือโอนถ่ายข้อมูลเกิดความผิดพลาด หรือ เกิดจากข้อจำกัดของเทคโนโลยีที่ใช้ ดังนั้น ข้อมูลที่สูญหายไปจะทำให้การประมวลผลเกิดความผิดพลาดได้ จึงมีผู้สนใจศึกษาและวิจัยคิดค้นวิธีการประมาณค่าสูญหายของข้อมูลด้วยวิธีการต่างๆ ที่มีประสิทธิภาพมาใช้ประมาณค่าสูญหายของข้อมูลให้เหมาะสมกับประเภทของข้อมูลที่เกิดการสูญหาย เริ่มตั้งแต่วิธีพื้นฐานไปจนถึงวิธีที่มีความยุ่งยากซับซ้อน ได้แก่ วิธีหาค่าเฉลี่ย (Mean) และค่าฐานนิยม (Mode) ซึ่งเป็นวิธีที่ง่ายในการแทนค่าข้อมูลสูญหาย แต่วิธีการนี้อาจทำให้เกิดความเบี่ยงเบนของข้อมูลได้หรืออาจจะเหมาะสมกับข้อมูลบางประเภทเท่านั้น วิธีการประมาณค่าพารามิเตอร์ (Parameter Estimation) ได้แก่ วิธีอีเอ็ม (Expectation Maximization: EM) วิธีเอ็มแอล (Maximum Likelihood: ML) หรือวิธีทางสถิติอื่นๆ เช่น การวิเคราะห์การถดถอย (Regression Analysis) การวิเคราะห์การถดถอยเชิงพหุคูณ (Multiple Regression Analysis) การวิเคราะห์จำแนกประเภท (Discriminant Analysis) ต้นไม้การตัดสินใจ (Decision Tree) การประยุกต์ใช้ฟัสซี่เซต (Application of Fuzzy Sets) โครงข่ายประสาทเทียม (Neural Networks) การรู้จำรูปแบบ (Pattern Recognition) เซตอย่างหยาบ (Rough Set) เป็นต้น จากวิธีข้างต้นมีความเหมาะสมสำหรับการประมาณค่าสูญหายของข้อมูลที่มีความแตกต่างกันไปขึ้นอยู่กับประเภทของข้อมูลที่เกิดการสูญหายนั้นๆ

ดังนั้น เพื่อให้ผู้ใช้งานสามารถนำข้อมูลที่เกิดการสูญหายนำกลับมาใช้ได้เกือบทั้งหมด จึงต้องจำเป็นต้องนำข้อมูลที่เกิดการสูญหายนั้นมาทดแทนด้วยค่าใหม่ด้วยวิธีการประมาณค่าใหม่ให้มีความใกล้เคียงกับค่าที่สูญหายไปให้มากที่สุด เพื่อที่จะได้นำมาใช้ในการวิเคราะห์และพยากรณ์ผลลัพธ์ได้ หากเกิดการสูญหายมากๆ อาจต้องตัดข้อมูลชุดนั้นทิ้งไป

ในการนี้ ผู้วิจัยได้ตระหนักถึงความสำคัญและความจำเป็นของการประมาณค่าสูญหายของข้อมูลดังกล่าว จึงมีความสนใจที่จะศึกษาและวิจัยเกี่ยวกับเทคนิคการประมาณค่าสูญหายของข้อมูลระหว่างระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท เพื่อหาวิธีที่มีความแม่นยำที่สุด โดยใช้วิธีการวัดความแม่นยำจากค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (Mean Magnitude of Relative Error: MMRE) หากมีค่าน้อยๆ แสดงว่ามีค่าความแม่นยำสูงมากๆ

2. วัตถุประสงค์ของการวิจัย

เพื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของข้อมูลระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท

3. ขอบเขตของการวิจัย

1.) ข้อมูลที่ใช้ในการศึกษาวิจัย คือ ฐานข้อมูลคุณลักษณะที่พึงประสงค์ของอาจารย์ผู้สอนคอมพิวเตอร์ จำนวน 285 ชุด (Case) เมื่อปี พ.ศ. 2550 จากคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

2.) งานวิจัยนี้กำหนดให้ค่าของแอทริบิวต์ K411 คือผู้สอนมีความสามารถ อธิบายพื้นฐานของระบบ ปฏิบัติการ (Operating System) ให้เป็นที่เข้าใจได้ เท่านั้นที่เป็นค่าสูญหายเพื่อใช้ในการเปรียบเทียบผลลัพธ์

4. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยที่ได้รศึกษาถึงวิธีการประมาณค่าสูญหายของข้อมูลมีดังนี้

4.1. วิธีเซตอย่างหยาบ (Rough Set)

ผู้คิดค้น คือ Pawlak เมื่อ ปี ค.ศ. 1982 [1] เป็นวิธีทางคณิตศาสตร์ในการวิเคราะห์ข้อมูลที่มีความคลุมเครือ และ ความไม่แน่นอน ซึ่งอยู่บนสมมุติฐานที่ว่า สามารถหาฐานความรู้ (ข้อมูล) จากทุกวัตถุในเอกภพสัมพันธ์ วิธีการทำงาน คือ 1) ลดคุณสมบัติที่ไม่จำเป็นออกไป 2) หาค่าความสำคัญของคุณสมบัติ 3) หา Reduct ที่ดีที่สุด Discernibility Matrix และ Core สุดท้ายจะได้กฎความรู้สำหรับนำไปใช้งาน

4.2 วิธีการวิเคราะห์จำแนกประเภท (Discriminant Analysis: DA)

เป็น เทคนิค การวิเคราะห์ความสัมพันธ์ หรือ การหาสาเหตุที่สอดคล้องกัน [2],[3],[4], [5] เพื่อเปรียบเทียบลักษณะของกลุ่มประชากร 2 กลุ่มขึ้นไป โดยค่ามาก อยู่ที่กลุ่มใดจะถือว่าข้อมูลใหม่จัดอยู่ในกลุ่มนั้น ตามสูตร

$$\hat{d} = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \quad (1)$$

โดยที่ $\hat{d}$ = ค่าอำนาจการจำแนกประเภท

a = ค่าคงที่ (ส่วนตัดแกน y, เมื่อ x = 0)

$x_1, \ldots, x_n$ = ค่าตัวแปรอิสระ, ตัวแปรจำแนกกลุ่ม

$b_1, \ldots, b_n$ = ค่าสัมประสิทธิ์ของสมการจำแนกกลุ่ม

4.3 วิธีการวิเคราะห์องค์ประกอบ (Principle Component Analysis: PCA)

เป็นการนำความสัมพันธ์เชิงเส้นระหว่างตัวแปรที่ใช้เป็นข้อมูลกับองค์ประกอบหลัก เพื่ออธิบายความแปรปรวนของกลุ่มข้อมูลที่มีองค์ประกอบปัจจัยหลักมีความแปรปรวนมาก-น้อยตามลำดับจนได้องค์ประกอบหลัก (ปัจจัย) ที่สามารถเป็นตัวแทนอธิบายการผันแปรของทุกตัวแปรได้ครบถ้วน [6] สำหรับสมการที่ใช้ในการประมาณค่าปัจจัยที่ j ตามสูตร

$$F_j = w_{j1} x_1 + w_{j2} x_2 + \ldots + w_{jp} x_p + e \quad (2)$$

โดยที่ F_j = ปัจจัยหรือองค์ประกอบหลักตัวที่ j

x_p = ตัวแปรที่ p

w_{jp} = สัมประสิทธิ์ของตัวแปรที่ jp

e = ค่าความคาดเคลื่อน

4.4 วิธีการประมาณความแม่นยำ (Evaluation Criterion)

จากการใช้วิธีต่างๆ ที่สร้างขึ้น [7],[8] ต้องมีค่าความแม่นยำเข้ากันได้ (Model Best Fit) จะถูกนำไปทดสอบกับกลุ่มข้อมูลอีกชุดหนึ่งที่ทราบค่าจริง (Actual Data) ผลจากการพยากรณ์ข้อมูลชุดใหม่ (Predicted Data) จะถูกนำมาคำนวณหาค่าความคลาดเคลื่อนสัมพัทธ์ (Magnitude of Relative Error: MRE) ตามสูตร

$$MRE_i = \frac{\left| actual_i - predicted_i \right|}{actual_i} \quad (3)$$

หากข้อมูลมีจำนวนมากต้องนำมาหาค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (Mean-MRE: MMRE) โดยที่ MMRE มีค่ามาก แสดงว่า มีค่าที่ได้จากการประมาณค่าสูญหายมีความแม่นยำน้อย แต่ถ้า MMRE มีค่าเข้าใกล้หรือเท่ากับศูนย์ (0) แสดงว่าค่าที่ได้มีความแม่นยำสูงหรือใกล้เคียงกับค่าจริงมากที่สุด ตามสูตร

$$MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{\left| actual_i - predicted_i \right|}{actual_i} x100 \quad (4)$$

$$Accuracy = 100 - MMRE \quad (5)$$

4.5. งานวิจัยที่เกี่ยวข้อง

วราภรณ์ [9] ได้นำทฤษฎีเซตอย่างหยาบมาใช้ในการวิเคราะห์ระบบสารสนเทศของนักศึกษาเพื่อค้นหากฎที่สามารถนำมาใช้ช่วยคาดหมายสถานภาพของนักศึกษา โดยพิจารณาจากประวัติและผลการเรียนในรายวิชาต่างๆ พบว่า วิธีนี้เหมาะสมและให้ผลลัพธ์ที่ใกล้เคียงที่สุด

จริยา ประสิทธิ์ และ อำไพ [10] ได้ทำงานวิจัยเกี่ยวกับการเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ พบว่าให้ความแม่นยำสูงกว่าการแทนค่าด้วยวิธีหาค่าเฉลี่ย

ฉรงค์ และ คณะ [11] ได้ทำงานวิจัยเกี่ยวกับการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ พบว่า วิธีการวิเคราะห์จำแนกประเภทมีความแม่นยำมากที่สุด

ฉรงค์ และ คณะ [12] ได้ทำงานวิจัยเกี่ยวกับการเปรียบเทียบวิธีการประมาณค่าสูญหาย โดยวิธีการวิเคราะห์จำแนกประเภท ต้นไม้การตัดสินใจและค่าเฉลี่ย พบว่า วิธีการวิเคราะห์จำแนกประเภทให้ความแม่นยำสูงสุด

สมชาย [7] ได้ศึกษาวิธีการประมาณการเวลาในการพัฒนาซอฟต์แวร์ประยุกต์เชิงโครงข่าย โดยวิธีแบบจำลองสมการโครงสร้าง มีความแม่นยำมากที่สุด

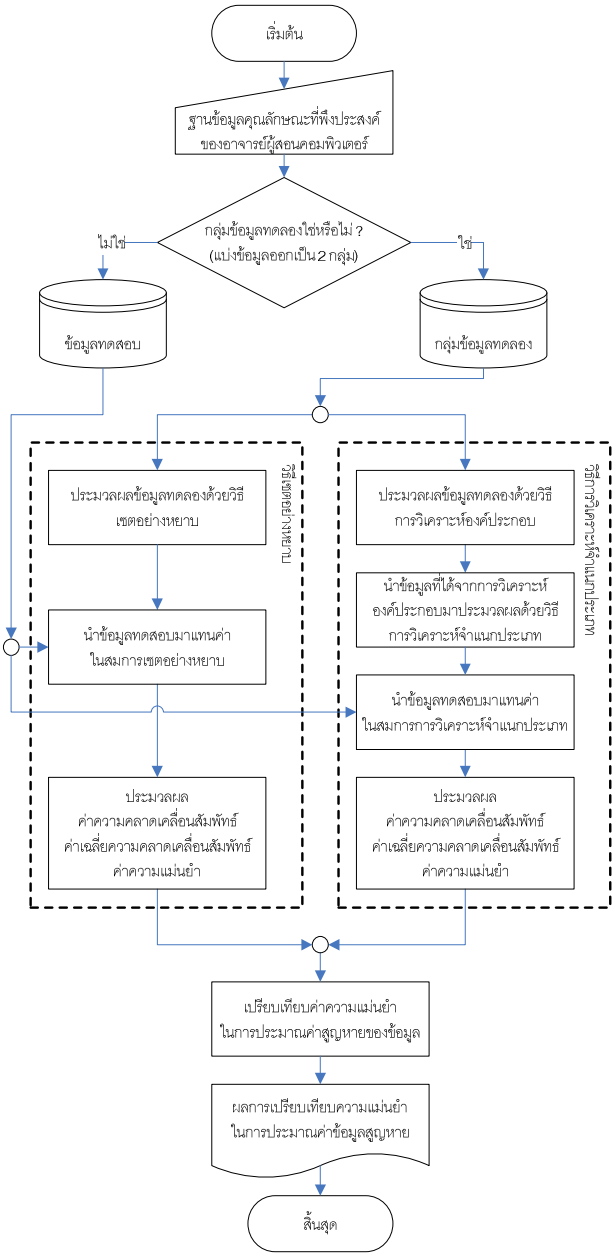
ฉรงค์ และ สมชาย [13] ได้วิจัยการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย โดยวิธีแบบจำลองสมการโครงสร้าง ระหว่างแบบผ่านและไม่ผ่านวิธีการวิเคราะห์จำแนกประเภท พบว่า การนำข้อมูลไปผ่านกระบวนการวิเคราะห์จำแนกประเภทก่อนนำมาแทนค่าในสมการแบบจำลองสมการโครงสร้างให้ความแม่นยำสูงสุด

Hidetomo [14] ได้ทำงานวิจัยเกี่ยวกับวิธีการจัดกลุ่มข้อมูลจากค่าข้อมูลเริ่มต้นและค่าข้อมูลที่สูญหายด้วยวิธี Fuzzy C-Means (FCM) พบว่าให้ความแม่นยำมากที่สุด

สรุปงานวิจัยที่ได้ศึกษาข้างต้นเป็นการใช้เทคนิคการประมาณค่าสูญหายของข้อมูลที่มีความแตกต่างกันซึ่งขึ้นอยู่กับประเภทและลักษณะของข้อมูลที่นำมาใช้งาน สำหรับงานวิจัยนี้ ผู้วิจัยจึงได้ทดลองการประมาณค่าสูญหายกับฐานข้อมูลคุณลักษณะที่พึงประสงค์ของอาจารย์ผู้สอนคอมพิวเตอร์ จำนวน 285 ชุด ที่ได้จากแบบสอบถาม เมื่อปี พ.ศ. 2550 ของคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ เพื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของข้อมูลระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท

5. กรอบแนวคิดสำหรับการประมาณค่าสูญหายของงานวิจัย

กรอบแนวคิดสำหรับการประมาณค่าสูญหายของงานวิจัย ดังรูปที่ 1



รูปที่ 1. กรอบแนวคิดสำหรับการประมาณค่าสูญหายของงานวิจัย

6. วิธีการดำเนินการวิจัย

วิธีการประมาณค่าสูญหายของข้อมูลระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท ผู้วิจัยมีขั้นตอนและวิธีการทำงานดังนี้

1.) นำฐานข้อมูลคุณลักษณะที่พึงประสงค์ของอาจารย์ผู้สอนคอมพิวเตอร์ ซึ่งมีแอทริบิวต์ทั้งหมด 46 ตัว ที่ได้จากแบบสอบถามดังตารางที่ 1 จำนวน 285 ชุด แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มข้อมูลทดลอง (Training Data) 200 ชุด และกลุ่มข้อมูลทดสอบ (Testing Data) 85 ชุด

ตาราง 1. ข้อมูล 46 แอทริบิวต์ของฐานข้อมูลที่ใช้ในการวิจัย

ชื่อข้อมูล	คำอธิบายความหมายของข้อมูล	ช่วงข้อมูล
M101	ผู้สอนให้คำแนะนำเมื่อนักศึกษาทำงานผิดพลาด	1-5
M102	ผู้สอนมีพฤติกรรมที่แสดงออกด้วยวาจาหรือการกระทำที่ให้การช่วยเหลือต่อนักศึกษาโดยไม่หวังผลตอบแทน	1-5
M103	ผู้สอนให้โอกาสนักศึกษาแก้ตัวเมื่อมีการทำงานส่งไม่ถูกต้อง	1-5
M104	ผู้สอนตอบแทนบุญคุณของผู้มีอุปการะคุณ หน่วยงาน สถานที่ ตามโอกาสอันสมควร	1-5
M105	ผู้สอนมีความมั่นคงทางอารมณ์ สุขุม รอบคอบ ใช้เหตุผลตัดสินปัญหา	1-5
M106	ผู้สอนมีความร่วมมือและประสานงานกับนักศึกษา ผู้มีส่วนได้ส่วนเสีย บุคคลต่างๆ และชุมชน	1-5
M107	ผู้สอนมีความเที่ยงตรง เปิดเผย จริงใจ กับทุกๆ คน	1-5
M108	ผู้สอนปฏิบัติตนตามกฎระเบียบของหน่วยงานและประเทศชาติ	1-5
M109	ผู้สอนให้ความเท่าเทียมกันกับนักศึกษาทุกคนเหมือนกัน	1-5
M110	ผู้สอนมีความเป็นกัลยาณมิตร และเอื้ออาทรต่อนักศึกษา เช่น รักและเมตตา ช่วยเหลือ เอาใจใส่ให้กำลังใจ โดยเสมอภาค	1-5
T211	ผู้สอนอธิบายตรงประเด็นและยกตัวอย่างประกอบให้เข้าใจชัดเจน	1-5
T212	ผู้สอนใช้อุปกรณ์/เครื่องมือ/สถานที่/มีความเหมาะสมกับจำนวนนักศึกษาและเนื้อหาวิชา	1-5
T213	ผู้สอนให้คำแนะนำที่ดี สามารถนำมาปฏิบัติได้	1-5
T214	ผู้สอนแนะนำนักศึกษาให้ใช้เครื่องมือ หรือข้อมูลใหม่ๆ ที่สามารถค้นคว้าเพิ่มเติมได้	1-5
T215	ผู้สอนใช้ภาษาที่ง่าย ชัดเจน ในการถามคำถาม เพื่อให้ผู้ฟังเข้าใจง่าย ไม่สับสนกับคำถาม	1-5
T221	ผู้สอนมีเอกสารแผนการสอน พร้อมทั้งอธิบายจุดประสงค์ขอบเขต เนื้อหาของวิชา แนะนำวิธีการเรียนการสอนและการประเมินผลวิชาเรียนให้นักเรียนทราบ ก่อนการเรียนการสอนจะเริ่มต้น	1-5
T222	ผู้สอนมีการกำหนดเวลาในกิจกรรมการเรียนการสอนได้อย่างเหมาะสม	1-5
T223	ผู้สอนยอมรับข้อมูลใหม่ที่มีเหตุผลถูกต้องหรือดีกว่าเดิม	1-5
T224	ผู้สอนสามารถกระตุ้นผู้เรียนให้คิดหาความเชื่อมโยงระหว่างเนื้อหาวิชาอื่นๆ	1-5
T225	ผู้สอนมีช่องทางในการตอบปัญหาหรือข้อสงสัยของนักศึกษาหลายช่องทาง เช่น MSN, E-mail, Web board	1-5

ตาราง 1. ข้อมูล 46 แอทริบิวต์ของฐานข้อมูลที่ใช้ในการวิจัย (ต่อ)

ชื่อข้อมูล	คำอธิบายความหมายของข้อมูล	ช่วงข้อมูล
T231	ผู้สอนสอบถามความรู้และความเข้าใจของนักศึกษาก่อนเรียนทุกครั้ง	1-5
T232	ผู้สอนสอบถามความรู้และความเข้าใจของนักศึกษาหลังเรียนทุกครั้ง	1-5
T233	ผู้สอนมีการสอบตรงตามเนื้อหา วัตถุประสงค์ที่กำหนดไว้ในแผนการสอน และควรมีการเฉลยข้อคำถาม	1-5
B311	ผู้สอนแต่งกาย สะอาด เรียบร้อย เหมาะสมกับกาลเทศะ	1-5
B312	ผู้สอนแสดงออกทางวาจาและกิริยาด้วยความเป็นมิตรและให้เกียรตินักศึกษา	1-5
B313	ผู้สอนมีกำลังกาย และกำลังใจในการสอนอย่างเต็มเปี่ยม	1-5
B321	ผู้สอนสามารถรับรู้พฤติกรรมของนักศึกษาทุกคนได้	1-5
B322	ผู้สอนมีการติดตามความเคลื่อนไหวและการเปลี่ยนแปลงของความรู้ใหม่ๆ	1-5
B323	ผู้สอนสามารถควบคุมอารมณ์ตนเอง โดยใช้เหตุผลในการแก้ไขปัญหา	1-5
B324	ผู้สอนแลกเปลี่ยนและยอมรับฟังความคิดเห็นของผู้อื่นเกี่ยวกับแนวทางการทำงานหรือแก้ปัญหาด้านการสอน	1-5
B325	ผู้สอนปฏิบัติตนเป็นตัวอย่างที่ดีกับนักศึกษาและบุคคลรอบข้าง	1-5
B326	ผู้สอนสามารถนำทักษะความรู้มาปรับใช้ให้เกิดประโยชน์สูงสุด	1-5
K411	ผู้สอนมีความสามารถอธิบายพื้นฐานของระบบปฏิบัติการ (Operating System) ให้เป็นที่เข้าใจได้	1-5
K412	ผู้สอนมีความสามารถอธิบายพื้นฐานของการประมวลผลข้อมูล (Data Processing) ให้เป็นที่เข้าใจได้	1-5
K413	ผู้สอนมีความสามารถอธิบายพื้นฐานของการเขียนโปรแกรม (Computer Programming) ให้เป็นที่เข้าใจได้	1-5
K414	ผู้สอนมีความสามารถอธิบายพื้นฐานของโครงสร้างข้อมูลและการวิเคราะห์อัลกอธิธึม (Data Structure and Algorithm Analysis) ให้เป็นที่เข้าใจได้	1-5
K415	ผู้สอนมีความสามารถอธิบายพื้นฐานของระบบสื่อสารข้อมูลและเครือข่ายคอมพิวเตอร์ (LAN, WAN) ให้เป็นที่เข้าใจได้	1-5
K416	ผู้สอนมีความสามารถอธิบายพื้นฐานของการวิเคราะห์และออกแบบระบบ (System Analysis and Design) ให้เป็นที่เข้าใจได้	1-5
K417	ผู้สอนมีความสามารถอธิบายพื้นฐานของระบบฐานข้อมูล (Database System) ให้เป็นที่เข้าใจได้	1-5
K421	ผู้สอนมีการศึกษาค้นคว้าเพื่อพัฒนาตนเองในด้านต่างๆอย่างต่อเนื่อง	1-5
K422	ผู้สอนนำผลงานการวิจัยมาใช้ประโยชน์ในการเรียนการสอน	1-5
S511	ผู้สอนชมเชยและให้กำลังใจเมื่อนักศึกษาให้ความสนใจและปฏิบัติตนในการรักษามรดกของชุมชน อาทิ การจัดทำเว็บไซต์ เพื่อถ่ายทอดมรดกของชุมชน	1-5
S512	ผู้สอนมีการประยุกต์ใช้เทคโนโลยีกับศิลปะและวัฒนธรรมไทย อาทิ การสร้างตัวอักษร หรือการส่งเสริมการออกแบบด้วยวัฒนธรรมไทย	1-5

ตาราง 1. ข้อมูล 46 แอทริบิวต์ของฐานข้อมูลที่ใช้ในการวิจัย (ต่อ)

ชื่อข้อมูล	คำอธิบายความหมายของข้อมูล	ช่วงข้อมูล
S521	ผู้สอนมีส่วนร่วมในกิจกรรมต่าง ๆ ที่ทางชุมชนจัดขึ้นมา	1-5
S522	ผู้สอนนำเอาอุปกรณ์ด้าน ICT มาช่วยเสริมความรู้ ทักษะให้กับชุมชน เช่น การใช้อินเทอร์เน็ต และโปรแกรมสำเร็จรูป เป็นต้น	1-5
S523	ผู้สอนเป็นผู้ประสานความร่วมมือระหว่างบ้านและชุมชน	1-5

2.) วิธีการประมาณค่าสูญหายของข้อมูลด้วยวิธีการวิเคราะห์จำแนกประเภท มีขั้นตอนดังต่อไปนี้

2.1) นำข้อมูลกลุ่มทดลองมาวิเคราะห์หาองค์ประกอบได้จำนวน 9 องค์ประกอบ (ปัจจัย) ได้ดังสมการที่ 6-14 ประกอบด้วย

G1 = {S523, S522, S521, S512, S511, K422, K421} (6)

G2 = {K411, K412, K413, K414, K415, K416, K417} (7)

G3 = {B323, B322, B321, B324, B313, B325, B326} (8)

G4 = {M19, M18, M17, T212, M110} (9)

G5 = {T214, T232, T231, T225, T233} (10)

G6 = {T211, T213, T224, T222, T223, T215} (11)

G7 = {M11, M13, M12, M15} (12)

G8 = {B311, T221, B312} (13)

G9 = {M14, M16} (14)

2.2) นำข้อมูลองค์ประกอบที่มีความสัมพันธ์กันกับกลุ่มของค่าที่เกิดการสูญหาย (K411) ในที่นี้ คือ G2 มาคำนวณหาสมการทดแทนค่าสูญหายด้วยวิธีการวิเคราะห์จำแนกประเภท ได้สมการประมาณค่า 4 สมการ ดังสมการ (บางส่วน) ที่ 15-18 และค่าสัมประสิทธิ์ของสมการวิเคราะห์จำแนกประเภท ดังตารางที่ 2

ตาราง 2. ค่าสัมประสิทธิ์ของสมการวิเคราะห์จำแนกประเภท

Classification Function Coefficients				
	K411			
	น้อย	ปานกลาง	มาก	มากที่สุด
K412	6.101	7.567	9.125	11.479
K413	3.306	5.214	5.389	6.119
K414	2.197	1.849	2.310	2.335
K415	0.525	1.272	1.481	1.843
K416	1.149	2.011	2.553	2.827
K417	1.573	2.740	3.334	3.666
(Constant)	-19.712	-36.330	-49.269	-66.983
Fisher's linear discriminant functions				

$$\hat{d}_1 = -19.712+(6.101*K412) + (3.306*K413) + ... + (1.573*K417) \quad (15)$$

$$\hat{d}_2 = -36.330 + (7.567*K412) + (5.214*K413) + ... + (2.740*K417) \quad (16)$$

$$\hat{d}_3 = -49.269 + (9.125*K412) + (5.389*K413) + ... + (3.334*K417) \quad (17)$$

$$\hat{d}_4 = -66.983 + (11.479*K412) + (6.119*K413) + ... + (3.666*K417) \quad (18)$$

2.3.) นำข้อมูลกลุ่มทดสอบมาแทนค่าในสมการที่ 15-18 แล้วทำการคำนวณหาค่าความคลาดเคลื่อน (MRE) ในแต่ละข้อมูลทดสอบ จากนั้นนำผลลัพธ์ที่ได้มาคำนวณหาค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) เท่ากับ 7.06% แสดงว่ามีค่าความแม่นยำ (Accuracy) เท่ากับ 92.94%

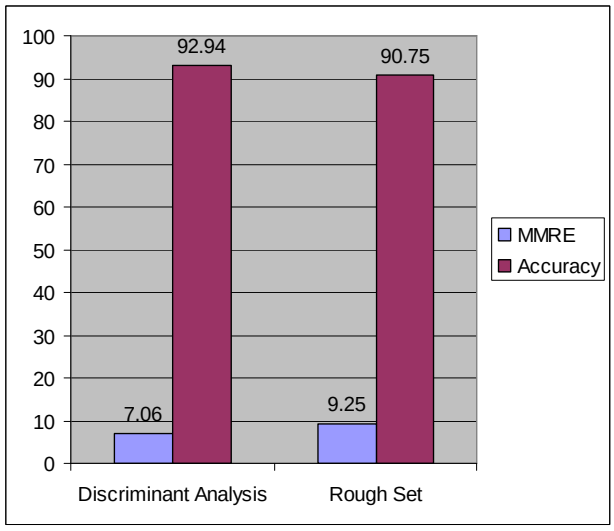
3.) วิธีการประมาณค่าสูญหายของข้อมูลด้วยวิธีเซตอย่างหยาบ โดยคำนวณผลลัพธ์ด้วยโปรแกรม RSES (Rough Set Exploration System) พบว่า มีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) เท่ากับ 9.25% แสดงว่ามีค่าความแม่นยำ (Accuracy) เท่ากับ 90.75%

7. ผลการวิจัย

ผลการวิจัยเพื่อเปรียบเทียบความแม่นยำในการประมาณค่าสูญหายของข้อมูล ระหว่างวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภท โดยข้อมูลที่ใช้วิจัย คือ ข้อมูลคุณลักษณะที่พึงประสงค์ของอาจารย์ผู้สอนคอมพิวเตอร์ จากคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ (ปี พ.ศ. 2550) จำนวน 285 ชุด มาแบ่งเป็น 2 กลุ่ม คือ ข้อมูลกลุ่มทดลอง 200 ชุด และข้อมูลกลุ่มทดสอบ 85 ชุด

ผลลัพธ์จากการประมาณค่าสูญหายของข้อมูลด้วยวิธีการวิเคราะห์จำแนกประเภท โดยการนำข้อมูลกลุ่มทดลองไปผ่านการวิเคราะห์องค์ประกอบก่อนแล้วจึงนำข้อมูลปัจจัยที่มีความสัมพันธ์กันกับข้อมูลแอทริบิวต์ที่เกิดการสูญหายมาประมวลผลด้วยวิธีการวิเคราะห์จำแนกประเภท เพื่อหาสมการที่เหมาะสมในการประมาณค่าสูญหายของข้อมูล จากนั้นจึงนำข้อมูลกลุ่มทดสอบมาแทนค่าในสมการเพื่อประมาณค่าใหม่ของข้อมูลที่สูญหายไป ซึ่งการวัดความแม่นยำจากค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ของข้อมูล พบว่า มีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์เท่ากับ 7.06% หรือ แสดงว่ามีความแม่นยำของการประมาณค่าสูญหายถูกต้อง เท่ากับ 92.94%

ผลลัพธ์จากการประมาณค่าสูญหายของข้อมูลด้วยวิธีเซตอย่างหยาบ โดยการนำข้อมูลมาประมวลผลด้วยโปรแกรม RSES (Rough Set Exploration System) พบว่า มีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ เท่ากับ 9.25% หรือแสดงว่ามีค่าความแม่นยำของการประมาณค่าสูญหายถูกต้อง เท่ากับ 90.75% ซึ่งสรุปได้ดังรูปที่ 2 และตารางที่ 3



รูปที่ 2. กราฟเปรียบเทียบค่า MMRE และ Accuracy ของวิธี DA และ Rough Set

ตาราง 3. ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ของวิธี DA และ Rough Set

	Discriminant Analysis	Rough Set
MMRE	22.21	9.25
Accuracy	77.79	90.75

8. บทสรุป วิจารณ์ผล และข้อเสนอแนะ

สรุปผลการวิจัยการประมาณค่าสูญหายของข้อมูลด้วยวิธีเซตอย่างหยาบและวิธีการวิเคราะห์จำแนกประเภทจะเป็นอีกทางเลือกหนึ่งในการทดแทนค่าสูญหายของข้อมูลที่จะนำไปใช้ในงานวิจัย การพยากรณ์ หรือ ประมวลผล เพื่อให้ได้ผลลัพธ์ในการทดแทนค่าสูญหายของข้อมูลใหม่ที่มีความคลาดเคลื่อนจากข้อมูลจริงให้น้อยที่สุด จากวิธีการข้างต้นพบว่ามีความแม่นยำมากกว่าวิธีพื้นฐานทั่วๆ ไป แต่ทั้งนี้ต้องขึ้นอยู่กับลักษณะของข้อมูลที่นำมาใช้ในประมาณค่าสูญหายด้วย เพราะข้อมูลบางประเภทไม่เหมาะสมที่จะใช้วิธีการนี้

วิจารณ์ผลลัพธ์ที่ได้จากการประมาณค่าสูญหายของข้อมูลด้วยวิธีการวิเคราะห์จำแนกประเภทจะให้ค่าความแม่นยำมากกว่าวิธีเซตอย่างหยาบ เนื่องจากก่อนที่จะนำข้อมูลมาประมวลผลด้วยวิธีการนี้จะมีการนำข้อมูลไปผ่านการวิเคราะห์หาองค์ประกอบเพื่อจัดกลุ่มข้อมูลที่มีความใกล้เคียงกันมาอยู่ด้วยกันก่อนนำมาประมวลผลด้วยวิธีการวิเคราะห์จำแนกประเภท จึงทำให้ผลลัพธ์ออกมามีความแม่นยำมากยิ่งขึ้น แต่บางครั้งอาจจะไม่เหมาะกับข้อมูลบางประเภท จึงอาจจะทำให้ผลลัพธ์ออกมาไม่มีความแม่นยำเท่าที่ควร ซึ่งอาจจะเกิดจากข้อมูลหรือแอทริบิวต์ในฐานข้อมูลไม่มีความสัมพันธ์กัน วิธีการประมาณค่าสูญหายด้วยวิธีการนี้จะใช้ไม่ได้ผล เพราะการจะคำนวณหาค่าข้อมูลที่สูญหายไปของข้อมูลในฐานข้อมูลต้องอาศัยการประมาณค่าจากแอทริบิวต์ที่มีความสัมพันธ์กันจึงจะให้ความแม่นยำมากที่สุด

ข้อเสนอแนะการทำวิจัยครั้งต่อไปเพื่อให้ได้ผลลัพธ์ที่แม่นยำมากยิ่งขึ้น มีดังนี้

1.) ควรมีการจัดกลุ่มข้อมูล (Data Grouping) แบบหลัก (Column) และแบบแถว (Row) ก่อนนำข้อมูลมาใช้งาน เนื่องจากวิธีการทั้งสองอย่างนี้จะทำให้ได้ข้อมูลที่เป็นกลุ่มเดียวกัน การประมาณค่าทดแทนค่าสูญหายจึงจะได้ค่าความแม่นยำสูงสุด

2.) เทคนิคจากผลการวิจัยนี้ เป็นการเปรียบเทียบความแม่นยำของวิธีการประมาณค่าสูญหายเชิงตัวเลข ดังนั้น ในกรณีที่ข้อมูลเป็นเชิงคุณภาพ ควรใช้เทคนิคและวิธีการอื่นๆ ที่เหมาะสมกับประเภทของข้อมูลเข้ามาเปรียบเทียบผลลัพธ์ เพื่อให้ได้ความแม่นยำที่อยู่ในระดับเดียวกัน

3.) ข้อมูลที่ใช้ในการวิจัยครั้งนี้ สร้างจากกลุ่มข้อมูลทดลอง จำนวน 200 ชุด และใช้กลุ่มข้อมูลทดสอบ จำนวน 85 ชุด จะเห็นได้ว่ากลุ่มข้อมูลทดลองยังมีจำนวนน้อยอยู่ หากมีการเพิ่มกลุ่มข้อมูลทดลองให้มากขึ้นจะทำให้ผลการวิจัยมีความมั่นใจและได้รับผลจากการประมาณค่าทดแทนค่าสูญหายที่มีความใกล้เคียงกับข้อมูลจริงมากกว่านี้ หากข้อมูลมีจำนวนจำกัดอาจต้องใช้เทคนิค Bootstrap หรือ Jackknife เข้ามาช่วยในการสร้างกลุ่มข้อมูลทดลองให้มากขึ้นตามความเหมาะสมหรือตามที่ผู้วิจัยต้องการ

4.) ข้อมูลที่ใช้ในการวิจัยครั้งต่อไปควรใช้อย่างน้อย 2 ชุดข้อมูลขึ้นไป ที่เป็นลักษณะเดียวกันเพื่อยืนยันผลลัพธ์และความแม่นยำของวิธีการที่ใช้งานว่ามีความเหมาะสม

5.) การวิจัยครั้งนี้กำหนดให้ค่าของแอทริบิวต์ในฐานข้อมูล จำนวน 1 แอทริบิวต์ เท่านั้นที่เป็นค่าสูญหาย ดังนั้น การพัฒนาและวิจัยครั้งต่อไปควรเพิ่มจำนวนของแอทริบิตว์ที่สูญหายด้วย

เอกสารอ้างอิง

[1] Z. Pawlak. "Rough Sets". *Internat. J. Comput Inform Sci*. vol. 11, no. 5, pp 341-356. 1982

[2] กัลยา วานิชย์บัญชา. **การวิเคราะห์ข้อมูลด้วย SPSS for Windows**. กรุงเทพฯ : ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย, 2542.

[3] กัลยา วานิชย์บัญชา. **การวิเคราะห์ข้อมูลหลายตัวแปร**. พิมพ์ครั้งที่ 2. กรุงเทพฯ : ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย, 2550.

[4] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. USA, 2005.

[5] มนต์ชัย เทียนทอง. **สถิติและวิธีการวิจัยทางเทคโนโลยีสารสนเทศ**. พิมพ์ครั้งที่ 1. กรุงเทพฯ : สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2550.

[6] Miguel A. Crreira-Perpinan. "A Review of Dimension Reduction Techniques". *Technical Report CS-96-09*, Dept of Computet Science. University of Sheffield, 1997.

[7] สมชาย ปราการเจริญ. "การประมาณการเวลาในการพัฒนาซอฟต์แวร์ประยุกต์เชิงโครงข่าย โดยวิธีแบบจำลองสมการโครงสร้าง". **วารสารเทคโนโลยีสารสนเทศ**. ปีที่ 4, ฉบับที่ 7 (ม.ค.- มิ.ย. 2551).

[8] Martin Sheppard. *Effort estimating using analogy*. USA : Bournemouth University,. 1996.

[9] วราภรณ์ โปตะวัฒน์. "การใช้ทฤษฎี Rough Set เพื่อการวิเคราะห์ข้อมูลอัตโนมัติ". วิทยานิพนธ์วิศวกรรมมหาบัณฑิต คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีสุรนารี, 2548.

[10] จริยา แสงสุวรรณ ประสิทธิ์ พยัคฆพงษ์ และ อำไพ ทองธีรภาพ. "การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ". **วารสารวิทยาศาสตร์ประยุกต์**. ปีที่ 7, ฉบับที่ 1 (มิ.ย. 2551). หน้า 1-7.

[11] ฉรงค์ โพธิ สมชาย ปราการเจริญ เดช ธรรมศิริ ภรัณยา อำมฤครัตน์ และ วาทินี นุ้ยเพียร. "การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ". **การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5**. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2552.

[12] ฉรงค์ โพธิ สมชาย ปราการเจริญ เดช ธรรมศิริ ภรัณยา อำมฤครัตน์ และ วาทินี นุ้ยเพียร. "การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการวิเคราะห์จำแนกประเภท ต้นไม้การตัดสินใจ และ ค่าเฉลี่ย". **การประชุมทางวิชาการระดับชาติด้านวิทยาศาสตร์และเทคโนโลยี ครั้งที่ 2**. สกลนคร: มหาวิทยาลัยราชภัฏสกลนคร, 2552.

[13] ฉรงค์ โพธิ และ สมชาย ปราการเจริญ. "การเปรียบเทียบความแม่นยำวิธีการประมาณค่าสูญหาย โดยวิธีแบบจำลองสมการโครงสร้างระหว่างแบบผ่านและไม่ผ่านวิธีการวิเคราะห์จำแนกประเภท. **การประชุมวิชาการวิทยาศาสตร์และเทคโนโลยี ครั้งที่ 8**. ปทุมธานี: มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต, 2553.

[14] Hidetomo Ichihashi. "Fuzzy c-Means Classifier with Deterministic Initialization and Missing Value Imputation". *Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007),* 2007.