การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย โดยวิธีการวิเคราะห์จำแนกประเภท ต้นไม้การตัดสินใจ และ ค่าเฉลี่ย

- *ณรงค์ โพธิ*
- *สมชาย ปราการเจริญ*

การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหาย โดยวิธีการวิเคราะห์จำแนกประเภท ต้นไม้การตัดสินใจ และ ค่าเฉลี่ย

A Comparative Imputation Methods Study by Discriminant Analysis, Decision Tree and Mean

ณรงค์ โพธิ และสมชาย ปราการเจริญ [1]

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ ได้แก่ 1) การวิเคราะห์จำแนกประเภท 2) ต้นไม้การตัดสินใจ 3) ค่าเฉลี่ย เกณฑ์เปรียบเทียบประสิทธิภาพของการประมาณค่าสูญหายใช้ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) ข้อมูลที่ใช้ในการศึกษาวิจัย คือ ฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านม (Mammographic mass) ของ UCI Machine learning repository data set จำนวน 800 ชุด (Case) วิธีการวิจัยมีดังต่อไปนี้ 1) แบ่งฐานข้อมูลเป็น 2 กลุ่ม ที่ระดับความเชื่อมั่น 95% ด้วยวิธีของ Taro Yamane คือ กลุ่มข้อมูลเรียนรู้ (Training data) 533 ชุด และ กลุ่มข้อมูลทดสอบ (Testing data) 267 ชุด 2) นำข้อมูลเรียนรู้มาหาสมการทดแทนค่าสูญหายของแต่ละวิธี 3) นำข้อมูลทดสอบมาแทนค่าสมการที่ได้เพื่อหาค่าทดแทนค่าสูญหาย 4) คำนวณหาค่าความคลาดเคลื่อนสัมพัทธ์และค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ของแต่ละวิธี ผลการวิจัยพบว่าวิธีที่ดีที่สุดเรียงตามลำดับจากมากไปน้อยเป็นดังนี้ 1) การวิเคราะห์จำแนกประเภท (MMRE=26.56%) 2) ต้นไม้การตัดสินใจ (MMRE=33.30%) และ 3) ค่าเฉลี่ย (MMRE=63.26%) ดังนั้น การวิเคราะห์จำแนกประเภทจึงมีความเหมาะสมกับข้อมูลที่มีความสัมพันธ์กันและมีการกระจายออกเป็นกลุ่มที่มีความแตกต่างกันอย่างชัดเจน การประมาณค่าสูญหายจึงจะมีความใกล้เคียงกับค่าของข้อมูลจริงมากที่สุด

Abstract

The objective of this research was to compare statistical imputation methods: 1) Discriminant Analysis (DA), 2) Decision Tree, and 3) Mean. The criteria for the efficiency comparison was estimated by Mean Magnitude of Relative Error (MMRE). The data used for the study were extracted from 800 sets of the mammographic mass database of UCI Machine Learning Repository Data Set. The research processes included: 1) Divided the data of 800 sets into two groups (95% Reliability based on Yamane's formula, resulting in 533 sets of learning data and 267 sets of testing data), 2) Used the data of learning sets to form an

[1] มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

equation for finding the substitution of incomplete values provided, 3) Replaced the missing values of the equation with the data of testing sets, and 4) Computed Magnitude of Relative Error (MRE) and Mean Magnitude of Relative Error (MMRE) for each of 3 methods. In summary, it was found that the best approaches for calculating MMRE ranged from 1) Discriminant Analysis, 2) Decision Tree, and 3) Mean (26.56%, 33.30%, 63.27%) respectively. Discriminant Analysis, therefore, was appropriate approach for data that could be identified correlation and had the high degree of dispersion in order to predict the best results for missing values.

บทนำ

ความเป็นมาและความสำคัญของปัญหา

โดยทั่วไปงานวิจัยต้องการข้อมูลที่มีความครบถ้วนสมบูรณ์ในการวิเคราะห์และพยากรณ์ เพื่อให้ได้ผลลัพธ์ที่แม่นยำและถูกต้องมากที่สุด แต่เนื่องจากการเก็บข้อมูลอาจมีบางส่วนที่ขาดหายไปหรือไม่สมบูรณ์เกิดขึ้นอยู่เสมอ ดังนั้น เพื่อให้สามารถนำข้อมูลทั้งหมดมาใช้งานได้ ผู้วิจัยจึงต้องนำชุดข้อมูลที่สูญหายมาแทนค่าด้วยวิธีประมาณค่าที่ใกล้เคียงที่สุด เพื่อนำมาใช้ในการวิเคราะห์และพยากรณ์ผลลัพธ์ หากข้อมูลเกิดการสูญหายมากๆ ผู้วิจัยอาจต้องตัดข้อมูลชุดนั้นทิ้งไป

สำหรับการเปรียบเทียบประสิทธิภาพของการประมาณค่าสูญหายที่ได้จากสมการ ใช้วิธีการทางสถิติ 3 วิธี คือ 1) การวิเคราะห์จำแนกประเภท 2) ต้นไม้การตัดสินใจและ 3) ค่าเฉลี่ย ผู้วิจัยใช้ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (Mean magnitude of relative error: MMRE) หากมีค่าน้อย แสดงว่ามีความแม่นยำมาก

ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้ในการศึกษาวิจัย คือ ฐานข้อมูลจากภาพถ่ายก้อนเนื้อเต้านม (Mammographic mass) ของ UCI Machine learning repository data Set จำนวน 800 ชุด (Case)

2. การวิจัยนี้กำหนดให้ค่าของแอทริบิวต์ Shape เท่านั้นที่เป็นค่าสูญหาย เพื่อใช้ในการศึกษาเปรียบเทียบ

3. ลักษณะของการสูญหายของข้อมูลจัดอยู่ในประเภทการสูญหายสมบูรณ์เชิงสุ่ม (Missing Completely Random: MCAR) (Garson G. David. 2005)

วัตถุประสงค์ของการวิจัย

1. เพื่อเปรียบเทียบวิธีการประมาณค่าทดแทนค่าสูญหายของข้อมูลโดยวิธีทางสถิติ(Discriminant Analysis, Decision Tree, Mean)

2. เพื่อประมาณความแม่นยำในการประมาณค่าทดแทนของค่าสูญหายของแอทริบิวต์ Shape ของฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านมในแต่ละวิธี

ทฤษฎีและเทคโนโลยีที่เกี่ยวข้อง

1. การวิเคราะห์จำแนกประเภท (Discriminant Analysis: DA)เป็นเทคนิคการวิเคราะห์ความสัมพันธ์หรือการหาสาเหตุที่สอดคล้องกัน (กัลยา วานิชย์บัญชา. 2542; กัลยา วานิชย์บัญชา. 2550; มนต์ชัย เทียนทอง. 2550; Geoffrey J. McLachlan. 2005) เพื่อเปรียบเทียบลักษณะของกลุ่มประชากร 2 กลุ่มขึ้นไป โดยค่ามากอยู่ที่กลุ่มใดจะถือว่าข้อมูลใหม่จัดอยู่ในกลุ่มนั้นทันที โดยใช้สูตร

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

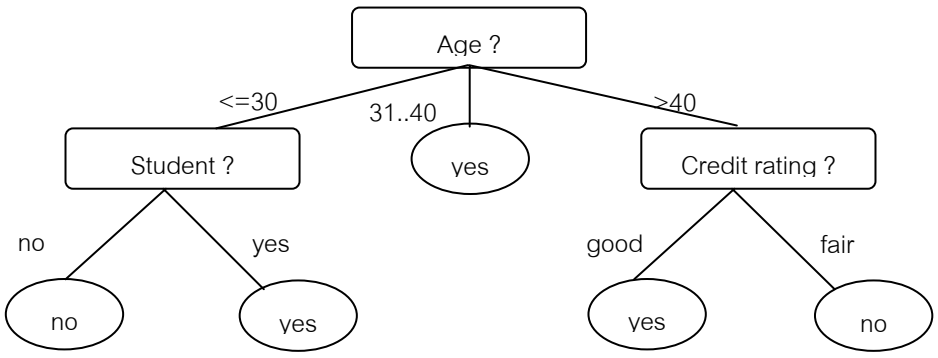
โดยที่ $\hat{d}$ = ค่าอำนาจการจำแนกประเภท

a = ค่าคงที่ (ส่วนตัดแกน Y, เมื่อ X = 0)

$x_1, ..., x_n$ = ค่าตัวแปรอิสระ, ตัวแปรจำแนกกลุ่ม

$b_1, ..., b_n$ = ค่าสัมประสิทธิ์ของสมการจำแนกกลุ่ม

2. ต้นไม้การตัดสินใจ (Decision Tree) เป็นส่วนหนึ่งของกระบวนการทางด้านเหมืองข้อมูล (ณรงค์ โพธิ และคณะ. 2551; พยูน พาณิชย์กุล. 2548 ; สกลนันท์ หุ่นเจริญ. 2548) ที่ได้นำเอาการตัดสินใจแบบโครงสร้างต้นไม้มาใช้เป็นตัวช่วยในการทำงานเกี่ยวกับการช่วยในการตัดสินใจต่าง ๆ ไม่ว่าจะเป็นด้านระบบธุรกิจหรืออื่นๆ โดยปกติมักประกอบด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” ดังภาพที่ 1



ภาพที่ 1 กระบวนการของต้นไม้การตัดสินใจ

3. ค่าเฉลี่ย (Mean) เป็นเทคนิคที่ใช้ในการหาค่าเฉลี่ยแบบกลางๆ(Darlington, R.B. 1990; ชูศรี วงศ์รัตนะ. 2541; กัลยา วานิชย์บัญชา. 2542; กัลยา วานิชย์บัญชา. 2550; มนต์ชัย เทียนทอง. 2550) จากชุดข้อมูลที่มีข้อมูลสมบูรณ์ โดยใช้สูตร

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} \qquad (2)$$

โดยที่ $\overline{x}$ = ค่าเฉลี่ยกลาง

$x_1, ..., x_n$ = ข้อมูลสมบูรณ์ตัวที่ 1, 2, ..., n

n = จำนวนข้อมูลสมบูรณ์ทั้งหมด

4. การประมาณความแม่นยำ (Evaluation Criterion) จากการใช้วิธีต่างๆ ที่สร้างขึ้น (Martin Sheppard. 1996 ; สมชาย ปราการเจริญ. 2551) ต้องมีความแม่นยำเข้ากันได้ (Model best fit) จะถูกนำไปทดสอบกับกลุ่มข้อมูลอีกชุดหนึ่งที่ทราบค่าจริง (Actual Data) ผลจากการพยากรณ์ข้อมูลชุดใหม่ (Predicted Data) จะถูกนำมาคำนวณหาค่าความคลาดเคลื่อนสัมพัทธ์ (Magnitude of Relative Error: MRE) โดยใช้สูตร

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

หากมีข้อมูลหลายชุด การคำนวณให้นำค่าความคลาดเคลื่อนสัมพัทธ์ (MRE) ของแต่ละชุดข้อมูลมารวมกันและหาค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (Mean-MRE: MMRE) หาก MMRE มีค่ามากๆ แสดงว่ามีความคลาดเคลื่อนสูง แต่ถ้า MMRE = 0 (ศูนย์) แสดงว่าค่าของการพยากรณ์เท่ากับค่าจริงทุกค่า นั่นคือ MMRE ยิ่งมีค่าน้อย แสดงว่ามีความแม่นยำมาก ในการประมาณค่าสูญหายของข้อมูล โดยใช้สูตร

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

5. การกำหนดขนาดตัวอย่างของทาโร ยามาเน่ (Taro Yamane) เป็นการกำหนดขนาดกลุ่มตัวอย่าง (Sample Size) เพื่อให้เกิดความเชื่อมั่นว่าทุกหน่วยประชากรได้มีโอกาสรับเลือกเป็นตัวแทนของประชากร งานวิจัยนิยมกำหนดขนาดกลุ่มตัวอย่างตามวิธีของ ทาโร ยามาเน่ (Taro Yamane. 1967) ประโยชน์ของการเลือกศึกษากลุ่มตัวอย่าง คือ 1) ประหยัดเวลา ค่าใช้จ่ายและแรงงาน 2) มีความสมบูรณ์และถูกต้องมากกว่าเพราะจำนวนน้อย 3) ควบคุมความคลาดเคลื่อนได้ง่าย โดยใช้สูตร

$$n = \frac{N}{1 + Ne^2} \qquad (5)$$

โดยที่ n = ขนาดของตัวอย่าง

N = ขนาดของประชากร

e = ระดับความคลาดเคลื่อน

6. งานวิจัยที่เกี่ยวข้อง

Hidetomo Ichihashi (2007) ได้ศึกษาวิธีการจัดกลุ่มข้อมูลจากค่าข้อมูลเริ่มต้นและค่าข้อมูลที่สูญหายด้วยวิธี Fuzzy C-Means (FCM) มีความแม่นยำมากที่สุด

ไก้รุ่ง เฮงพระพรหม และพยุง มีสัจ (2551) ได้วิจัยการเลือกคุณลักษณะด้วยวิธีสมาชิกที่ใกล้ที่สุดสำหรับการแทนที่ค่าข้อมูลที่ขาดหายด้วยวิธีการที่ใกล้ที่สุดแบบให้น้ำหนักของข้อมูลไมโครอาร์เรย์ พบว่าใช้วิธี KNNFSW Impute มีความแม่นยำมากที่สุด

จริยา แสงสุวรรณ, ประสิทธิ์ พยัคฆพงษ์ และอำไพ ทองธีรภาพ (2551) ได้ศึกษาการเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์ความถดถอยพหุคูณ วิธีนี้ให้ความแม่นยำสูงกว่าการแทนค่าด้วยวิธีอื่นๆ

ณรงค์ โพธิ และคณะ (2551) ได้วิจัยการเปรียบเทียบผลการแยกประเภทของก้อนเนื้อจากภาพแมมโมแกรม โดยการใช้เทคนิคการทำเหมืองข้อมูลในการทำนายผลโรคมะเร็งเต้านม พบว่าการใช้วิธีโครงข่ายประสาทเทียมให้ความแม่นยำสูงสุด

สมชาย ปราการเจริญ (2551) ได้ศึกษาวิธีการประมาณเวลาในการพัฒนาซอฟต์แวร์ประยุกต์เชิงโครงข่ายโดยวิธีแบบจำลองสมการโครงสร้าง มีความแม่นยำมากที่สุด

ณรงค์ โพธิ และคณะ (2552) ได้วิจัยการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ พบว่าการใช้วิธีการวิเคราะห์จำแนกประเภท มีความแม่นยำมากที่สุด

จากงานวิจัยที่ได้ศึกษาเป็นการใช้เทคนิคการประมาณค่าสูญหายและจำแนกข้อมูล ซึ่งมีความเหมาะสมแตกต่างกันไปขึ้นอยู่กับประเภทและลักษณะของข้อมูลที่นำมาใช้ สำหรับงานวิจัยนี้ผู้วิจัยได้ทดลองการประมาณค่าสูญหายกับฐานข้อมูลมาตรฐานสากลจากฐานข้อมูลของ UCI

วิธีดำเนินงานวิจัย

การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ จำนวน 3 วิธี ผู้วิจัยดำเนินการตามขั้นตอนและวิธีการดังต่อไปนี้

1. ศึกษาหลักการ ข้อมูล และเทคนิคต่างๆ จากหนังสือเอกสาร งานวิจัย ที่เกี่ยวข้อง

2. แบ่งฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านม (Mammographic Mass) ของ UCI Machine Learning Repository Data Set จำนวน 800 ชุด (Case) ออกเป็น 2 กลุ่มที่ระดับความเชื่อมั่น 95% (หรือระดับความคลาดเคลื่อน 0.05) ตามวิธีการกำหนดขนาดตัวอย่างของ Taro Yamane คือ กลุ่มข้อมูลเรียนรู้ (Training Data) 533 ชุด และกลุ่มข้อมูลทดสอบ (Testing Data) 267 ชุด

ตารางที่ 1 แอทริบิวต์ฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านมที่ใช้ในการวิจัย

ชื่อข้อมูล	คำอธิบาย	ข้อมูล
BI-RADS	ลักษณะภาพที่พบจากการตรวจ	**1-5**
Age	อายุของผู้ป่วยโรคมะเร็งเต้านม	**18-96**
Shape	รูปร่างของโรคมะเร็งเต้านม	**1-4**
Margin	ขอบเขตของก้อนเนื้อ	**1-5**
Density	ความหนาแน่นของก้อนเนื้อ	**1-4**
Severity	**ผลการตรวจของโรคมะเร็งเต้านม**	**0-1**

3. นำกลุ่มข้อมูลเรียนรู้มาหาสมการที่ดีที่สุดในแต่ละวิธี

4. นำกลุ่มข้อมูลทดสอบมาแทนค่าในสมการที่ได้ในข้อ 3 เพื่อประมาณหาค่าทดแทนค่าสูญหายของข้อมูลแอทริบิวต์ Shape

5. นำค่าประมาณการที่ได้ในข้อ 4 แทนค่าลงในชุดข้อมูลที่สูญหายแล้วนำข้อมูลทั้งหมดไปทำซ้ำตามวิธีในข้อ 3 และ 4 อีกครั้งเพื่อให้ได้ค่าทดแทนของค่าสูญหายที่ดีที่สุด ผลจากการทดลองโดยใช้โปรแกรมสำเร็จรูปทางสถิติที่ระดับความเชื่อมั่น 95% มีดังต่อไปนี้

5.1 การประมาณค่าสูญหายจากสมการวิเคราะห์จำแนกประเภท ได้สมการประมาณค่า จำนวน 4 สมการ ดังนี้

$$\hat{d}_1 = -72.697+(17.048*BI)+(0.120*A)-(1.508*M)+(23.751*D) \quad (6)$$

$$\hat{d}_2 = -67.179+(16.857*BI)+(0.125*A)-(0.741*M)+(21.613*D) \quad (7)$$

$$\hat{d}_3 = -77.571+(17.512*BI)+(0.140*A)+(0.802*M)+(22.713*D) \quad (8)$$

$$\hat{d}_4 = -87.661+(19.688*BI)+(0.118*A)+(2.027*M)+(21.781*D) \quad (9)$$

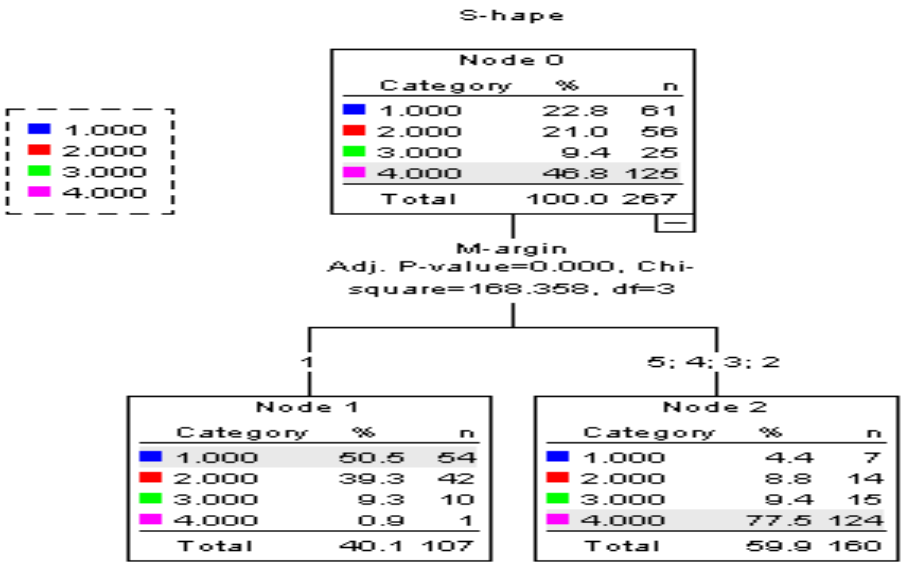
ตารางที่ 2 ค่าสัมประสิทธิ์ของสมการวิเคราะห์จำแนกประเภท

Classification Function Coefficients				
	S-hape			
	1	**2**	**3**	**4**
BI-RADS	17.048	16.857	17.512	19.688
A-ge	0.120	0.125	0.140	0.118
M-argin	-1.508	-0.741	0.802	2.027
D-ensity	23.751	21.613	22.713	21.781
(Constant)	-72.697	-67.179	-77.571	-87.661
Fisher's linear discriminant functions				

5.2 การประมาณค่าสูญหายจากต้นไม้การตัดสินใจ แสดงดังตารางที่ 3 และภาพที่ 2

ตารางที่ 3 ผลการประมวลผลด้วยต้นไม้การตัดสินใจ

Classification					
Observed	**Predicted**				
	1	**2**	**3**	**4**	**Percent Correct**
1	54	0	0	7	88.5%
2	42	0	0	14	0.0%
3	10	0	0	15	0.0%
4	1	0	0	124	99.2%
Overall Percentage	40.1%	0.0%	0.0%	59.9%	66.7%
Growing Method: CHAID **Dependent Variable: S-hape**					



ภาพที่ 2 ผลการประมวลผลด้วยต้นไม้การตัดสินใจ

5.3 การประมาณค่าสูญหายด้วยค่าเฉลี่ย

ตารางที่ 4 ผลการประมวลผลด้วยค่าเฉลี่ย

Descriptive Statistics					
	N	**Range**	**Mean**		**Std. Dev.**
	Statistic	**Statistic**	**Statistic**	**Std. Error**	**Statistic**
BI-RADS	800	3	4.34	0.021	0.582
A-ge	800	78	55.65	0.523	14.784
S-hape	**800**	**3.00**	**2.7503**	**0.03599**	**1.01787**
M-argin	800	4	2.79	0.055	1.567
D-ensity	800	3	2.91	0.013	0.356
Se-verity	800	1	0.48	0.018	0.500
Valid N (listwise)	800				

6. ตรวจสอบความแม่นยำของผลลัพธ์ที่ได้จากสมการของแต่ละวิธีด้วยการหาค่าความคลาดเคลื่อนสัมพัทธ์ (MRE) และค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) ดังตารางที่ 5

ตารางที่ 5 ค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ (MMRE) จากผลการวิจัย

	DA	**Decision Tree**	**Mean**
MMRE	26.56 %	33.30 %	63.26 %
Accuracy	73.44 %	66.70 %	36.74 %

7. ผลการวิจัยเมื่อเปรียบเทียบวิธีทางสถิติที่ใช้ในการทดแทนค่าข้อมูลสูญหายของแอทริบิวต์ Shape จากฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านม พบว่า วิธีการวิเคราะห์จำแนกประเภทให้ผลการประมาณค่าทดแทนค่าสูญหายที่มีค่าใกล้เคียงกับค่าข้อมูลจริงมากที่สุด

สรุปผลการวิจัยและข้อเสนอแนะ

ผลการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ สรุปได้ดังนี้

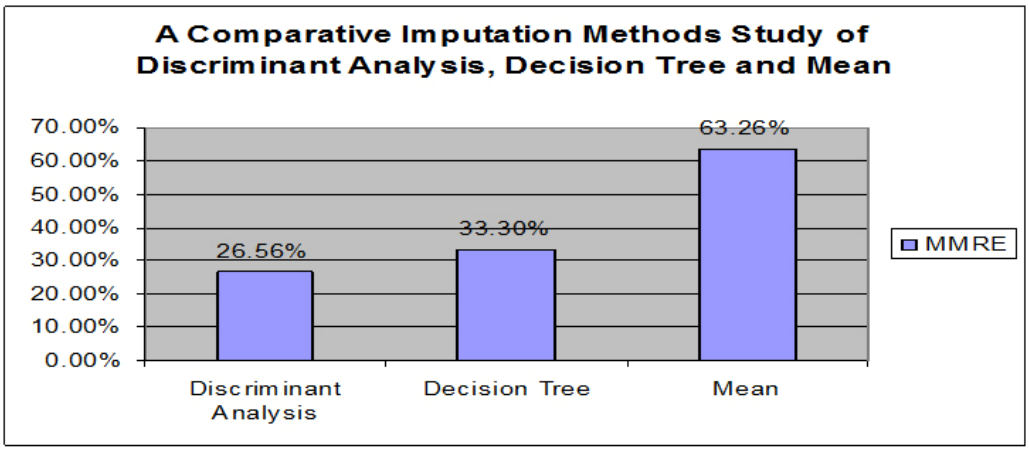
1. สรุปผลการศึกษาเปรียบเทียบวิธีการประมาณค่าทดแทนค่าสูญหายของข้อมูลจากสมการที่ดีที่สุดโดยเรียงตามลำดับจากมากไปน้อย ดังนี้

ลำดับที่ 1 การประมาณค่าสูญหายจากสมการวิเคราะห์จำแนกประเภท เป็นวิธีที่ดีที่สุด เนื่องจากมีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ 26.56% ซึ่งเป็นค่าน้อยที่สุด หรือแสดงว่าความแม่นยำในการประมาณค่าทดแทนค่าสูญหายมีความถูกต้อง 73.44%

ลำดับที่ 2 การประมาณค่าสูญหายจากต้นไม้การตัดสินใจ เป็นวิธีที่ดีลำดับที่ 2 เนื่องจากมีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ 33.30% หรือแสดงว่าความแม่นยำในการประมาณค่าทดแทนค่าสูญหายมีความถูกต้อง 66.70%

ลำดับที่ 3 การประมาณค่าสูญหายด้วยค่าเฉลี่ย เป็นวิธีที่ดีเป็นลำดับสุดท้าย เนื่องจากมีค่าเฉลี่ยความคลาดเคลื่อนสัมพัทธ์ 63.26% หรือแสดงว่าความแม่นยำในการประมาณค่าทดแทนค่าสูญหายมีความถูกต้อง 36.74%

สรุปผลการวิจัยเป็นกราฟเปรียบเทียบค่า MMRE ของแต่ละวิธีที่ได้ทำการทดสอบได้ดังภาพที่ 3



ภาพที่ 3: กราฟเปรียบเทียบค่า MMRE ของแต่ละวิธี

2. ผลลัพธ์ที่ได้จากการศึกษาเปรียบเทียบ

2.1 ได้วิธีการประมาณค่าทดแทนค่าสูญหายของข้อมูลที่เหมาะสมโดยใช้วิธีทางสถิติ (Discriminant Analysis, Decision Tree, Mean)

2.2 ได้ผลลัพธ์จากการประมาณค่าทดแทนค่าสูญหายของข้อมูลที่มีความใกล้เคียงกับค่าจริงและมีความน่าเชื่อถือของข้อมูลมากยิ่งขึ้น

ข้อเสนอแนะในการวิจัย

จากการศึกษาเปรียบเทียบวิธีการประมาณค่าทดแทนค่าสูญหายของแอทริบิวต์ Shape จากฐานข้อมูลภาพถ่ายก้อนเนื้อเต้านมที่กำหนดให้เป็นข้อมูลสูญหายด้วยวิธีทางสถิติ เพื่อใช้ทดสอบความแม่นยำในการประมาณค่าทดแทนค่าสูญหาย พบว่า วิธีการวิเคราะห์จำแนกประเภทมีความเหมาะสมกับข้อมูลที่มี

ความสัมพันธ์กันและมีการกระจายออกเป็นกลุ่มที่แตกต่างกันอย่างชัดเจน การประมาณค่าทดแทนค่าสูญหายจึงจะมีความแม่นยำและใกล้เคียงกับค่าของข้อมูลจริงมากที่สุด หากนำวิธีการนี้ไปใช้ทดสอบกับข้อมูลที่ไม่มีความสัมพันธ์กันและมีความเป็นอิสระต่อกันมากๆ จะทำให้การแทนค่าในสมการด้วยวิธีการข้างต้นได้ค่าที่มีความคลาดเคลื่อนมากจากค่าจริงของข้อมูล ดังนั้น ข้อมูลที่มีความเป็นอิสระต่อกันและมีการกระจายออกเป็นกลุ่มที่แตกต่างกันมากๆ ผู้วิจัยต้องแบ่งช่วงข้อมูลให้เหมาะสมและจัดการกับข้อมูลที่ไม่เกี่ยวข้องออกไปก่อนที่จะนำมาใช้งาน

เอกสารอ้างอิง

กัลยา วานิชย์บัญชา. (2542). **การวิเคราะห์ข้อมูลด้วย SPSS for Windows**. กรุงเทพฯ: ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย.

_____. (2550). **การวิเคราะห์ข้อมูลหลายตัวแปร**. พิมพ์ครั้งที่ 2. กรุงเทพฯ: ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย,.

ไก้รุ่ง เฮงพระพรหม และพยุง มีสัจ. (2551). การเลือกคุณลักษณะด้วยวิธีสมาชิกที่ใกล้ที่สุดสำหรับการแทนที่ค่าข้อมูลที่ขาดหายด้วยวิธีการที่ใกล้ที่สุดแบบให้น้ำหนักของข้อมูลไมโครอาร์เรย์. **การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 4**. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระ จอมเกล้าพระนครเหนือ.

จริยา แสงสุวรรณ, ประสิทธิ์ พยัคฆพงษ์ และอำไพ ทองธีรภาพ. (2551). การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยพหุคูณ. **วารสารวิทยาศาสตร์ประยุกต์**. 7(1), 1-7.

ชูศรี วงศ์รัตนะ. (2541). **เทคนิคการใช้สถิติเพื่อการวิจัย**. พิมพ์ครั้งที่ 7. กรุงเทพฯ: เทพเนรมิตการพิมพ์.

ณรงค์ โพธิและคณะ. (2551). การเปรียบเทียบผลการแยกประเภทของก้อนเนื้อจากภาพแมมโมแกรมโดยการใช้เทคนิคการทำเหมืองข้อมูลในการทำนายผลของโรคมะเร็งเต้านม. **การประชุมทางวิชาการ NCIT2008 ครั้งที่ 2**. กรุงเทพฯ: มหาวิทยาลัยรังสิต.

_____. (2552). การศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายโดยวิธีการทางสถิติ. **การประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 5**. กรุงเทพ : มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

พยูน พาณิชย์กุล. (2548). **การพัฒนาระบบดาต้าไมน์นิ่งโดยใช้ Decision Tree**. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง. (อัดสำเนา).

มนต์ชัย เทียนทอง. (2550). **สถิติและวิธีการวิจัยทางเทคโนโลยีสารสนเทศ**. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

สกลนันท์ หุ่นเจริญ. (2548). **การค้นหาเมนูอาหารไทยเพื่อสุขภาพ โดยวิธีกฎความสัมพันธ์และแผนผังต้นไม้เพื่อการตัดสินใจ**. วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ. (อัดสำเนา).

สมชาย ปราการเจริญ. (2551). การประมาณการเวลาในการพัฒนาซอฟต์แวร์ประยุกต์เชิงโครงข่ายโดยวิธี แบบจำลองสมการโครงสร้าง. **วารสารเทคโนโลยีสารสนเทศ**. 4(7).

Darlington, R.B. (1990). **Regression and Linear Models**. New York: McGraw-Hill.

Garson G. David. (2005). **Data Imputation for Missing Values**.USA: North Carolina University.

Geoffrey J. McLachlan. (2005). **Discriminant Analysis and Statistical Pattern Recognition**. USA.

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

Martin Sheppard. (1996). **Effort Estimating Using Analogy**. USA : Bournemouth University.

Yamane, T. (1967). **Elementary Sampling Theory**. New Jersey: Prentice-Hall.