การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลการเลือกอาชีพโดยใช้เทคนิคเหมืองข้อมูล
Comparative Efficiency of Classification Choosing a Career with Data Mining Techniques.

ชัชชฎา วันดี (Chatchada Wandee)[1] จิรัฏฐา ภูบุญอบ (Jiratta Phuboon-ob)[2]
และฉัตรเกล้า เจริญผล (Chatklaw Jareanpon)[3]
สำนักวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
[1]Emails: chatchada.wd@gmail.com, [2]jiratta.p@msu.ac.th, [3]chatklaw.j@msu.ac.th

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา โดยในงานวิจัยนี้ได้ใช้ชุดข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างปี พ.ศ. 2550-2554 จำนวน 12 คุณลักษณะ และ 2,515 ระเบียน ซึ่งได้นำเทคนิคแบบจำลองต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม และการเรียนรู้แบบเบย์มาทำการเปรียบเทียบประสิทธิภาพ จากนั้นจึงทำการเลือกตัวอย่างโดยใช้การแบ่งตามสัดส่วนข้อมูลเรียนรู้ตามสัดส่วน 60,70,80 และ 90 ตามลำดับ ผลจากการศึกษาพบว่าประสิทธิภาพในการจำแนกข้อมูลแบบต้นไม้ตัดสินใจ มีประสิทธิภาพในการจำแนกสูงสุดด้วยค่าเฉลี่ย 89.45% ซึ่งแบบจำลองโครงข่ายประสาทเทียม การเรียนรู้แบบเบย์ ได้ให้ค่าความความถูกต้องเท่ากับ 86.85%, 80.52%, ตามลำดับ

คำสำคัญ: ข้อมูลภาวะการมีงานทำของบัณฑิต ข้อมูลระเบียนประวัติของนิสิต ประสิทธิภาพในการจำแนกข้อมูลการเลือกอาชีพ

Abstract

This research aims to study the comparative of classification choosing the career in undergraduate students after graduated. In this research used the data set of the status having the job of graduates, which used the data set of students for undergraduate after graduation from Faculty of Informatics in Mahasarakham university from 2007-2011 years. The total of data set have 12 attributes and 2,515 record. This research using techniques Decision tree, Artificial neural network and Naive bayes. The results from the model was according to the following ratio 60,70,80 and 90, which showed that the highest accuracy is Decision tree with 89.45% but Artificial neural network and Naive bayes is 86.85% and 80.52% respectively.

Keyword: status having the job of graduates, students data of undergraduate after graduation, Choosing a career

1. บทนำ

การที่จะพัฒนาประเทศให้เจริญก้าวหน้าจะต้องมีการผลิตประชากรที่มีความรู้ความสามารถให้ครบทุกด้าน ตรงกับตลาดแรงงานและสภาพเศรษฐกิจตามความต้องการของประเทศที่มีการเปลี่ยนแปลงอยู่ตลอดเวลา ซึ่งตลาดแรงงานจะขึ้นอยู่กับสภาพการเจริญเติบโตของเศรษฐกิจ ณ ช่วงเวลานั้น ทำให้การเลือกอาชีพนับว่าเป็นเรื่องสำคัญอย่างยิ่งในชีวิตมนุษย์ ในการเลือกอาชีพจะต้องมีการเริ่มต้นด้วยการวางแผน ตั้งแต่วัยเรียน โดยเป็นการวางแผนระยะยาวที่ต้องใช้เวลานาน ซึ่งคนเรามีความถนัด ความสามารถ และความสนใจในงานอาชีพแตกต่างกัน ดังนั้นหลายคนอาจต้องตัดสินใจเลือกอาชีพที่อาจตรงหรือไม่ตรงกับสาขาที่เรียน ด้วยเหตุผลที่แตกต่างกันออกไป ดังนั้นทุกมหาวิทยาลัยจึงได้จัดทำแบบสำรวจภาวะการมีงานทำของบัณฑิต เพื่อใช้ในการสำรวจปัญหา/อุปสรรคในการหางานทำ และเพื่อให้ทราบว่านิสิตที่สำเร็จการศึกษาออกไปเลือกที่จะประกอบอาชีพอะไร โดยบัณฑิตจะต้องบันทึกข้อมูลให้กับมหาวิทยาลัยหลังจากสำเร็จการศึกษา ผู้วิจัยจึงได้นำข้อมูลภาวะการมีงานทำของบัณฑิต [1] และข้อมูล

ระเบียนประวัติของนิสิต [2] ที่สำเร็จการศึกษาตั้งแต่ปี พ.ศ. 2550-2554 ของคณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม จำนวน 2,515 ระเบียน 12 คุณลักษณะ มาใช้ในงานวิจัยนี้ และในการวิจัยจะต้องเลือกใช้เทคนิควิธีในการจำแนกข้อมูลที่เหมาะสมกับข้อมูล เพื่อเป็นการเพิ่มประสิทธิภาพในการจำแนกให้มีความถูกต้องมากที่สุด

จากปัญหาดังกล่าวข้างต้น พบว่าเทคนิคการจำแนกข้อมูลที่มีประสิทธิภาพมีหลากหลายวิธี ดังนั้นในงานวิจัยนี้จึงได้เสนอเทคนิคในการจำแนกข้อมูล 3 เทคนิควิธี ได้แก่ แบบจำลองต้นไม้ตัดสินใจ โครงข่ายประสาทเทียม การเรียนรู้แบบเบย์ เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูล

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ต้นไม้ตัดสินใจ (Decision tree: DT)

ต้นไม้ตัดสินใจ [3, 4] คือ แบบจำลองที่มีลักษณะคล้ายกับต้นไม้ จะมีการสร้างกฎต่างๆ ขึ้นเพื่อใช้ในการตัดสินใจ ซึ่งแต่ละโหนด (Node) จะแสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล รูปแบบของต้นไม้จะประกอบด้วยโหนดแรกสุดที่เรียกว่า Root node จาก Root node จะแยกออกเป็นโหนดลูก และที่โหนดลูกก็จะมีลูกของตัวเองซึ่งโหนดในระดับสุดท้ายจะเรียกว่า Leaf node ซึ่งแสดงกลุ่มหรือคลาส (Class) ที่กำหนดไว้ สำหรับในงานวิจัยนี้ใช้อัลกอริทึม C4.5 [5]

ถ้าให้ชุดข้อมูล C ประกอบด้วยค่าที่เป็นไปได้ คือ $\{c_1,c_2...c_n\}$และให้ค่าความน่าจะเป็นที่เกิดขึ้นเป็นค่า c_i ซึ่งมีค่าเท่ากับ $P(c_i)$ จะได้ค่า Information Gain *ของ* C เขียนแทนด้วย $I(C)$ ดังสมการที่ (1)

$$I(C) = \sum_{i=1}^{n} -P(c_i)\ log_2\ P(c_i) \qquad (1)$$

ถ้าให้ข้อมูลสอน คือ T โดยคุณลักษณะที่เป็นโหนด เช่นค่า x และมีค่าทั้งหมดที่เป็นไปได้เท่ากับ n ค่า ซึ่งโหนดปัจจุบันจะแบ่งตัวอย่างค่า T ออกตามกิ่งเป็น $\{t_1,\ t_2...\ t_n\}$ ตามค่าที่เป็นไปได้ของค่า x ดังนั้น จึงสามารถคำนวณค่า Information Gain หลังจากแบ่งตามคุณลักษณะ ดังสมการที่ (2)

$$I_x(T) = \sum_{i=1}^{n} \frac{t_i}{T} I(t_1) \qquad (2)$$

ค่า Gain ของคุณลักษณะ x ดังสมการที่ (3)

$$Gain(x)=I(T) - I_x(T) \qquad (3)$$

จากนั้นคำนวณค่า Information Gain ของ Split Information ตามคุณลักษณะแต่ละตัว ถ้าให้ T คือชุดของตัวอย่างเมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1,\ t_2,...,\ t_n\}$ จำนวน n ชุดตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่า Split Information ได้ ดังสมการที่ (4)

$$Split\ Information= \sum_{i=1}^{n} \frac{t_i}{T} \log_2 \frac{t_i}{T} \qquad (4)$$

ค่า Gain Ratio สามารถคำนวณได้จาก *Gain Ratio = Gain - Split Information* ค่า Gain Ratio สูงสุดจะถูกเลือกเป็นคุณลักษณะเริ่มต้น และเลือกคุณลักษณะถัดไป ตามค่า Gain Ratio น้อยลงตามลำดับ

2.2 โครงข่ายประสาทเทียม (Artificial neural network: ANN)

โครงข่ายประสาทเทียม [6, 7] มีพื้นฐานมาจากการจำลองการทำงานของสมองมนุษย์ ด้วยโปรแกรมคอมพิวเตอร์ ซึ่ง Back Propagation Algorithm เป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของโครงข่ายประสามเทียมวิธีหนึ่งที่นิยมใช้ใน Multilayer Perceptron เพื่อปรับค่าน้ำหนักสำหรับข่ายงานไปข้างหน้าหลายชั้น โดยเคลื่อนจากข้อมูลชั้นนำเข้า ชั้นซ่อน ไปจนถึงชั้นแสดงผล ดังภาพที่ 1 ซึ่งจะคำนวณค่าผิดพลาดระหว่างเอาต์พุตของข่ายงานและค่าจริง เพื่อใช้ในการปรับค่าเวกเตอร์น้ำหนัก และจะทำซ้ำไปซ้ำมาจนได้ค่าเวกเตอร์น้ำหนักที่ให้ค่าผิดพลาดน้อยที่สุด ดังสมการที่ (5)

$$n = \sum_{i=1}^{z} x_i\ w_i + b \qquad (5)$$

โดยที่ n คือ ผลรวมที่ได้จากฟังก์ชันผลรวม

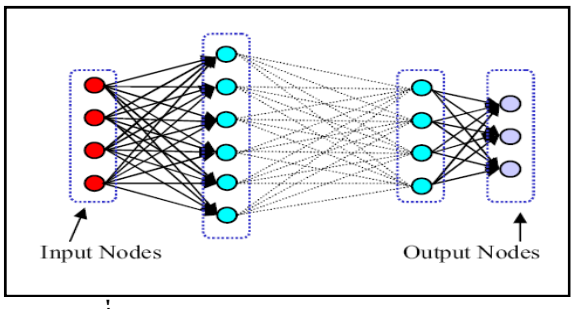
x_i คือ ค่าข้อมูลเข้าตัวที่ i

w_i คือ ค่าน้ำหนักของนิวรอนตัวที่ i

z คือ จำนวนนิวรอนชั้นข้อมูลเข้า

b คือ ค่าความโน้มเอียง

i คือ มีค่าตั้งแต่ 1 ถึง z



ภาพที่ 1: โครงข่ายประสาทเทียมแบบ Back Propagation

2.3 การเรียนรู้แบบเบย์ (Naive bayes: NB)

การเรียนรู้แบบเบย์ [8, 9] เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพอีกวิธีหนึ่ง ซึ่งใช้งานได้ดีและเหมาะสมกับกรณีของเซตตัวอย่างที่มีจำนวนมากและมี Attribute ของตัวอย่างไม่ขึ้นต่อกัน เป็นการเรียนรู้ที่ใช้หลักการของความน่าจะเป็นซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างอัลกอริทึมที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำอัลกอริทึมมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ข้อดีของวิธีการเรียนรู้แบบนี้คือเราสามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior knowledge) เข้ามาช่วยในการเรียนรู้ได้ ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดีไม่ด้อยกว่าวิธีการเรียนรู้ประเภทอื่น ดังสมการที่ (6)

$$P(H|E) = \frac{[P(E|H) \times P(H)]}{P(E)} \quad (6)$$

โดยที่ $P(H|P)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ H เมื่อเกิดเหตุการณ์ E

$P(H)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ H

เมื่อศึกษาข้อมูลถ้าแอททริบิวต์ที่ต้องการนำมาวิเคราะห์มีหลาย แอททริบิวต์ สามารถแสดงการจำแนกประเภทข้อมูล ที่มีแอททริบิวต์มากกว่าหนึ่ง แอททริบิวต์ดังสมการที่ (7)

$$P(H|E_1, E_2..E_N) = \frac{[P(E_1|H) \times P(E_2|H) \times ..\times P(E_n|H) \times P(H)]}{P(E_1,E_2..E_N)} \quad (7)$$

จากสมการที่ (7) ค่า $E_1,E_2,...E_N$ คือ ข้อมูลเหตุการณ์ใดๆที่มีหลายเหตุการณ์ ซึ่งในการคำนวณจะทำให้ มีค่าที่สูงจนเกินจริง ดังนั้นจึงสามารถตัดพจน์ $P(E_1,E_2,....,E_n)$ ที่เป็นเศษส่วนออกได้เพราะเป็นค่าคงที่

2.4 งานวิจัยที่เกี่ยวข้อง

พัสการ สิงห์โต อัครวุฒิ ปรมะปุญญา และ ปฐมาภรณ์ เถาว์พัน [10] ใช้เทคนิค Naive bayes, KNN, Rule base และ Decision tree กับชุดข้อมูล 4 ชุด จากฐานข้อมูล UCI ได้ทำการสร้างโมเดลการจำแนกข้อมูลด้วยอัลกอริทึมต่างๆ โดยการปรับค่าพารามิเตอร์ที่เหมาะสม และเลือกแอทริบิวต์ที่มีผลต่อการเปลี่ยนแปลงคลาสมากที่สุด โดยวัดประสิทธิภาพการจำแนกข้อมูลด้วยโมเดลที่แตกต่างกัน จากค่าความถูกต้อง พบว่า Decision tree, Rule base, KNN และ Naive bayes ให้ประสิทธิภาพดีที่สุดในการจำแนกข้อมูลตามลำดับ

กฤตยา ทองผาสุก [8] ได้เสนอการเปรียบเทียบ ด้วยเทคนิคต้นไม้ตัดสินใจ กฎนาอีฟเบย์ และ เคเนียรเรสเนเบอร์ เพื่อการสร้างอัลกอริทึมที่แตกต่างกันจำนวน 10 อัลกอริทึม โดยเทคนิคเคเนียรเรสเนเบอร์ เป็นเทคนิคที่ดีที่สุดโดยให้ค่าความถูกต้อง 76.9231%

Aitkenhead M.J. [11] ได้นำเสนอการพัฒนาการจำแนกร่วมกับต้นไม้ตัดสินใจ จากปัญหาการจำแนกและจัดหมวดหมู่นั้นมีลักษณะที่แตกต่างกันหรือคุณสมบัติของระบบต่างกันและข้อมูลยังมีการสูญหายหรือเกิดการรบกวนทำให้ข้อมูลไม่มีคุณภาพ จึงวิเคราะห์ข้อมูลด้วยเทคนิคต้นไม้ตัดสินใจ โดยเลือกใช้อัลกอริทึม C4.5 ซึ่งผลการวิจัยพบว่า อัลกอริทึมต้นไม้ตัดสินใจแสดงให้เห็นถึงวิวัฒนาการหรือโครงสร้างของข้อมูลได้ง่าย และสามารถจัดการกับช่วงของค่าและชนิดของมูลได้ และอัลกอริทึมนี้ยังมีความเข้าใจกว่าวิธีอื่นๆ

Lawrence O. Hall และคณะ [12] ได้นำเสนองานวิจัยที่ทดสอบเทคนิค Pruning สำหรับ อัลกอริทึม C4.5 คือ Error Based Pruning (EBP) ผลที่ได้คือต้นไม้มีขนาดใหญ่แต่ความถูกต้องไม่เพิ่มขึ้น ทั้งนี้เพราะงานวิจัยส่วนใหญ่ใช้ค่าพารามิเตอร์ Certainty Factor เท่ากับ 0.25 โดยงานวิจัยนี้ได้ทำการทดลองกับข้อมูลหลายชุด เพื่อทดลองว่ามีผลต่อขนาดของต้นไม้และความถูกต้องหรือไม่ ผลที่ได้พบว่ามีค่าความถูกต้องใกล้เคียงกัน

S.chen และคณะ [13] ได้นำเสนอการจำแนกข้อมูลด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนที่มีการกำหนดค่าคอร์เนลฟังก์ชันรวมทั้งค่าพารามิเตอร์ที่หลากหลาย ทำให้ตัวแทนที่สร้างขึ้น

สามารถนำไปแก้ปัญหาที่หลากหลายเช่นเดียวกัน รวมทั้งยังเป็นการเพิ่มประสิทธิภาพโดยรวมของโมเดลอีกด้วย

ภัทร์พงศ์ พงศ์ภัทรกานต์ [14] ใช้ชุดข้อมูลนักศึกษาที่เข้าศึกษาระหว่างปี พ.ศ. 2546-2549 ของมหาวิทยาลัยราชภัฏเลย ซึ่งคอมมิตตีแมชชีนเป็นการทำงานระหว่าง SVM ร่วมกับ C5.0 โดยได้ทำการทดลองวัดประสิทธิภาพความถูกต้องเปรียบเทียบกับนิวรอลเน็ตเวิร์ก ซึ่งแบบจำลองแบบคอมมิตตีแมชชีน มีประสิทธิภาพในการจำแนกข้อมูลสูงที่สุด มีค่าเท่ากับ 75.32 %

3. วิธีการดำเนินการวิจัย

3.1. ข้อมูลที่ใช้ในการวิจัย

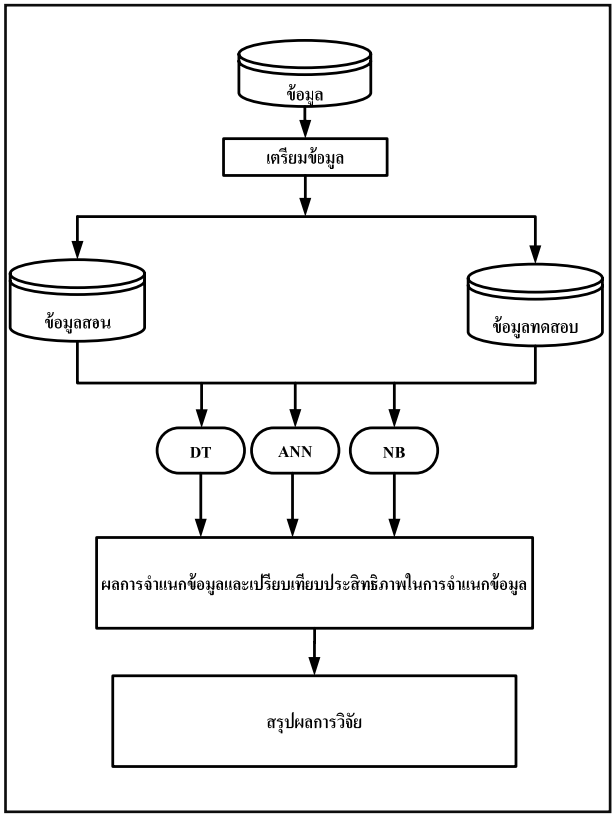
งานวิจัยนี้นำข้อมูลภาวะการมีงานทำของบัณฑิต และข้อมูลระเบียนประวัติของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม โดยใช้ชุดข้อมูลระหว่าง ปี พ.ศ. 2550 – 2554 มีจำนวน 12 คุณลักษณะ และ 2,515 ระเบียน ดังตารางที่ 1

ตารางที่ 1: แอตทริบิวต์ ที่นำมาใช้ในการวิจัย

ที่	รายละเอียด	ที่	รายละเอียด
1.	สาขาที่เรียน	7.	อาชีพบิดา
2.	เพศ	8.	รายได้บิดา/ปี
3.	โรงเรียนเดิม	9.	อาชีพมารดา
4.	เกรดโรงเรียนเดิม	10.	รายได้มารดา/ปี
5.	เกรดเฉลี่ยรวม	11.	ตำแน่งงาน
6	เกรดเฉพาะวิชาสาขา	12.	ตรงหรือไม่ตรงสาขา

3.2. ขั้นตอนในการดำเนินงานวิจัย

การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล โดยเริ่มจากการเตรียมข้อมูล เนื่องจากข้อมูลที่ได้มาอาจมีความไม่สมบูรณ์ เช่น ข้อมูลแปลกปลอม หรือ ข้อมูลขาดหาย จึงต้องทำการกลั่นกรองข้อมูล (Data Cleaning) และการแปลงข้อมูล (Data Trasformation) [15] จากนั้นจะแบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดข้อมูลสำหรับสอน และชุดข้อมูลสำหรับ ดังภาพที่ 2



ภาพที่ 2: ขั้นตอนเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล

จากนั้นทำการเลือกตัวอย่างโดยใช้การแบ่งตามสัดส่วนข้อมูลเรียนรู้ตามสัดส่วน 60,70,80 และ 90 ตามลำดับ โดยใช้เทคนิคการสุ่มอย่างง่าย (Simple random sampling) สมาชิกทั้งหมดของประชากรเป็นอิสระซึ่งกันและกัน แล้วสุ่มหน่วยของการสุ่ม (Sampling unit) จนกว่าจะได้จำนวนตามที่ต้องการ โดยแต่ละครั้งที่สุ่ม สมาชิกแต่ละหน่วยของประชากรมีโอกาสถูกเลือกเท่ากัน ซึ่งก่อนที่จะทำการสุ่มนั้น จะต้องนิยามประชากรให้ชัดเจน ทำรายการสมาชิกทั้งหมดของประชากรสุ่มตัวอย่างโดยใช้วิธีที่ทำให้โอกาสของสมาชิกแต่ละหน่วยในการถูกเลือกมีค่าเท่ากัน[16] โดยจะใช้การสุ่มด้วยวิธีการใช้ตารางเลขสุ่ม (table of random number) ซึ่งตัวเลขในตารางได้มาจากการอาศัยคอมพิวเตอร์กำหนดค่า

เมื่อได้ชุดข้อมูลสำหรับการเรียนรู้ โดยจะนำชุดข้อมูลมาทำการจำแนกข้อมูลด้วยอัลกอริทึม DT, ANN, และ NB จากนั้นจะนำชุดข้อมูลสำหรับการทดสอบ มาทำการทดสอบประสิทธิภาพในการจำแนกข้อมูล เพื่อวิเคราะห์หาค่าความถูกต้องของการจำแนกข้อมูล โดยวัดค่าความถูกต้องแม่นยำ

จากการค่าสัมบูรณ์ของค่าคลาดเคลื่อนเฉลี่ย (Mean Absolute Error: MAE) [17] ของชุดข้อมูลทดสอบ ดังสมการที่ (8)

$$MAE = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad (8)$$

โดยที่ e_i คือ ผลต่างระหว่างค่าข้อมูลจริงและค่าพยากรณ์

n คือ ข้อมูลในการพยากรณ์

4.ผลการดำเนินงานวิจัย

ผลการวิเคราะห์ประสิทธิภาพวิธีการจำแนกข้อมูลจากฐานข้อมูล ซึ่งในการทดสอบแต่ละครั้งได้ผลลัพธ์ความถูกต้องไม่เท่ากัน เพราะในแต่ละครั้งของการเลือกตัวอย่างข้อมูลจะได้ข้อมูลไม่เหมือนกัน ดังนั้นความถูกต้องจะขึ้นอยู่กับข้อมูลที่เลือกมาได้ด้วย และจะพบว่าการแบ่งสัดส่วนข้อมูลจะมีผลกับความถูกต้องด้วย คือ ถ้าสัดส่วนการแบ่งข้อมูลมากจะมีถูกต้องมากขึ้นผลการทดสอบแสดงดังตารางที่ 2

ตารางที่ 2: ผลการเปรียบเทียบประสิทธิภาพการจำแนกข้อมูล

อัลกอริทึม / ช่วงข้อมูล	ค่าคลาดเคลื่อนเฉลี่ย		
	DT	ANN	NB
60:40	88.53	85.96	80.035
70:30	89.97	87.21	80.81
80:20	89.21	86.95	79.91
90:10	90.48	87.30	81.33
ค่าเฉลี่ย	89.45	86.85	80.52

จากตารางที่ 2 จะเห็นว่าอัลกอริทึม DT มีค่าความถูกต้องสูงสุดให้ค่าความถูกต้องเฉลี่ยเท่ากับ 89.45% อัลกอริทึม ANN มีค่าความเฉลี่ยเท่ากับ 86.85% และอัลกอริทึม NB มีค่าความเฉลี่ยเท่ากับ 80.52%

5.สรุปผลการวิจัยและข้อเสนอแนะ

การวิจัยในครั้งนี้สามารถสรุปได้ว่า ในการทดสอบแต่ละครั้งได้ผลลัพธ์ความถูกต้องไม่เท่ากัน เพราะในแต่ละครั้งของการเลือกตัวอย่างข้อมูลจะได้ข้อมูลไม่เหมือนกัน ดังนั้นค่าความถูกต้องจะขึ้นอยู่กับข้อมูลที่เลือกมาได้ด้วย และจะพบว่าการแบ่งสัดส่วนข้อมูลจะมีผลกับความถูกต้องด้วย คือ ถ้าสัดส่วนการแบ่งข้อมูลมากจะมีค่าความถูกต้องมาก ในการจำแนกข้อมูลได้ทำการเลือกตัวอย่างโดยใช้การแบ่งตามสัดส่วนข้อมูลเรียนรู้ตามสัดส่วน 60,70,80 และ 90 ตามลำดับ พบว่าค่าเฉลี่ยความถูกต้องของอัลกอริทึม DT ให้ค่าความถูกต้องสูงสุดโดยมีค่าเฉลี่ย 89.45% รองลงมาคือ ANN และNB มีค่าความถูกต้องสูงสุดโดยมีค่าเฉลี่ย 86.85% และ 80.52% ดังนั้นจึงสรุปได้ว่าอัลกอริทึม DT ให้ประสิทธิภาพในการจำแนกข้อมูลมากที่สุดและเหมาะสมกับข้อมูลที่ใช้งานวิจัยนี้มากที่สุด

ในอนาคตผู้วิจัยมีจะนำประโยชน์จากการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลในงานวิจัยนี้ ไปพัฒนาเพื่อทำการวิเคราะห์ปัจจัยว่าเหตุใดนิสิตจึงได้ตัดสินใจเลือกอาชีพที่ตรงหรือไม่ตรงกับสาขาที่ตนเรียนมา เพื่อให้สามารถนำไปประยุกต์ใช้กับคณะหรือหน่วยงานที่เกี่ยวข้อง ไปวางแผนพัฒนาโครงสร้างหลักสูตรหรือวางแผนการศึกษาให้กับนิสิตได้

เอกสารอ้างอิง

[1] กองแผนงาน. มหาวิทยาลัยมหาสารคาม. . ระบบภาวะการมีงานทำของบัณฑิต. [ออนไลน์]. 2554 [สืบค้นเมื่อ 28 พฤศจิกายน 2554]; http://www.survey.msu.ac.th/.

[2] กองทะเบียนและประมวลผล. มหาวิทยาลัยมหาสารคาม. งานทะเบียนและประมวลผล. [ออนไลน์]. 2554 [สืบค้นเมื่อ 28 พฤศจิกายน 2554]; http:// www.regpr.msu.ac.th.

[3] มลธิดา ฤทธิ์สมบูรณ์, สุชา สมานชาติ. การพัฒนาระบบสนับสนุนการพิจารณาอนุมัติให้สินเชื่อเพื่อการเช่าซื้อโดยใช้เทคนิคต้นไม้ตัดสินใจ. วารสารเทคโนโลยีสารสนเทศ 2551; 4[7]: 9-14.

[4] กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ, ธนาวินท์ รักธรรมานนท์. การใช้เทคนิคการทำเหมืองข้อมูล (Data Mining)เพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์. วารสารเทคโนโลยีสารสนเทศ 2544; 3[11]: 134-142.

[5] Quinlan John Ross. C4.5: programs for machine learning. 1st ed. London: England; 1988.

[6] เรวดี ศกดิ์ดุลยธรรม. การใช้เทคนิคดาต้าไมน์นิ่ง ในการสร้างฐานความรู้ เพื่อการทำนายสัมฤทธิ์ ผลทางการเรียนของนักศึกษา. นนทบุรี: วิทยาลัยราชพฤกษ์; 2552.

[7] นรินทร์ พนาวาส, นิเวศ จิระวิชิตชัย. การจำแนกมะเร็งเม็ดเลือดขาวโดยใช้โครงข่ายประสาทเทียมเพอร์เซ็ปตรอนหลายชั้น. งานประชุมสัมมนาวิชาการมหาวิทยาลัยเทคโนโลยีราชมงคลตะวันออก ครั้งที่ 4; 27 พฤษภาคม 2554; ชลบุรี.

[8] กฤตยา ทองผาสุข. การเปรียบเทียบเทคนิคต้นไม้ตัดสินใจกฎนาอีฟเบย์ และเคเนียรเรสเนเบอร์ เพื่อการจำแนกข้อมูล. National Conference on Computer Informastion Technolgies 2011; 26-28 มกราคม 2554; นครปฐม. pp. 30-35.

[9] ธัญญาภรณ์ บุญยัง, เอกรัฐ หล่อพิเชียร. การเปรียบเทียบข้อมูลของแบบจำสองเคเนียรเรสเนเบอร์ นาอีฟเบย์ ต้นไม้ตัดสินใจ และกฎพื้นฐาน. National Conference on Computer Informastion Technolgies 2011; 26-28 มกราคม 2554; นครปฐม. pp. 19-23.

[10] พัสกร สิงห์โต, อัครวุฒิ ปรมะปุญญา, เถาว์พัน ป. การเปรียบเทียบประสิทธิภาพการจำแนกข้อมูลด้วยเทคนิคการทำเหมืองข้อมูล. National Conference on Computer Informastion Technolgies 2011; 26-28 มกราคม 2554; นครปฐม. pp. 42-48.

[11] Aitkenhead. MJ. A co-evolving decision tree classification method. ScienceDirect 2008 2008; pp. 18-25.

[12] Lawrence O. Hall, Richard Collins, Bowyer KW. Error-Based Pruning of Decision Trees Grown on Very Large Data Sets Can Work. International Conference on Tools for Artificial Intelligence 2002; pp. 233 - 238.

[13] S.Chen, W.Wang, H.V.Zuylen. Construct support vector machine ensemble to detect traffic incidnt. Expert Systems with Applications 2009; 36pp. 10976-10986.

[14] ภัทร์พงศ์ พงศ์ภัทรกานต์. การวิเคราะห์ปัจจัยทีDส่งผลต่อการพ้นสภาพของนักศึกษา ระดับปริญญาตรีโดยใช้คอมมิตตีแมชชีน. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต. เลย. : มหาวิทยาลัยราชภัฏเลย; 2553.

[15] กิตติ ภักดีวัฒนะกุล. การออกแบบและพัฒนาคลังข้อมูล = Data warehouse. 1st ed. กรุงเทพฯ: วี.ซี.พี; 2552.

[16] สิรินทร์ นิยมางกูร. เทคนิคการสุ่มตัวอย่าง. 1st ed. กรุงเทพฯ สานักพิมพ์มหาวิทยาลัยเกษตรศาสตร์; 2541.

[17] ธานินทร์ ศิลป์จารุ. การวิเคราะห์ข้อมูลทางสถิติ SPSS 11st ed. กรุงเทพฯ ธรรมสาร 2553.