บทที่ 5

บทสรุป

การวิจัยครั้งนี้มีจุดมุ่งหมายเพื่อเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มของข้อมูล 2 วิธี คือ วิธีนาอีฟ เบส์ และวิธีถดถอยโลจิสติก เมื่อตัวแปรอิสระมีลักษณะเป็นแบบไบนารี่ โดยพิจารณาอัตราความผิดพลาดของการจำแนกเป็นเกณฑ์ในการเปรียบเทียบ ภายใต้ขอบเขตของการศึกษาดังนี้

1. ตัวแปรอิสระมีลักษณะเป็นแบบไบนารี่ หรือมีค่าเพียง 2 ค่า (0 กับ 1)
2. ขนาดตัวอย่างที่ใช้ในการศึกษาแบ่งเป็น 5 ระดับ คือ 30, 50, 70, 100 และ 500
2. กำหนดจำนวนตัวแปรอิสระเป็น 4, 6 และ 10 ตัว ตามลำดับ
3. กำหนดจำนวนกลุ่มที่ใช้ในการจำแนกเป็น 2 กลุ่ม (กลุ่ม $Y = 0$ และกลุ่ม $Y = 1$)
4. การวิเคราะห์ข้อมูลจะแบ่งตัวอย่างออกเป็น 2 ชุด คือ ชุดตัวอย่างที่ใช้สร้างเกณฑ์ในการจำแนก (Training data) และชุดตัวอย่างที่ใช้ตรวจสอบเกณฑ์การจำแนก (Validation data) ในอัตราส่วน Training : Validation เป็น 90:10, 80:20, 70:30, 60:40 และ 50:50
5. กำหนดสัดส่วนค่าของตัวแปรอิสระ (0 และ 1) ของทั้งสองกลุ่มให้มีค่าใกล้เคียงกัน (ความน่าจะเป็นอยู่ในช่วงระหว่าง 0.4 ถึง 0.6) และแตกต่างกัน (ความน่าจะเป็นอยู่ในช่วงระหว่าง 0.2 ถึง 0.4 และ 0.7 ถึง 0.9)

การวิจัยครั้งนี้ผู้วิจัยทำการจำลองข้อมูลและวิเคราะห์ข้อมูล โดยใช้โปรแกรม MATLAB เพื่อสร้างข้อมูลในแต่ละสถานการณ์ตามที่กำหนด และกระทำซ้ำทั้งหมด 1,000 ครั้งในแต่ละสถานการณ์ แล้วจึงเปรียบเทียบกันอัตราความผิดพลาดของการจำแนกกลุ่มด้วยวิธีนาอีฟ เบส์กับวิธีถดถอยโลจิสติก สามารถสรุปผลการวิจัย อภิปรายผล รวมทั้งข้อเสนอแนะ ได้ดังนี้

1. สรุปผลการวิจัย

จากการเปรียบเทียบอัตราความผิดพลาดของการจำแนกกลุ่มด้วยวิธีนาอีฟ เบส์และวิธีถดถอยโลจิสติก สามารถสรุปได้ดังนี้

1.1 เมื่อพิจารณาอัตราส่วนของข้อมูลระหว่าง Training : Validation ที่กำหนดเป็น 50:50, 60:40, 70:30, 80:20, และ 90:10 โดยใช้วิธีการจำแนกกลุ่มข้อมูลทั้งสองวิธี โดยส่วนใหญ่พบว่า ถ้าอัตราส่วนของข้อมูลใน Training มีค่าเพิ่มขึ้น อัตราการจำแนกกลุ่มข้อมูลผิดพลาดมีแนวโน้มลดลง นั่นคือ การจำแนกกลุ่มข้อมูลมีความถูกต้องมากขึ้น

1.2 เมื่อพิจารณาจากขนาดตัวอย่างที่กำหนดให้เป็น 30, 50, 70, 100 และ 500 พบว่า ถ้าขนาดตัวอย่างเพิ่มขึ้น อัตราการจำแนกกลุ่มข้อมูลผิดพลาดของทั้งสองวิธีมีแนวโน้มลดลง ดังนั้นเมื่อตัวอย่างมีขนาดใหญ่ขึ้น จะทำให้ผลของการจำแนกกลุ่มข้อมูลมีความถูกต้องมากขึ้นตามไปด้วย

1.3 เมื่อพิจารณาเปรียบเทียบอัตราการจำแนกกลุ่มข้อมูลผิดพลาดระหว่างวิธีนาอีฟ เบส์กับวิธีถดถอยโลจิสติก ในแต่ละสถานการณ์ที่กำหนด พบว่าส่วนใหญ่แล้ววิธีนาอีฟ เบส์สามารถจำแนกกลุ่มข้อมูลได้ถูกต้องมากกว่าวิธีถดถอยโลจิสติก แต่อย่างไรก็ตาม ประสิทธิภาพในการจำแนกกลุ่มของทั้งสองวิธีไม่แตกต่างกันมากนัก

1.4 เมื่อพิจารณาเปรียบเทียบอัตราการจำแนกกลุ่มข้อมูลผิดพลาดระหว่างสัดส่วนค่าของตัวแปรอิสระที่มีค่าใกล้เคียงกันและแตกต่างกันของทั้งสองวิธี ในแต่ละสถานการณ์ที่กำหนด โดยส่วนใหญ่พบว่า เมื่อสัดส่วนค่าของตัวแปรอิสระของข้อมูลทั้งสองกลุ่มมีค่าใกล้เคียงกัน วิธีจำแนกกลุ่มข้อมูลทั้งสองวิธีสามารถจำแนกข้อมูลได้ถูกต้องมากกว่าเมื่อสัดส่วนค่าของตัวแปรอิสระในข้อมูลทั้งสองกลุ่มมีค่าแตกต่างกัน

2. อภิปรายผล

จากการศึกษาเปรียบเทียบอัตราการจำแนกกลุ่มข้อมูลผิดพลาดระหว่างวิธีนาอีฟ เบส์กับวิธีถดถอยโลจิสติกเมื่อตัวแปรอิสระมีลักษณะแบบไบนารี่ พบว่าควรกำหนดให้ขนาดตัวอย่างของข้อมูลมีขนาดใหญ่ เพราะหากจำนวนข้อมูลมากขึ้น อัตราความผิดพลาดของการจำแนกกลุ่มข้อมูลจะมีค่าลดลง และเป็นจริงตามผลการวิเคราะห์ข้อมูลในสองวิธี ซึ่งสอดคล้องกับงานวิจัยของ McCallium, A. และ Nigam, K. (1998) ที่ได้ศึกษาเปรียบเทียบการจำแนกกลุ่มข้อความด้วยวิธีนาอีฟ เบส์ใน 2 ตัวแบบ ซึ่งพบว่าเมื่อคำในข้อความมีจำนวนมากขึ้น วิธีนาอีฟ เบส์สามารถจำแนกกลุ่มคำในข้อความได้ถูกต้องมากยิ่งขึ้น และยังสอดคล้องกับ**ทฤษฎีการลู่เข้าในเชิงความน่าจะเป็น** (Convergence in probability) นั่นคือ ถ้าขนาดตัวอย่างเพิ่มขึ้น หรือทำการทดลองซ้ำ ๆ กันมากขึ้น ค่าประมาณที่ได้จากตัวอย่างจะใกล้เคียงกับค่าที่แท้จริงมากขึ้น (สุชาดา กีระนันทน์, 2545. หน้า 2)

สำหรับอัตราส่วนของข้อมูลระหว่าง Training : Validation พบว่าควรกำหนดให้อัตราส่วนของข้อมูลใน Training มีค่ามาก ๆ เพราะจะทำให้อัตราความผิดพลาดของการจำแนกกลุ่มข้อมูลมีค่าลดลง ซึ่งสอดคล้องกับงานวิจัยของ Ratanamahatana, C. และ Gunopulos, D. (n.d.) ที่ศึกษาการจำแนกกลุ่มด้วยวิธีนาอีฟ เบส์ ซึ่งพบว่าเมื่ออัตราส่วน ของข้อมูลใน Training เพิ่มขึ้น ความถูกต้องในการจำแนกกลุ่มข้อมูลมีความถูกต้องมากขึ้นเช่นกัน และจากงานวิจัยของ

Provost, J. (n.d.) ที่ศึกษาวิธีนาอีฟ เบส์เปรียบเทียบกับกฎการเรียนรู้ในการจำแนกกลุ่มของจดหมายอิเล็กทรอนิกส์ และพบว่าเมื่อจำนวนของข้อมูลใน Training เพิ่มขึ้น การจำแนกกลุ่มจดหมายอิเล็กทรอนิกส์มีความถูกต้องมากขึ้นตามไปด้วย โดยมีเปอร์เซ็นต์ในการจำแนกถูกต้องประมาณ 90% แต่อย่างไรก็ตามในการศึกษาครั้งนี้พบว่า บางกรณีที่อัตราส่วน ของข้อมูลใน Training เพิ่มขึ้น อัตราความผิดพลาดในการจำแนกกลุ่มข้อมูลมีค่าสูงขึ้นตามไปด้วย ทั้งนี้เนื่องจากการสร้างข้อมูลเพื่อแบ่งเป็น Training data และ Validation data ไม่สามารถตรวจสอบได้ว่าชุดข้อมูลทั้งสองกลุ่มมีลักษณะของข้อมูลคล้ายคลึงกันหรือไม่ ซึ่งถ้าชุดข้อมูลใน Training กับ Validation มีลักษณะของข้อมูลแตกต่างกันมาก ทำให้ข้อมูลที่ใช้ในการสร้างเกณฑ์การจำแนกกับข้อมูลที่ใช้ในการตรวจสอบเกณฑ์การจำแนกมีความแตกต่างกันมาก จึงส่งผลให้อัตราความผิดพลาดในการจำแนกกลุ่มข้อมูลมีค่าสูงขึ้นในบางกรณี

3. ข้อเสนอแนะ

ผลของการวิจัยนี้ สามารถนำไปประยุกต์ใช้ได้ดังนี้

3.1 การนำผลการวิจัยไปใช้

ควรเลือกใช้วิธีนาอีฟ เบส์ในการจำแนกกลุ่มเมื่อตัวแปรอิสระมีลักษณะเป็นแบบไบนารี่ เนื่องจากให้ผลในการจำแนกกลุ่มถูกต้องมากกว่าวิธีถดถอยโลจิสติกเป็นส่วนใหญ่ นอกจากนี้เมื่ออัตราส่วนของข้อมูลใน Training มีค่ามากขึ้น ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มข้อมูลมีแนวโน้มลดลง ดังนั้นควรกำหนดให้อัตราส่วนของข้อมูลใน Training มีค่ามาก ๆ เพื่อสร้างเกณฑ์การจำแนกที่มีความถูกต้องมากยิ่งขึ้นและทำให้จำแนกกลุ่มข้อมูลได้ถูกต้องมากขึ้นด้วย นอกจากนี้ควรเลือกใช้ตัวอย่างที่มีขนาดปานกลางถึงใหญ่ เพราะจะทำให้อัตราการจำแนกกลุ่มผิดพลาดมีค่าต่ำ

ยกตัวอย่างข้อมูลในงานวิจัยทางด้านการแพทย์ เช่น การศึกษาข้อมูลของผู้ป่วยโรคหัวใจ ซึ่งเป็นข้อมูลจาก UCI Repository จำนวน 597 ข้อมูล โดยตัวแปรที่แสดงคุณลักษณะหรือตัวแปรแบ่งกลุ่มที่ใช้ในการจำแนกกลุ่มผู้ป่วยโรคหัวใจประกอบด้วยคุณลักษณะที่เป็นปัจจัยในการ จำแนกกลุ่มข้อมูลผู้ป่วยโรคหัวใจจำนวน 7 คุณลักษณะ ได้แก่ เพศ ชนิดของอาการเจ็บหน้าอก ความดันโลหิต (มม.ปรอท) ระดับคอเลสเตอรอลในซีรัม (มก./เดซิลิตร) ระดับน้ำตาลในกระแสเลือด (มก./เดซิลิตร) การตรวจสอบคลื่นหัวใจด้วยกระแสไฟฟ้า และความเจ็บปวดที่เกิดจากการออกกำลังกาย

3.2 การศึกษาหัวข้อวิจัยสืบเนื่อง

3.2.1 เป็นแนวทางในการวิเคราะห์การจำแนกกลุ่ม เมื่อข้อมูลประกอบด้วยตัวแปรอิสระที่มีลักษณะอื่น ๆ เช่น เป็นตัวแปรเชิงคุณภาพแบบเรียงลำดับทั้งหมด หรือเป็นตัวแปรเชิงปริมาณผสมกับตัวแปรเชิงคุณภาพ เป็นต้น

3.2.2 เป็นแนวทางในการศึกษาการจำแนกกลุ่มของข้อมูลมากกว่า 2 กลุ่มขึ้นไป

3.2.3 เป็นแนวทางในการศึกษาและเปรียบเทียบวิธีการจำแนกกลุ่มแบบอื่น เช่น วิธีเค-เนียเลสเนเบอร์ (k-Nearest neighbor method) วิธีเคอร์เนล (Kernel method) และวิธีนิวรอลเน็ตเวิร์ก (Neural network method) เป็นต้น