**บทที่ 5**

**ผลการวิเคราะห์การทดลอง**

เป้าหมายหลักของบทนี้คือการวิเคราะห์ผลการทดลองจากการนำชุดข้อมูลฝึกสอนที่สร้างจากกฎของไฟร์วอลล์ที่ออกแบบให้ตรงตามจุดประสงค์ของสมมติฐาน เพื่อหาชุดข้อมูลฝึกสอนที่สามารถทำให้โมเดลมีประสิทธิภาพในด้านความแม่นยำในการทำนายและเวลาที่ใช้ได้ดีที่สุด จึงจำเป็นต้องมีการวิเคราะห์ในเชิงเปรียบเทียบ ปรับรูปแบบกราฟเพื่อหาความสัมพันธ์ของตัวแปรต่างๆ

**5.1. วิเคราะห์กลไกการทำงานโดยรวมของโมเดล**

5.1.1 สรุปการวิเคราะห์ความสัมพันธ์ เวลาที่ใช้ในการฝึกสอนโมเดลและจำนวนข้อมูลฝึกสอน

จากผลการทดลองในบทก่อนหน้าพบว่า ทั้ง 2 ผลการทดลอง เมื่อมีจำนวนชุดข้อมูลฝึกสอนในระบบมากขึ้น โมเดลจะใช้เวลาในการเรียนรู้ชุดข้อมูลฝึกสอนในอัตราคงที่ สังเกตได้จากกราฟที่ออกมามีลักษณะใกล้เคียงกับกราฟเส้นตรงมาก ซึ่งเมื่อลองคำนวณจากทั้ง 2 การทดลองพบว่าจะเป็นไปตามสูตรดังนี้

รูปภาพที่ 5.1 กราฟผลลัพธ์ เวลาที่ใช้ในการฝึกสอนโมเดล : จำนวนชุดข้อมูลฝึกสอนที่ใช้

5.1.2. สรุปการวิเคราะห์เวลาที่ใช้ในการฝึกสอนโมเดลและจำนวนข้อมูลฝึกสอน

5.1.2. สรุปการวิเคราะห์ความสัมพันธ์ เวลาที่ใช้ในการประมวลผลและจำนวนข้อมูลที่ใช้ฝึกสอน

จากผลการทดลองในบทก่อนหน้าพบว่า ทั้ง 2 ผลการทดลอง จะเห็นได้ว่าจำนวนชุดข้อมูลฝึกสอนแทบไม่ส่งผลกับเวลาที่ใช้ในการประมวลผล นั่นหมายความว่าถ้าหากเราใช้ชุดกฎไฟร์วอลล์ที่มีเงื่อนไขมากขึ้นก็ยังใช้เวลาในการประมวลผลเท่าเดิม

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

รูปภาพที่ 5.2 เปรียบเทียบกราฟผลลัพธ์เวลาที่ใช้ในการประมวลของ N Sample และ Ratio

จะเห็นได้ว่าทั้ง 2 รูป ถึงจะใช้เวลามากขึ้นหรือน้อยลงบ้าง แต่ค่าความแตกต่างจะอยู่ในเสี้ยววินาทีเท่านั้น

5.1.3. สรุปการวิเคราะห์ความสัมพันธ์ ความแม่นยำการทำนายต่อจำนวนของข้อมูลฝึกสอน

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

รูปภาพที่ 5.3 เปรียบเทียบกราฟผลลัพธ์ความแม่นยำของการแบ่งชุดข้อมูลฝึกสอนแต่ละแบบ

จะเห็นได้ว่า ถ้าหากเป็น N Sample ค่าความแม่นยำที่ได้จะค่อยๆเพิ่มขึ้น และขึ้นเกือบจุดสูงสุดที่จุดหนึ่ง ซึ่งตัวแปรที่ใช้ในการทดลองจุดนั้นคือ 600 ชุดข้อมูลฝึกสอนต่อหนึ่งกฎไฟร์วอลล์ แต่เมื่อให้จำนวนข้อมูลฝึกสอนที่มากกว่านั้น ความแม่นยำในทุกชุดกฎไฟร์วอลล์จะเริ่มตกลงเล็กน้อย โดยเฉพาะชุดกฎไฟร์วอลล์ที่มีเงื่อนไขที่มากกว่า ยกเว้นชุด 2 กฎไฟร์วอลล์ที่มี Default Rule ด้วย

จึงสรุปได้ว่า ความซับซ้อนและเงื่อนไขของชุดข้อมูลฝึกสอนส่งผลต่อ การเรียนรู้ของโมเดล หมายความว่า ถ้าหากเงื่อนไขกฎไฟร์วอลล์ที่ใช้นั้นมีจำนวนเงื่อนไขและกฎการประเมินที่มากขึ้น จำเป็นต้องหาจำนวนของข้อมูลที่จะใช้ฝึกสอนที่เหมาะสมที่ทำให้โมเดลมีประสิทธิภาพมากที่สุด

**5.2. วิเคราะห์โมเดลในเชิงประสิทธิภาพ**

5.2.1. อัตราการเรียนรู้ของโมเดล

ตัวแปรสำคัญในการวัดผลในเชิงประสิทธิภาพ ได้แก่ ความแม่นยำในการทำนายผล เวลาที่ใช้ในการฝึกโมเดล และ จำนวนชุดข้อมูลทดสอบ ซึ่งทั้ง 3 ค่านี้ให้ความหมายในเชิงประสิทธิภาพดังนี้

* โมเดลที่มีความแม่นยำสูง เป็นส่วนสำคัญของโมเดลที่ได้ประสิทธิภาพ
* โมเดลที่ใช้เวลาในการเรียนรู้น้อยกว่าย่อมดีกว่าโมเดลที่ใช้เวลาในการเรียนรู้มากกว่า ถ้าหากโมเดลทั้งสองให้ผลลัพธ์ความถูกต้องในการทำนายผลเท่ากัน
* โมเดลที่ใช้จำนวนชุดข้อมูลฝึกสอนน้อยกว่าดีกว่าโมเดลที่ใช้จำนวนชุดข้อมูลฝึกสอนมากกว่า ถ้าหากโมเดลทั้งสองให้ผลลัพธ์ความถูกต้องในการทำนายผลเท่ากัน ซึ่งจำนวนชุดข้อมูลฝึกสอนจะมีผลโดยตรงกับเวลาที่ใช้ นั่นหมายความว่า เราจะต้องใช้เวลาในการสร้างชุดข้อมูลและฝึกฝนนานขึ้น

ซึ่ง 3 ตัวแปรนี้ ทำให้ได้กราฟอีกตัวหนึ่งคือกราฟ ความแม่นยำ / เวลาฝึกโมเดล : จำนวนชุดข้อมูลฝึกสอน

Chart, line chart

Description automatically generated

รูปที่ 5.4 กราฟความแม่นยำ / เวลาฝึกโมเดล : จำนวนชุดข้อมูลฝึกสอนของ N Sample

Chart, line chart

Description automatically generated

รูปที่ 5.5 กราฟความแม่นยำ / เวลาฝึกโมเดล : จำนวนชุดข้อมูลฝึกสอนของอัตราส่วน Ratio

Chart, line chart

Description automatically generated

รูปที่ 5.6 กราฟความแม่นยำ / เวลาฝึกโมเดล : จำนวนชุดข้อมูลฝึกสอนของ N Sample แบบปรับตามผลลัพธ์

จะเห็นได้ว่ากราฟทั้งสองรูปแบบทั้ง N Sample และอัตราส่วน Ratio มีอัตราการเรียนรู้ที่มีลักษณะคล้ายกันคือ โมเดลที่มีการเรียนรู้จากชุดข้อมูลฝึกสอนน้อยยังสามารถคำนวณหาผลลัพธ์ได้ถูกต้องบ้างอยู่ แต่การที่จะเพิ่มความแม่นยำได้นั้นจะต้องเพิ่มจำนวนชุดข้อมูลฝึกสอนไปอีกเกือบเท่าตัวหรือหลายเท่า นั่นหมายความว่าอัตราการเรียนรู้จะเริ่มน้อยลงเรื่อยๆแปรผกผันกับจำนวนชุดข้อมูลฝึกสอนที่ป้อนเข้าไป

5.2.2. การหาจุดจำนวนชุดข้อมูลฝึกสอนที่เหมาะสมที่สุดในการพัฒนาโมเดล

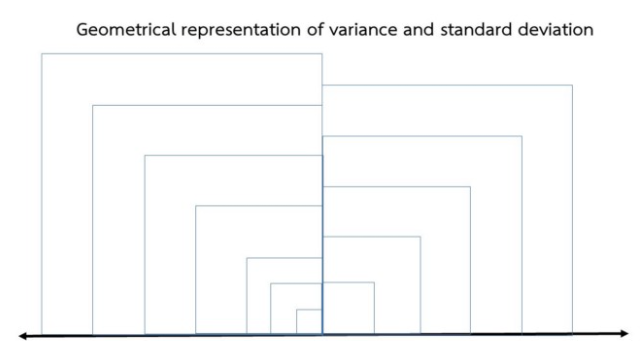
ในการเลือกจำนวนชุดข้อมูลฝึกสอนมาใช้พัฒนาโมเดล ผลลัพธ์สุดท้ายที่สามารถตอบโจทย์ประสิทธิภาพทั้ง 3 ค่าได้ โดยประกอบไปด้วย จำนวนชุดข้อมูลฝึกสอน เวลาที่ใช้ในการฝึกสอน และความแม่นยำในการทำนายผล จะต้องให้ผลลัพธ์ได้ดีที่สุด ซึ่งในหัวข้อวิเคราะห์นี้เราได้ใช้ผลลัพธ์การทดลองและหลักการมาอ้างอิง ดังนี้

5.2.2.1. การอ้างอิงความสัมพันธ์ของเวลาฝึกสอนและจำนวนชุดข้อมูลฝึกสอน

จากการนำกราฟที่ 5.1 มาวิเคราะห์ พบว่าจำนวนชุดข้อมูลฝึกสอนโมเดลและเวลาที่ใช้ในการฝึกสอนโมเดลมีความสัมพันธ์แปรผันตรงด้วยอัตราการเพิ่มค่าที่คงที่ หมายความว่าเราสามารถหาเวลาที่ใช้ในการฝึกสอนโมเดลได้ถ้าหากเราทราบจำนวนชุดข้อมูลฝึกสอนที่ใช้ ซึ่งจุดจำนวนชุดข้อมูลฝึกสอนที่ให้ค่าความถูกต้องได้เยอะที่สุด เหมาะสมที่จะเป็นจำนวนที่ควรหยิบมาใช้มากที่สุดนั่นเอง

5.2.2.2. การใช้หลักการหาค่าความแปรปรวนในการเปลี่ยนมิติการเปรียบเทียบ

ความแปรปรวนคือความแกว่งของข้อมูลหรือความห่างของการกระจายตัวของชุดข้อมูล โดยในทางสถิติหรือการเปรียบเทียบเชิงคณิตศาสตร์มักมีการนำความแปรปรวนและส่วนเบี่ยงเบนมาใช้ในการหาค่าเฉลี่ยของกราฟ ซึ่งหลักการหนึ่งที่ทำให้เห็นผลชัด คือการนำค่ามายกกำลังสองตามสูตรความแปรปรวน เราจะสามารถหาจุดแตกต่างหรือมีความแปรปรวณได้ชัดเจนขึ้น โดยสมการมีรูปแบบสูตร ดังนี้

  
รูปที่ 5.7 เปรียบเทียบเมทริกซ์ Geometrical ก่อนและหลังยกกำลังสอง

จาก 2 ประเด็นพิจารณา 2 ข้อข้างต้นแล้วทำให้เราตัดสินใจสรุปเป็นกราฟใหม่คือ ค่าความแม่นยำในการทำนายที่ยกกำลังสองต่อจำนวนข้อมูลฝึกสอนที่ใช้ เนื่องจากเวลาที่ใช้ในฝึกโมเดลเพิ่มขึ้นด้วยอัตราคงที่มีผลโดยตรงต่อจำนวนที่ป้อนเข้าไปในระบบอยู่แล้ว จึงไม่จำเป็นต้องนำค่านี้มาคิด

จากกราฟที่ได้ทำการปรับค่าตัวแปรใหม่ จะเห็นได้ว่า N Sample ที่มีการแบ่งจำนวนชุดข้อมูลฝึกสอนต่อกฎด้วยจำนวนที่เท่ากันจะได้จุดที่เหมาะสมที่อยู่ในช่วงจำนวน 300 ถึง 1,000 ชุดข้อมูลฝึกสอน และค่อนข้างได้ผลดีสุดที่ช่วงจำนวนข้อมูลฝึกสอน 600 ชุด แต่สำหรับการแบ่งเป็นอัตราส่วน Ratio ที่มีการแบ่งจำนวนชุดข้อมูลฝึกสอนด้วยอัตราส่วนที่เท่ากันจะให้ค่าที่ไม่แน่นอน ในบางการทดลองสามารถหาจุดที่เหมาะสมได้ แต่ส่วนใหญ่ค่าความแม่นยำจะตกลง อาจเนื่องจากปัญหา Overfitting ที่มีการป้อนข้อมูลไม่จำเป็นมากเกินไปจนทำให้โมเดลคำนวณได้ผิดพลาด ดังนั้นจากผลการทดลองจึงสรุปได้ว่าการแบ่งจำนวนข้อมูลฝึกสอนควรจะใช้แบบ N Sample จะได้ประสิทธิภาพมากกว่าการแบ่งชุดข้อมูลฝึกสอนโดยใช้อัตราส่วน Ratio