บทที่ 1

บทนำ

เนื้อหาในบทนี้จะกล่าวถึงความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย และรายละเอียดของวิทยานิพนธ์ มีรายละเอียดดังต่อไปนี้

ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันนี้เป็นที่ยอมรับกันโดยทั่วไปว่าคอมพิวเตอร์เป็นส่วนหนึ่งของวิถีชีวิตมนุษย์ บางคนใช้งานคอมพิวเตอร์มากกว่าวันละ 8 ชั่วโมง และบางคนใช้งานคอมพิวเตอร์ไม่ถึง 1 ชั่วโมงต่อวัน ซึ่งแต่ละคนนั้นมีจุดมุ่งหมายในการใช้คอมพิวเตอร์แตกต่างกันออกไป บ้างใช้เพื่อการทำงาน พิมพ์รายงาน บ้างใช้เพื่อการบริโภคข้อมูลข่าวสาร เช่น การรับส่งจดหมายอิเล็กทรอนิกส์ (e - mail) อ่านข่าวจากหนังสือพิมพ์ที่อยู่บนอินเตอร์เน็ต วิเคราะห์ข้อมูลตลาดหุ้นผ่านเว็บไซต์ที่แสดงไว้บนอินเตอร์เน็ต การค้นหาข้อมูลต่าง ๆ บนอินเตอร์เน็ต บ้างก็ใช้เพื่อความบันเทิงเช่น การเล่นเกมส์ ดูหนัง ฟังเพลง หรือสื่อสารในรูปแบบต่าง ๆ เป็นต้น นั่นก็หมายความว่าผู้ใช้คอมพิวเตอร์จะมีการใช้แป้นพิมพ์ที่แตกต่างกัน เช่นหากใช้เพื่อการทำงานพิมพ์รายงาน หรือพิมพ์สนทนาโต้ตอบกันในโปรแกรมประเภทข้อความด่วน (Instant Messaging) ซึ่งเป็นที่แน่นอนว่าจะต้องมีการสัมผัสกับแป้นพิมพ์อยู่เสมอ แต่ถ้าใช้เพื่อเล่นเกมส์บางประเภทอาจใช้เพียงแค่เมาส์ก็สามารถเล่นเกมส์ได้ ทำให้การสัมผัสกับแป้นพิมพ์น้อยลง การพิมพ์บนแป้นพิมพ์ของผู้ใช้แต่ละคนนั้นย่อมแตกต่างกัน

งานวิจัยนี้จึงได้นำแนวคิดในส่วนของความแตกต่างของเวลาที่ผู้ใช้ทำการพิมพ์เพื่อมาใช้ในการจำแนกผู้ใช้ ซึ่งความแตกต่างของเวลาที่ใช้ในการพิมพ์เกิดจากบางคนพิมพ์เร็ว บางคนพิมพ์ช้า บางคนพิมพ์สัมผัสโดยใช้มือทั้งสองมือในการพิมพ์ แต่บางคนอาจใช้เพียงแค่นิ้วชี้ในการพิมพ์เท่านั้น ขึ้นอยู่กับทักษะในด้านการพิมพ์ของผู้ใช้แต่ละคน ดังนั้นพลวัตการพิมพ์ (Keystroke Dynamic) ของแต่ละคนก็ย่อมไม่เท่ากัน นั่นย่อมแสดงให้เห็นว่าคนแต่ละคนมีเอกลักษณ์ของการพิมพ์เฉพาะบุคคล และสำหรับงานวิจัยนี้ได้นำเอาพลวัตของจังหวะการพิมพ์ของแต่ละบุคคล มาเป็นข้อมูลที่นำเข้า (Input) โดยการตรวจสอบจังหวะการพิมพ์นี้ เป็นวิธีการ

ตรวจสอบในรูปแบบหนึ่งของ ไบโอเมตริก (biometric) ซึ่งอยู่บนพื้นฐานของการวัดค่าเวลาระหว่างแป้นพิมพ์ ค่าเวลานี้ได้ถูกจัดเก็บตั้งแต่ขณะที่ผู้ใช้พิมพ์รหัสผ่านไปจนถึงการพิมพ์ข้อมูลต่าง ๆ เพื่อใช้สำหรับการจำแนกผู้ใช้จากกลุ่มผู้ใช้ทั้งหมดด้วยข้อความอิสระ (Free – Text)

1.2 วัตถุประสงค์ของการวิจัย

เพื่อศึกษาวิธีการที่สามารถนำมาใช้จำแนกความแตกต่างของผู้ใช้ออกจากกลุ่มผู้ใช้ทั้งหมดได้ โดยใช้เวลาในการพิมพ์ของคู่อักขระทั้งประโยคมาใช้ในการจำแนก

1.3 ขอบเขตของการวิจัย

1. กลุ่มผู้ทดลองเป็นบุคลากรของสำนักงานผู้บังคับทหารอากาศดอนเมือง จำนวน 20 คน เป็นกรณีศึกษา
2. โปรแกรมภาษาจาวาที่ใช้ในการเก็บเวลาที่ใช้พิมพ์ระหว่างคู่อักขระ ซึ่งอยู่บนพื้นฐานของการวัดค่าเวลาการกดแป้นพิมพ์โดยค่าเวลาที่ได้นั้นมีหน่วยวัดเป็นมิลลิวินาที (millisecond) ซึ่งค่าเวลานี้ได้ถูกจัดเก็บระหว่างที่ผู้ใช้พิมพ์ข้อความที่กำหนด โดยผู้ใช้ทุกคนพิมพ์ด้วยแป้นพิมพ์เดียวกันคือแป้นพิมพ์ของเครื่องคอมพิวเตอร์โน้ตบุก IBM รุ่น R51 หน่วยประมวลผล (CPU) Intel Pentium 1.50 MH_Z หน่วยความจำสำรอง (RAM) 512 MB
3. ผู้ใช้ทุกคนจะต้องไม่พิมพ์ต่อเนื่องกัน คือสามารถพิมพ์ได้เพียง 1 - 2 ครั้ง เท่านั้นใน 1 วัน และในระหว่างวันควรเว้นระยะในการพิมพ์ครั้งต่อไปอย่างน้อย 1 – 2 ชั่วโมง จึงจะสามารถพิมพ์การทดลองครั้งต่อไปได้
4. ตัวอย่างประโยคที่นำมาให้ผู้ใช้พิมพ์นั้นเป็นคำ หรือประโยคที่ผู้ใช้ได้พิมพ์อยู่เป็นประจำ โดยประโยคที่ใช้พิมพ์มีอยู่ด้วยกันทั้งหมด 10 ประโยค ซึ่งผู้ใช้นั้นจะต้องพิมพ์ทั้งหมดคนละ 10 ครั้ง

1.4 ประโยชน์ที่ได้จากงานวิจัย

สามารถนำเทคนิคที่ได้นี้มาใช้เพื่อทำการจำแนกผู้ที่ทำการพิมพ์ว่าเป็นผู้ใช้คนใด ทำให้สามารถจำแนกผู้ใช้ได้อย่างถูกต้อง

1.5 รายละเอียดของงานวิจัย

บทที่ 1 บทนำ ความสำคัญของปัญหา วัตถุประสงค์ของงานวิจัย ขอบเขตของงานวิจัย ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย และรายละเอียดของวิทยานิพนธ์

บทที่ 2 ทฤษฎีที่นำมาใช้ ผลงานวิจัยและงานเขียนอื่น ๆ ที่เกี่ยวข้อง

บทที่ 3 ระเบียบการดำเนินงาน การควบคุมการทดลอง รูปแบบการทดลอง ผู้เข้าร่วมการทดลอง ขั้นตอนการทดลอง เครื่องมือในการทดลอง วิธีการประเมินหรือการวัดผล การวิเคราะห์ข้อมูลทางสถิติ

บทที่ 4 แนวทางการพัฒนา สถาปัตยกรรมของระบบ ขั้นตอนการทำงานของระบบ วิธีการพัฒนา

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ วิธีการวิจัย ผลของการทดสอบโปรแกรมสรุปผลการทดลอง ปัญหาและอุปสรรค ข้อจำกัดของระบบ ข้อเสนอแนะ

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

บทนี้ได้นำเสนอทฤษฎีและงานวิจัยต่างๆ ที่เกี่ยวข้องเพื่อใช้ในการจำแนกผู้พิมพ์แต่ละคน โดยได้แบ่งการนำเสนอดังนี้

2.1 การจำแนกประเภทข้อมูล (Data Classification)

2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

2.1.2 โครงข่ายงานประสาทเทียม (Artificial Neural Network)

2.2 การประเมินตัวแบบ

2.2.1 วิธีการประเมินตัวแบบด้วยวิธีการไขว้ข้าม 10 กลุ่ม (10 - Fold Cross Validation)

2.2.2 วิธีการประเมินตัวแบบโดยการใช้การรวมตัวจำแนก (Ensemble of Classifiers)

- แบ็กกิง (Bagging)
- บูสต์ (Boosting Classifier)

2.3 งานวิจัยที่เกี่ยวข้อง

2.1 การจำแนกประเภทข้อมูล (Data Classification)

การจำแนกประเภทข้อมูลเป็นเทคนิคที่ใช้ในการจัดกลุ่มข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นถึงความแตกต่างของคลาสข้อมูล (Class) เพื่อทำนายว่าข้อมูลนี้ควรจัดอยู่ในคลาสใด โดยจะทำการแบ่งข้อมูลออกเป็น 2 กลุ่ม คือ กลุ่มของชุดข้อมูลสำหรับเรียนรู้ (Training Data) และกลุ่มของชุดข้อมูลทดสอบ (Testing Data) โดยชุดของข้อมูลสำหรับเรียนรู้เพื่อให้ระบบทำการเรียนรู้ว่ามีข้อมูลใดอยู่ในกลุ่มเดียวกัน และนำข้อมูลตั้งต้นส่วนที่เหลือจากชุดข้อมูลสำหรับเรียนรู้เป็นข้อมูลที่ใช้ทดสอบ ซึ่งข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากรูปแบบเพื่อทดสอบความถูกต้องและปรับปรุงแบบจำลองจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามาจะนำข้อมูลมาผ่านแบบจำลอง โดยแบบจำลองสามารถทำนายกลุ่มของข้อมูลได้ ซึ่งแบบจำลองที่ได้อาจแสดงในรูปของ ต้นไม้ตัดสินใจ

(Decision Tree) หรือ โครงข่ายงานประสาทเทียม (Artifical Neural Network) เป็นต้น (Han and Kamber,2000)

2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Quinlan, J.R.,1993) เป็นวิธีหนึ่งที่สำคัญในการจำแนกประเภทข้อมูล และเป็นที่รู้จักกันอย่างกว้างขวาง คำตอบที่ได้จากต้นไม้ตัดสินใจนั้นถือว่าเป็นวิธีในการประเมินคำตอบที่ทำงานได้ดีแม้ว่าจะมีข้อมูลรบกวนมาก มีอัลกอริทึมที่ใช้ในการนำมาสร้างต้นไม้ตัดสินใจหลายแบบด้วยกัน เช่น ต้นไม้จำแนกและถดถอย (Classification and Regression Tree:CART), การอนุมานต้นไม้ตัดสินใจ (Induction of Decision Tree:ID3) และ C4.5 ซึ่งวิทยานิพนธ์ฉบับนี้ได้เลือกใช้อัลกอรึทึมของ C4.5 มาใช้ในการทำงาน รูปแบบของเอาต์พุตที่ได้อยู่ในรูปของแผนภูมิต้นไม้ที่ง่ายต่อการเรียนรู้และการทำความเข้าใจ

การเรียนรู้ต้นไม้ตัดสินใจ (Decision Tree Learning)

ต้นไม้ตัดสินใจเป็นวิธีการเรียนรู้จากตัวอย่างที่อาศัยการจำแนกประเภทจากตัวอย่างที่เรียกว่าข้อมูลเรียนรู้ และสร้างเป็นต้นไม้ตัดสินใจ โดยการเรียนรู้ต้นไม้ตัดสินใจจะถูกใช้ในการอนุมานแบบอุปนัย (Inductive inference) เช่น การวินิจฉัยโรค, การประเมินเสี่ยง หรือการจำแนกลักษณะเฉพาะบุคคล เป็นต้น

ข้อมูลเรียนรู้ประกอบไปด้วยแถวแสดงถึงตัวอย่าง และสดมภ์แสดงถึงคุณลักษณะ ซึ่งมี 2 ชนิดดังตัวอย่างในตารางที่ 2.1 ดังนี้

1. คุณลักษณะแบบมีค่าเป็นกลุ่ม (Category attribute) เป็นคุณลักษณะที่กำหนดว่าตัวอย่างถูกจัดอยู่ในคลาสใด โดยจะมีเพียงคุณลักษณะเดียวเท่านั้นในแต่ละตัวอย่างในชุดข้อมูลสำหรับการเรียนรู้

2. คุณลักษณะแบบมีค่าไม่เป็นกลุ่ม (Non-Category attribute) เป็นคุณลักษณะที่บอกถึงลักษณะต่างๆ ของตัวอย่าง ซึ่งสามารถมีได้หลายลักษณะในแต่ละชุดของข้อมูล

ตารางที่ 2.1

ตัวอย่างชุดข้อมูลสำหรับเรียนรู้ของคลาสเล่นกีฬาเทนนิส

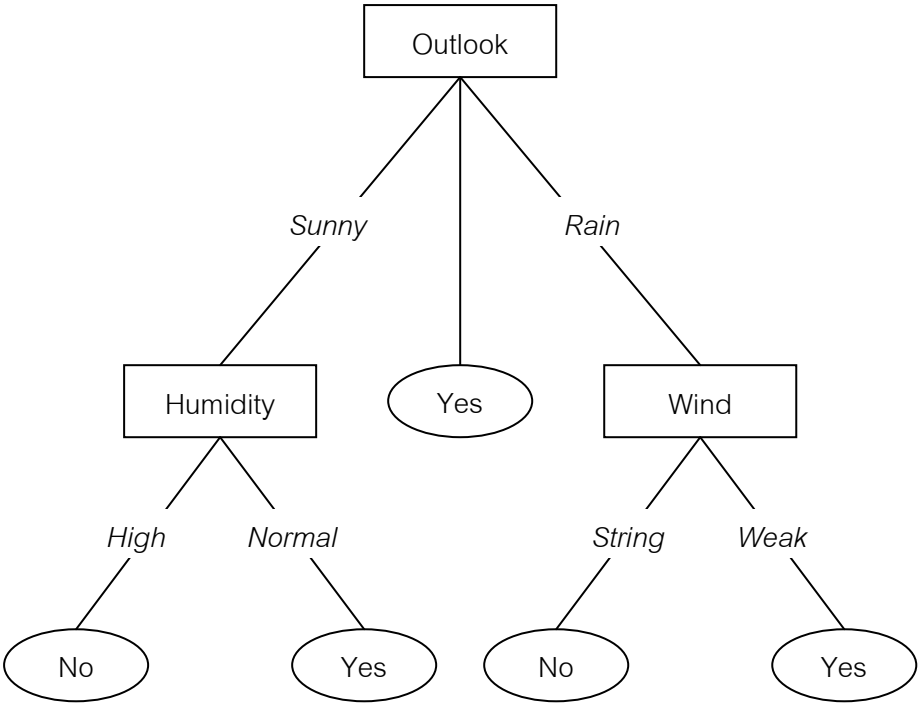
Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Hot	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

จากตารางที่ 2.1 แสดงถึงคุณลักษณะต่างๆ ที่มีผลต่อการเล่นเทนนิส โดยมีคุณลักษณะแบ่งกลุ่มเป็นเล่นเทนนิส (Play Tennis) ซึ่งมีค่าที่เป็นไปได้คือ เล่น (Yes) กับ ไม่เล่น (No) และมีคุณลักษณะไม่แบ่งกลุ่มคือ สภาพอากาศ (Outlook), อุณหภูมิ (Temperature), ความชื้นสัมพัทธ์ (Humidity) และ กำลังลม (Wind) ตามลำดับ ซึ่งต้นไม้ตัดสินใจเป็นโครงสร้างข้อมูลแบบกราฟ โดยมีส่วนประกอบดังภาพที่ 2.1 ดังนี้

1. โหนด(Node) แสดงถึงคุณลักษณะไม่แบ่งกลุ่มและเรียกโหนดแรกของต้นไม้ตัดสินใจว่าราก (Root Node)
2. กิ่ง แสดงถึงค่าที่เป็นไปได้ตามคุณลักษณะไม่แบ่งกลุ่ม
3. ใบ แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะแบ่งกลุ่ม ซึ่งเป็นค่าคำตอบของการตัดสินใจ

4. วิถี คือ เส้นทางที่เชื่อมต่อกันของโหนดแต่ละกิ่งจากรากถึงใบ
5. ขนาด คือจำนวนทั้งหมดของต้นไม้ตัดสินใจ

ภาพที่ 2.1
ตัวอย่างโครงสร้างต้นไม้เพื่อคาดเดาว่าจะเล่นกีฬาเทนนิสหรือไม่



ค่าการเพิ่มสารสนเทศ (Information Gain)

การทำงานของอัลกอริทึมนี้ พยายามสร้างต้นไม้ตัดสินใจที่มีขนาดเล็กกว่าต้นไม้ตัดสินใจ เดิมที่มีขนาดใหญ่ โดยทำการเลือกค่าคุณลักษณะที่ดีที่สุดในการจำแนกตัวอย่าง จากนั้นใช้ค่าการเพิ่มสารสนเทศ (Information Gain) ช่วยในการวัดค่าคุณลักษณะที่ดีที่สุด โดยเริ่มต้นอธิบายการวัดที่ใช้ในทฤษฎีสารสนเทศ (Information Theory) ที่เรียกว่า เอนโทรปี (Entropy) ซึ่งค่าเอนโทรปีแสดงถึงการวัดของการปนเปื้อนของตัวอย่างเรียนรู้ โดยกำหนดให้

S คือ กลุ่มของตัวอย่างที่บรรจุทั้งตัวอย่างบวกและตัวอย่างลบ

$Entropy(S)$ คือ จำนวนบิตที่ใช้ในการอธิบายเหตุการณ์ที่ต้องการเข้ารหัสคลาสตัวอย่างบวกและตัวอย่างลบ

p_+ คือ สัดส่วนตัวอย่างบวกใน S

p_- คือ สัดส่วนตัวอย่างลบใน S

$$Entropy(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

โดยค่าเอนโทรปีจะมีค่าเท่ากับศูนย์ “0” ก็ต่อเมื่อสมาชิกทั้งหมดใน S มีคลาสเหมือนกันทั้งหมด ดังนั้นถ้าตัวอย่างทั้งหมดเป็นตัวอย่างบวกจะได้ว่า

$$Entropy(S) \equiv -1\log_2(1) - 0\log_2(0)$$

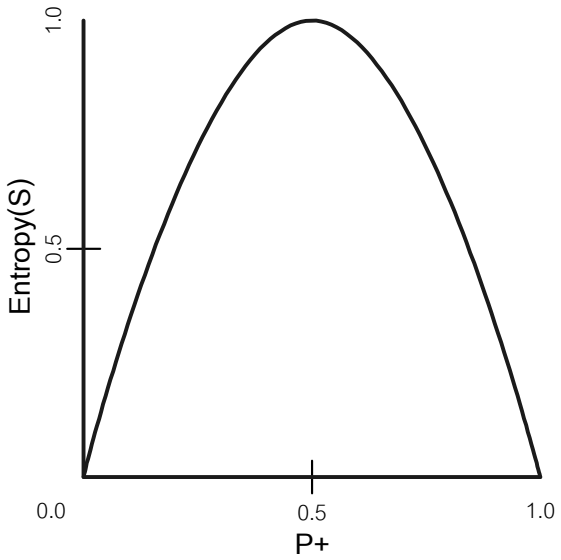
จากตาราง 2.1 มีตัวอย่างทั้งหมด 14 ตัวอย่าง เป็นตัวอย่างบวก 9 ตัวอย่างและตัวอย่างลบ 5 ตัวอย่างดังนั้น ค่าเอนโทรปีของตัวอย่างทั้งหมดคำนวณได้ ดังนี้

$$Entropy(S) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$$
$$= 0.940 \text{ บิต (bits)}$$

กรณีที่เป็นตัวอย่างบวกทั้งหมดหรือตัวอย่างลบทั้งหมดค่าเอนโทรปีมีค่าเท่ากับ 1 ดังนั้นสรุปได้ว่าค่าเอนโทรปีจะมีค่าตั้งแต่ 0 ถึง 1 ดังภาพที่ 2.2

ภาพที่ 2.2

ความสัมพันธ์ของค่าเอนโทรปี



การเรียนรู้ต้นไม้ตัดสินใจอาศัยหลักการเพิ่มสารสนเทศ ซึ่งค่าเอนโทรปีแสดงถึงการวัดของการปนเปื้อนของตัวอย่างเรียนรู้ โดยค่าการเพิ่มสารสนเทศสามารถอธิบายการวัดประสิทธิภาพของค่าคุณลักษณะในข้อมูลการเรียนรู้ เพื่อให้ทราบว่าค่าคุณลักษณะไหนดีกว่า โดยในการคำนวณค่าความสามารถในการแยกตัวอย่าง และเรียกค่าที่คำนวณได้ว่าค่าการเพิ่มสารสนเทศ มีค่าอยู่ระหว่าง 0 กับ 1 ค่าการเพิ่มสารสนเทศเท่ากับ 0 แสดงถึงความสามารถในการแยกตัวอย่างเป็นกลุ่มเดียวกันทั้งหมด และความสามารถในการแยกตัวอย่างลดลงเมื่อค่าการเพิ่มสารสนเทศเพิ่มขึ้น เมื่อค่าเอนโทรปีแสดงการปนเปื้อนของตัวอย่างการเรียนรู้ เราใช้ค่าการเพิ่มสารสนเทศอธิบายการวัดของประสิทธิภาพของคุณลักษณะในข้อมูลสำหรับเรียนรู้เพื่อให้ทราบว่าคุณลักษณะใดดีกว่า โดยให้ $Gain(S, A)$ คือ ค่าการเพิ่มสารสนเทศของค่าคุณลักษณะ A

กำหนดให้

S คือ ตัวอย่างทั้งหมด

$Values(A)$ คือ กลุ่มของค่าที่เป็นไปได้ทั้งหมดสำหรับค่าคุณลักษณะ A

S_v คือ กลุ่มย่อยของ S ของค่าคุณลักษณะ A

A มีค่า v เช่น $S_v = s \in S | A(s) = v$

จะได้

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

or

$$Gain(S,A) \equiv Entropy(S) - E(A)$$

$$E(A) \equiv \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

จากสมการ $Gain(S,A)$ ในพจน์แรกเป็นค่าเอนโทรปีของกลุ่มตัวอย่างเริ่มต้นทั้งหมด ในส่วนพจน์ที่สอง คือ ค่าที่คาดหวังของเอนโทรปีหลังจากที่ S ถูกแบ่งโดยใช้ค่าคุณลักษณะ A ยกตัวอย่างเช่น การหาค่าการเพิ่มสารสนเทศของคุณลักษณะ Wind ในตาราง 2.1 ประกอบไปด้วยตัวอย่างบวก 9 ตัวอย่าง ตัวอย่างลบ 5 ตัวอย่าง และค่าคุณลักษณะของ Wind มี 2 ค่าคือ Weak และ Strong ดังนั้นค่าการเพิ่มสารสนเทศมีค่าเท่ากับ

$$Values(Wind) = Weak, Strong$$
$$S = [9+,5-]$$
$$S_{Wind=Weak} \leftarrow [6+,2-]$$
$$S_{Wind=Strong} \leftarrow [3+,3-]$$

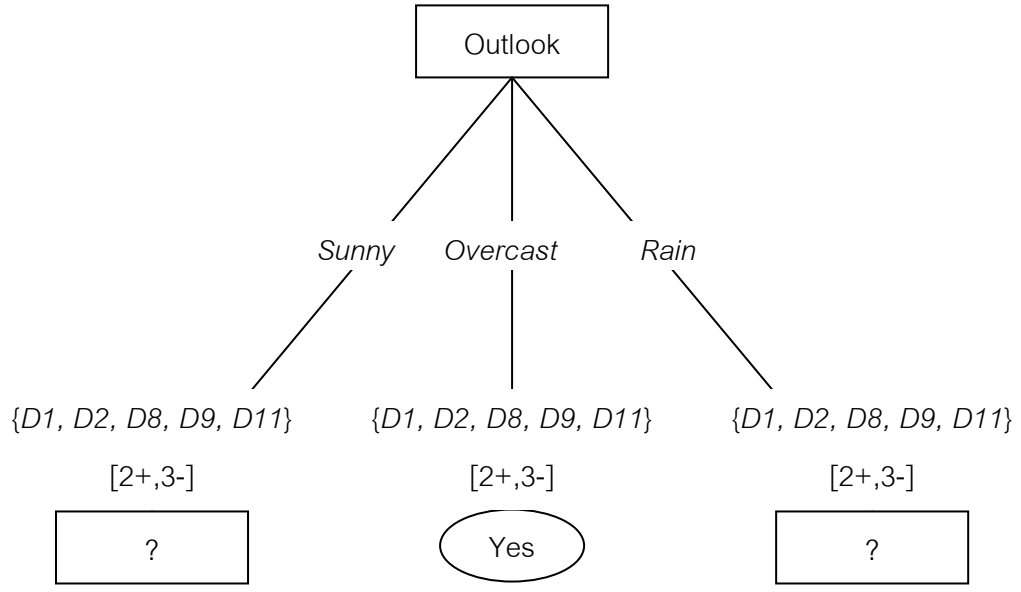
$$Gain(S,Wind) = Entropy(S) - \sum_{v \in \{Weak,Strong\}} \frac{|S_v|}{[S[} Entropy(S_v)$$
$$= Entropy(S) - \left(\frac{8}{14}\right) Entropy(S_{Wind=Weak}) - \left(\frac{6}{14}\right) Entropy(S_{Wind=Strong})$$
$$= Entropy(S) - \left(\frac{8}{14}\right) 0.811 - \left(\frac{6}{14}\right) 1.00$$
$$= 0.940 - \left(\frac{8}{14}\right) 0.811 - \left(\frac{6}{14}\right) 1.00$$
$$= 0.048$$

คำนวณ $Gain(S, A)$ จนครบทุกคุณลักษณะ จากนั้นเลือกค่าคุณลักษณะที่มีค่า $Gain(S, A)$ ที่ดีที่สุด จากตัวอย่างนี้คำนวณค่าการเพิ่มสารสนเทศจากค่าคุณลักษณะทั้งหมดได้ดังนี้

$$Gain(S, Outlook) = 0.246$$
$$Gain(S, Temperature) = 0.029$$
$$Gain(S, Humidity) = 0.151$$
$$Gain(S, Wind) = 0.048$$

ดังนั้นเมื่อเลือกค่าคุณลักษณะ Outlook ซึ่งมีค่าการเพิ่มสารสนเทศที่ดีที่สุดเป็นคุณลักษณะแรกในการจำแนกตัวอย่างดังภาพที่ 2.3

ภาพที่ 2.3
เอาต์พุตต้นไม้ตัดสินใจจากการหาค่าการเพิ่มสารสนเทศขั้นแรก



จากนั้นคำนวณหาค่าการเพิ่มสารสนเทศที่ดีที่สุดไปเรื่อยๆ จนกระทั่งไม่สามารถเพิ่มโหนดลงในต้นไม้ได้ เช่น

$$S_{Sunny} = \{D1, D2, D8, D9, D11\}$$

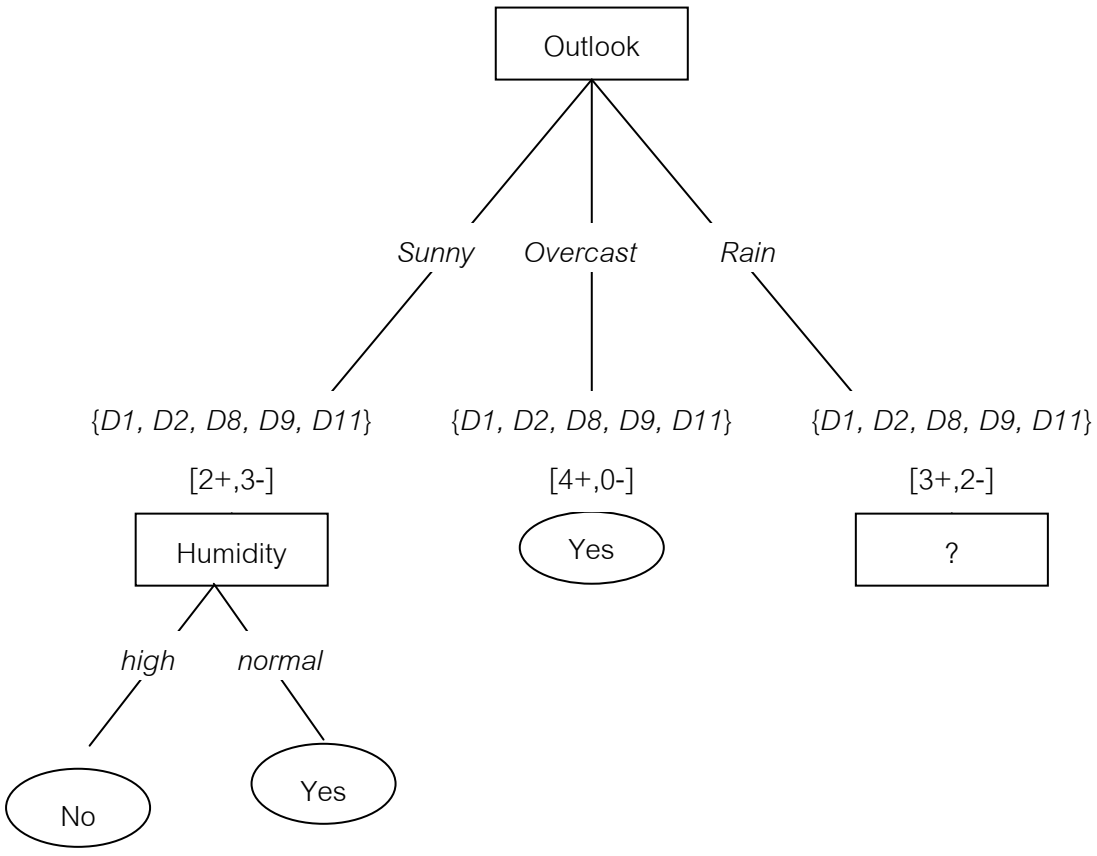
$$Gain(S_{Sunny}, Humidity) = 0.970 - \left[\left(\frac{3}{5}\right)0.0 - \left(\frac{2}{5}\right)0.0\right] = 0.970$$

$$Gain(S_{Sunny}, Temperature) = 0.970 - \left[\left(\frac{2}{5}\right)0.0 - \left(\frac{2}{5}\right)1.0 - \left(\frac{1}{5}\right)0.0\right] = 0.570$$

$$Gain(S_{Sunny}, Wind) = 0.970 - \left[\left(\frac{2}{5}\right)1.0 - \left(\frac{3}{5}\right)0.918\right] = 0.019$$

จะเห็นได้ว่าโหนดต่อมาจาก Outlook ที่มาทางกิ่ง Sunny นั้น เราเลือกได้จากค่าการเพิ่มสารสนเทศที่มีค่าสูงสุด ดังตัวอย่างการคำนวณด้านบนค่าคุณลักษณะ Humidity ซึ่งมีค่าการเพิ่มสารสนเทศที่ดีที่สุดเป็นคุณลักษณะต่อมาในการจำแนกตัวอย่างดังภาพที่ 2.4

ภาพที่ 2.4
เอาต์พุตต้นไม้ตัดสินใจจากการหาค่าการเพิ่มสารสนเทศในขั้นที่ 2



การคำนวณค่าการเพิ่มสารสนเทศที่ดีที่สุดไปเรื่อยๆ จนกระทั่งกลุ่มของข้อมูลที่ได้อยู่ในคลาสเดียวกันทั้งหมดจึงจะหยุดสร้าง เช่นในภาพที่ 2.4 ข้อมูลของกิ่ง Overcast นั้นมีข้อมูลอยู่ในคลาสเดียวกันทั้งหมดจึงหยุดทำการสร้างโหนดต่อไป แต่ในกิ่งของ Sunny และ Rain นั้นข้อมูลยังมีการปนเปื้อนอยู่ดังนั้นจึงต้องทำการสร้างการจำแนกด้วยคุณลักษณะต่อมา จึงต้องทำการสร้างต้นไม้ส่วนย่อย (Subtree) ต่อมา โดยเลือกค่าคุณลักษณะต่อมาเพื่อมาทำหน้าที่เป็นโหนดรากของต้นไม้ส่วนย่อย และทำซ้ำจนกระทั่งค่าการปนเปื้อนของข้อมูลมีค่าเป็นศูนย์ “0” นั่นก็แสดงว่าไม่สามารถเพิ่มโหนดลงในต้นไม้ได้อีก จึงได้ต้นไม้ที่สมบูรณ์ตามภาพที่ 2.1

การเปลี่ยนต้นไม้เป็นกฎ

ระบบปัญญาประดิษฐ์ส่วนใหญ่ใช้การแทนความรู้ในรูปของกฎ ดังนั้นเมื่อเราสร้างต้นไม้ตัดสินใจแล้วเราสามารถเปลี่ยนต้นไม้ให้อยู่ในรูปของกฎเพื่อใช้กับในกรณีที่ระบบของเรา ใช้การแทนความรู้ของกฎเป็นหลัก วิธีการแปลงต้นไม้เป็นกฎ “IF THEN” ทำได้โดยแสดงทุกเส้นทางเริ่มต้นจากโหนดรากไปยังโหนดใบ และทุกครั้งที่พบโหนดทดสอบก็ให้เพิ่มโหนดทดสอบกับค่าของการทดสอบไว้ในส่วนของ IF และเมื่อพบโหนดใบก็ให้ใส่คลาสไว้ในส่วนของ THEN จากต้นไม้ เช่น

ถ้า อากาศ “ร้อน” และ อุณหภูมิ “สูง” และ ความชื้น “สูง” และ ลม “ไม่แรง” เอาต์พุตคือ ไม่เล่น

IF อากาศ “ร้อน” and อุณหภูมิ “สูง” and ความชื้น “สูง” and ลม “ไม่แรง” THEN ไม่เล่น

การทำให้เป็นค่าวิยุต (Discretization)

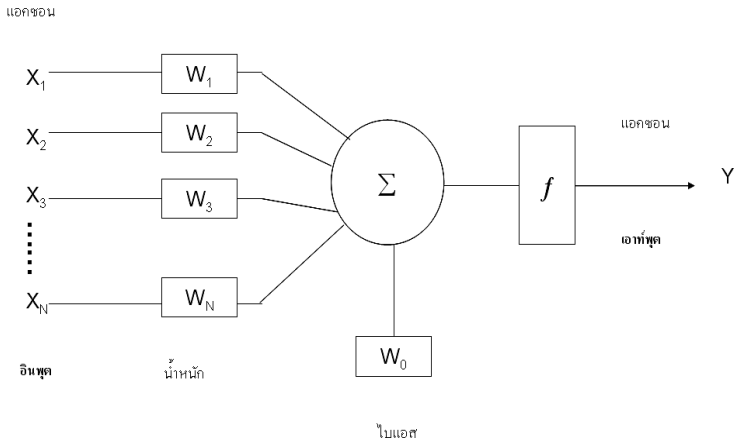
ในทางคณิตศาสตร์การทำให้เป็นค่าวิยุต (Discretization) คือ แบบจำลองของกระบวนการถ่ายโอนข้อมูลอย่างต่อเนื่องด้วยสมการเพื่อแยกส่วนที่คล้ายหรือเหมือนกันออกเป็นส่วนๆ ซึ่งขั้นตอนแรกของกระบวนการนี้ คือ การจัดเตรียมข้อมูลเชิงตัวเลขเพื่อการประมวลผลบนเครื่องคอมพิวเตอร์ นอกเหนือจากวิธีการทำให้เป็นค่าวิยุต แล้วสมการเชิงเส้นของการทำให้เป็นค่าวิยุตสามารถเปลี่ยนรูปแบบของสมการที่ซึ่งมีความแตกต่างอย่างต่อเนื่อง (Continuous differential Equations) มาเป็นสมการที่ไม่ต่อเนื่องได้ (Discrete Difference Equation) เพื่อความเหมาะสมต่อการคำนวณ

2.1.2 **โครงข่ายงานประสาทเทียม** (Artificial Neural Network **หรือ** ANN)

ทฤษฎีโครงข่ายงานประสาทเทียมเป็นความรู้แขนงหนึ่งของปัญญาประดิษฐ์ ซึ่งเลียนแบบการทำงานและคุณสมบัติเซลล์สมองหรือระบบประสาทของมนุษย์ เมื่อหลักการโครงข่างงานประสาทเทียมผนวกกับความสามารถของวิทยาการทางคอมพิวเตอร์ในปัจจุบันทำให้ได้ระบบที่มีศักยภาพในการทำงาน ซึ่งมีคุณลักษณะและคุณสมบัติที่น่าสนใจ เช่น สามรถจำลองปัญหาได้โดยไม่จำเป็นต้องทราบลักษณะรูปแบบการกระจายของข้อมูล ระบบการทำงานมีการใช้ฟังก์ชันทางคณิตศาสตร์อย่างง่ายแทนที่จะใช้กลไกทางกายภาพ และไม่ได้ทำงานตามชุดคำสั่งแต่อย่างเดียวดังเช่นกับโปรแกรมคอมพิวเตอร์ทั่วไป คำตอบหรือเอาต์พุตมีความน่าเชื่อถือด้วยเหตุผลดังกล่าว โครงข่ายงานประสาทเทียมจึงสามารถแก้ปัญหาได้ใกล้เคียงกับเซลล์สมองหรือระบบประสาทของสิ่งมีชีวิตโดยเฉพาะมนุษย์ ระบบการเรียนรู้จากตัวอย่าที่มีจำนวน และความหลากหลาย แหล่งที่มาของตัวอย่างอาจได้จากข้อมูลในอดีต (Historical Record) หรือกระบวนการการจำลอง (Simon,1998)

ระบบการทำงานของโครงข่ายงานประสาทเทียม เป็นการจำลองในรูปแบบฟังก์ชันทางคณิตศาสตร์โดยจำลองการทำงานจากเซลล์ประสาท (Neuron) ในสมองมนุษย์ซึ่งเซลล์ประสาทรับข้อมูลอินพุตจากเซลล์ประสาทอื่น โดยผ่านทางจุดเชื่อมต่อหรือที่เรียกว่า จุดประสานประสาท (Synapse) สัญญาณข้อมูลอินพุตจะได้รับการประมวลผลภายใน จากนั้นสัญญาณข้อมูลเอาต์พุตจากเซลล์ประสาท จะถูกส่งออกมาทางส่วนของแกนประสาทนำออก (Axon) ซึ่งจะส่งไปยังแกนประสาทนำเข้า (Dendrites) ต่อไป ส่วนวิธีการประมวลผลภายในโดยเซลล์ประสาทแต่ละเซลล์มีจุดเชื่อมโยงระหว่างการทำงานใน 2 ลักษณะคือ การกระตุ้น (Excitatory) เป็นการทำให้สัญญาณที่ผ่านมามีความถี่สูงขึ้น และยับยั้ง (Inhibitory) ซึ่งเป็นการทำให้สัญญาณที่ผ่านมามีความถี่ลดลง แบบจำลองของโครงข่ายใยประสาทเทียมมีอัตราการขยายหรือลดลงจะถูกกำหนดโดยค่าถ่วงน้ำหนัก (Weight) รูปแบบการจำลองระบบการทำงานของเซลล์ประสาทเทียมแสดงดังภาพที่ 2.5 (Looney, 1997)

ภาพที่ 2.5
แสดงหลักการทำงานขั้นพื้นฐานของเซลล์ประสาทเทียม



หลักการทำงานพื้นฐานของเซลล์ประสาทเทียมสามารถพิจารณาในรูปแบบของสมการได้ดังนี้

$$h_k = \sum_{i=1}^{n} w_{ik} x_i + b_k = w_{1k} x_1 + w_{2k} x_2 + ... + w_{nk} x_n + b_k$$

โดยที่ x_i เป็นค่าตัวแปรอินพุตของโหนดที่ i

w_{ik} เป็นค่าถ่วงน้ำหนักระหว่างโหนดที่ i และ k

b_k เป็นค่าไบแอส (Bias) ของโหนดที่ k

h_k เป็นค่าผลรวมของตัวแปรอินพุตของโหนดที่ k

สัญญาณ h_k ถูกส่งผ่านค่าฟังก์ชันกระตุ้น (Activation Function) เพื่อคำนวณค่าตัวแปรเอาต์พุตเป้าหมายที่ออกจากเซลล์ประสาทเทียมหรือโหนด y_k ตามสมการ $y_k = f(h_k)$ ตามหลักการทำงานของเซลล์ประสาท นักวิทยาศาสตร์พยามยามสร้างแบบจำลองของเซลล์ซึ่งเรียกว่าเซลล์ประสาทเทียม (Artificial Neural) แต่เซลล์ประสาทในสมองของมนุษย์จะสามารถทำงานหลายงานที่ซับซ้อนพร้อม ๆ กันได้เร็วกว่าเซลล์ประสาทเทียมที่จำลองโดยคอมพิวเตอร์ เนื่องจากโครงสร้างของเซลล์ประสาทมีโครงสร้างที่ขนานกันอย่างมาก (Massively Parallel

Structure) ความสัมพันธ์ระหว่างเซลล์ประสาทกับเซลล์ประสาทเทียม โดยสรุปแสดงได้ดังตาราง 2.2

ตารางที่ 2.2

ความสัมพันธ์ระหว่างเซลล์ประสาทกับเซลล์ประสาทเทียม

เซลล์ประสาท	เซลล์ประสาทเทียม
ตัวเซลล์ (Cell Body)	โหนด (Unit)
แกนประสาทนำเข้า (Dendrites)	ตัวแปรอินพุต (Input)
แกนประสาทนำออก (Axon)	ตัวแปรเอาต์พุต (Output)
จุดประสานประสาท (Synapse)	ค่าถ่วงน้ำหนัก (Weight)
มีเซลล์จำนวนมาก (ประมาณ 10^9 ยูนิต)	มีเซลล์จำนวนน้อย (ประมาณ 100 ยูนิต)

โครงข่ายประสาทเทียมแบ่งตามลักษณะการเรียนรู้หรือการสอนได้เป็น 2 กลุ่มใหญ่ ๆ คือ โครงข่ายประสาทเทียมแบบมีผู้สอนหรือผู้ดูแล และโครงข่ายประสาทเทียมแบบไม่มีผู้สอน หรือแบบเรียนรู้ด้วยตนเอง (Simon Haykin, 1998)

การเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithm) การเรียนรู้ด้วยวิธีนี้จะกำหนดข้อมูลตัวอย่างของการเรียนรู้ให้กับโครงข่าย ข้อมูลตัวอย่างนี้ประกอบด้วยตัวอย่างข้อมูลอินพุต และตัวอย่างข้อมูลเอาต์พุตที่ต้องการ เมื่อป้อนข้อมูลตัวอย่างให้โครงข่าย โครงข่ายจะคำนวณค่าผิดพลาด เพื่อนำมาใช้ปรับค่าถ่วงน้ำหนัก จนกว่าจะได้เอาต์พุตตามต้องการจึงจะหยุดการเรียนรู้ (Ethem Alpaydin, 2005)

การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning Algorithm หรือ Self-Oranizing) วิธีนี้จะป้อนข้อมูลตัวอย่างให้โครงข่าย ภายในโครงข่ายจะมีโหนดเอาต์พุตอยู่หลายโหนด แต่ละโหนดแทน กลุ่มของข้อมูลตัวอย่างที่มีคุณสมบัติเหมือนกัน เมื่อป้อนข้อมูลตัวอย่างเข้าสู่โครงข่าย โครงข่ายจะคำนวณความสัมพันธ์ที่มีภายในกลุ่มของตัวอย่าง โดยอาศัยค่าถ่วงน้ำหนักเป็นตัวแยกแยะชนิดของข้อมูลตัวอย่างไปเก็บไว้ในโหนดเอาต์พุตของโครงข่าย การเรียนรู้ด้วยวิธีนี้จะไม่สามารถระบุได้ว่าโหนดเอาต์พุตใดเป็นของกลุ่มข้อมูลไหน ผู้ใช้จะต้องเป็นผู้กำหนด (Hinton, Geoffrey , Sejnowski, Terrence , 1999)

ทิศทางการป้อนภายในของโครงข่างใยประสาทเทียม

ทิศทางการป้อนภายในของโครงข่ายประสาทเทียมมี 2 ลักษณะ คือ

1. **เน็ตเวิร์กแบบป้อนไปข้างหน้า** (Feed forward Network)

ทิศทางการเคลื่อนที่ของข้อมูลจะเป็นลักษณะเคลื่อนที่ไปข้างหน้าจากชั้นอินพุต (Input Laryer) ผ่านชั้นแฝง (Hidden Layer) ไปสู่ชั้นเอาต์พุต (Output Layer) โดยการเชื่อมต่อกันเป็นลักษณะจากชั้นหนึ่งไปยังชั้นถัดไป และไม่มีการเชื่อมต่อกันภายในชั้นเดียวกัน ตัวอย่างของโครงข่ายชนิดนี้คือ อะดาไลน์ (Adaline)

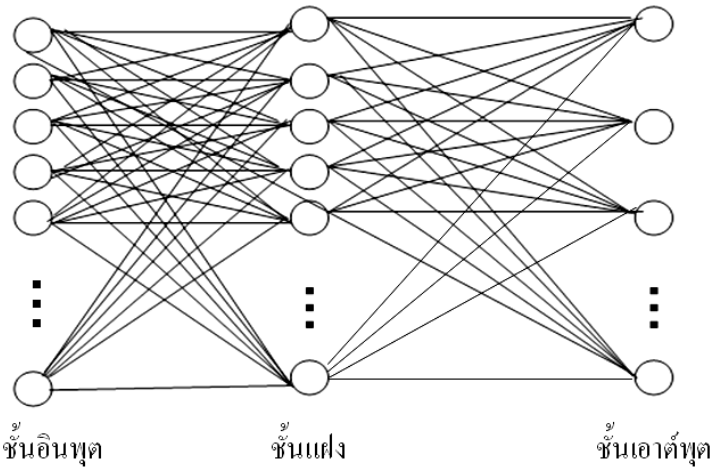
2. **เน็ตเวิร์กแบบรีเคอร์เรนต์** (Recurrent Network)

ทิศทางการเคลื่อนที่ของข้อมูลจะแบ่งเป็น 2 ทิศทางคือทิศทางการเคลื่อนที่ไปข้างหน้า (Feed forward) จากชั้นอินพุตผ่านชั้นแฝงไปสู่ชั้นเอาต์พุต และทิศทางการเคลื่อนที่ย้อนกลับ (Feedback) จากชั้นเอาต์พุตไปยังชั้นอินพุต หรือจากชั้นเอาต์พุตผ่านชั้นแฝงไปยังชั้นอินพุต ตัวอย่างของโครงข่ายชนิดนี้คือ โครงข่ายจอร์แดน (Jordan Network) และโครงข่าย ฮอปฟิลด์ (Hopfield Network)

เพอร์เซปตรอนแบบหลายชั้น (Multi – Layer Perceptron)

เพอร์เซปตรอนแบบหลายชั้นเป็นโครงข่ายใยประสาทเทียมแบบหนึ่งที่สำคัญ ซึ่งประกอบด้วยส่วนรับข้อมูล (Sensory Units or Source Nodes) โดยทั่วไปเรียกว่าชั้นอินพุต ชั้นที่มีการคำนวณภายใน (Computation Node) ที่เรียกว่าชั้นแฝง ซึ่งอาจจะมีหนึ่งชั้นหรือหลายชั้นก็ได้ และส่วนแสดงผลข้อมูลที่คำนวณได้เรียกว่าชั้นเอาต์พุต โดยมีรูปแบบเป็นเน็ตเวิร์กแบบป้อนไปข้างหน้า ดังแสดงในภาพ 2.6 (Rosenblatt,pp. 386 – 408)

ภาพที่ 2.6
เพอร์เซปตรอนแบบหลายชั้น



สถาปัตยกรรมของโครงข่ายประสาทเทียมแบบมีผู้สอนชนิดแพร่กระจายย้อนกลับ

โครงข่ายประสานเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับจะประกอบด้วยชั้นต่าง ๆ คือ ชั้นอินพุต กับชั้นเอาต์พุต อย่างละ 1 ชั้น และชั้นแฝง ซึ่งจะอยู่ระหว่างชั้นอินพุตกับชั้นเอาต์พุต สามารถมีกี่ชั้นก็ได้

การเชื่อมต่อระหว่างชั้นต่าง ๆ ทุก ๆ โหนดในชั้นอินพุตจะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นแฝงชั้นแรก และทุก ๆ โหนดในชั้นแฝงชั้นแรกจะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นถัดไปจนในที่สุดทุก ๆ โหนดในชั้นแฝงชั้นสุดท้าย จะส่งสัญญาณไปยังทุก ๆ โหนดในชั้นเอาต์พุต

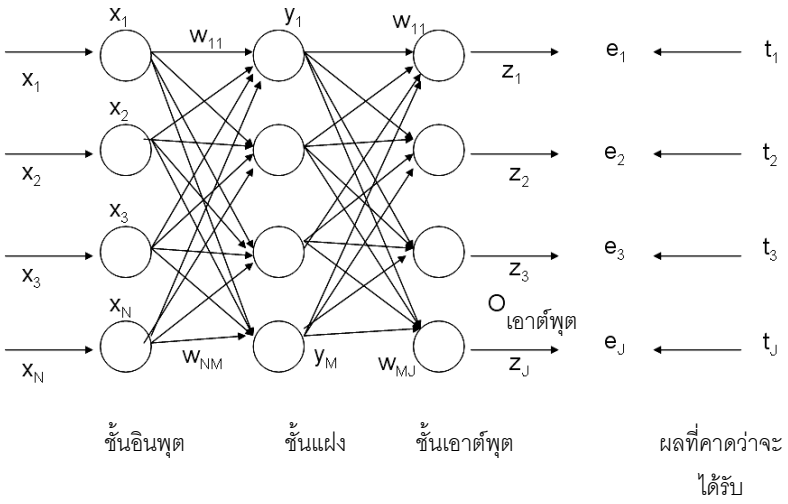
การทำงานของชั้นต่าง ๆ ชั้นอินพุตจะไม่มีการประมวลผลทำหน้าที่รับสัญญาณเข้าแล้วกระจายออกยังแต่ละโหนด ในชั้นถัดไปเท่านั้น ส่วนชั้นแฝงและชั้นเอาต์พุต นั้นมีการประมวลผล ภาพที่ 2.7 แสดงสถาปัตยกรรมของโครงข่ายประสาทเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับ ซึ่งจะประกอบไปด้วย ชั้นอินพุต ชั้นแฝง และชั้นเอาต์พุต จำนวนชั้นละ 1 ชั้นและแต่ละโหนดจะถูกเชื่อมต่อกันเป็นโครงข่ายด้วยเส้นเชื่อมซึ่งมีค่าน้ำหนักติดอยู่

วิธีการเรียนรู้แบบแพร่กระจายย้อนกลับ

อธิบายขั้นตอนวิธีการเรียนรู้ของโครงข่ายประสาทเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับแบบพื้นฐานที่มีชั้นแฝง 1 ชั้น ภาพของสถาปัตยกรรมของโครงข่ายประสาทเทียมที่จะใช้อธิบายประกอบพร้อมด้วยตัวแปรต่าง ๆ แสดงดังภาพที่ 2.7

ภาพที่ 2.7

การทำงานของโครงข่ายประสาทเทียมแบบแพร่กระจายย้อนกลับ



ตารางที่ 2.3 ใช้อธิบายความหมายของตัวแปรต่าง ๆ ที่จะใช้ประกอบการอธิบายขั้นตอนการทำงานของโครงข่ายประสาทเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับ ขั้นพื้นฐานที่มีชั้นข้อมูลอินพุต ชั้นแฝง และชั้นเอาต์พุตอย่างละ1 ชั้น

ตารางที่ 2.3

ตัวแปรต่าง ๆ พร้อมความหมาย

ตัวแปร	ความหมาย
x_n	ข้อมูลอินพุตโหนดที่ n มีทั้งหมด N โหนด
S_m	เอาต์พุตของชั้นแฝง ก่อนทำการปรับค่า (Activation) เป็น y_m
y_m	เอาต์พุตของชั้นแฝง หลังทำการปรับค่าของโหนดที่ m มีทั้งหมด M โหนด
V_j	เอาต์พุตของชั้นเอาต์พุต ก่อนทำการปรับค่าเป็น Z_j
z_j	ค่าเอาต์พุตที่ได้ทำการปรับค่าแล้วของชั้นเอาต์พุตโหนดที่ j มีทั้งหมด J โหนด
t_j	ค่าเอาต์พุตที่ต้องการที่ชั้นเอาต์พุตโหนดที่ j มีทั้งหมด J โหนด
w_{nm}	น้ำหนักของเส้นเชื่อมระหว่างชั้นอินพุต กับชั้นแฝง
w_{mj}	น้ำหนักของเส้นเชื่อมระหว่างชั้นแฝง กับชั้นเอาต์พุต
η	อัตราการเรียนรู้มีค่าอยู่ระหว่าง 0 ถึง 1
R	จำนวนรอบที่ทำการเรียนรู้ มี R เป็นจำนวนรอบที่กำหนด
Q	จำนวนชุดของข้อมูลตัวอย่าง มี Q เป็นตัวกำหนด
$e^{(q)}$	ค่าผิดพลาดของข้อมูลตัวอย่าง
E	ค่าผิดพลาดรวมเฉลี่ยของข้อมูลตัวอย่าง

ขั้นตอนการเรียนรู้ของโครงข่ายประสาทเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับ อธิบายเป็นขั้นตอนต่าง ๆ 8 ขั้นตอนดังนี้

1. กำหนดจำนวนโหนดรับข้อมูล (N), จำนวนโหนดเอาต์พุต (J), จำนวนโหนดของชั้นแฝง (M), ข้อมูลตัวอย่างอินพุตและข้อมูลเอาต์พุต จากนั้นรับจำนวนรอบสูงสุดที่จะทำการเรียนรู้ (R) และค่าผิดพลาดที่ยอมรับได้

จำนวนโหนดของแต่ละชั้นจะมีผลกับการเรียนรู้ของโครงข่ายประสาทเทียม กล่าวคือหากจำนวนโหนดมากเกินไป จะทำให้โครงข่ายงานประสาทเทียมเรียนรู้ได้ช้า แต่ถ้าน้อยเกินไปแล้วข้อมูลที่จะสอนให้โครงข่ายมีเป็นจำนวนมากก็จะส่งผลให้โครงข่ายมีเป็นจำนวนมากก็จะส่งผลให้โครงข่ายประสาทเทียมไม่สามารถเรียนรู้ข้อมูลชุดนั้นได้

ค่าผิดพลาดที่ยอมรับได้ มีผลกับคำตอบของโครงข่ายงานประสาทเทียมและการเรียนรู้เช่นกัน หากค่าผิดพลาดที่ยอมรับได้ต่ำ หรือเข้าใกล้ค่าศูนย์ “0” จะทำให้โครงข่าย

เรียนรู้ได้ช้า แต่ความแม่นยำในการรู้จำก็จะสูงขึ้น ในทางตรงกันข้ามถ้าค่าผิดพลาดที่ยอมรับได้สูง หรือออกห่างค่าศูนย์ “0” โครงข่ายประสาทเทียมจะสามารถรู้จำได้เร็วกว่า แต่ผลที่ได้จะมีความแม่นยำต่ำเช่นกัน

2. กำหนดอัตราการเรียนรู้ ค่าอัตราการเรียนรู้ (η) เป็นพารามิเตอร์ที่สำคัญตัวหนึ่งของโครงข่ายประสาทเทียม โดยทั่วไปจะอยู่ในช่วง [0,1] อัตราการเรียนรู้จะมีความสัมพันธ์กันความเร็วในการเรียนรู้ กล่าวคือ ถ้าอัตราการเรียนรู้เข้าใกล้ 1 จะมีผลให้การเรียนรู้เป็นไปอย่างรวดเร็วในช่วงแรก แต่เมื่อเข้าสู่ช่วงปรับคำตอบให้เป็นไปตามผลที่ต้องการจะทำได้ไม่ดีเท่าการตั้งค่าอัตราการเรียนรู้เข้าใกล้ค่าศูนย์ “0”

3. สุ่มน้ำหนักให้เส้นเชื่อม ในการเรียนรู้ของโครงข่ายประสาทเทียมในครั้งแรก จะต้องสุ่มน้ำหนักให้เส้นเชื่อมต่อแต่ละเส้น เพื่อให้โครงข่ายสามารถคำนวณผลและแพร่เอาต์พุตสู่โหนดชั้นถัดไป การสุ่มน้ำหนักโดยทั่วไปจะนิยมอยู่ 2 ช่วง คือช่วง [-0.5, 0.5] และ [-1, 1]

4. รับข้อมูลตัวอย่างที่โหนดอินพุตให้รับข้อมูลที่ต้องการเรียนรู้ เพื่อใช้ในการคำนวณหาค่าเอาต์พุตของโครงข่ายประสาทเทียมในการรู้จำ โดยครั้งแรกจะรับข้อมูลอินพุตชุดแรกนำมาประมวลผลเพื่อส่งให้ชั้นถัดไป และจะคอยรับข้อมูลอินพุตชุดต่าง ๆ ที่ต้องการเรียนรู้วนไปเรื่อย ๆ จนกว่าจะจบการเรียนรู้

5. คำนวณและปรับเอาต์พุตที่ชั้นแฝง โดยนำข้อมูลอินพุตของชุด ที่จะคำนวณหาค่าเอาต์พุตของชั้นแฝงมาคำนวณ แล้วปรับเอาต์พุตของชั้นแฝงก่อนทำการปรับค่า

$$s_m = \sum_{nn=1}^{N} x_n * w_{nm} \qquad (1)$$

สมการที่ 2 ใช้เอาต์พุตของชั้นแฝงหลังทำการปรับค่า มีสมการดังนี้

$$y_m = f(s_m) \qquad (2)$$

สมการปรับค่าเอาต์พุต $f(x)$ จะกล่าวในหัวข้อถัดไป

6. คำนวณและปรับเอาต์พุตที่ชั้นเอาต์พุต หลังจากนั้นปรับค่าเอาต์พุตให้อยู่ในช่วง [0, 1] สำหรับแต่ละโหนดของชั้นเอาต์พุต ค่าผลลัพทธ์ของชั้นเอาต์พุตก่อนการปรับค่าให้อยู่ระหว่าง [0, 1] แสดงสมการที่ 3

$$v_j = \sum_{m=1}^{M} y_m * w_{mj} \qquad (3)$$

สมการที่ 4 ใช้ปรับค่าเอาต์พุตชั้นเอาต์พุต มีสมการดังนี้

$$z_j = f(v_j) \qquad (4)$$

7. คำนวณค่าผิดพลาดและปรับน้ำหนัก นำเอาต์พุตที่ได้กับเอาต์พุตที่ได้กำหนดไว้มาคำนวณหาความผิดพลาดของเอาต์พุต ถ้าค่าผิดพลาดของเอาต์พุตน้อยกว่าค่าผิดพลาดที่ยอมรับได้ทำการรับข้อมูลชุดต่อไป ถ้าไม่ใช่ปรับน้ำหนักของเส้นเชื่อม แล้วกลับไปทำข้อ 4 เพื่อรับข้อมูลสำหรับการเรียนรู้ในชุดถัดไป แต่ถ้าเป็นข้อมูลชุดสุดท้ายทำข้อ 8 สมการคำนวณค่าความผิดพลาดในแต่ละชุดของข้อมูลตัวอย่าง สมการที่ 5 ใช้คำนวณค่าผิดพลาดของแต่ละอย่าง

$$e^{(q)} = \frac{1}{2}\sum_{j=1}^{J}\left(t_{j-}z_j\right)^2 \qquad (5)$$

การปรับน้ำหนักของชั้นแฝงกับชั้นเอาต์พุต และชั้นอินพุตชั้นเอาต์พุต แสดงดังสมการที่ 6 และ 7

$$w_{mj}^{(r+1)} = w_{mj}^{(r)} + \eta\left\{\left(t_j^{(q)} - z_j^{(q)}\right) * \left[z_j^{(q)}\left(1 - z_j^{(q)}\right)\right] * y_m^{(q)}\right\} \qquad (6)$$

$$w_{nm}^{(r+1)} = w_{nm}^{(r)} + \eta\left\{\sum_{j=1}^{J}\left(t_j^{(q)} - z_j^{(q)}\right)\left[z_j^{(q)}\left(1 - z_j^{(q)}\right)\right]w_{mj}^{(r)}\right\} * \left[y_m^{(q)}\left(1 - y_m^{(q)}\right)\right]\left[x_n^{(q)}\right] \qquad (7)$$

8. คำนวณค่าผิดพลาดรวมเฉลี่ย โดยนำค่าผิดพลาดของชุดข้อมูลแต่ละชุดมารวมกันแล้ว หาค่าเฉลี่ยเพื่อใช้ในการตรวจสอบว่าเอาต์พุตของทุก ๆ ข้อมูลในแต่ละรอบนั้นมีค่าน้อยกว่าค่าผิดพลาดที่ยอมรับได้ในทุก ๆ ข้อมูลหรือไม่ถ้าใช่แสดงว่าโครงข่ายประสาทเทียมสามารถเรียนรู้ข้อมูลชุดที่สอนจนได้เอาต์พุตที่น่าพอใจแล้ว ให้จบการเรียนรู้ ถ้าไม่ใช่กลับไปทำข้อ 4 เพื่อรับข้อมูลชุดแรกใหม่ สมการที่ 8 คำนวณหาค่าผิดพลาดรวมเฉลี่ย

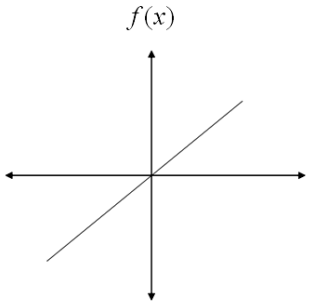
$$E = \frac{1}{Q}\sum_{q=1}^{Q} e^{(q)} \qquad (8)$$

สมการปรับค่าเอาต์พุตด้วยค่าฟังก์ชันกระตุ้น (Activation Functions) ของชั้นแฝง และชั้นเอาต์พุตที่นิยมใช้ในปัจจุบัน มีหลากหลายชนิดขึ้นอยู่กับลักษณะการนำไปใช้ตัวอย่างข้อมูล และวิธีการว่าจะเลือกวิธีให้สมการปรับค่าเอาต์พุตที่นิยมใช้มีอยู่ 3 แบบดังนี้

1. สมการเชิงเส้น (Identity or Linear Function) สมการและกราฟคุณสมบัติแสดงในภาพที่ 2.8

ภาพที่ 2.8

สมการและกราฟแสดงการปรับเอาต์พุตเชิงเส้น



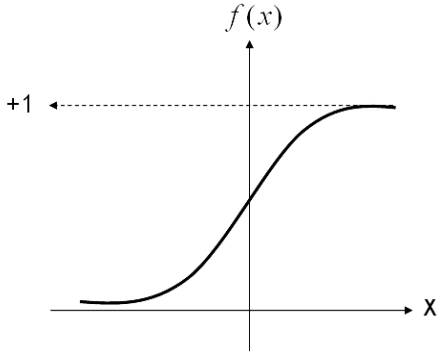
for all x

$$f(x) = x$$
$$f'(x) = 1$$

2 สมการที่ให้ผลระหว่าง 0 ถึง 1 (Binary Sigmoid Function หรือ Logistic Function) สมการที่จะให้เอาต์พุตอยู่ระหว่าง [0,1] สมการและกราฟคุณสมบัติแสดงดังภาพที่ 2.9

ภาพที่ 2.9

สมการและกราฟแสดงการปรับเอาต์พุตอยู่ระหว่าง 0 ถึง 1

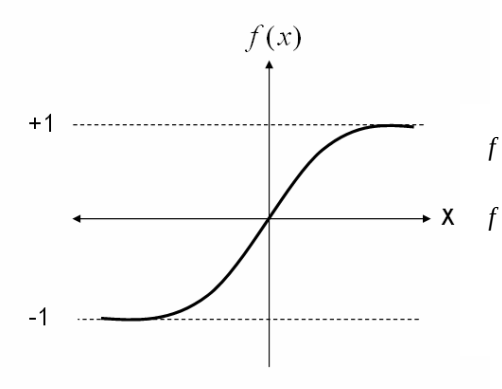


$$f(x) = \frac{1}{1+e^{-x}}$$
$$f'(x) = f(x)[1 - f(x)]$$

3 สมการที่ให้ผลระหว่าง -1 ถึง 1 (Bipolar Sigmoid Function) สมการที่จะให้เอาต์พุตอยู่ระหว่าง [-1, 1] สมการและกราฟคุณสมบัติแสดงดังภาพที่ 2.10

ภาพที่ 2.10

สมการและกราฟแสดงการปรับเอาต์พุตอยู่ระหว่าง -1 ถึง 1



$$f(x) = \frac{2}{1+e^{-x}} - 1$$

$$f'(x) = \frac{1}{2}[1+f(x)][1-f(x)]$$

ขั้นตอนการรู้จำของโครงข่ายงานประสาทเทียมแบบมีผู้สอนชนิดแพร่ย้อนกลับ อธิบายเป็นหัวข้อย่อยต่าง ๆ ได้ 3 ข้อดังนี้

1. อ่านค่าตัวแปรต่าง ๆ ของโครงข่ายงานประสาทเทียม เพื่อสร้างโครงข่าย ดังนี้ จำนวนโหนดอินพุต(N) จำนวนโหนดเอาต์พุต (J) จำนวนโหนดของชั้นแฝง (M) ข้อมูลอินพุตที่ต้องการค่าน้ำหนักของเส้นเชื่อม

2. คำนวณและปรับเอาต์พุตที่ชั้นแฝง โดยนำข้อมูลขาข้าวของชุดที่จะคำนวณหาค่าเอาต์พุตของชั้นแฝงมาคำนวณ แล้วปรับเอาต์พุตของชั้นแฝง ให้อยู่ในช่วง [0, 1] สำหรับแต่ละโหนดของชั้นแฝง

3. คำนวณและปรับเอาต์พุตที่ชั้นเอาต์พุต แล้วปรับค่าเอาต์พุตให้อยู่ในช่วง [0, 1] สำหรับแต่ละโหนดของชั้นเอาต์พุต ก็จะได้เอาต์พุตตามที่ต้องการ (Alpaydin, 2005)

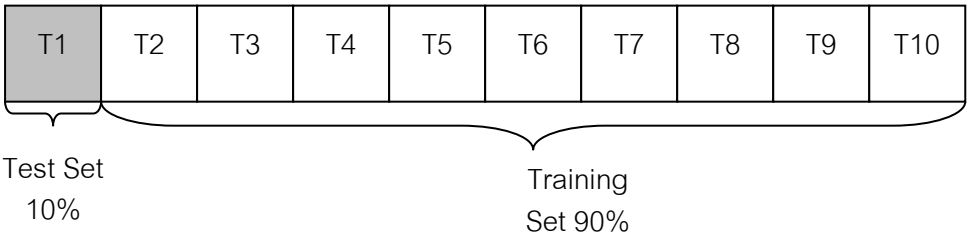
2.2 การประเมินตัวแบบ

2.2.1 วิธีการประเมินตัวแบบด้วยวิธีการไขว้ข้าม 10 กลุ่ม (10 - Fold Cross Validation)

ในกลไกการเรียนรู้นิยมที่จะใช้เทคนิคที่มีพื้นฐานมาจากเทคนิคเชิงสถิติ (Statistical Techniques) โดยเรียกเทคนิคนี้ว่า “วิธีการประเมินตัวแบบด้วยวิธีการไขว้ข้าม 10 กลุ่ม” หรือ 10 Fold Cross – Validation ซึ่งมีแนวคิดในการแบ่งข้อมูลที่มีอยู่ออกเป็นส่วน ๆ เรียกว่า “โฟลด์” (Fold) ข้อมูลถูกแบ่งออกเป็น 10 ส่วน ในครั้งแรกจะเก็บส่วนแรกไว้เพื่อเป็นชุดข้อมูลทดสอบ และส่วนที่ 2 ถึง 10 เป็นส่วนที่ใช้สำหรับเป็นชุดข้อมูลสำหรับเรียนรู้ ทำเช่นนี้ไปเรื่อย ๆ จนทุก ๆ ส่วน

ถูกใช้สำหรับทดสอบจนครบ จากนั้นจึงหาค่าเฉลี่ยของความผิดพลาดของแต่ละตัวแบบที่ได้ ซึ่งสามารถทำให้มอบเห็นภาพรวมของความผิดพลาดที่เกิดขึ้น หรือประสิทธิภาพของตัวแบบที่ได้ ดังแสดงในภาพที่ 2.11 แสดงให้เห็นถึงการจัดเตรียมข้อมูลทั้งหมดเพื่อใช้ในการเรียนรู้โดยวิธีการที่ใช้ในการจำแนกข้อมูลนั้น ข้อมูลได้ถูกแบ่งออกเป็นชุด2 ชุดคือ ชุดข้อมูล T1 สำหรับเป็นชุดข้อมูลทดสอบ (Test Set) มีจำนวนร้อยละสิบของจำนวนข้อมูลทั้งหมด และ T2 – T10 ชุดข้อมูลสำหรับเรียนรู้ (Training Set) เป็นจำนวนร้อยละเก้าสิบของข้อมูลทั้งหมด

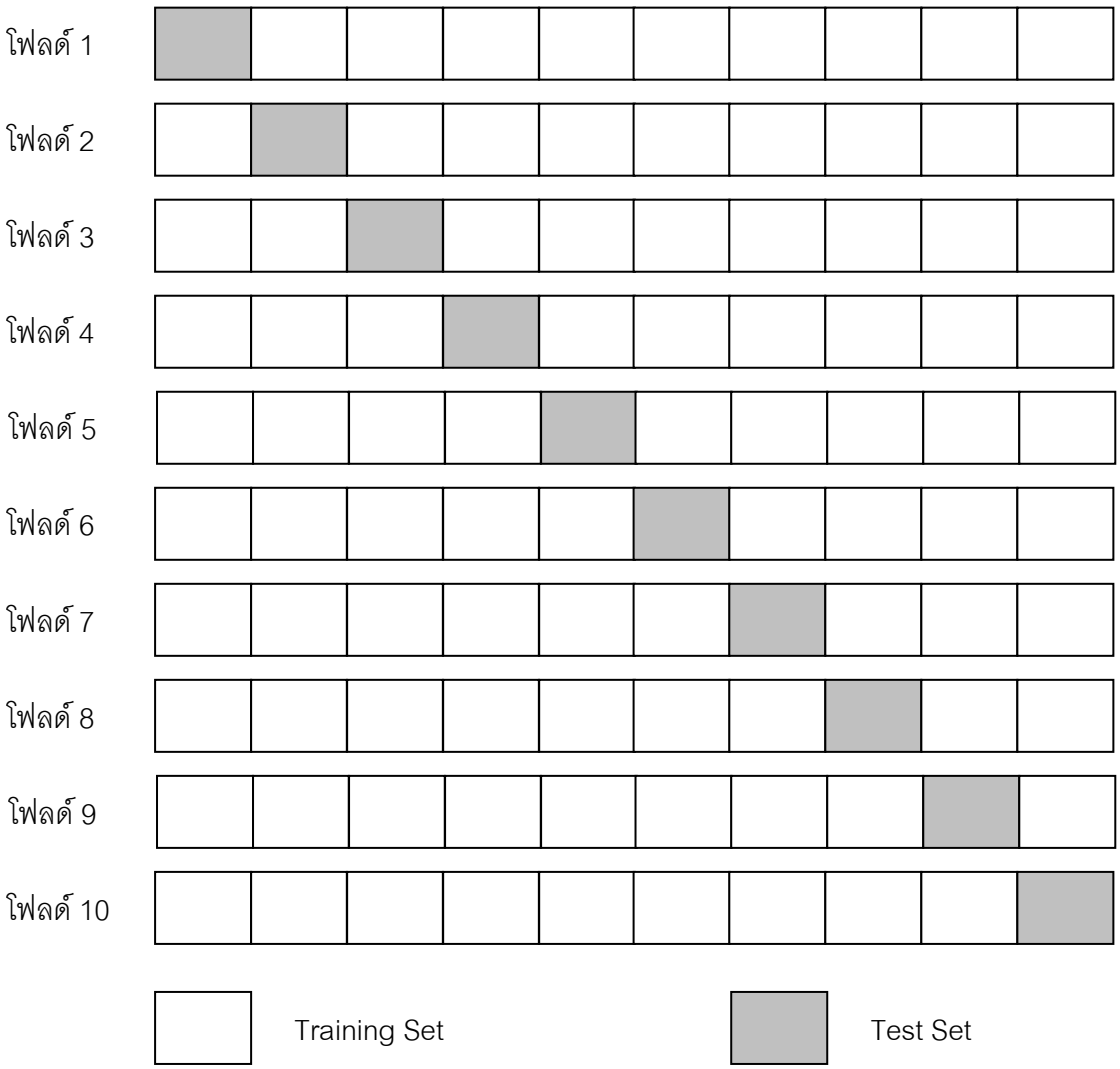
ภาพที่ 2.11
การแบ่งข้อมูลสำหรับการเรียนรู้และทดสอบ



งานวิจัยฉบับนี้ได้นำวิธีทดสอบความถูกต้องแบบไขว้ข้าม 10 กลุ่ม (10-Fold Cross Validation) คือ แบ่งข้อมูลที่ใช้ในการทดสอบเป็นชุดข้อมูลย่อยจำนวน 10 ชุดข้อมูลย่อย ทำเช่นนี้ไปเรื่อย ๆ จนครบ 10 ครั้ง ซึ่งจะได้ตัวแบบการเรียนรู้ทั้งสิ้น 10 ตัวแบบ และข้อมูลทุก ๆ ส่วนถูกใช้สำหรับการทดสอบ ซึ่งแสดงให้เห็นได้ดังภาพที่ 2.12 เมื่อทำการทดสอบตัวแบบในแต่ละครั้ง จะได้ค่าความผิดพลาดของแต่ละตัวแบบ เมื่อทดสอบจนครบทั้ง 10 ครั้ง ก็จะนำค่าความผิดพลาดทั้งหมดมาเฉลี่ย เพื่อให้ได้ภาพรวมของค่าความผิดพลาดทั้งหมด (Kohavi, 1995, pp. 1137-1145)

ภาพที่ 2.12

การแบ่งข้อมูลจำนวน 10 ข้อมูลย่อย

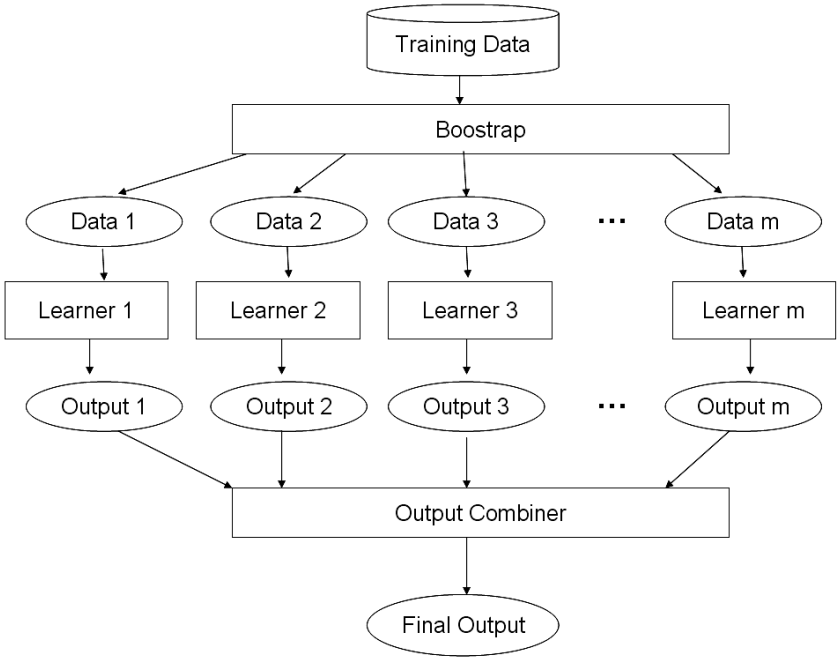


2.2.2 **วิธีทดสอบความถูกต้องโดยการใช้การรวมตัวจำแนก** (Ensemble Classifier)

เป็นวิธีการรวมคำตอบซึ่งคำตอบที่ได้จากการรวมคำตอบของตัวจำแนกหลาย ๆ ตัว เป็นคำตอบเดียว และการรวมตัวจำแนกที่เป็นที่รู้จักกันอย่างแพร่หลายได้แก่ วิธีแบ็กกิง (Breiman, 1996) และวิธีบูสต์ โดยภาพที่ 2.13 แสดงวิธีการรวมตัวจำแนก โดยนำเอาต์พุต์ที่ได้จากตัวจำแนกแต่ละตัวรวมเป็นเอาต์พุตของคำตอบเดียว

ภาพที่ 2.13

การรวมตัวจำแนก



วิธีการบูตสแตรป (Bauer and Kohavi, 1999, pp 105 – 139) เป็นการเตรียมชุดข้อมูลของตัวอย่างการเรียนรู้สำหรับตัวจำแนก โดยการสุ่มตัวอย่างข้อมูลแต่ละชุดข้อมูลเพื่อนำชุดข้อมูลที่ได้นั้นมาใช้เป็นข้อมูลอินพุตของกระบวนการเรียนรู้ของตัวจำแนก โดยที่จำนวนชุดข้อมูลที่นำมาใช้มีจำนวนเท่ากับจำนวนชุดข้อมูลในชุดข้อมูลเริ่มต้น

ตัวจำแนกแบบแบ็กกิง (Bagging)

วิธีการแบ็กกิง (**B**ootstrap **agg**regat**ing**) (Brieman, 1996a, pp 123-140) ใช้วิธีบูตสแตรปในการเตรียมชุดตัวอย่างเรียนรู้ วิธีการแบ็กกิงทำการสร้างการจำแนกตามชุดข้อมูลการเรียนรู้ที่ได้รับมาจากวิธีการของบูตสแตรป การสร้างการจำแนกสุดท้ายสร้างจากการนำคลาสที่ถูกทำนายบ่อยที่สุดมาใช้สร้างการจำแนกสุดท้ายหรือคำตอบสุดท้าย โดยการสร้างการจำแนกสุดท้ายนี้วิธีแบ็กกิงให้ค่าน้ำหนักในแต่ละตัวอย่าง และแต่ละกฎเท่ากันทั้งหมด ซึ่งโดยวิธีการเตรียมเตรียมชุดทดสอบ ด้วยวิธีบูตสแตรปสำหรับต้นไม้แต่ละต้น ด้วยการสุ่มตัวอย่าง *m* ตัวอย่างจากชุดตัวอย่างเรียนรู้ตั้งต้นขนาด *m* ดังนั้นชุดตัวอย่างเรียนรู้ใหม่ที่ได้อาจมีตัวอย่างจากชุดตัวอย่างเรียนรู้ตั้งต้นจำนวนมากตัวอย่างในขณะที่บางตัวอย่างอาจไม่ปรากฏเลย

ตารางที่ 2.4
ชุดข้อมูลที่ได้ของวิธีแบ็กกิง

ตัวอย่างชุดข้อมูลเริ่มต้น	1	2	3	4	5	6	7	8	9	10
ชุดข้อมูลแบ็กกิง รอบที่ 1	7	8	10	8	2	5	10	10	5	9
ชุดข้อมูลแบ็กกิง รอบที่ 2	1	4	9	1	2	3	2	7	3	2
ชุดข้อมูลแบ็กกิง รอบที่ 3	1	8	5	10	5	4	9	6	3	7

จากตัวอย่างที่แสดงในตารางที่ 2.4 ชุดข้อมูลเริ่มต้นประกอบด้วยหมายเลข 1 ถึงหมายเลข 10 เมื่อเตรียมชุดเรียนรู้ด้วยวิธีบูตสแตรป ผลปรากฏว่า

ชุดเรียนรู้ที่ 1 มีตัวอย่างหมายเลข 10 ซ้ำกัน 3 ครั้ง ในขณะที่ไม่มีตัวอย่างหมายเลข 1, 3, 4 และ 6

ชุดเรียนรู้ที่ 2 มีตัวอย่างหมายเลข 2 ซ้ำกัน 3 ครั้ง และหมายเลข 3 ซ้ำกัน 2 ครั้ง ในขณะที่ไม่มีตัวอย่างหมายเลข 5, 6, 8 และ 10

ชุดเรียนรู้ที่ 3 มีตัวอย่างหมายเลข 5 ซ้ำกัน 2 ครั้ง ในขณะที่ไม่มีตัวอย่างหมายเลข 2

ตัวจำแนกแบบบูสต์ (Boosting Classifier)

การรวมของการจำแนกด้วยวิธีบูสต์เน้นที่การสร้างชุดต้นไม้ตัดสินใจ โดยชุดเรียนรู้ที่ใช้ในแต่ละต้นไม้ตัดสินใจถูกเลือกอิงกับประสิทธิภาพของต้นไม้ตัดสินใจก่อนหน้าคือ ตัวอย่างที่คาดเดาผิดจากการจำแนกก่อนหน้าจะถูกเลือกบ่อยครั้งมากกว่าตัวอย่างที่ คาดเดาถูกต้อง ดังนั้น หลักการสำคัญของวิธีบูสต์คือ พยายามสร้างต้นไม้ตัดสินใจให้ได้ค่าคาดเดาที่ดีกว่าประสิทธิภาพการรวมของ ต้นไม้ตัดสินใจปัจจุบัน

ตารางที่ 2.5

ชุดข้อมูลที่ได้ของวิธีบูสต์

ตัวอย่างชุดข้อมูลเริ่มต้น	1	2	3	4	5	6	7	8	9	10
ตัวอย่างบูสต์รอบที่ 1	7	3	2	8	7	9	4	10	6	3
ตัวอย่างบูสต์รอบที่ 2	5	4	9	4	2	5	1	7	4	2
ตัวอย่างบูสต์รอบที่ 3	4	4	8	10	4	5	4	6	3	4

จากตารางที่ 2.5 จะเห็นว่าชุดข้อมูลเริ่มต้นมีหมายเลข 1 ถึง หมายเลข 10 เมื่อสุ่มชุดข้อมูลพบว่าชุดเรียนรู้ที่ 4 มีตัวอย่างหมายเลข 4 ซ้ำกัน 5 ครั้ง แสดงว่าตัวอย่างหมายเลข 4 เป็นตัวอย่างที่มีการคาดเดาผิดจากการจำแนกก่อนหน้าจะเห็นได้ว่าวิธีบูสต์ไม่สามารถใช้สร้างต้นไม้ตัดสินใจแต่ละ ต้นพร้อมๆ กันได้เนื่องจากในการสุ่มข้อมูลสำหรับแต่ละต้นต้องอิงกับข้อมูลที่ได้จากต้นไม้ก่อนหน้าต่างกับแบ็กกิงที่สามารถทำได้เนื่องจากไม่ต้องอิงกับข้อมูลที่ได้จากต้นไม้ก่อนหน้า วิธีบูสต์มีสองรูปแบบที่เป็นที่รู้จักกันดีคือ อาร์คกิงเอ็กโฟร์ (Arcing-x4) และเอดาบูสต์ (AdaBoost)

1. วิธีอาร์คกิงเอ็กโฟร์ (มาจากคำว่า Adaptively resample and combine) คล้ายวิธีแบ็กกิง เริ่มจากเลือกชุดเรียนรู้ขนาด n สำหรับการจำแนกครั้งที่ $k+1$ โดยเลือกตัวอย่างเรียนรู้ต้นฉบับจำนวน n ชุด ต่างจากวิธีแบ็กกิงตรงที่ค่าความน่าจะเป็นในการเลือกตัวอย่างแต่ละตัวอย่างของวิธีอาร์คกิงเอ็กโฟร์ไม่เท่ากัน โดยค่าความเป็นไปได้ขึ้นกับจำนวนการจำแนกผิดในการจำแนกก่อนหน้า ซึ่งถ้ามีจำนวนมาก ค่าความน่าจะเป็นที่จะถูกเลือกจะมีค่ามากด้วย

2. วิธีเอดาบูสต์ (ย่อมาจาก Adaptive Boosting หรือเรียกว่าอาร์คกิงเอฟเอส Arcing-fs โดย Breiman) สามารถทำได้สองรูปแบบคือ

ก) เลือกชุดตัวอย่างโดยอิงกับความเป็นไปได้ของตัวอย่าง

ข) พิจารณาตัวอย่างทั้งหมดและค่าน้ำหนักแต่ละตัวอย่าง (โดยตัวอย่างที่มีค่าน้ำหนักสูงกว่ามีผลกับความผิดพลาดมาก)

2.3 งานวิจัยที่เกี่ยวข้อง

ได้มีการนำพลวัตการพิมพ์มาทำการตรวจสอบผู้ใช้ด้วยวิธีการต่าง ๆ กันมีดังนี้

การตรวจสอบผู้ใช้ด้วยรหัสผ่าน

ในปี 1991 ได้มีการพัฒนานิวรอลเน็ตเวิร์กเพื่อจดจำรูปแบบเฉพาะตัวในการกดแป้นพิมพ์ของบุคคล เพื่อที่ระบบคอมพิวเตอร์จะยอมรับหรือปฏิเสธความพยายามล็อกอินที่เกิดขึ้นภายหลัง โดยประเมินจากรูปแบบการล็อกอินที่เคยเกิดขึ้นก่อน โดยต้องการให้ระบบมีจำนวนครั้งการยอมรับผู้ใช้ตัวจริง (ความสำเร็จของผู้ใช้) และจำนวนครั้งการปฏิเสธผู้บุกรุก (ความล้มเหลวของผู้บุกรุก) มากที่สุด พร้อมลดจำนวนครั้งการปฏิเสธผู้ใช้ตัวจริง (การเตือนผิดพลาด) และจำนวนครั้งของการยอมรับผู้บุกรุกเข้าระบบ (การบุกรุก) ให้เหลือน้อยที่สุด ขั้นตอนของการทดลองผู้วิจัยทำให้ระบบเรียนรู้โดยการป้อนรหัสผ่านหลายครั้ง และทำให้แต่ละนิวรอลเน็ตเวิร์กจดจำลักษณะของพาสเวิร์ด เมื่อมีการพิมพ์พาสเวิร์ดซ้ำหลายๆครั้ง ระบบจะบันทึกเวลาระหว่างการพิมพ์ตัวอักษรที่อยู่ติดกัน การวิจัยการพิมพ์พาสเวิร์ดความยาวปานกลาง พบว่า มีอัตราความสำเร็จ (ความสำเร็จของผู้ใช้และความล้มเหลวของ ผู้บุกรุก) 75% โดยจำนวนความผิดพลาด 25% มีเพียง 3% ที่เป็นการบุกรุก ขณะที่ 22% เป็นการเตือนผิดพลาด) จากผลการวิจัย สามารถสรุปได้ว่า สิ่งที่แฝงอยู่ในจังหวะการพิมพ์สามารถใช้เป็นวิธีการรักษาความปลอดภัย โดยการตรวจสอบรหัสผ่านและมาตรการรักษาความปลอดภัยอื่นๆ (แอลเจลล่า เลมเมอร์ และชารอน แลงเจนเฟล,1991)

ในปี 1997 ได้เสนอเทคนิคเพื่อพิสูจน์ตัวผู้ใช้คอมพิวเตอร์ จากจังหวะการกดแป้นพิมพ์ล็อกอิน โดยใช้เทคนิคนิลรอลเน็ตเวิร์ก และการจดจำรูปแบบ จากการทดลอง และได้ใช้เวลาที่ใช้ในการกดแป้นพิมพ์ตัวอักษร (key hold times) เพื่อเปรียบเทียบการประสิทธิภาพการพิสูจน์ตัวผู้ใช้คอมพิวเตอร์ โดยอาศัยระยะเวลาระหว่างการพิมพ์ตัวอักษรที่อยู่ติดกัน (interkey times) ก่อนนำไปเปรียบเทียบกับการจำแนกตัวผู้ใช้ โดยอาศัยทั้งเวลาในการกดแป้นพิมพ์ และระยะเวลาระหว่างการพิมพ์ตัวอักษรที่อยู่ติดกัน เมื่อมีการใช้ทั้งเวลากดพิมพ์ตัวอักษรหนึ่งๆ และระยะเวลา

ระหว่างการกดพิมพ์ตัวอักษรที่อยู่ติดกัน เพื่อจำแนกการล็อกอินขนาดสั้น การจำแนกจะมีประสิทธิภาพสูงสุดเมื่อใช้แบบโครงข่ายประสาทเทียมชนิดตรรกศาสตร์คลุมเครือ,อาร์ตแมป (ARTMAP), อาร์บีเอ็ฟเอ็น (RBFN) , และแอลวีคิว (LVQ) โดยให้ผลการจำแนกผิดพลาด 0% และหากใช้โครงข่ายประสาทเทียมแบบอื่นๆ เช่นแบ็กพร็อพพาเกชัน (BP) ค่าน้ำหนักของซิกมอยด์ (sigmoid) และตรีโกณแทน เอช (Tan H) และรูปแบบเอชเอสโอพี HSOP และเอสโอพี (SOP) จะมีประสิทธิภาพในการจำแนกปานกลาง โดยมีความผิดพลาดของการจำแนกผิดแบบชนิดที่ 1 และชนิดที่ 2 เฉลี่ย .05%, .75% และ 3.25% ตามลำดับ ทั้งนี้สำหรับเทคนิคการจดจำแบบดั้งเดิม พบว่าอัลกอริทึมเป็นรูปแบบที่มีประสิทธิภาพสูงสุด โดยมีความผิดพลาดของการจำแนกผิดเฉลี่ย .08% แต่ประสิทธิภาพยังต่ำกว่าโครงข่ายประสาทเทียมชนิด, อาร์ทแมพ, อาร์บีเอ็ฟเอ็น และแอลวีคิว (เอ็ม เอส โอไบแดท และ เบลไคว์ ซูดาน, 1997)

ในปี 1998 ได้มีการเสนอแนวคิดว่าการเคลื่อนไหวของจังหวะการพิมพ์ชื่อผู้เข้าใช้ของผู้ใช้คอมพิวเตอร์ทำให้เกิดรูปแบบเฉพาะที่สามารถใช้เพื่อยืนยันตัวผู้ใช้คอมพิวเตอร์ได้ งานวิจัยนี้ได้เก็บข้อมูลโดยการติดตามการพิมพ์ชื่อผู้เข้าใช้ของชื่อผู้ใช้และรหัสผ่าน (user id) ของนักเรียน 140 ราย ก่อนที่จะคัดเลือกผู้ที่ใช้งานบ่อยที่สุด 10 คน (ผู้ใช้ตัวจริง) เพื่อนำมาวิเคราะห์ โดยความยาวเฉลี่ยของรหัสผ่านคือ 6.4 ตัวอักษร ส่วนข้อมูลของผู้ใช้ปลอมนำมาจากผู้ใช้ปลอม 10 คนที่พยายามใช้รหัสผ่านคนละ 10 ครั้ง จากนั้นจึงมีการใช้ตัวแยกประเภท 3 แบบ คือ ใช้ระยะเวลาระหว่างรอการกดแป้นพิมพ์เพียงอย่างเดียว, ใช้ระยะเวลาระหว่างการกดแป้นพิมพ์สองอักขระเพียงอย่างเดียว และใช้ร่วมกันระหว่างระยะเวลาระหว่างรอการกดแป้นพิมพ์ และระยะเวลาระหว่างการกดแป้นพิมพ์สองอักขระเพื่อแยกประเภทผู้ใช้ จากการศึกษาพบว่าเวลาที่ผู้ใช้กดแป้นอักษรแต่ละตัว เป็นปัจจัยที่มีประสิทธิภาพมากกว่าช่วงเวลาระหว่างการกดแป้นอักษร 2 ตัว ในการระบุความแตกต่างของกลุ่มตัวอย่างที่เป็นผู้ใช้ตัวจริงและผู้ใช้ตัวปลอม อย่างไรก็ตามหากต้องการผลที่ดีที่สุดต้องประมวลผลจากทั้งสองปัจจัยร่วมกัน ทั้งนี้ได้พิจารณาประสิทธิภาพของตัวแยกประเภท โดยพิจารณาความผิดพลาดแบบ Type I ซึ่งหมายถึงการปฏิเสธผู้ใช้ตัวจริง และความผิดพลาดแบบ Type II ซึ่งหมายถึงการยอมรับผู้บุกรุกหรือผู้ใช้ตัวปลอม พบว่าภาพรวมของตัวแยกประเภทที่ดีที่สุด (ใช้ทั้งใช้ระยะเวลาระหว่างรอการกดแป้นพิมพ์ และ ระยะเวลาระหว่างการกดแป้นพิมพ์สองอักขระ) เกิดความผิดพลาดแบบชนิดที่ 1 เฉลี่ย 10% และความผิดพลาดแบบชนิดที่ 2 เฉลี่ย 9% การทดลองพบว่าจำนวนข้อผิดพลาดที่เกี่ยวกับการพิมพ์ล็อกอินอยู่ในระดับสูง ซึ่งในทางปฏิบัติเป็นข้อผิดพลาดที่สามารถแก้ไขได้ โดยการใช้แป้น

backspace เพื่อแก้ไขข้อผิดพลาด และแนวทางดังกล่าวทำให้สามารถเข้าถึงระบบคอมพิวเตอร์ได้สำเร็จ (จอห์น เอ. โรบินสันและคณะ, 1998)

ในปี 2000 ได้มีการนำเสนอเทคนิคการพิสูจน์ตัวผู้พิมพ์ด้วยรูปแบบเออาร์ (AR Model) ซึ่งผู้วิจัยเห็นว่ามนุษย์แต่ละคนมีจังหวะการพิมพ์แตกต่างกัน และคิดว่ากระบวนการของเออาร์เป็นรูปแบบที่ดีสำหรับจังหวะการพิมพ์ ในงานวิจัยนี้ ผู้วิจัยได้ใช้วิธีสัมประสิทธิ์ของรูปแบบเออาร์เป็นลักษณะสำคัญในการแบ่งกลุ่มตัวอย่างข้อมูลจังหวะการพิมพ์ 432 ครั้งของผู้พิมพ์ 24 คน ซึ่งผลการทดลองพิสูจน์ว่าผู้วิจัยคาดการณ์ถูกต้อง ในการทดลองผู้พิมพ์แต่ละคนต้องพิมพ์เนื้อหาที่มีความยาว 1,100 ตัวอักษรเป็นจำนวน 18 ครั้ง ดังนั้นจึงมีการพิมพ์เนื้อหาทั้งหมด 432 ครั้ง ซึ่งผู้วิจัยได้ใช้วิธี ยูล – วอร์คเกอร์ (Yule-Walker) และวิธี เบิร์ก (Burg) เพื่อประเมินความถูกต้องของการแยกประเภทโดยใช้ตัวจำแนกค่าที่ใกล้เคียง (nearest neighborhood classifier) แล้วตรวจสอบความถูกต้องโดยวิธียูล – วอร์คเกอร์ พบว่ารูปแบบเออาร์สามารถแยกประเภทข้อมูลได้ถูกต้อง 60.88% และเมื่อประเมินโดยใช้วิธีเบิร์ก พบว่ารูปแบบเออาร์สามารถแยกประเภทข้อมูลได้ถูกต้อง 61.34% แสดงให้เห็นว่าการหาลักษณะสำคัญเพื่อจดจำความแตกต่างของการเคาะแป้นพิมพ์มีความสำคัญอย่างยิ่ง และจากการทดลองพบว่าสามารถใช้รูปแบบเออาร์เพื่อใช้ยืนยันตัวผู้พิมพ์ได้ (เซง ชางซุย และ ซัน ยันฮัว, 2000)

ในปี 2000 ได้มีการนำเสนอเทคนิคเพื่อตรวจสอบรับรองรหัสผ่าน โดยการใช้เครือข่ายประสาทเทียม (neural networks) ตรรกศาสตร์คลุมเครือ (fuzzy logic) วิธีการเชิงสถิติ (statistical methods) และการผสมผสานของแนวทางดังกล่าว (hybrid combination) ในการทดลองผู้ใช้ใหม่จะมีชื่อผู้เข้าใช้เป็นของตนเอง และสามารถกำหนดรหัสผ่านได้เอง โดยระบบจำกัดความยาวของรหัสผ่านไว้ที่ 7 ตัวอักษร ซึ่งถือเป็นความยาวมาตรฐานของรหัสผ่าน ระหว่างช่วงเรียนรู้ ผู้ใช้จะต้องใส่รหัสผ่าน 15 ครั้ง โดยระบบจะไม่สามารถควบคุมความผิดพลาดทางการพิมพ์ได้ ดังนั้นผู้ใช้จะต้องใส่รหัสผ่านโดยไม่ให้มีข้อผิดพลาดใดๆ ทั้งนี้จะมีการบันทึกเวลารอคอย (delays) ระหว่างการเคาะแป้นพิมพ์ตัวอักษรแต่ละตัวไว้ในระบบและเมื่อสิ้นสุดช่วงการเรียนรู้ ระบบจะมี 6 เวกเตอร์ของการพิมพ์ 15 ครั้ง ที่แทนเวลารอคอยระหว่างการพิมพ์ตัวอักษรแต่ละตัว จากนั้นจะส่งค่าดังกล่าวต่อไปยังหน่วยโครงข่ายประสาทเทียม, หลักการคำนวณทางสถิติ และตรรกศาสตร์คลุมเครือ ซึ่งจะใช้ค่าเหล่านี้เพื่อตั้งค่าพารามิเตอร์ หรือตัวแปรเสริม ซึ่งจะมีการบันทึกตัวแปรเสริมในข้อมูลของผู้ใช้ ผลการทดลองพบว่าในการทำงานจริง เมื่อผู้ใช้ต้องการเข้าสู่ระบบ พวกเขาจะมีโอกาส 2 ครั้งที่จะใส่รหัสผ่าน จากการสังเกตพบว่าผู้ใช้ตัวจริงที่ถูกปฏิเสธจากระบบในครั้งแรก จะสามารถใส่รหัสผ่านได้ถูกต้องในครั้งที่ 2 ซึ่งผู้วิจัยได้ประเมิน

ความพยายามทั้ง 2 ครั้งเป็นหน่วยเดียวกันในการวิเคราะห์ผล ในช่วงเริ่มต้น ระบบให้โอกาสผู้ใช้เข้าสู่ระบบเพียงครั้งเดียว พบว่าเกิดความผิดพลาดแบบชนิดที่ 1 (ความเป็นไปได้ที่ผู้ใช้ตัวจริงถูกปฏิเสธ) สูงมาก แต่เมื่อนับรวมโอกาสครั้งที่ 2 ความผิดพลาดแบบชนิดที่ 2 ลดลงมาก แต่จำนวนโอกาสในการใส่รหัสผ่าน ไม่ได้ส่งผลกระทบหรือส่งผลกระทบน้อยต่อความผิดพลาดแบบชนิดที่ 2 (ความเป็นไปได้ที่จะยอมรับผู้บุกรุกหรือผู้ใช้ตัวปลอม) จากการทดลอง ผู้วิจัยสรุปว่าการใช้เวลาระยะเวลาระหว่างรอการกดแป้นพิมพ์ ร่วมกับระยะเวลาระหว่างการกดแป้นพิมพ์สองอักขระ จะทำให้การตรวจสอบผู้ใช้จริงได้อย่างมีประสิทธิภาพดีขึ้น และการใช้เทคนิคทุกๆอย่างร่วมกันจะทำให้ประเมินผลได้ดีขึ้น (ซาจจัด ไฮเดอร์และคณะ, 2000)

ในปี 2005 ได้นำเสนอวิธีอัลกอริทึมมอนติคาร์โล (Monte Carlo) ในการสร้างข้อมูลสำหรับการเรียนรู้ และใช้วิธีการแบ่งข้อมูล (Splitting) ปรับปรุงประสิทธิภาพของข้อมูลสำหรับการเรียนรู้ จากนั้นนำมาประกอบเป็นต้นไม้ตัดสินใจแบบขนาน (Parallel Decision Tree) เพื่อนำไปใช้ในการพิสูจน์ตัวจริงของผู้ใช้ (user authentication) ด้วยการเคาะแป้นพิมพ์ ซึ่งการเก็บข้อมูลเพื่อนำมาสร้างต้นไม้การตัดสินใจที่มีประสิทธิภาพเพียงพอ ทำโดยสร้างข้อมูลจำลองขึ้นมา 19 ครั้งโดยสร้างจากข้อมูลที่ผู้ใช้แต่ละคนพิมพ์เข้ามา จากนั้นนำมาประกอบกับเวกเตอร์ของข้อมูลดิบจำนวน 387 เวกเตอร์เพื่อรวมเป็นข้อมูลสำหรับการเรียนรู้ และจะแบ่งข้อมูลออกเป็น 4 สับเซต (subset) โดยในแต่ละ ส่วนย่อย จะใช้วิธีการแปลงเวฟเล็ต (Wavelet Transform) ในการสร้างข้อมูลสำหรับการเรียนรู้ย่อย (training subset) จำนวน 8 ชุดสำหรับผู้ใช้แต่ละคน จากนั้นต้นการตัดสินใจจำนวน 8 ชุด จะถูกทำการเรียนรู้โดยใช้ข้อมูลสำหรับการเรียนรู้ย่อยที่ได้มา ต้นไม้การตัดสินใจแบบขนานจะถูกสร้างสำหรับผู้ใช้แต่ละคน โดยต้นไม้ตัดสินใจแต่ละต้นจะประกอบไปด้วยกฎเกณฑ์สำหรับใช้ในการพิสูจน์ตัวจริงของผู้ใช้ โดยถ้ามีอย่างน้อย 3 ต้น ที่ตรงกับกฎเกณฑ์ ผู้ใช้คนนั้นคือตัวจริงแต่ถ้าไม่ตรงตามเงื่อนไขนี้ก็จะไม่ยอมรับผู้ใช้คนนั้น จากการทดลองโดยเก็บชุดข้อมูลเรียนรู้ (ไม่รวมข้อมูลที่จำลองขึ้นมา) จากผู้ใช้ 43 คน โดยให้ผู้ใช้พิมพ์ข้อความเดียวกันที่มีความยาว 37 ตัว (รวมช่องว่าง) ต่อกัน 9 ครั้ง และเก็บชุดข้อมูลทดสอบโดยให้ผู้ใช้ป้อนค่าใด ๆ โดยไม่มีข้อกำหนด เมื่อผู้ใช้พิมพ์ข้อความเดียวกัน ที่เวลาต่าง ๆ กัน ในช่วงเดือนพฤศจิกายน ถึงเดือนธันวาคม 2545 ได้ค่าเฉลี่ยของอัตราการปฏิเสธที่ผิด (average false reject rate) ที่ 9.62% และ ค่าเฉลี่ยของอัตราการยอมรับที่ผิด (average false accept rate) ที่ 0.88% (ยัง เซง และคณะ, 2005)

การตรวจสอบผู้ใช้ด้วยข้อความอิสระ

ในปี 2003 ได้เสนอว่าพฤติกรรมในการพิมพ์ เป็นลักษณะการตรวจสอบบุคคลจากพฤติกรรมการพิมพ์ที่เป็นเอกลักษณ์เฉพาะของแต่ละบุคคล และยากต่อการลอกเลียนแบบ เพราะลักษณะในการพิมพ์เกิดจากประสบการณ์ที่สั่งสมมา และความชำนาญในการพิมพ์ของแต่ละบุคคล สิ่งที่เราใช้ในการจำแนกความแตกต่างของบุคคล คือ เวลาที่ใช้ในการพิมพ์ แต่เวลาที่เราสนใจนั้นไม่ใช่ความเร็วของการพิมพ์ แต่เป็นเวลาที่แต่ละคนใช้ในการกดค้างไว้ในแต่ละตัวอักษร และ ระยะเวลาที่ใช้ในการพิมพ์ระหว่างตัวอักษรหนึ่งไปอีกตัวอักษรหนึ่ง สำหรับการพิสูจน์บุคคลจากพฤติกรรมการพิมพ์เป็นระบบที่ได้รับการเชื่อถือและยอมรับเนื่องจากมีคุณสมบัติที่สำคัญ 2 ประการคือ สามารถตรวจสอบตัวบุคคลได้ (Verification) และสามารถชี้ตัวแยกแยะ บุคคลที่แตกต่างกันได้ ระบุตัวบุคคลได้ (Identification) การพิสูจน์บุคคลจากพฤติกรรมการพิมพ์ เป็นระบบหนึ่งของไบโอเมตริก ซึ่งเป็นการใช้กระบวนการในการระบุตัวบุคคลหรือ ตรวจสอบตัวบุคคลโดยอัตโนมัติ โดยใช้ลักษณะทางกายภาพ ที่แตกต่างกันแต่ละบุคคล หรือใช้ลักษณะทางพฤติกรรมของแต่ละบุคคล โดยการตรวจวัดคุณลักษณะทางกายภาพ และลักษณะทางพฤติกรรม ที่เป็นลักษณะเฉพาะของแต่ละคนมาใช้ในการระบุตัวบุคคลนั้นๆ แล้วนำสิ่งเหล่านั้นมาเปรียบเทียบกับคุณลักษณะที่ได้มีการบันทึกไว้ในฐานข้อมูลก่อนหน้านี้ เพื่อใช้แยกแยะบุคคลนั้นจากบุคคลอื่นๆ ระบบไบโอพาสเวิร์ด (Biopassword) สามารถติดตั้งใน ระบบวินโดว์ 2000 (Windows 2000) เพื่อป้องการการใช้งานจากบุคคลอื่น และเป็นระบบที่ถูกตรวจสอบแล้วว่า ไม่รบกวนระบบการทำงานของคอมพิวเตอร์ ไม่ต้องใช้อุปกรณ์เสริมอื่นๆ และมีราคาไม่แพงเกินไป อย่างไรก็ตาม ระบบตรวจสอบบุคคลจากพฤติกรรมการพิมพ์ ยังมีปัญหาในบางสถานการณ์ เช่นผู้ใช้งานอยู่ในสภาวะง่วงซึม หรือป่วย นิ้วมือได้รับบาดเจ็บทำให้รูปแบบลายเซ็นเปลี่ยนไป หรือเซนต์ชื่อในขณะที่อีกมือหนึ่งถือกาแฟก็ทำให้ลักษณะลายเซ็นเปลี่ยนแปลงไปได้ การเปลี่ยนแปลง แป้นพิมพ์ หรือเครื่องคอมพิวเตอร์ก็สามารถมีผลกระทบต่อการตรวจสอบบุคคลในระบบพฤติกรรมการเคาะแป้นพิมพ์ด้วยเช่นกัน (จาโม อิโลนิน, 2003)

ในปี 2005 ได้นำเสนอวิธีการแบบใหม่ จุดประสงค์เพื่อปรับปรุงระบบการรักษาความปลอดภัยของโปรแกรมสำคัญต่างๆและสิ่งอื่นๆที่สำคัญในระบบไว้ การค้นคว้านี้เป็นประโยชน์ในการที่จะสืบค้นหาผู้ที่อาจเข้าสู่ระบบเพื่อหาผลประโยชน์ เพื่อที่จะจัดการผู้คนเหล่านั้นให้ออกจากระบบไป ถึงแม้ว่าเราอาจจะรู้วิธีการและทราบถึงผู้ไม่ประสงค์ดีกับระบบโดยตรวจสอบจากจังหวะการพิมพ์บนแป้นพิมพ์ วิธีการนี้มีความแตกต่างจากวิธีการอื่นคือเนื่องจากวิธีการนี้ไม่จำเป็นต้องใช้ส่วนฮาร์ดแวร์ (Hardware) หรือ โปรแกรมคอมพิวเตอร์ใดๆ

เป็นเครื่องมือในการตรวจสอบ นอกจากการใช้เครื่องคอมพิวเตอร์เท่านั้นที่ผู้ใช้ต้องใช้ในการเข้าสู่ระบบ วิธีการนี้อาศัยการเข้าใช้งานระบบของผู้ใช้งานบนเครื่องคอมพิวเตอร์และการใส่ชื่อและรหัสผ่านเท่านั้น แต่ในกรณีที่กลับกัน มีความขัดแย้งกันอยู่เกี่ยวกับวิธีการนี้เนื่องจากผู้ที่ไม่ประสงค์ดีต่อระบบสามารถใช้วิธีการเดียวกันเพื่อค้นหาและนำข้อมูลดังที่กล่าวมาไปหาประโยชน์ได้เช่นกัน ส่วนพฤติกรรมในการพิมพ์สามารถนำไปใช้ประโยชน์ในการสืบค้นข้อมูลส่วนตัวถึงแม้ว่าผู้ที่ประสงค์จะเข้าไปในระบบสามารถผ่านขั้นตอนในการรับรองจากระบบแล้วก็ตาม และแม้ว่าเราจะสามารถรับมือกับวิธีการโดยการใช้จังหวะการพิมพ์นั้นๆได้ หรือการเลือกที่จะเข้าสู่ระบบของผู้ไม่ประสงค์ดีโดยไม่มีการระบุหรือการบังคับในการใส่ตัวอักษรใดๆเป็นรหัสผ่าน ในบทนี้ผู้เขียนขอเสนอวิธีการในการเปรียบเทียบตัวอย่างการพิมพ์ที่ไม่กำหนดตัวอักษรที่สามารถระบุรูปพรรณของผู้พิมพ์ได้ มีการทดสอบเทคนิคต่างๆนี้กับชุดการทดสอบเป็นจำนวนมากและพบว่าระดับในการพิมพ์ที่ผิดพลาดซึ่งเป็นการผิดพลาดโดยทั่วไปคิดเฉลี่ยได้เป็น 5% และ 0.005% เป็นการพิมพ์ปลอมเข้ามาเพื่อเข้าสู่ระบบ วิธีการที่ใช้จากการศึกษาและทดสอบนี้จะอาศัยการพิมพ์จากบุคคลทั่วไปซึ่งใช้ในการทำงานปกติประจำวัน และอาศัยจากตัวอักษรในประโยคต่างๆกัน หรือจากการไปเก็บข้อมูลทำงานของภาคอื่นๆ ซึ่งข้อมูลที่ได้มาจากแหล่งต่างๆเหล่านี่มีประสิทธิ์ภาพสูงในการนำมาใช้ประโยชน์ ซึ่งจากข้อมูลเหล่านี้ได้มีการปรับปรุงเป็นส่วนให้ดีขึ้นจนกลายมาเป็นข้อมูลที่นำมาใช้ประโยชน์ได้ จากผลการศึกษายังมีข้อโต้แย้งที่ว่าวิธีการที่ได้ศึกษานี้มีประโยชน์เพียงหรือไม่ในการที่จะเป็นหนทางหนึ่งในการช่วยหรือเป็นเพียงแค่ส่วนประกอบของเรื่องความปลอดภัยของระบบหรือเป็นเพียงแค่ทางเลือกอีกทางหนึ่งเพื่อการสืบค้นผู้ที่ไม่ประสงค์ดีและหาผลประโยชน์กับระบบ (แดนเนียล กูเน็ตติ และ คลาวเดียร์ พิคาร์ดี, 2005)

บทที่ 3

วิธีการดำเนินงานวิจัย

งานวิจัยนี้เป็นการนำเทคนิคของต้นไม้ตัดสินใจ มาใช้ในวิธีการรวมตัวจำแนก เพื่อให้ได้ ผลลัพธ์ที่มีประสิทธิภาพดีกว่าต้นไม้ในการตัดสินใจแบบเดิม ในส่วนนี้อธิบายถึงโครงสร้างของระบบและวิธีการวัดประสิทธิภาพ

โครงสร้างของระบบ

โครงสร้างของระบบการจำแนกผู้ใช้โดยใช้เวลาในการกดแป้นพิมพ์ประกอบด้วยข้อมูลที่ใช้ในการทดลอง การเก็บข้อมูลตัวอย่างของเวลาที่ใช้ในการทดลอง การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ และการเรียนรู้ของต้นไม้ตัดสินใจ

3.1 ข้อมูลที่ใช้ในการทดลอง

ในหัวข้อนี้อธิบายถึงข้อมูลที่ใช้ในการทดลอง และลักษณะของข้อมูลที่ใช้ในการทดลอง สำหรับข้อมูลที่นำมาใช้ในการทดลองนี้ได้ทำการเก็บข้อมูลเวลาที่ใช้ในการพิมพ์จากกลุ่มผู้ใช้ที่สำนักงานผู้บังคับทหารอากาศดอนเมือง โดยข้อมูลที่ทำการเก็บเวลาที่ใช้ในการพิมพ์มาได้คือเวลาที่ผู้ใช้ทำการกดแป้นพิมพ์ระหว่างสองแป้นพิมพ์ โดยมีตัวอย่างประโยคที่ใช้ในการพิมพ์ในครั้งนี้ได้คัดเลือกมาจากประโยคที่ใช้ในการพิมพ์หนังสือราชการ และเป็นประโยคที่ทำการพิมพ์เป็นประจำที่สำนักงานผู้บังคับทหารอากาศดอนเมืองในช่วงระยะเวลา 10 ปีที่ผ่านมา ดังตารางที่ 3.1

ตารางที่ 3.1

ประโยคตัวอย่างที่ใช้สำหรับการพิมพ์เพื่อเก็บข้อมูลเวลาการพิมพ์ของผู้ใช้

บรรทัดที่	ตัวอย่างประโยค
1	ตามหนังสือที่
2	รายละเอียดตามความทราบแล้วนั้น
3	สน.ผบ.ดม.ตรวจสอบและพิจารณาแล้ว เห็นสมควรให้
4	จึงเรียนมาเพื่ออนุมัติตามข้อ ๓ ให้ต่อไป
5	จึงเรียนมาเพื่อโปรดพิจารณาดำเนินการให้ต่อไป
6	จึงเรียนมาเพื่อพิจารณา หากเห็นชอบด้วย ขอได้ลงชื่อในร่างคำสั่ง ฯ ที่แนบให้ต่อไป
7	จึงเสนอมาเพื่อพิจารณา หากเห็นชอบด้วย ขอได้ลงชื่อในหนังสือเสนอ กพ.ทอ.ให้ต่อไป
8	กระผมพิจารณาแล้ว เห็นสมควรลงชื่อในร่างคำสั่ง ฯ ที่แนบให้ต่อไป
9	กพ.บก.สน.ผบ.ดม.ดำเนินการในส่วนที่เกี่ยวข้อง และเก็บรวบรวมไว้เป็นหลักฐานต่อไป
10	สำนักงานผู้บังคับทหารอากาศดอนเมือง

3.2 การเก็บข้อมูลตัวอย่างของเวลาที่ใช้ในการทดลอง

งานวิจัยนี้ได้ทำการเก็บตัวอย่างเวลาของผู้ใช้จำนวน 20 คนจากสำนักงานผู้บังคับทหารอากาศดอนเมือง โดยเวลาที่ทำการเก็บนี้จะนำไปใช้สำหรับเป็นข้อมูลในการเรียนรู้ เพื่อเข้าสู่กระบวนการจำแนกต่อไป

ตารางที่ 3.2

ตารางข้อมูลตัวอย่างเวลาระหว่าการกดแป้นพิมพ์สองอักขระ (Keystroke Interval Time)

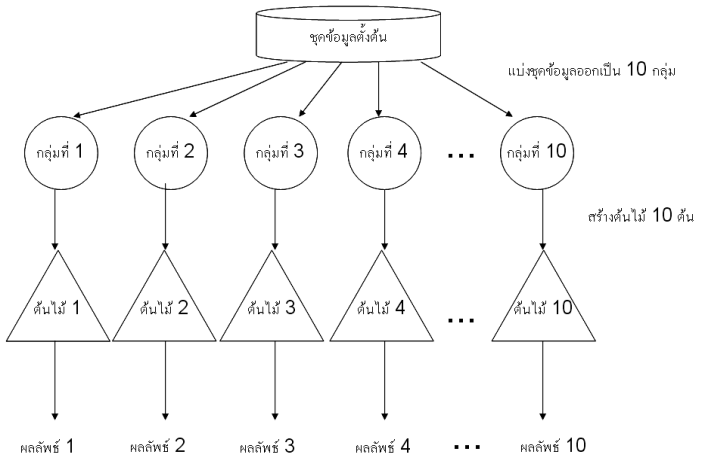
ผู้ใช้	ต	สระอา	ม	ห	น	ไม้หันอากาศ	ง	ส	สระอือ	อ	ท	สระอี	ไม้เอก
คนที่1 (U1)	0	330	862	430	221	430	391	330	461	370	321	300	191
คนที่ 2 (U2)	0	390	231	1642	351	380	241	340	611	200	1542	431	301
⋮													
คนที่ 20 (U20)	0	250	211	270	180	551	291	220	861	341	240	250	251

จากตารางที่ 3.2 เป็นตัวอย่างตารางที่ใช้ในการเก็บค่าของเวลาระหว่างการกดแป้นพิมพ์ระหว่างอักขระ (Keystroke Time Interval) ของผู้ใช้ 20 คน ที่ได้จากการพิมพ์ประโยคตัวอย่างที่ 1 คือ “ตามหนังสือที่” ซึ่งผู้ใช้ทั้ง 20 คนจะต้องทำการพิมพ์ประโยคนี้ทั้งหมด 10 ครั้ง นั่นคือผู้ใช้จะต้องทำการพิมพ์ประโยคตัวอย่างทั้ง 10 ประโยค เป็นจำนวนประโยคละ 10 ครั้ง เวลาที่ได้นี้จะนำไปเป็นกลุ่มข้อมูลสำหรับเรียนรู้ และกลุ่มข้อมูลสำหรับทดสอบเพื่อทำการจำแนกผู้ใช้ต่อไป

3.3 การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ

ภาพที่ 3.1

การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ



จากภาพที่ 3.1 เป็นการสุ่มเลือกแบ่งกลุ่มข้อมูลสำหรับการเรียนรู้ที่ได้มาจากตารางที่ 3.2 ซึ่งเป็นชุดข้อมูลตั้งต้นด้วยวิธีบูตสแตรป โดยข้อมูลที่ได้อาจจะซ้ำกัน หรือไม่ซ้ำกันก็ได้ในกลุ่มข้อมูลที่แบ่งออกมาเป็นกลุ่มย่อย 10 กลุ่ม

3.4 การเรียนรู้ของต้นไม้ตัดสินใจ

วิธีการจำแนกข้อมูลตัวอย่าง

1. กำหนดให้ในปัญหา D มีการแบ่งตัวอย่างออกเป็นคลาส

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$$

2. ให้ $D_1, D_2, ..., D_m$ เป็นตัวอย่างของชุดข้อมูลเรียนรู้ และให้ m เป็นจำนวนชุดของการเรียนรู้

3. สร้างแบบจำลองการเรียนรู้ $G_m(X)$ ด้วยวิธีการสุ่มตัวอย่างจากบูตสแตรปเป็นชุดข้อมูลการเรียนรู้ D_m

4. หาค่าการคาดเดาจากสมการ $\hat{y} = \frac{1}{M}\sum_{m=1}^{M} G_m(X)$

หาคำตอบสุดท้ายได้จากค่าการลงคะแนน (Vote) ของ $G_1(X), ..., G_M(X)$

Pseudo Code Bagging

Require : L, T ≥ 1, A

for $t \leftarrow 1...T$ do

{draw uniformly and with replacement |L| ofbjects form L}

$L_t \leftarrow$ Uniform (L, |L|)

$h_t \leftarrow A(L_t)$

end for

{the output is the majority – vote of the learned hypotheses}

out ← H: $\vec{x} \rightarrow \arg\max_{y \in Y} \left| \{t | h_t(\vec{x}) = y\} \right|$

(ที่มาจาก : Roberto Esposito, <http://www.di.unito.it> และ Leo Breiman, 1994)

Pseudo Code AdaboostM1

Require: L, T $\geq$ 1,A

t $\leftarrow$ 1

for each i = 1 . . . |L| do D1(i) = $\frac{1}{|L|}$

for t 1 . . . T do

$h_t \leftarrow A(L, D_t)$

{$\in_t$ is evaluated as a function of the weighted training error}

$\in_t = E_i \sim I\,(h_t(\vec{x}_i) \neq y_i)$

$\alpha_t \leftarrow \frac{1}{2}\log\left(\frac{1-\in_t}{\in_t}\right)$

for each i = 1 . . . |L| do Dt+1(i) = $\frac{\mathrm{Dt(i)}\exp(-\alpha_t \mathrm{y_i h_t(x_1)})}{\mathrm{Z_t}}$

end for

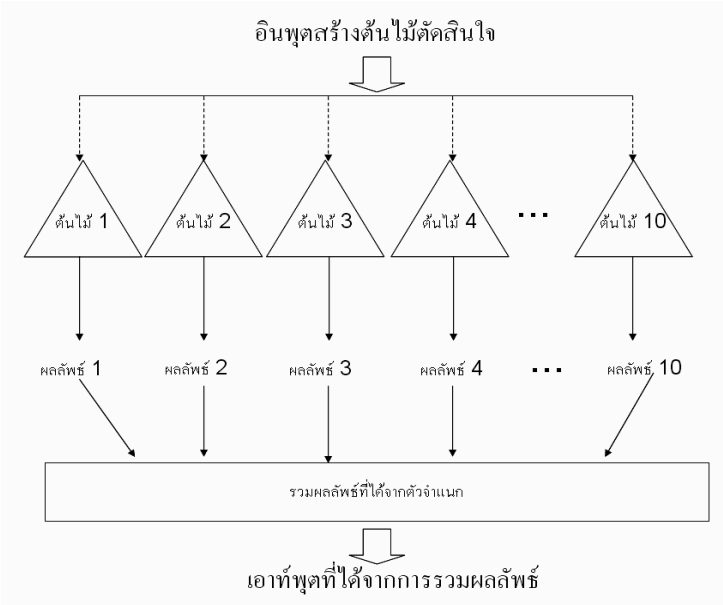
{the output is a weighted-majority-vote of the learned hypotheses }

out H : $\sim\vec{x} \rightarrow sign\left(\sum_{t\in 1\ldots T} h_t(\vec{x})\right)$

(ที่มาจาก : Roberto Esposito, <http://www.di.unito.it> และ Leo Breiman, 1994)

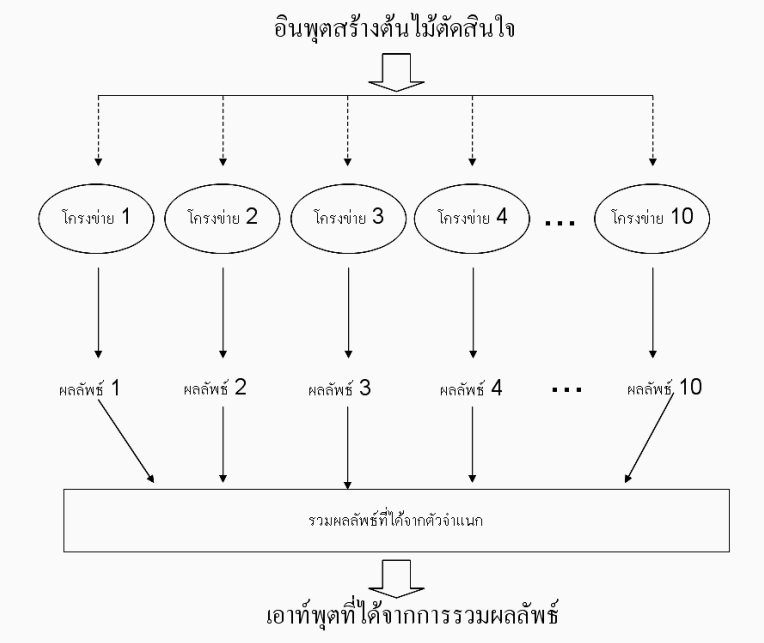
ภาพที่ 3.2

การรวมคำตอบที่ได้จากต้นไม้ตัดสินใจ



จากภาพที่ 3.2 เป็นการนำคำตอบที่ได้จากต้นไม้ตัดสินใจแต่ละต้นเพื่อนำมาคำนวณหาคำตอบที่ดีที่สุดในแต่ละวิธี ซึ่งวิธีการทำงานวิจัยนี้ได้เลือกนำมาใช้ในการประเมินตัวแบบโดยการใช้การรวมตัวจำแนกคือ แบ็กกิง และบูสต์

ภาพที่ 3.3
การรวมคำตอบที่ได้จากเพอร์เซปตรอนแบบหลายชั้น



จากภาพที่ 3.3 เป็นการนำคำตอบที่ได้จากเพอร์เซปตรอนแบบหลายชั้น ค่าน้ำหนักที่ได้จากแต่ละโครงข่ายนำมาคำนวณเพื่อหาคำตอบที่ดีที่สุดในแต่ละวิธี ซึ่งวิธีการทำงานวิจัยนี้ได้เลือกนำมาใช้ในการประเมินตัวแบบโดยการใช้การรวมตัวจำแนกคือ แบ็กกิง และบูสต์

บทที่ 4

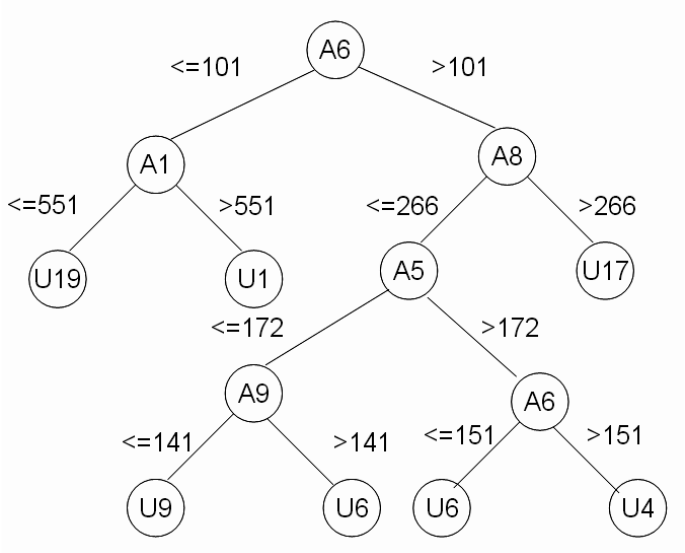
วิธีการวัดประสิทธิภาพ และผลการทดลอง

เนื้อหาในบทนี้จะกล่าวถึง วิธีการวัดประสิทธิภาพ และข้อมูลผลการทดลองที่แสดงการเปรียบเทียบให้เห็นถึงอัตราความถูกต้องจากตัวอย่างที่ใช้ทดลองในแต่ละบรรทัดของต้นไม้ตัดสินใจต้นเดียว กับการรวมคำตอบของต้นไม้ตัดสินใจหลายต้นคือวิธีแบ็กกิง (Bagging) และวิธีเอด้าบูสต์เอ็มวัน (AdaboostM1) และระบบเพอร์เซปตรอนแบบหลายชั้น กับการรวมคำตอบของเพอร์เซปตรอนแบบหลายโครงข่าย โดยใช้เครื่องมือ WEKA (Written and Frank, 2005) ในส่วนของผลการทดลองแสดงตารางการเปรียบเทียบอัตราความถูกต้องในแต่ละวิธีการ

4.1 การแทนต้นไม้ตัดสินใจต้นเดียวด้วยกฎ

นำชุดข้อมูลทดสอบมาทดสอบในกฎที่สร้างจากต้นไม้ตัดสินใจต้นเดียว ผลลัพธ์ที่ได้จะ แสดงอัตราความถูกต้องของการจำแนกผู้ใช้ ดังภาพที่ 4.1

ภาพที่ 4.1
ตัวอย่างต้นไม้ตัดสินใจ



จากภาพที่ 4.1 เป็นตัวอย่างต้นไม้ย่อยที่ได้มาจากการสร้างต้นไม้ตัดสินใจต้นเดียวของข้อมูลชุดตัวอย่างการเรียนรู้ โดยนำมาสร้างเป็นกฎของต้นไม้ต้นไม้ตัดสินใจต้นเดียวได้ดังนี้

กฎที่ 1 A6 <= 101 and A1 <= 551 THEN U19

กฎที่ 2 A6 <= 101 and A1 > 551 THEN U1

กฎที่ 3 A6 > 101 and A8 <= 266 and A5 <= 172 and A9 <= 141 THEN U9

กฎที่ 4 A6 > 101 and A8 <= 266 and A5 <= 172 and A9 > 141 THEN U6

กฎที่ 5 A6 > 101 and A8 <= 266 and A5 > 172 and A6 <= 151 THEN U6

กฎที่ 6 A6 > 101 and A8 <= 266 and A5 > 172 and A6 > 151 THEN U4

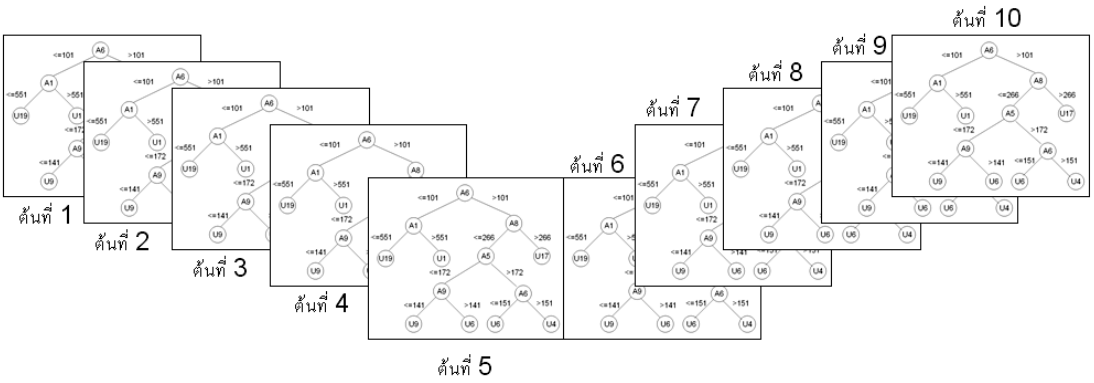
กฎที่ 7 A6 > 101 and A8 > 266 THEN U17

4.2 การแทนต้นไม้โดยใช้การรวมตัวจำแนกด้วยกฎ

นำชุดข้อมูลทดสอบมาทดสอบในกฎที่สร้างจากต้นไม้ตัดสินใจแบบการรวมตัวจำแนกจะได้ต้นไม้ทั้งหมด 10 ต้นผลลัพธ์ที่ได้จะแสดงอัตราความถูกต้องของการจำแนกผู้ใช้ในแต่ละต้นดังภาพที่ 4.2

ภาพที่ 4.2

การแทนต้นไม้ตัดสินใจแบบการรวมตัวจำแนกด้วยกฎ



จากภาพที่ 4.2 สามารถนำต้นไม้ตัดสินใจทั้ง 10 ต้นมาทำการสร้างเป็นกฎ วิธีการสร้างต้นไม้ตัดสินใจวิธีแบบการรวมตัวจำแนกเป็นกฎในแต่ละต้นนั้นเหมือนกันกับต้นไม้ตัดสินใจต้นเดียว จำนวน 10 ต้น

กฎจากต้นไม้ตัดสินใจวิธีแบบการรวมตัวจำแนกแบบวิธีแบ็กกิงต้นที่ 1

IF A12 <= 161 and A4 <= 250 and A3 <= 511 and A6 <= 181 and A6 <= 101 THEN U19

IF A12 <= 161 and A4 <= 250 and A3 <= 511 and A6 <= 181 and A6 > 101 and A2 <= 190 THEN U4

IF A12 <= 161 and A4 <= 250 and A3 <= 511 and A6 <= 181 and A6 > 101 and A2 > 190 THEN U6

IF A12 <= 161 and A4 <= 250 and A3 <= 511 and A6 > 181 and A12 <= 150 THEN U3

IF A12 <= 161 and A4 <= 250 and A3 <= 511 and A6 > 181 and A12 > 150 THEN U17

IF A12 <= 161 and A4 <= 250 and A3 > 511 and A5 <= 251 THEN U9

IF A12 <= 161 and A4 <= 250 and A3 > 511 and A5 > 251 THEN U12

IF A12 <= 161 and A4 > 250 and A1 <= 401 THEN U5

IF A12 <= 161 and A4 > 250 and A1 > 401 THEN U8

IF A12 > 161 and A11 <= 231 and A6 <= 251 and A8 <= 271 THEN U6

IF A12 > 161 and A11 <= 231 and A6 <= 251 and A8 > 271 THEN U17

IF A12 > 161 and A11 <= 231 and A6 > 251 and A8 <= 180 THEN U11

IF A12 > 161 and A11 <= 231 and A6 > 251 and A8 > 180 and A9 <= 230 and A7 <= 350 THEN U18

IF A12 > 161 and A11 <= 231 and A6 > 251 and A8 > 180 and A9 <= 230 and A7 > 350 and A8 <= 491 THEN U10

IF A12 > 161 and A11 <= 231 and A6 > 251 and A8 > 180 and A9 <= 230 and A7 > 350 and A8 > 491 THEN U12

IF A12 > 161 and A11 <= 231 and A6 > 251 and A8 > 180 and A9 > 230 THEN U20

จากกฎข้างต้นเป็นตัวอย่างการแทนต้นไม้ตัดสินใจวิธีแบ็กกิงด้วยกฎของต้นไม้เพียง 1 ต้น จากทั้งหมด 10 ต้น เมื่อได้กฎทั้งหมดแล้วจึงนำกฎต่าง ๆ เหล่านี้ไปทำการทดสอบกับชุดข้อมูลสำหรับทดสอบเพื่อทำการจำแนกผู้ใช้ต่อไป

กฎจากต้นไม้ตัดสินใจวิธีแบบการรวมตัวจำแนกแบบวิธีเอด้าบูสต์เอ็มวันต้นที่ 1

IF A6 <= 200 and A6 <= 101 and A1 <= 551 THEN U19

IF A6 <= 200 and A6 <= 101 and A1 > 551 THEN U1

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 <= 172 and A9 <= 141 THEN U9

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 <= 172 and A9 > 141 THEN U6

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 > 172 and A6 <= 151 THEN U6

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 > 172 and A6 > 151 and A4 <= 191 THEN U4

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 > 172 and A6 > 151 and A4 > 191 THEN U6

IF A6 <= 200 and A6 > 101 and A8 > 266 THEN U17

IF A6 > 200 and A8 <= 265 and A11 <= 271 and A6 <= 251 and A12 <= 161 THEN U3

IF A6 > 200 and A8 <= 265 and A11 <= 271 and A6 <= 251 and A12 > 161 THEN U16

IF A6 > 200 and A8 <= 265 and A11 <= 271 and A6 > 251 and A11 <= 200 THEN U20

IF A6 > 200 and A8 <= 265 and A11 <= 271 and A6 > 251 and A11 > 200 and A7 <= 361 THEN U11

IF A6 > 200 and A8 <= 265 and A11 <= 271 and A6 > 251 and A11 > 200 and A7 > 361 THEN U10

จากกฎข้างต้นเป็นตัวอย่างการแทนต้นไม้ตัดสินใจวิธีเอดาบูสต์เอ็มวันด้วยกฎของต้นไม้เพียง 1 ต้นจากทั้งหมด 10 ต้น เมื่อได้กฎทั้งหมดแล้วจึงนำกฎต่าง ๆ เหล่านี้ไปทำการทดสอบกับชุดข้อมูลสำหรับทดสอบเพื่อทำการจำแนกผู้ใช้ต่อไป

4.3 อธิบายวิธีการทดสอบกฎ

เพื่อแสดงเห็นว่าวิธีแบ็กกิง และวิธีเอดาบูสต์เอ็มวัน สามารถจำแนกตัวอย่างได้ดีกว่าวิธีการของต้นไม้ตัดสินใจต้นเดียว จึงได้แสดงตัวอย่างจากการทดลองด้วยการสร้างเป็นกฎการเรียนรู้ ซึ่งตัวอย่างที่ใช้ในการทดสอบผู้ใช้จริงคือ U9 ดังตารางที่ 4.1

ตารางที่ 4.1

ตัวอย่างข้อมูลที่นำมาทดสอบจากชุดข้อมูลสำหรับการทดสอบ

ผู้ใช้	A0	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
	ต	สระอา	ม	ห	น	ไม้หันอากาศ	ง	ส	สระอือ	อ	ท	สระอี	ไม้เอก
คนที่ 9 (U9)	0	210	211	480	81	160	180	441	210	120	251	190	150

จากตารางที่ 4.1 เมื่อนำตัวอย่างข้อมูลเวลาที่ใช้ในการพิมพ์ของผู้ใช้คนที่ 9 มาทำการทดสอบกับกฎที่สร้างจากต้นไม้ตัดสินใจต้นเดียว และต้นไม้ตัดสินในแบบวิธีแบ็กกิง และวิธีเอดาบูสต์เอ็มวัน ได้ผลดังนี้

ตัวอย่างผลการทดสอบโดยกฎที่ได้จากการสร้างต้นไม้ตัดสินใจต้นเดียว จากตัวอย่างนี้แสดงถึงตัวอย่างที่ต้นไม้ตัดสินใจต้นเดียวทำการจำแนกผู้ใช้ผิด

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 > 172 and A6 > 151 and A4 <= 191 THEN U4

ตัวอย่างผลการทดสอบโดยกฎที่สร้างได้จากวิธีแบ็กกิง จากตัวอย่างนี้แสดงถึงตัวอย่างที่ทำให้วิธีแบ็กกิงทำการจำแนกผู้ใช้ถูก

IF A12 <= 161 and A4 <= 250 and A3 > 511 and A5 <= 251 THEN U9

ตัวอย่างผลการทดสอบโดยกฎที่สร้างได้จากวิธีเอดาบูสต์จากตัวอย่างนี้แสดงถึงตัวอย่างที่ทำให้วิธีเอดาบูสต์ทำการจำแนกผู้ใช้ถูก

IF A6 <= 200 and A6 > 101 and A8 <= 266 and A5 <= 172 and A9 <= 141 THEN U9

เมื่อทำการทดสอบโดยใช้วิธีการของต้นไม้ต้นเดียวนั้น ผลลัพธ์ที่ได้คือการจำแนกผู้ใช้ผิด เพราะผู้ใช้ที่ถูกนำไปเป็นข้อมูลชุดทดสอบคือ U9 แต่เมื่อนำไปทดสอบกับต้นไม้ตัดสินใจต้นเดียว ผลลัพธ์ที่ได้ของการจำแนกผู้ใช้นั้นตอบเป็น U4

วิธีการแบ็กกิง ผลลัพธ์ที่ได้จากการรวมผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจ คือการจำแนกผู้ใช้ถูกเพราะผู้ใช้ที่ถูกนำไปเป็นข้อมูลชุดทดสอบคือ U9 แต่เมื่อนำไปทดสอบกับต้นไม้ตัดสินใจต้นเดียวผลลัพธ์ที่ได้ของการจำแนกผู้ใช้นั้นตอบเป็น U9

วิธีการเอดาบูสต์ ผลลัพธ์ที่ได้จากการรวมผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจ คือ การจำแนกผู้ใช้ถูกเพราะผู้ใช้ที่ถูกนำไปเป็นข้อมูลชุดทดสอบคือ U9 แต่เมื่อนำไปทดสอบกับต้นไม้ตัดสินใจต้นเดียวผลลัพธ์ที่ได้ของการจำแนกผู้ใช้นั้นตอบเป็น U9

ดังแสดงให้เห็นว่าผลลัพธ์ที่ได้จากวิธีแบ็กกิง และเอด้าบูสต์ ตรงกันกับคลาสของตัวอย่างทดสอบดีกว่า ผลลัพธ์ที่ได้จากต้นไม้ตัดสินใจต้นเดียว

4.2 ผลการทดลอง

ผลการทดสอบการหาอัตราความถูกต้องเพื่อแสดงให้เห็นความสามารถในการจำแนกผู้ใช้จากกลุ่มผู้ใช้อื่น ๆ

ตารางที่ 4.2

ตารางผลการทดสอบหาอัตราความถูกต้องในชุดข้อมูลสำหรับการทดสอบด้วยต้นไม้ตัดสินใจ

ประโยค	ต้นไม้ตัดสินใจต้นเดียว	รวมต้นไม้ตัดสินใจ		โครงข่ายประสาทเทียมแบบหลายชั้น	รวมเพอร์เซปตรอน แบบหลายชั้น	
		แบ็กกิ้ง	บูสต์		แบ็กกิ้ง	บูสต์
ประโยคที่ 1	48.50	65.00	61.00	55.00	62.50	59.00
ประโยคที่ 2	45.50	67.00	63.00	49.00	58.00	55.50
ประโยคที่ 3	47.50	70.00	76.50	57.00	65.50	64.00
ประโยคที่ 4	50.00	75.00	72.00	64.50	71.50	72.00
ประโยคที่ 5	53.00	72.00	70.00	50.50	54.50	55.00
ประโยคที่ 6	51.50	72.00	73.50	67.50	63.50	63.50
ประโยคที่ 7	51.00	70.00	76.50	62.50	63.00	65.00
ประโยคที่ 8	56.50	71.00	72.50	62.50	68.00	66.00
ประโยคที่ 9	50.00	71.00	73.00	58.00	61.00	62.50
ประโยคที่ 10	56.00	73.00	76.00	59.00	64.50	61.50
อัตราเฉลี่ยค่าความถูกต้อง	50.95	70.60	71.40	58.55	63.20	62.40
ระดับนัยสำคัญ		0.01	0.01		0.01	0.01

จากตาราง 4.2 เป็นตารางแสดงผลการทดลองการจำแนกผู้ใช้ด้วยเวลาที่ใช้ในการพิมพ์จากประโยคตัวอย่างทั้ง 10 ประโยค จากผลการทดลองที่ได้อัตราค่าความถูกต้องเฉลี่ยในการจำแนกผู้ใช้ที่ได้ในคอลัมน์ของต้นไม้ตัดสินใจต้นเดียว การรวมคำตอบของต้นไม้ตัดสินใจ

หลายต้นแบบแบ็กกิง และการรวมคำตอบของต้นไม้ตัดสินใจหลายต้นแบบบูสต์นั้นให้คำตอบคิดเป็นร้อยละ 50.95, 70.60 และ 71.40 ตามลำดับ โดยพบว่าอัตราค่าความถูกต้องดังกล่าวเมื่อเปรียบเทียบกันระหว่างต้นไม้ตัดสินใจต้นเดียว กับวิธีการรวมต้นไม้ตัดสินใจแบบแบ็กกิงและบูสต์ให้ค่าความถูกต้องที่เพิ่มมากขึ้นที่ระดับนัยสำคัญเท่ากับ 0.01

จากผลการทดลองที่ได้อัตราค่าความถูกต้องเฉลี่ยในการจำแนกผู้ใช้ที่ได้ในคอลัมน์ของเพอร์เซปตรอนหลายชั้นโครงข่ายเดียว การรวมคำตอบของเพอร์เซปตรอนหลายชั้นหลายโครงข่ายแบบแบ็กกิง และการรวมคำตอบของการรวมหลายโครงข่ายแบบบูสต์นั้นให้คำตอบคิดเป็นร้อยละ 58.55, 63.20 และ 62.40ตามลำดับ โดยพบว่าอัตราค่าความถูกต้องดังกล่าวเมื่อเปรียบเทียบกันระหว่างเพอร์เซปตรอนหลายชั้นโครงข่ายเดียว กับวิธีการรวมเพอร์เซปตรอนหลายชั้นหลายโครงข่ายแบบแบ็กกิง และบูสต์ให้ค่าความถูกต้องที่มากขึ้นที่ระดับนัยสำคัญเท่ากับ 0.01

นั่นย่อมแสดงให้เห็นว่าวิธีการจำแนกผู้ใช้ด้วยวิธีต้นไม้ตัดสินใจแบบรวมตัวจำแนกนั้นสามารถจำแนกผู้ใช้ได้ดีกว่าต้นไม้ตัดสินใจต้นเดียว และในวิธีการของเพอร์เซปตรอนแบบหลายชั้นก็ให้ผลลัพธ์เช่นเดียวกัน

บทที่ 5

สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษาและวิจารณ์

งานวิจัยฉบับนี้ เป็นการเปรียบเทียบระหว่างการสร้างต้นไม้ตัดสินใจต้นเดียว โดยใช้อัลกอริทึม C4.5 และการรวมผลลัพธ์ที่ได้จากตัวจำแนกในสร้างต้นไม้ตัดสินใจหลายต้น โดยทำการสร้างรูปแบบการเรียนรู้แบบการรวมต้นไม้ตัดสินใจจำนวน 10 ต้นแล้วทำการหาคำตอบสุดท้ายเพื่อหาอัตราความถูกต้องที่ใช้ในการจำแนกผู้ใช้ด้วยเวลาที่ใช้ในการพิมพ์ และทำการเปรียบเทียบความถูกต้องกับต้นไม้ตัดสินใจต้นเดียว การรวมต้นไม้ตัดสินใจวิธีแบ็กกิง และการรวมต้นไม้ตัดสินใจวิธีเอดาบูสต์เอ็มวัน รวมไปถึงได้นำไปทำการเปรียบเทียบกับนิวรอลเน็ตเวิร์กคือโครงข่ายประสาทเทียมแบบหลายชั้น เพื่อหาอัตราความถูกต้องที่ดีที่สุดในการทดสอบ ดังนั้นผลของวิทยานิพนธ์แสดงให้เห็นว่าการนำวิธีการแบ็กกิง และวิธีเอดาบูสต์เอ็มวันสามารถนำมาประยุกต์ใช้ในการทำการจำแนกได้

ในส่วนของผลการทดลองอัตราความถูกต้องเปรียบเทียบระหว่างต้นไม้ตัดสินใจต้นเดียว วิธีแบ็กกิง และวิธีเอดาบูสต์เอ็มวัน และได้ทำการเปรียบเทียบระหว่างเพอร์เซปตรอนแบบหลายชั้น และเพอร์เซปตรอนแบบหลายชั้นที่ทำการแบ่งข้อมูลด้วยวิธีแบ็กกิง และวิธีเอดาบูสต์เอ็มวันเช่นเดียวกันกับต้นไม้ตัดสินใจนั้น จากตัวอย่างชุดข้อมูลเวลาที่ใช้ในการพิมพ์ของผู้ใช้แต่ละคนแสดงให้เห็นว่า

- อัตราความถูกต้องของตัวจำแนกด้วยต้นไม้ตัดสินใจ วิธีการของแบ็กกิง และวิธีการของเอดาบูสต์เอ็มวัน มีอัตราความถูกต้องดีกว่าต้นไม้ตัดสินใจต้นเดียว
- อัตราความถูกต้องของตัวจำแนกด้วยเพอร์เซปตรอนหลายชั้น วิธีการของแบ็กกิง และวิธีการของเอด้าบูสต์เอ็มวัน มีอัตราความถูกต้องดีกว่าเพอร์เซปตรอนแบบหลายชั้น

อัตราความถูกต้องของตัวจำแนกไม่ว่าจะเป็นต้นไม้ตัดสินใจต้นเดียว และเพอร์เซปตรอนแบบหลายชั้น ซึ่งให้อัตราความถูกต้องน้อยกว่าการจำแนกด้วยวิธีการของแบ็กกิง และบูสต์ทั้งคู่ เป็นเพราะวิธีการแบ่งชุดข้อมูลสำหรับการเรียนรู้นั้นแตกต่างกัน วิธีการของแบ็กกิง และบูสต์ใช้วิธีการสุ่มเลือกซึ่งยอมให้มีการซ้ำกันของข้อมูลได้ จากนั้นนำไปให้เครื่องทำการเรียนรู้ แล้วนำชุดข้อมูลทดสอบเข้าไปทดสอบซึ่งวิธีการที่หาคำตอบนั้นวิธีของต้นไม้ตัดสินใจคิดคำนวณหาค่าเฉลี่ยของอัตราความถูกต้องจากต้นไม้ต้นเดียว แต่ในวิธีการของแบ็กกิงนั้นใช้การลงคะแนน

เสียงข้างมากทำให้มีโอกาสได้คำตอบที่ถูกต้องมากขึ้น ส่วนวิธีการของบูสต์นั้นผลลัพธ์คำนวณจากค่าความผิดพลาดของกระบวนการสร้างรูปแบบก่อนหน้าในกระบวนการเรียนรู้ทำให้สามารถจำแนกผู้ใช้ได้ถูกต้องสูงเช่นกัน แต่สำหรับเพอร์เซปตรอนแบบหลายชั้นนั้นโดยปกติแล้วจะให้ค่าความถูกต้องได้ดีว่าต้นไม้ตัดสินใจเนื่องด้วยกระบวนการในการหาคำตอบนั้นมีความซับซ้อนมากกว่า แต่ในชุดข้อมูลนี้เมื่อนำมาเพอร์เซปตรอนแบบหลายชั้นมาทำการคำนวณหาค่าความถูกต้อง ได้ให้ผลลัพธ์ของอัตราค่าความถูกต้องน้อยกว่าต้นไม้ตัดสินใจ เนื่องมาจากข้อมูลชุดนี้ได้เลือกอัลกอริทึมของต้นไม้ตัดสินใจคือ C4.5 จึงมีการนำข้อมูลนั้นไปผ่านการหาค่าวิยุตก่อนที่จะนำมาทำการคำนวณ แต่ในนิวรอนเน็ตเวิร์กนั้นใช้การปรับค่าน้ำหนักโดยตรงจึงทำให้ได้ค่าความถูกต้องน้อยกว่า

5.2 **ข้อเสนอแนะ**

จากการทดสอบวิธีการเห็นของของต้นไม้ตัดสินใจ พบว่าสิ่งที่เป็นแนวทางการศึกษาต่อไปของวิทยานิพนธ์ มีดังต่อไปนี้

5.2.1 จากผลการทดลองจะเห็นได้ว่าวิธีการของต้นไม้ตัดสินใจไม่ได้มีความถูกต้องดีที่สุดในทุกๆ ชุดข้อมูล ในบางชุดข้อมูลวิธีต้นไม้ตัดสินใจต้นเดียวดีกว่า บางกรณีวิธีแบ็กกิง หรือวิธีเอดาบูสต์เอ็มวันได้อัตราความถูกต้องที่มากกว่า ดังนั้นควรที่จะมีการวิเคราะห์รูปแบบของข้อมูลว่ารูปแบบของข้อมูลแบบใดที่เหมาะสมกับวิธีการที่เรานำมาเลือกใช้ยกตัวอย่างเช่น กรณีที่ไม่มีค่าคุณลักษณะ จำนวนคุณลักษณะ หรือกรณีที่ค่าของคุณลักษณะเป็นค่าต่อเนื่อง เป็นต้น

5.2.2 ถ้ามีการทำการวิเคราะห์จังหวะของการพิมพ์ของผู้ใช้ที่ละเอียดลงไปถึงการค้นหาค่าของเวลาการพิมพ์ระหว่างคู่อักขระสองตัว (Bi-gram) ที่ผู้ใช้ทำการพิมพ์เพื่อทำการระบุตัวตน โดยใช้เครื่องมือต่างๆ เพื่อให้ได้ผลการวิเคราะห์ที่ดียิ่งขึ้น