บทที่ 1

บทนำ

1.1 ความเป็นมา และความสำคัญของงานวิจัย

กระบวนการโคแอกกูเลชัน (coagulation) ถือเป็นกระบวนการหนึ่งที่มีความสำคัญในการผลิตน้ำประปาซึ่งเป็นการทำให้อนุภาคคอลลอยด์ในน้ำเกิดการรวมตัวและตกตะกอนโดยการเติมสารเคมีที่เรียกว่าโคแอกกูแลนต์ (coagulant) เช่น สารส้ม ลงในน้ำดิบ โดยทั่วไปแล้ววิธีการหาปริมาณสารโคแอกกูแลนต์ที่เหมาะสมที่ทำให้เกิดการตกตะกอนได้ดีที่สุด จะใช้วิธีเก็บน้ำตัวอย่างไปวิเคราะห์ในห้องปฏิบัติการด้วยวิธีการทำจาร์เทสต์ (jar test) อย่างไรก็ตามข้อเสียของการทำจาร์เทสต์คือเป็นวิธีการที่ใช้เวลานาน ทำให้ไม่สามารถตอบสนองต่อลักษณะของน้ำได้อย่างทันท่วงที และเป็นวิธีที่ต้องใช้ผู้เชี่ยวชาญเฉพาะด้านในการดำเนินการ

ถึงแม้ในปัจจุบันมีระบบเติมสารเคมีแบบอัตโนมัติซึ่งทำการวัดประจุจากคอลลอยด์ที่ทำให้เกิดสีกับอนุภาคที่ทำให้เกิดความขุ่น โดยอนุภาคเหล่านี้จะมีประจุเป็นลบ ในขณะที่สารเคมีที่ใช้เติมลงในน้ำมีประจุเป็นบวก ปริมาณสารโคแอกกูแลนต์ที่เติมลงในน้ำจะวัดจากเครื่องหมายรวมของประจุ โดยหากประจุรวมเป็นลบก็จะทำการเติมสารเคมี และหากว่าประจุรวมเป็นศูนย์หรือบวกก็จะหยุดเติมสารเคมี แต่วิธีการนี้ก็มีข้อเสียนั่นคือมีค่าอุปกรณ์ และค่าดำเนินการที่ค่อนข้างสูง

งานวิจัยนี้จึงได้นำวิธีด้านการเรียนรู้ของเครื่อง (machine learning) มาใช้ในการทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปาของสำนักงานประปาสาขาบางเขน การประปานครหลวง โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) ต้นไม้ตัดสินใจ (decision tree) และ ป่าของต้นไม้ตัดสินใจ (decision tree forest) เปรียบเทียบกับวิธีการแบบนิวรอลเน็ตเวิร์ก (neural networks) ที่ถูกนำมาใช้ในงานวิจัยส่วนใหญ่เพื่อทำนายปริมาณการเติมสารโคแอกกูแลนต์ สำหรับข้อมูลที่นำมาใช้ในการฝึกสอนให้เครื่องเกิดการเรียนรู้เป็นข้อมูลจากสำนักงานประปาสาขาบางเขน การประปานครหลวง ที่เก็บในช่วงระยะเวลา 1 ปี ตั้งแต่วันที่ 1 มกราคม 2549 ถึง 31 ธันวาคม 2549 โดยในการทดลองส่วนแรก ตัวแปรอินพุตที่นำมาใช้เป็นค่าคุณลักษณะของน้ำซึ่งมีความสัมพันธ์กับปริมาณการเติมสารส้ม ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน การทดลองส่วนที่สองทำการปรับเพิ่มตัวแปรอินพุตเพื่อช่วยในการทำนาย สำหรับการเพิ่มตัวแปรอินพุตในงานวิจัยนี้ได้

เสนอเทคนิคในการสร้างตัวแปรอินพุตเพิ่มเติมด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมเพื่อใช้ในการทำนายร่วมกับวิธีการทำนายแบบต่างๆ การทดลองส่วนที่สามทำการลดตัวแปรอินพุตโดยใช้เฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม ได้แก่ ความขุ่น พีเอช และความเป็นด่าง ทั้งนี้เพื่อศึกษาถึงผลของตัวแปรอินพุตหลักๆ ที่มีต่อการทำนายปริมาณการเติมสารส้ม

ผลการทดลองที่มีความแม่นยำในระดับที่น่าพอใจจะสามารถนำไปใช้ในการทำนายปริมาณสารส้มได้ เป็นการช่วยประหยัดเวลา และค่าใช้จ่ายได้จำนวนมาก รวมทั้งสามารถตอบสนองต่อลักษณะของน้ำดิบได้อย่างรวดเร็ว และยังสามารถนำมาใช้สนับสนุนและตรวจสอบผลการทำจาร์เทสต์ได้อีกด้วย

1.2 วัตถุประสงค์

1.2.1 เพื่อนำความรู้ด้านการเรียนรู้ของเครื่องมาประยุกต์ใช้กับกระบวนการผลิตน้ำประปา

1.2.2 เพื่อหาวิธีในการทำนายปริมาณการเติมสารส้มที่แม่นยำที่สุด

1.3 ขอบเขตของงานวิจัย

งานวิจัยนี้เป็นการนำนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ มาทำการทดลองเปรียบเทียบ โดยทำการปรับค่าพารามิเตอร์ รูปแบบของวิธีการทำนายต่างๆ และตัวแปรอินพุตที่ใช้ เพื่อหาวิธีการทำนายที่สามารถทำนายปริมาณการเติมสารส้มในน้ำดิบได้แม่นยำที่สุดบนเซตตัวอย่างซึ่งเป็นข้อมูลจริงจากสำนักงานประปาสาขาบางเขน การประปานครหลวง ที่เก็บในช่วงระยะเวลา 1 ปี ตั้งแต่วันที่ 1 มกราคม 2549 ถึง 31 ธันวาคม 2549 การทดลองแบ่งออกเป็นสามส่วนตามรูปแบบตัวแปรอินพุตที่นำมาใช้ ส่วนที่หนึ่งตัวแปรอินพุตที่นำมาใช้เป็นค่าคุณลักษณะของน้ำซึ่งมีความสัมพันธ์กับปริมาณการเติมสารส้ม ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน การทดลองส่วนที่สองทำการปรับเพิ่มตัวแปรอินพุตเพื่อช่วยในการทำนาย สำหรับการเพิ่มตัวแปรอินพุตในงานวิจัยนี้ได้เสนอเทคนิคในการสร้างตัวแปรอินพุตเพิ่มเติมด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมเพื่อใช้ในการทำนายร่วมกับวิธีการทำนายแบบต่างๆ การทดลองส่วนที่สามทำการ

ลดตัวแปรอินพุตโดยใช้เฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม ทั้งนี้เพื่อศึกษาถึงผลของตัวแปรอินพุตหลักๆ ที่มีต่อการทำนายปริมาณการเติมสารส้ม

1.4 ผลที่คาดว่าจะได้รับ

1.4.1 ได้วิธีในการทำนายปริมาณการเติมสารส้มที่มีความแม่นยำมากที่สุด

1.4.2 ได้ค่าของตัวแปรที่เหมาะสมที่ทำให้สามารถทำนายปริมาณส้มได้แม่นยำที่สุด

1.4.3 ได้ตัวแปรอินพุตที่เหมาะสมสำหรับใช้ในการทำนายปริมาณสารส้ม

1.5 รายละเอียดของวิทยานิพนธ์

ในส่วนของวิทยานิพนธ์แบ่งออกเป็นส่วนต่างๆ ดังนี้

บทที่1 บทนำ ประกอบด้วย ความเป็นมาและความสำคัญของงานวิจัยนี้ วัตถุประสงค์ ขอบเขตของงานวิจัย ผลที่คาดว่าจะได้รับ และรายละเอียดของวิทยานิพนธ์

บทที่ 2 ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ประกอบด้วย ความรู้พื้นฐานที่เกี่ยวข้องกับวิทยานิพนธ์ฉบับนี้ ทฤษฎีนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ ตัวอย่างงานวิจัยที่มีการนำวิธีในการทำนายต่างๆ ไปใช้ในการแก้ปัญหาแบบถดถอย และทฤษฎีของโปรแกรมเชิงพันธุกรรม

บทที่ 3 การดำเนินการวิจัย ประกอบไปด้วย ลักษณะของข้อมูลที่นำมาใช้ในการทดลอง การจัดเตรียมข้อมูลเพื่อใช้ในการทดลอง การจัดเตรียมนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ สำหรับทำการทดลอง การสร้างข้อมูลอินพุตเพิ่มเติมด้วยโปรแกรมเชิงพันธุกรรม การเลือกใช้เฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม วิธีการคำนวณหาความแม่นยำของวิธีการทำนาย และแนวทางการทดลอง

บทที่ 4 ผลการทดลอง ประกอบไปด้วย ผลการทดลองของการทำนายด้วยนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ ป่าของต้นไม้ตัดสินใจ ซึ่งแบ่งออกเป็นสามส่วน ส่วนที่หนึ่งเป็นผลการทดลองของกรณีอินพุตแบบ 7 ตัว ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน ส่วนที่สองเป็นผลการทดลองของกรณีอินพุตแบบ 8 ตัว ซึ่งอินพุตที่เพิ่มเติมขึ้นมาเป็นตัวแปรอินพุตที่ได้จากการประยุกต์ใช้โปรแกรม

เชิงพันธุกรรม ส่วนที่สามเป็นผลการทดลองของกรณีอินพุต 3 ตัว ซึ่งตัวแปรอินพุตที่ใช้เป็นตัวแปรอินพุตหลักที่มีผลต่อปริมาณการเติมสารส้ม ได้แก่ ความขุ่น พีเอช และความเป็นด่าง

บทที่ 5 สรุปผลการศึกษา วิเคราะห์ และข้อเสนอแนะ ประกอบไปด้วย สรุปผลการศึกษา วิเคราะห์ผลการทดลอง และข้อเสนอแนะ

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึง กระบวนการผลิตน้ำประปา และกระบวนการโคแอกกูเลชันซึ่งเป็นกระบวนการที่สำคัญกระบวนการหนึ่งในการผลิตน้ำประปา การทำจาร์เทสต์ซึ่งเป็นการหาปริมาณการเติมสารส้มในปัจจุบันที่นิยมใช้กันโดยทั่วไป การเลือกตัวแปรอินพุตเพื่อใช้ในการทดลอง ทฤษฎีนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ ตัวอย่างงานวิจัยที่มีการนำวิธีในการทำนายต่างๆ ไปใช้ในการแก้ปัญหาแบบถดถอย และทฤษฎีของโปรแกรมเชิงพันธุกรรม

2.1 กระบวนการโคแอกกูเลชันในกระบวนการผลิตน้ำประปา

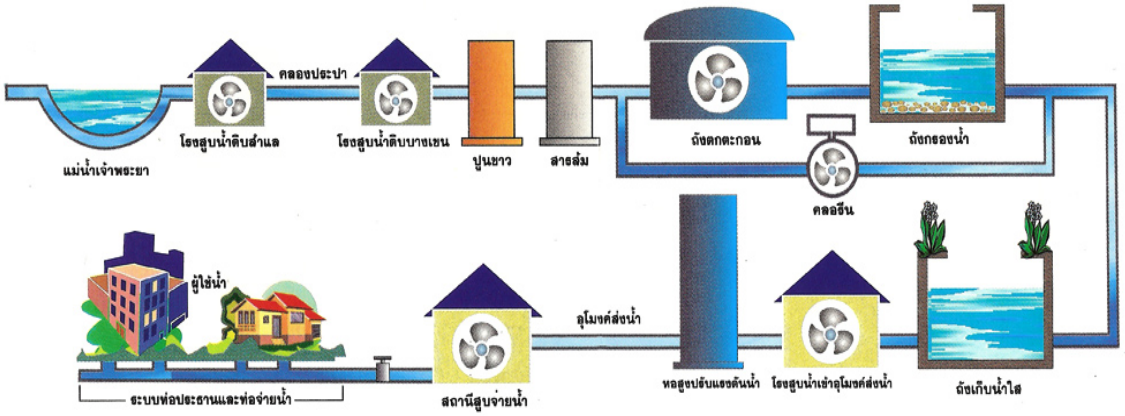
กระบวนการโคแอกกูเลชันเป็นกระบวนการที่สำคัญกระบวนการหนึ่งในการผลิตน้ำประปา ซึ่งในที่นี้จะอธิบายถึงกระบวนการและขั้นตอนการผลิตน้ำประปาของสำนักงานประปาสาขาบางเขน การประปานครหลวง หรือโรงงานผลิตน้ำประปาบางเขนเป็นหลัก

2.1.1 โรงงานผลิตน้ำประปาบางเขน

โรงงานผลิตน้ำประปาบางเขนมีพื้นที่ 1.136 ตารางกิโลเมตร หรือ 710 ไร่ ตั้งอยู่ที่แขวงทุ่งสองห้อง เขตหลักสี่ ปัจจุบันสามารถผลิตน้ำประปาได้สูงสุดวันละ 3,600,000 ลูกบาศก์เมตร ซึ่งในอนาคตสามารถเพิ่มปริมาณการผลิตได้สูงสุดถึง 4,800,000 ลูกบาศก์เมตรต่อวัน เพื่อรองรับปริมาณความต้องการใช้น้ำประปาที่เพิ่มมากขึ้น ถือเป็นโรงงานผลิตน้ำประปาที่ใหญ่ที่สุดในเขตกรุงเทพมหานคร โดยผลิตน้ำประปาเพื่อบริการเขตพื้นที่จังหวัดกรุงเทพมหานคร นนทบุรี และสมุทรปราการ

ภาพที่ 2.1

ไดอะแกรมกระบวนการผลิตน้ำประปาของโรงงานผลิตน้ำประปาบางเขน



ที่มา : แผ่นพับประชาสัมพันธ์โรงงานผลิตน้ำประปาบางเขน

2.1.2 กระบวนการผลิตน้ำประปา

กระบวนการผลิตน้ำประปาของโรงงานผลิตน้ำประปาบางเขนแบ่งออกเป็น 7 ขั้นตอน แสดงดังภาพที่ 1 ด้วยกัน ได้แก่ การปรับปรุงคุณภาพน้ำดิบ การเติมสารเคมี การตกตะกอน การกรอง การฆ่าเชื้อโรค การปรับปรุงคุณภาพน้ำประปา การสูบจ่ายน้ำประปา

1. การปรับปรุงคุณภาพน้ำดิบ น้ำดิบที่นำมาใช้ในการผลิตเป็นน้ำดิบจากแม่น้ำเจ้าพระยาที่สูบขึ้นมาจากสถานีสูบน้ำดิบสำแล ต.บางกระแชง อ.เมือง จ.ปทุมธานี น้ำดิบจะถูกส่งมาตามคลองประปาเป็นระยะทางกว่า 18 กิโลเมตร ซึ่งตะกอนในน้ำดิบบางส่วนจะเกิดการตกตะกอนตามธรรมชาติ ทำให้คุณภาพน้ำดีขึ้นและก่อนที่จะสูบน้ำดิบขึ้นมาเพื่อเข้ากระบวนการถัดไป จะมีการกำจัดเศษขยะ เศษไม้ สาหร่าย และวัสดุต่างๆ ที่ไม่ต้องการไม่ให้เข้าสู่ระบบ ด้วยตะแกรงหยาบ และตะแกรงละเอียด ซึ่งจะกั้นไว้หน้าสถานีสูบน้ำดิบ
2. การเติมสารเคมี ระหว่างที่น้ำดิบถูกลำเลียงไปยังถังตกตะกอนจะมีการเติมสารเคมีในท่อ ได้แก่ ปูนขาว (lime) สารส้ม และคลอรีน โดนปูนขาวจะปรับสภาพให้น้ำดิบมีความเป็นด่าง ทำให้สารส้มทำปฏิกิริยาในกระบวนการโคแอกกูเลชันได้ดีขึ้น ส่วนคลอรีนช่วยกำจัดตะไคร่ กลิ่น และสี ในน้ำดิบซึ่งเป็นกระบวนการที่เรียกว่าการเติมคลอรีนก่อนบำบัด (pre-chlorination)

3. การตกตะกอน หลังจากเติมสารเคมี น้ำดิบจะถูกนำเข้าสู่ถังตกตะกอน (clarifier) สารเคมีจะถูกกวนให้สัมผัสและทำปฏิกิริยากับตะกอนและอนุภาคคอลลอยด์ ทำให้เกิดกระบวนการโคแอกกูเลชันจนเหลือแต่น้ำใสไหลไปยังบ่อกรอง (filter) ระยะเวลาที่ใช้ในการตกตะกอนจะใช้เวลาประมาณ 2 ชั่วโมง ความขุ่นของน้ำที่ออกจากถังตกตะกอนจะมีค่าความขุ่นไม่เกิน 5 NTU (Nephelometric Turbidity Unit) ทั้งนี้ปริมาณการเติมสารส้มขึ้นกับคุณภาพของน้ำดิบ ซึ่งในแต่ละฤดูกาลก็จะมีคุณภาพที่แตกต่างกันออกไป สำหรับวิธีการหาปริมาณสารส้มที่เหมาะสมใช้วิธีการทำจาร์เทสต์
4. การกรอง น้ำที่ผ่านการตกตะกอนจะถูกส่งมายังบ่อกรองน้ำซึ่งใช้ผงถ่านแอนทราไซด์และทรายกรองเป็นสารกรอง มีหัวกรอง (nozzle) เพื่อกรองเอาตะกอนที่ละเอียดออก น้ำที่ผ่านการกรองจะมีความขุ่นไม่เกิน 2 NTU
5. การฆ่าเชื้อโรค น้ำที่ผ่านกระบวนการต่างๆ ดังกล่าวข้างต้นอาจยังมีแบคทีเรียหลงเหลืออยู่ ดังนั้นจึงต้องมีการฆ่าเชื้อโรคด้วยการเติมคลอรีนหลังบำบัด (post-chlorination) ซึ่งคลอรีนสามารถฆ่าเชื้อโรคได้เกือบทุกชนิด และสามารถกำจัดสารอินทรีย์ กลิ่น และสีได้ ทั้งนี้ในการเติมควรให้มีคลอรีนหลงเหลือ (free residual chlorine) เพื่อติดไปกับน้ำเพื่อฆ่าเชื้อโรคที่อาจเข้ามาปนเปื้อนน้ำในภายหลัง
6. การปรับปรุงคุณภาพน้ำประปา จะมีการเติมปูนขาวหลังบำบัด (post-lime) อีกเล็กน้อย เพื่อปรับพีเอชของน้ำให้มีฤทธิ์เป็นด่างเล็กน้อยเพื่อป้องกันการกัดกร่อนท่อประปาที่ใช้ในการลำเลียงน้ำ
7. การสูบจ่ายน้ำประปา น้ำประปาที่ผลิตได้จะถูกเก็บไว้ที่ถังเก็บน้ำใส หลังจากนั้นจะถูกสูบเข้าหอสูงปรับแรงดันน้ำ และส่งเข้าอุโมงค์ส่งน้ำและท่อส่งน้ำขนาดใหญ่ไปยังสถานีสูบจ่ายในย่านชุมชนต่างๆ เพื่อสูบจ่ายน้ำเข้าเส้นประธานและเส้นท่อจ่ายน้ำ เพื่อให้ประชาชนใช้อุปโภคบริโภคต่อไป

2.1.3 กระบวนการโคแอกกูเลชัน

อนุภาคคอลลอยด์ (colloidal particles) เป็นอนุภาคที่มีขนาดเล็กขนาดประมาณ 10^{-9} ถึง 10^{-6} เมตร มีคุณลักษณะคือสามารถลอดผ่านกระดาษกรองได้ แต่ไม่สามารถลอดผ่านกระดาษเซลโลเฟนได้ เมื่อผ่านลำแสงผ่านเข้าไปในคอลลอยด์ จะเกิดการกระเจิงของแสง ทำให้มองเห็นลำแสงได้อย่างชัดเจน เรียกว่าปรากฏการณ์ทินดอลล์ (Tyndall effect) อนุภาคคอลลอยด์เป็น

สาเหตุหนึ่งที่ทำให้น้ำขุ่นและสกปรกดังนั้นในกระบวนการผลิตน้ำประปาจึงจำเป็นต้องกำจัดคอลลอยด์ทิ้ง แต่เนื่องจากคอลลอยด์เป็นอนุภาคที่มีขนาดเล็ก การแยกอนุภาคออกจากน้ำด้วยวิธีการปล่อยให้ตกตะกอน และการกรองนั้นไม่สามารถกำจัดคอลลอยด์ได้ทั้งหมด

กระบวนการโคแอกกูเลชันเป็นกระบวนการที่ทำให้อนุภาคคอลลอยด์รวมตัวกันจนมีขนาดใหญ่ และเกิดการตกตะกอน เป็นวิธีที่ใช้ในการควบคุมปริมาณของโคแอกกูแลนต์และพีเอชให้เหมาะสม เพื่อให้เกิดกระบวนการโคแอกกูเลชันที่ดีที่สุด

2.1.4 สารเคมีที่นิยมใช้ในกระบวนการโคแอกกูเลชัน

สารเคมีที่ใช้ในกระบวนการโคแอกกูเลชัน เรียกว่า โคแอกกูแลนต์ ซึ่งมีอยู่หลายชนิดด้วยกัน เช่น พอลิอะลูมิเนียมคลอไรด์ (Poly Aluminum Chloride, PAC) พอลิอะลูมิเนียมซัลเฟต-ซิลิเกต (Poly Aluminum Sulfate Silicate, PASS) พอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟต (Poly Organic Aluminum Magnesium Sulfate, PSO-M) แต่สำหรับที่นิยมใช้กันโดยทั่วไปได้แก่ สารส้ม ซึ่งมีสูตรทางเคมีดังนี้ $Al_2(SO_4)_3 14.3H_2O$ (Aluminium Sulfate)

2.1.5 การทำจาร์เทสต์

จาร์เทสต์เป็นวิธีการที่นิยมมากที่สุดในการหาปริมาณโคแอกกูแลนต์ที่เหมาะสม กระบวนการทำจาร์เทสต์ประกอบด้วยบีคเกอร์ (beaker) 6 ใบ ซึ่งบรรจุตัวอย่างน้ำดิบ จากนั้นจะนำมาทดลองเติมสารโคแอกกูแลนต์ในปริมาณต่างๆ กัน แล้วกวนด้วยเครื่องซึ่งสามารถปรับความเร็วใบพัดได้ แล้วใช้วิธีการสังเกตดูว่าฟล็อกเกิดขึ้นได้ดีที่ความเข้มข้นเท่าใด จากนั้นจึงทำการทดลองใหม่อีกรอบโดยนำน้ำดิบมาปรับค่าพีเอชให้ต่างกันออกไป แล้วจึงเติมสารโคแอกกู-แลนต์ตามค่าที่หาไว้ในครั้งแรก ปรับใบพัดให้หมุนเร็วในช่วงแรก และลดความเร็วลงในช่วงหลัง เพื่อให้เกิดการจับตัวเป็นก้อนและตกตะกอน ทำการเปรียบเทียบผลของบีคเกอร์แต่ละใบ สุดท้ายก็จะทราบปริมาณสารโคแอกกูแลนต์ และค่าพีเอชที่เหมาะสมที่สุดสำหรับกระบวนการทำให้ตกตะกอน

2.1.6 การเลือกตัวแปรอินพุต

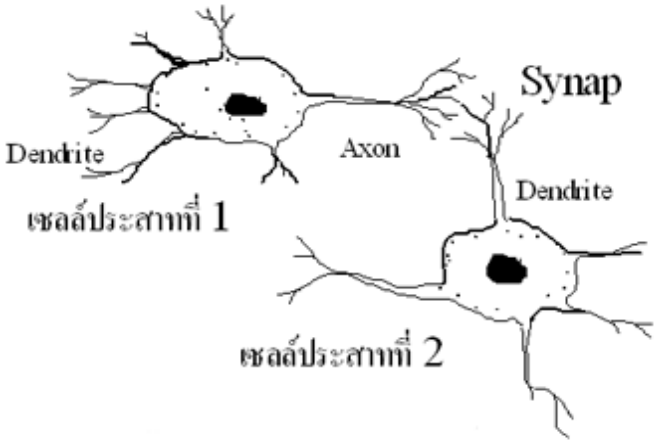
สำหรับการเลือกตัวแปรที่มีความสัมพันธ์กับปริมาณการเติมโคแอกกูแลนต์ จากงานวิจัยของ Mirsepaassi (1995) Han (1997) Chun (1999) Valentin (1999) และ Bae (2004) พบว่าตัวแปรหลักที่มีผลต่อปริมาณการเติมโคแอกกูแลนต์นั้นได้แก่ ความขุ่น ความเป็นด่าง อุณหภูมิ และพีเอช แต่การจะหาตัวแปรทั้งหมดที่มีผลต่อปริมาณการเติมโคแอกกูแลนต์นั้นเป็นสิ่งที่ไม่สามารถชี้ชัดลงไปทั้งหมด ดังนั้นในงานวิจัยนี้จึงได้เลือกตัวแปรที่มีความสัมพันธ์หลักๆ รวมกับตัวแปรอื่นๆ ที่คาดว่าจะมีผลต่อปริมาณการเติมสารส้ม และเป็นตัวแปรที่โรงงานผลิตน้ำประปาบางเขนมีการตรวจวัดและจัดเก็บข้อมูลไว้รวมทั้งสิ้น 7 ตัวแปร ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน โดยตัวแปรทั้ง 7 ตัวนี้จะถูกนำไปใช้ในการทดลองส่วนที่หนึ่ง ในการทดลองส่วนที่สองจะทำการปรับเพิ่มตัวแปรอินพุตเพิ่มเติมด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม และการทดลองส่วนที่สามทำการลดตัวแปรอินพุตโดยใช้เฉพาะตัวแปรอินพุตหลักๆ จำนวน 3 ตัว ได้แก่ ความขุ่น พีเอช และความเป็นด่าง โดยขาดค่าอุณหภูมิของน้ำ เนื่องจากไม่มีการจดบันทึกค่าไว้

2.2 นิวรอลเน็ตเวิร์ก

นิวรอลเน็ตเวิร์กหรือโครงข่ายประสาทเทียมเป็นศาสตร์แขนงหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence, AI) ซึ่งเลียนแบบการทำงานของเซลล์สมองหรือระบบประสาทของมนุษย์ สำหรับระบบประสาทมีลักษณะดังภาพที่ 2.2 โดยมีส่วนปลายประสาทรับรู้สัญญาณไฟฟ้า (Dendrite) ตัวรวบรวมข้อมูลที่ได้จากปลายประสาทรับรู้สัญญาณไฟฟ้า (Soma หรือ cell body) และท่อส่งข้อมูลเพื่อไปยังปลายประสาทรับรู้สัญญาณไฟฟ้าตัวอื่นๆ (Axon) โดย ส่วนปลายประสาทรับรู้สัญญาณไฟฟ้าเปรียบเสมือนอินพุต (input) ที่ใช้ในการรับข้อมูลเข้ามา ตัวรวบรวมข้อมูลที่ได้จากปลายประสาทรับรู้สัญญาณไฟฟ้าเปรียบเสมือนตัวรวบรวมน้ำหนักของสัญญาณไฟฟ้าที่ได้มาจากปลายประสาทรับรู้สัญญาณไฟฟ้าหลายๆ ตัวเข้ามา และท่อส่งข้อมูลเพื่อไปยังปลายประสาทรับรู้สัญญาณไฟฟ้าตัวอื่นๆ เปรียบเสมือนเอาต์พุต (output) ที่จะส่งผลลัพธ์ที่ได้ไปยังนิวรอนตัวอื่นๆ ในโครงข่ายต่อไป

ภาพที่ 2.2

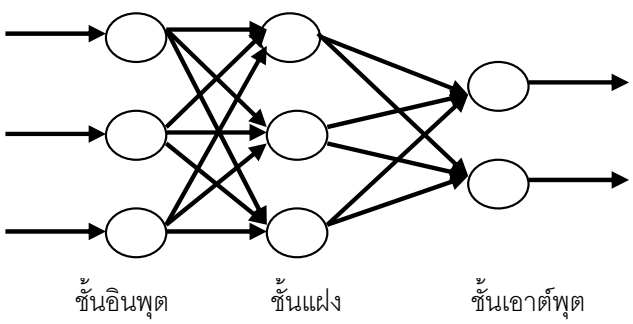
ระบบประสาท



ถึงแม้นิวรอลเน็ตเวิร์กจะมีความแตกต่างจากระบบประสาทในสมองทั้งในแง่ของโครงสร้าง ความซับซ้อน และปริมาณของหน่วยย่อย แต่อย่างไรก็ดีก็มีความเหมือนกัน ในแง่ที่นิวรอลเน็ตเวิร์ก คือการรวมกลุ่มแบบขนานของหน่วยประมวลผลย่อยๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญา รวมทั้งลักษณะการทำงานของนิวรอลเน็ตเวิร์กก็มีความใกล้เคียงกับเซลล์สมองหรือระบบประสาทโดยเฉพาะของมนุษย์ สำหรับแหล่งข้อมูลที่นำมาใช้เพื่อให้ระบบนิวรอลเน็ตเวิร์กเกิดการรู้จำอาจนำมาจากหลายแหล่ง ทั้งจากข้อมูลปัจจุบัน ข้อมูลในอดีต หรือข้อมูลที่ได้จากแบบจำลอง

ภาพที่ 2.3

โครงสร้างของนิวรอลเน็ตเวิร์ก



2.2.1 นิวรอลเน็ตเวิร์กแบบหลายชั้น (Multilayers Neural Networks)

นิวรอลเน็ตเวิร์กสามารถนำมาต่อกันเป็นโครงข่ายหลายชั้นโดยแต่ละชั้นมีหน่วยประสาทเทียมหรือโหนดมากกว่าหนึ่งตัวได้ สัญลักษณ์ของนิวรอลเน็ตเวิร์กแบบหลายชั้นสามารถแสดงได้ดังภาพที่ 2.3 โดยชั้นที่รับข้อมูลเข้ามาเรียกว่าชั้นอินพุต (input layer) ส่วนชั้นสุดท้ายที่ส่งข้อมูลออกไปเป็นผลลัพธ์เรียกว่าชั้นเอาต์พุต (output layer) และชั้นที่อยู่ระหว่างชั้นอินพุตและชั้นเอาต์พุตเรียกว่าชั้นแฝง (hidden layer) จำนวนโหนดของแต่ละชั้นไม่จำเป็นต้องเท่ากัน ฟังก์ชันการแปลงถ่ายทอดของแต่ละชั้นก็ไม่จำเป็นต้องเหมือนกัน นิวรอลเน็ตเวิร์กดังภาพที่ 2.3 เป็นแบบส่งผ่านข้อมูลไปข้างหน้า (feed-forward neural networks) ส่วนนิวรอลเน็ตเวิร์กที่มีการนำข้อมูลเอาต์พุตย้อนกลับมาเป็นข้อมูลอินพุตเรียกว่านิวรอลเน็ตเวิร์กแบบรีเคอร์เรนต์ (recurrent neural networks)

2.2.2 อัลกอริทึมแบ็กพรอพาเกชัน (Backpropagation Algorithm)

อัลกอริทึมแบ็กพรอพาเกชันเป็นอัลกอริทึมที่ใช้ในการเรียนรู้ของนิวรอลเน็ตเวิร์กวิธีหนึ่งที่นิยมใช้ในนิวรอลเน็ตเวิร์กแบบหลายชั้น ข้อมูลจากชั้นอินพุตจะถูกคำนวณและส่งผ่านฟังก์ชันจากชั้นแฝงไปยังชั้นเอาต์พุต ซึ่งหลักการสำคัญของการเรียนรู้คือ การเปลี่ยนแปลงค่าน้ำหนักของแต่ละเส้นเชื่อมระหว่างโหนด โดยในการปรับแก้ค่าน้ำหนักจะขึ้นกับความแตกต่างระหว่างค่าเอาต์พุตที่คำนวณได้กับค่าเอาต์พุตที่ต้องการ สำหรับขั้นตอนในการปรับแก้ค่าน้ำหนักมีขั้นตอนดังต่อไปนี้

1. กำหนดค่าอัตราการเรียนรู้ และค่าโมเมนตัม
2. ปรับแก้ค่าน้ำหนักของแต่ละเส้นเชื่อมระหว่างโหนดตามสมการที่ (2.1)

$$\Delta w_{ji}\left(n+1\right)=\eta\delta\left(n\right)\cdot y_{j}\left(n\right)+\alpha\Delta w_{ji}\left(n\right) \qquad (2.1)$$

โดยที่

x_i คือ ค่าข้อมูลด้านเข้าที่โหนด i

w_i คือ ค่าถ่วงน้ำหนักที่โหนด i

Δw_{ji} คือ ค่าปรับแก้ค่าถ่วงน้ำหนักระหว่างโหนด i และ j

η คือ ค่าอัตราการเรียนรู้

α คือ ค่าโมเมนตัม

δ_j คือ ค่าผลต่างระหว่างค่าจริงกับค่าที่ได้จากการคำนวณในรูปของอนุพันธ์ของฟังก์ชันถ่ายโอน (transfer function) ของโหนด j

y_i คือ ค่าผลลัพธ์ของแบบจำลองที่โหนด j

$n,\ n+1$ คือ ค่าที่แสดงถึงรอบของการปรับแก้ที่ n หรือ $n+1$

2.3 ซัพพอร์ตเวกเตอร์แมชชีน

2.3.1 การจำแนกเชิงเส้น (Linear Classification)

การจำแนกเชิงเส้นไม่ได้ถูกจำกัดด้วยลักษณะของข้อมูล กล่าวคือถ้าข้อมูลสามารถแสดงได้ในรูปของเวกเตอร์ n มิติ ข้อมูลนั้นจะสามารถพล็อตและถูกจำแนก (classified) ได้ด้วยไฮเปอร์เพลน n-1 มิติ ตัวอย่างเช่นเมืองที่อยู่บนตาราง 2 มิติ จะสามารถถูกแบ่งได้ด้วยเส้นตรง 1 มิติ เป็นต้น

การจำแนกเชิงเส้นนั้นสามารถทำได้หลายแบบ แต่การแบ่งที่มีระยะห่างระหว่าง 2 กลุ่มมากที่สุด หรือที่เรียกว่าขอบสูงสุด (maximum margin) คือสิ่งที่เราสนใจ

ถ้าข้อมูลที่นำมาใช้ในการเรียนรู้สามารถแบ่งได้เชิงเส้นโดยไม่มีข้อมูลใดอยู่ผิดระหว่างเส้นแบ่งเลย เราสามารถหาระยะระหว่างไฮเปอร์เพลน (hyperplane) ได้คือ 2|w| ซึ่งเราจะทำการเพิ่มค่า |w| ให้สูงสุด

2.3.2 ขอบแบบอ่อน (Soft Margin)

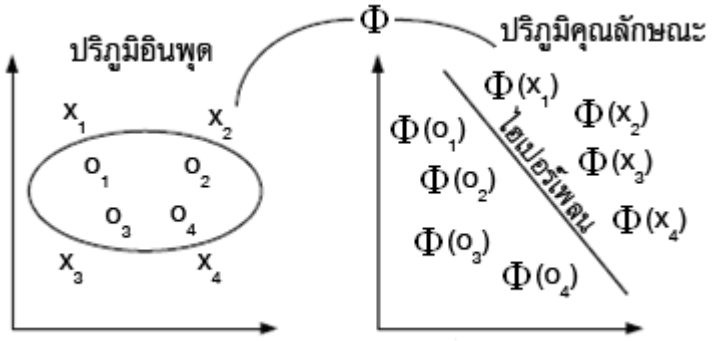
Cortes และ Vapnik (1995) ได้เสนอแนวทางในการปรับปรุงขอบสูงสุดซึ่งอนุญาตให้มีข้อมูลที่ถูกแบ่งผิดประเภทได้ ในกรณีที่ไม่มีไฮเปอร์เพลนใดๆ สามารถแบ่งแยกข้อมูลทั้งหมดได้อย่างถูกต้องทั้งหมด ซึ่งวิธีการนี้ถูกเรียกว่าขอบแบบอ่อน โดยวิธีการนี้จะทำการเลือกไฮเปอร์เพลนที่สามารถจำแนกประเภทของข้อมูลได้ถูกต้องมากที่สุดเท่าที่เป็นไปได้ โดยอนุญาตให้มีข้อมูลที่ถูกจำแนกผิดประเภทได้ ในขณะที่ยังคงเลือกระยะห่างของการจำแนกข้อมูลที่มากที่สุด วิธีการนี้ถูก

เรียกว่าซัพพอร์ตเวกเตอร์แมชชีน วิธีการนี้มีการใช้ตัวแปรสแล็ค (slack variable) ξ_i เพื่อวัดระดับความผิดพลาดในการจำแนกข้อมูล (degree of misclassification)

2.3.3 ตัวจำแนกแบบไม่เป็นเชิงเส้น (Non-Linear Classifiers)

ในปี 1963 Vladimir Vapnik ได้เสนอต้นแบบของอัลกอริทึมที่สามารถหาไฮเปอร์เพลนที่ดีที่สุดที่เป็นตัวจำแนกเชิงเส้น (linear classifier) ซึ่งต่อมาในปี 1992 Bernard Boser, Isabelle Guyon และ Vapnik ได้เสนอวิธีการสร้างตัวจำแนกแบบไม่เป็นเชิงเส้น โดยการประยุกต์ใช้เคอร์-เนล (kernel) (ต้นแบบของเคอร์เนลนำเสนอครั้งแรกโดย Aizerman (1964)) กับไฮเปอร์เพลนขอบสูงสุด (maximum margin hyperplane) โดยอัลกอริทึมที่ใช้นั้นเหมือนกันยกเว้นทุกๆ การคูณเวกเตอร์จะถูกแทนที่โดยฟังก์ชันเคอร์เนลแบบไม่เป็นเชิงเส้น (non-linear kernel function)

ภาพที่ 2.4
การจับคู่ข้อมูลจากปริภูมิอินพุตไปยังปริภูมิคุณลักษณะโดยใช้เคอร์เนล



ในการทำงานซัพพอร์ตเวกเตอร์จะถูกจับคู่ (map) จากปริภูมิอินพุต (input space) ไปยังปริภูมิคุณลักษณะ (feature space) ดังภาพที่ 2.4

2.3.4 ฟังก์ชันเคอร์เนล

ฟังก์ชันเคอร์เนลที่นิยมใช้ ได้แก่

พหุนาม (เอกพันธุ์)
polynomial (homogeneous) $k(X, X') = (X \bullet X')^d$

พหุนาม (ไม่เอกพันธุ์)
polynomial (inhomogeneous) $k(X, X') = (X \bullet X' + 1)^d$

ฟังก์ชันเรเดียลเบซิส
(radial basis function) $k(X, X') = \exp(-\gamma \| X - X' \|^2)$

ฟังก์ชันเรเดียลเบซิสแบบเกาส์
(gaussian radial basis function) $k(X, X') = \exp\left(-\frac{\| X - X' \|^2}{2\sigma^2}\right)$

ซิกมอยด์ (sigmoid) $k(X, X') = \tanh(kX \bullet X' + c)$

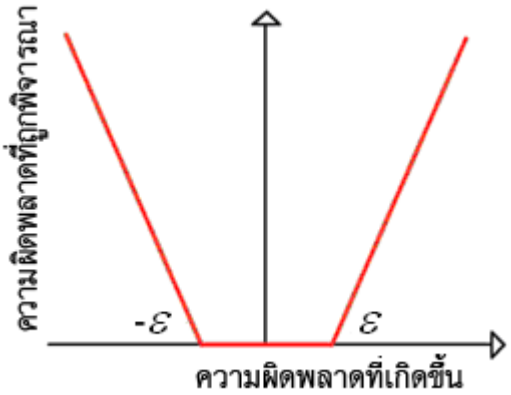
2.3.5 ซัพพอร์ตเวกเตอร์แบบถดถอย (Support Vector Regression, SVR)

ซัพพอร์ตเวกเตอร์แมชชีนสำหรับใช้แก้ปัญหาแบบถดถอย ถูกเสนอครั้งแรกโดย Vapnik, Drucker, Burges, Kaufman และ Smola (1997) ซึ่งถูกเรียกว่าซัพพอร์ตเวกเตอร์แบบถดถอย

ซัพพอร์ตเวกเตอร์แบบถดถอยมีหลักการสำคัญในการทำงานเหมือนกับซัพพอร์ตเวกเตอร์แมชชีน นั่นคือมีการลดความผิดพลาดให้น้อยที่สุด (minimizing error) จำแนกข้อมูลด้วยไฮเปอร์เพลนที่มีระยะขอบมากที่สุด (maximizing the margin) และยอมให้มีความผิดพลาดได้บางส่วน สำหรับสิ่งที่แตกต่างระหว่างซัพพอร์ตเวกเตอร์แบบถดถอยและซัพพอร์ตเวกเตอร์แมชชีน คือ เอาต์พุทของซัพพอร์ตเวกเตอร์แบบถดถอยมีค่าเป็นจำนวนจริง ในการที่จะนำซัพพอร์ตเวกเตอร์แมชชีนมาประยุกต์ใช้กับปัญหาแบบถดถอยสามารถทำได้โดยการเปลี่ยนแปลงฟังก์ชันสูญเสีย (loss function) ใหม่ สำหรับตัวอย่างของฟังก์ชันสูญเสียเช่น ฟังก์ชันสูญเสียแบบ ε-insensitive (ε-insensitive loss function) ซึ่งเสนอโดย Vapnik (1995) ดังภาพที่ 2.5

ภาพที่ 2.5

กราฟแสดงฟังก์ชันสูญเสียแบบ ε -insensitive



ให้ชุดข้อมูลสำหรับเรียนรู้คือ $(x_1, y_1),...,(x_N, y_N)$ โดยที่ $x_i \in R^n, y_i \in R$ ซึ่ง x_i คือเวกเตอร์อินพุต y_i คือค่าที่ต้องการ N คือขนาดของข้อมูลที่ใช้เรียนรู้ R คือจำนวนจริง เป้าหมายสำคัญคือการหาฟังก์ชัน $f(x)$ ซึ่งสามารถทำนายค่าเอาต์พุตของข้อมูลทั้งหมดได้แม่นยำ ซึ่งซัพพอร์ตเวกเตอร์แบบถดถอยเป็นวิธีการหนึ่งที่สามารถทำงานดังกล่าวนี้ได้

โดยทั่วไปฟังก์ชันที่ใช้ในการประมาณค่าเอาต์พุตของซัพพอร์ตเวกเตอร์แบบถดถอยจะอยู่ในรูป

$$y = f(x) = w\phi(x) + b \tag{2.2}$$

โดยที่

w คือ เวกเตอร์น้ำหนัก (weight vector)

x คือ เวกเตอร์อินพุต

b คือ ไบแอส และ $b \in R$

$\phi(x)$ คือ ปริภูมิคุณลักษณะซึ่งถูกจับคู่มาจากปริภูมิอินพุตของ x โดยใช้เคอร์เนล

ในการหาค่า w และ b ที่เหมาะสมเพื่อแทนในสมการที่ (2.2) สามารถหาได้ดังต่อไปนี้ แก้สมการที่ (2.3)

$$\min \frac{1}{2}(w \bullet w) \tag{2.3}$$

ตามเงื่อนไข

$$y_i - w \bullet \phi(x_i) - b \le \varepsilon$$
$$w \bullet \phi(x_i) + b - y_i \le \varepsilon$$

โดยที่

ε คือ ระยะของแถบ ที่ใช้ในการประมาณความถูกต้อง ดังภาพที่ 2.5

ในกรณีที่ยอมให้มีความผิดพลาดได้บ้างจะมีการใช้ตัวแปรสแล็ค ξ_i, ξ_i^* และสมการที่ (2.3) สามารถเขียนได้ใหม่ดัง (2.4) ซึ่งเป็นการหาไฮเปอร์เพลนที่ทำให้ระยะทางรวมของจุดทุกจุดและเส้นแบ่งมีค่าน้อยที่สุด

$$\min \frac{1}{2}(w \bullet w) + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{2.4}$$

ตามเงื่อนไข

$$y_i - w \bullet \phi(x_i) - b \le \varepsilon + \xi_i$$
$$w \bullet \phi(x_i) + b - y_i \le \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \ge 0, i = 1..N$$

โดยที่

ξ_i, ξ_i^* คือ ตัวแปรที่ใช้แทนระยะห่างของจุดของข้อมูลที่นำมาใช้เรียนรู้ทางด้านบนและด้านล่างจากเส้นแบ่งตามเงื่อนไข ซึ่งจะมีค่าเป็นศูนย์ในกรณีที่จุดตกอยู่ในช่วง $\pm\varepsilon$

C คือ ค่าคงที่ที่ต้องกำหนดล่วงหน้าสำหรับใช้ในการสร้างขอบแบบอ่อนซึ่งอนุญาตให้มีการจำแนกข้อมูลผิดประเภทได้ การเพิ่มค่า C จะเพิ่มการอนุญาตให้มีการจำแนกผิด ซึ่งก็จะทำให้ความถูกต้องของโมเดลลดลงด้วย

จากฟังก์ชันสูญเสียแบบ ε -insensitive ซึ่งเสนอโดย Vapnik (1995) ดังสมการ (2.5)

$$L_{\varepsilon}(y)=\begin{cases}0 & for \quad |f(x)-y|\le\varepsilon\\ |f(x)-y|-\varepsilon & otherwise\end{cases} \tag{2.5}$$

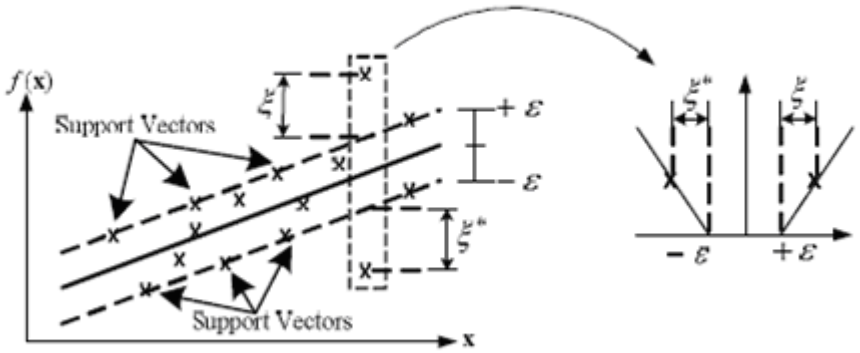
หรือ

$$L_{\varepsilon}(y)=\begin{cases}0 & for \quad \xi\le\varepsilon\\ |\xi|-\varepsilon & otherwise\end{cases}$$

จากสมการ (4) สามารถอธิบายได้ว่าถ้าตำแหน่งของข้อมูลอยู่ในช่วง $\pm\varepsilon$ แล้ว จะถือว่าไม่มีความผิดพลาดของเอาต์พุต โดยค่า ξ และ ε สามารถแสดงด้วยภาพได้ดังภาพที่ 2.6

ภาพที่ 2.6

ภาพแสดงตำแหน่งการตั้งค่า ε และ ξ ของซัพพอร์ตเวกเตอร์แบบถดถอย



ที่มา : Boonprasert

หาค่า a_i และ a_i^* หรือ Lagrange multiplier โดยการทำสมการที่ (2.6) ให้เป็นค่ามากที่สุดตามเงื่อนไข

$$\begin{aligned}\max \quad & -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(a_i-a_i^*)(a_j-a_j^*)\phi(x_i,x_j)\\ & -\varepsilon\sum_{i=1}^{N}(a_i+a_i^*)+\sum_{i=1}^{N}y_i(a_i-a_i^*)\end{aligned} \tag{2.6}$$

ตามเงื่อนไข

$$\sum_{i=1}^{N}(a_i-a_i^*)=0$$

$$a_i,a_i^*\in(0,C)$$

หลังจากแก้สมการที่ (2.6) ตามเงื่อนไข สุดท้ายจะสามารถหาค่า w ได้จากสมการ (2.7)

$$w = \sum_{i=1}^{N} (a_j - a_j^*)\phi(x_i) \tag{2.7}$$

และค่า b ได้จากสมการ

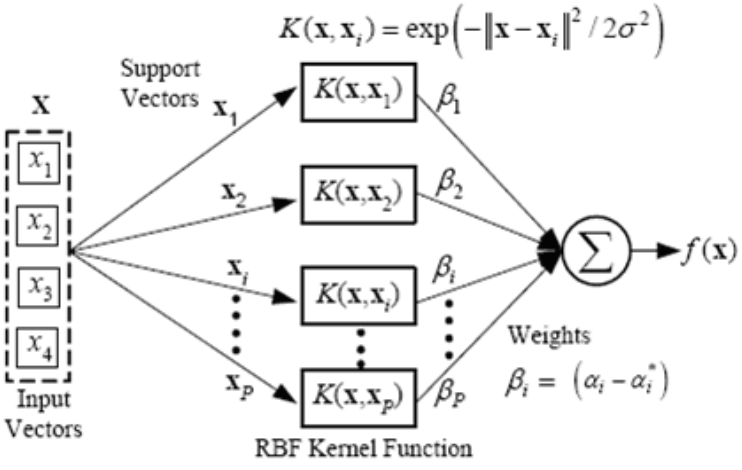
$$\begin{aligned} b &= y_i - w \bullet x_i - \varepsilon \quad for \quad a_i \in (0, C) \\ b &= y_i - w \bullet x_i + \varepsilon \quad for \quad a_i^* \in (0, C) \end{aligned} \tag{2.8}$$

และจากสมการ (2.2) (2.7) และ (2.8) จะได้ฟังก์ชันสำหรับใช้ในการทำนายคือ

$$y = f(x) = \sum_{i=1}^{N} (a_i - a_i^*)\phi(x_i, x) + b \tag{2.9}$$

ภาพที่ 2.7 แสดงวิธีการทำงานของซัพพอร์ตเวกเตอร์แบบถดถอย เริ่มต้นจากเวกเตอร์อินพุตจะถูกจับคู่ด้วยซัพพอร์ตเวกเตอร์ไปยังปริภูมิคุณลักษณะด้วยฟังก์ชันเคอร์เนลแบบเรเดียลเบซิสหลังจากนั้นจะถูกคูณด้วยค่าน้ำหนัก a_i และ a_i^* แล้วจึงนำผลที่ได้มารวมกันก็จะได้เป็นค่าเอาต์พุตสุดท้าย

ภาพที่ 2.7
วิธีการทำงานของซัพพอร์ตเวกเตอร์แบบถดถอย
ในการหาค่า f(x) โดยใช้ฟังก์ชันเคอร์เนลแบบเรเดียลเบซิส



ที่มา : Boonprasert

2.3.6 การนำซัพพอร์ตเวกเตอร์แมชชีนมาประยุกต์ใช้กับปัญหาแบบถดถอย

Wu, Wei, Su, Chang และ Ho (2003) ได้นำซัพพอร์ตเวกเตอร์แบบถดถอยมาใช้ในการทำนายเวลาในการเดินทาง เปรียบเทียบกับวิธีการทำนายแบบอื่นๆ ที่นิยมใช้ในปัจจุบัน โดยข้อมูลที่นำมาใช้ในการเรียนรู้เป็นข้อมูลการจราจรบนทางหลวง ซัพพอร์ตเวกเตอร์แบบถดถอยที่นำมาใช้ในการทดลองใช้ฟังก์ชันเคอร์เนลแบบเชิงเส้น และกำหนดค่า ε เท่ากับ 0.01 และ C เท่ากับ 1,000 อย่างไรก็ตามในการทดลองบางกรณีพบว่าการใช้เคอร์เนลฟังชันชนิดเรเดียลเบซิสให้ผลการทำนายแม่นยำกว่าฟังก์ชันเคอร์เนลชนิดเชิงเส้น ใช้ซอฟต์แวร์ที่ใช้ในการทดลองชื่อ mySVM ผลการทดลองแสดงให้เห็นว่าซัพพอร์ตเวกเตอร์แบบถดถอยให้ผลการทำนายที่แม่นยำที่สุด ตามด้วยวิธีการแบบตัวทำนายจากค่าเฉลี่ยในอดีต (Historical-mean Predictor) และวิธีตัวทำนายเวลาปัจจุบัน (Current-time Predictor) ซึ่งให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.0679, 0.1620 และ 0.2875 ตามลำดับสำหรับกรณีการเดินทางระยะทาง 45 กิโลเมตร กรณีระยะทาง 161 กิโลเมตรให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.0257, 0.0666 และ 0.0998 กรณีการเดินทางระยะทาง 350 กิโลเมตร ให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.0133, 0.0342 และ 0.0549

Ongsritrakul และ Soonthornphisaj (2003) ได้ประยุกต์ใช้ต้นไม้ตัดสินใจกับซัพพอร์ตเวกเตอร์แบบถดถอย เปรียบเทียบกับ วิธีการถดถอยเชิงเส้น (linear regression) และนิวรอลเน็ตเวิร์ก ในการทำนายราคาทองคำ ข้อมูลที่นำมาใช้ในการทดลองเป็นข้อมูลที่เก็บในช่วงระยะเวลา 1,000 วัน การทดลองเพื่อหาความแม่นยำของวิธีการทำนายต่างๆ จะใช้วิธีทดสอบแบบไขว้ข้ามสิบชุด (10 fold cross validation) ซัพพอร์ตเวกเตอร์แบบถดถอยที่ใช้ในการทดลองประยุกต์มาจากโปรแกรม SVMTorch ใช้ฟังก์ชันเคอร์เนลแบบเกาส์ นิวรอลเน็ตเวิร์กที่นำมาทดลองเป็นนิวรอลเน็ตเวิร์กแบบ 3 ชั้น ชั้นอินพุตมีจำนวนโหนดเท่ากับ 13 โหนด ชั้นแฝงมีจำนวนโหนดเท่ากับ 9 และชั้นเอาต์พุตมีจำนวนโหนดเท่ากับ 1 การทดลองแบ่งออกเป็นสองรูปแบบ แบบที่หนึ่ง ทำการเปรียบเทียบวิธีการระหว่างซัพพอร์ตเวกเตอร์แบบถดถอย วิธีการถดถอยเชิงเส้น และนิวรอลเน็ตเวิร์ก ตามปกติ แบบที่ 2 มีการประยุกต์ใช้ต้นไม้ตัดสินใจในการเลือกตัวแปรอินพุตที่เหมาะสม แล้วจึงนำตัวแปรที่ได้ไปเป็นอินพุตที่ใช้ในการเรียนรู้ แล้วจึงเปรียบเทียบความแม่นยำในการทำนายระหว่างวิธีการทั้ง 3 ผลการทดลองแบบที่ 1 พบว่าซัพพอร์ตเวกเตอร์แบบถดถอยและวิธีการถดถอยเชิงเส้นให้ผลการทำนายที่ดีกว่านิวรอลเน็ตเวิร์ก โดยให้ค่าเฉลี่ยความผิดพลาดกำลังสอง (MSE) เท่ากับ 47.356, 47.354 และ 122.189 ตามลำดับ ในขณะที่การทดลองแบบที่ 2 พบ

ว่าซัพพอร์ตเวกเตอร์แบบถดถอยให้ผลการทำนายดีที่สุด ตามมาด้วยวิธีการถดถอยเชิงเส้น และนิวรอลเน็ตเวิร์ก โดยให้ค่าเฉลี่ยความผิดพลาดกำลังสองเท่ากับ 43.434, 49.809 และ 84.493 ตามลำดับ จากผลของงานวิจัยพบว่านอกจากนิวรอลเน็ตเวิร์กให้ผลในการทำนายปัญหาแบบถดถอยได้ไม่แม่นยำแล้วยังพบว่าใช้เวลาในการคำนวณผลมากกว่าวิธีการอื่นๆ

Su, Wang, Celler, Ambikairajah และ Savkin (2005) นำซัพพอร์ตเวกเตอร์แบบถดถอยมาใช้ในการประมาณพลังงานที่ใช้ในการเดิน โดยเปรียบเทียบกับวิธีการทำนายโดยใช้วิธีถดถอยเชิงเส้น ผลการทดลองพบว่าซัพพอร์ตเวกเตอร์แบบถดถอย และวิธีถดถอยเชิงเส้น ให้ค่าความผิดพลาดกำลังสองเท่ากับ 0.3367 และ 0.4747 ตามลำดับ

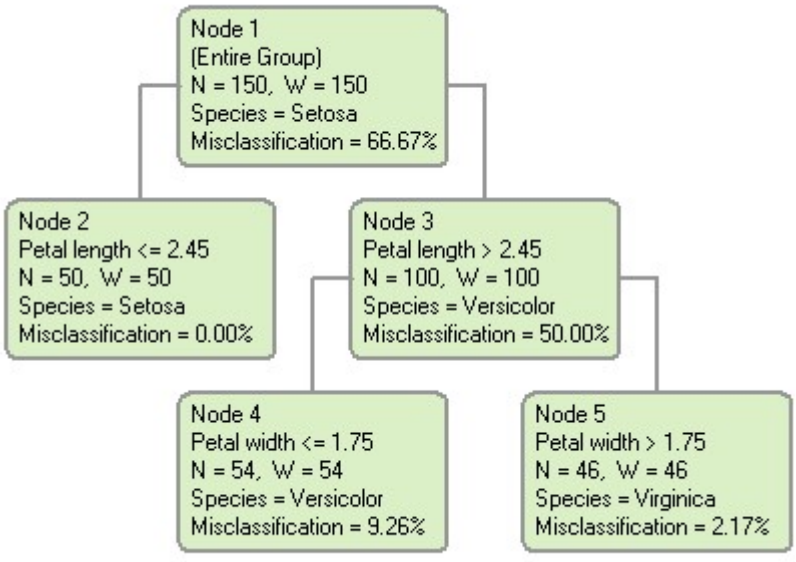
Hang, Lee, Lai, Wu และ Wang (2007) ได้นำวิธีการซัพพอร์ตเวกเตอร์แบบถดถอยกับอัลกอริทึมภูมิคุ้มกัน (Immune Algorithm) (SVRIA) ซัพพอร์ตเวกเตอร์แบบถดถอยกับอัลกอริทึมเชิงพันธุกรรม (SVMG) โมเดลแบบถดถอย (regression model) และนิวรอลเน็ตเวิร์กมาใช้ในการทำนายปริมาณการใช้ไฟฟ้าของประเทศไต้หวัน โดย SVRIA ที่นำมาทดลอง ใช้ฟังก์ชันเคอร์เนลแบบเกาส์ (gaussian kernel function) และนำอัลกอริทึมภูมิคุ้มกันมาประยุกต์ใช้เพื่อหาค่าของตัวแปร σ, C และ ε ที่เหมาะสม ผลการทดลองพบว่า SVRIA สามารถทำนายได้แม่นยำมากที่สุดเมื่อเทียบกับวิธีอื่นๆ โดยค่าเฉลี่ยเปอร์เซ็นต์ความคลาดเคลื่อนสัมบูรณ์ (Mean Absolute Percent Error, MAPE) ของวิธี SVRIA, SVMG, นิวรอลเน็ตเวิร์ก และโมเดลแบบถดถอย เท่ากับ 1.71, 1.97, 2.22 และ 5.84 ตามลำดับ

Ji, Han และ Zhai (2007) ได้เสนอการประยุกต์ใช้ซัพพอร์ตเวกเตอร์แบบถดถอยกับซัพพอร์ตเวกเตอร์แมชชีนร่วมกันเพื่อทำนายความเร็วลม เปรียบเทียบกับซัพพอร์ตเวกเตอร์แบบถดถอยแบบดั้งเดิม โดยวิธีที่เสนอเริ่มต้นจากแบ่งข้อมูลออกเป็น 3 ส่วน จากนั้นนำข้อมูลส่วนที่ 1 ไปใช้ในการเรียนรู้ของซัพพอร์ตเวกเตอร์แบบถดถอย ทดสอบซัพพอร์ตเวกเตอร์แบบถดถอยดังกล่าวด้วยข้อมูลส่วนที่ 2 และนำข้อมูลบางส่วนจากข้อมูลส่วนที่ 2 ไปใช้ในการเรียนรู้สำหรับซัพพอร์ตเวกเตอร์แมชชีนเพื่อใช้ในการหาค่าพารามิเตอร์ C และ γ ที่เหมาะสม จากนั้นนำข้อมูลส่วนที่ 3 มาทดสอบความแม่นยำในการทำนายระหว่างวิธีที่เสนอและซัพพอร์ตเวกเตอร์แบบถดถอยแบบดั้งเดิม ฟังก์ชันเคอร์เนลที่ใช้ในการทดลองใช้แบบเรเดียลเบซิส ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอให้ค่าเฉลี่ยความผิดพลาดกำลังสอง (MSE) เท่ากับ 0.0012 ในขณะที่วิธีที่เสนอให้ค่าความผิดพลาดกำลังสองเท่ากับ 0.0008

2.4 ต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเป็นการนำข้อมูลมาสร้างแบบจำลอง มีลักษณะเป็นผังงาน (flowchart) เหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (supervised learning) คือ สร้างแบบจำลองขึ้นมาจากข้อมูลที่นำมาใช้เรียนรู้ โดยต้นไม้ตัดสินใจสามารถนำมาใช้ในการทำนายค่าต่างๆ ได้ โดยผลการทำนายจะขึ้นอยู่กับตัวแปรต้น รูปแบบของต้นไม้ ดังภาพที่ 2.8 จะประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (root node) จากโหนดรากก็จะแตกออกเป็นโหนดลูกซึ่งมีกิ่งในการเชื่อมระหว่างโหนด และที่โหนดลูกก็จะมีลูกของตัวเองซึ่งที่โหนดระดับสุดท้ายจะเรียกว่าโหนดใบ (leaf node) แต่ละโหนดของโหนดรากและโหนดลูกจะแสดงค่าคุณลักษณะ (attribute) ที่ใช้ทดสอบข้อมูล ส่วนโหนดใบจะแสดงกลุ่ม (class) ที่กำหนดไว้ เมื่อมีข้อมูลที่ต้องการทำนาย การทำงานจะเริ่มต้นจากโหนดราก ซึ่งจะนำค่าคุณลักษณะต่างๆ ของข้อมูลนั้นไปเปรียบเทียบกับคุณลักษณะของโหนด และทำการตัดสินใจว่าจะเดินทางไปตามกิ่งใด หลังจากนั้นจะเดินทางผ่านโหนดลูก และทำการเปรียบเทียบคุณลักษณะไปเรื่อยๆ จนกระทั่งสุดท้ายไปถึงโหนดใบ ก็จะได้กลุ่มที่ถูกกำหนด

ภาพที่ 2.8
โครงสร้างของต้นไม้ตัดสินใจ



ที่มา : Sherrod (2007)

2.4.1 ลักษณะของปัญหาที่เหมาะกับการใช้ต้นไม้ตัดสินใจ

ถึงแม้ว่าวิธีการเรียนรู้ของต้นไม้ตัดสินใจจะถูกพัฒนาขึ้นมาในหลายรูปแบบเพื่อให้เหมาะสมกับรูปแบบของปัญหาต่างๆ แต่โดยทั่วไปแล้วลักษณะของปัญหาที่เหมาะกับการใช้ต้นไม้ตัดสินใจมีดังต่อไปนี้

1. ข้อมูลแสดงอยู่ในรูปของชุดของคุณลักษณะ (เช่น อุณหภูมิ) และค่าของมัน (เช่น ร้อน) สำหรับรูปแบบที่ง่ายที่สุดสำหรับการเรียนรู้ของต้นไม้ตัดสินใจคือเมื่อคุณลักษณะมีกลุ่มอยู่เพียงไม่กี่ตัว (เช่น ร้อน เย็น ปกติ) อย่างไรก็ตามมีอัลกอริทึมที่พัฒนาขึ้นมาเพื่อรองรับค่าของคุณลักษณะที่เป็นจำนวนจริง (เช่น ค่าของอุณหภูมิเป็นองศาเซลเซียส)

2. ฟังก์ชันเป้าหมาย (target function) มีเอาต์พุตเป็นแบบไม่ต่อเนื่อง (discrete output value) (เช่น ใช่ หรือ ไม่ใช่) แต่ต้นไม้ตัดสินใจก็มีอัลกอริทึมเสริมซึ่งอนุญาตให้เอาต์พุตของฟังก์ชันเป้าหมายสามารถเป็นจำนวนจริงได้

3. อนุญาตให้มีข้อมูลที่ผิดพลาด (error) อยู่ในข้อมูลที่ใช้ในการสอนได้ เนื่องจากวิธีการเรียนรู้ของต้นไม้ตัดสินใจนั้นทนทาน (robust) ต่อความผิดพลาด โดยยอมให้มีความผิดพลาดได้ทั้งค่าคุณลักษณะ และการแบ่งกลุ่ม

4. ข้อมูลที่ใช้ในการเรียนรู้ไม่จำเป็นต้องมีค่าของคุณลักษณะอย่างครบถ้วน นั่นคือวิธีการของต้นไม้ตัดสินใจสามารถใช้ได้ ถึงแม้ข้อมูลนั้นจะมีค่าบางค่าของคุณลักษณะบางตัวหายไป หรือไม่ทราบค่าก็ตาม

2.4.2 วิธีการสร้างต้นไม้ตัดสินใจ

มีขั้นตอนดังต่อไปนี้

1. หาคุณลักษณะที่สำคัญที่สุดมาแบ่งข้อมูลโดยคุณลักษณะนี้จะถูกตั้งให้เป็นโหนดราก

2. จากโหนดรากจะสร้างเส้นทางเชื่อมหรือกิ่งไปยังโหนดลูก โดยจำนวนเส้นทางเชื่อมจะเท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะของโหนดราก

3. ถ้าโหนดลูกเป็นกลุ่มของข้อมูลที่อยู่ในกลุ่มเดียวกันทั้งหมดให้หยุดการสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายกลุ่มปะปนกัน ต้องสร้างโหนดลูกเพื่อจำแนกข้อมูลต่อไป โดยวนกลับไปทำขั้นตอนที่ 1 ซ้ำเพื่อเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นตัวแบ่งข้อมูลต่อไป

สำหรับความสำคัญของคุณลักษณะสามารถหาได้จากการคำนวณค่าการเพิ่มสารสนเทศ (information gain) ดังต่อไป

<u>ทฤษฎีการเพิ่มสารสนเทศ</u>

ทฤษฎีการเพิ่มสารสนเทศ เป็นทฤษฎีหนึ่งที่ใช้ในการสร้างต้นไม้ตัดสินใจ ซึ่งมีหลักการทำงาน เริ่มต้นด้วยการนำค่าคุณลักษณะจากข้อมูลมาคำนวณหาค่าการเพิ่มสารสนเทศโดยค่าคุณลักษณะตัวใดมีค่าการเพิ่มสารสนเทศสูงสุด (มีค่าเอ็นโทรปี (entropy) น้อยที่สุด) จะถูกเลือกให้เป็นค่าคุณลักษณะของโหนดนั้น

สมการสำหรับการคำนวณหาค่าการเพิ่มสารสนเทศมีดังนี้

กำหนดให้ S เป็นเซตของข้อมูล

m แทนจำนวนกลุ่มที่แตกต่างกันข้อมูล ซึ่งแต่ละค่าที่แตกต่างกันแทนด้วย C_i (for i =1,…,m)

s_i เป็นกลุ่มข้อมูลย่อยของ S ในกลุ่ม C_i

สมการที่ใช้ในการแบ่งแยกกลุ่มของข้อมูลคือ

$$Gain(A) = I(s_1, s_2,\ldots,s_m) - E(A) \quad (2.10)$$

เมื่อ A คือ ค่าคุณลักษณะที่ใช้หาการเพิ่มสารสนเทศ

ค่าคุณลักษณะตัวใดที่มีค่า Gain(A) มากที่สุดจะใช้ค่าคุณลักษณะนั้นเป็นตัวแบ่งแยกข้อมูลในโหนดนั้น

จากสมการข้างต้น $I(s_1, s_2,\ldots,s_m)$ หาได้จากสมการที่ 2.11

$$I(s_1, s_2,\ldots,s_m) = - p_i \log2(p_i) \quad (2.11)$$

โดยที่

p_i คือ ความน่าจะเป็นที่ใช้ในการตัดสินข้อมูล S ใน class C_i

$p_i = s_i/S$

ให้ค่าคุณลักษณะ A มีค่าของข้อมูลที่แตกต่างกัน v ค่า แทนด้วย $a_1,a_2,\ldots,a_v$ ซึ่ง ค่าคุณลักษณะ A จะแบ่งข้อมูล S ออกเป็น v ซับเซต แทนด้วย $S_1,S_2,\ldots,S_v$

ให้ S_j แทนจำนวนข้อมูลในแต่ละค่าของค่าคุณลักษณะ A

ให้ S_{ij} แทนจำนวนข้อมูลในแต่ละกลุ่มของแต่ละค่าคุณลักษณะ A แทนด้วย $S_{1j},S_{2j},\ldots,S_{mj}$

สมการ ในการหาเอ็นโทรปีของกลุ่มข้อมูลย่อยนี้คือ

$$E(A) = S_{1j}+\ldots+S_{mj}\ I(S_{1j},\ldots,S_{mj}) \quad (2.12)$$

จากสมการข้างต้น $I(S_{1j},\ldots,S_{mj})$ หาได้จากสมการที่ 2.13

$$I(S_{1j},\ldots,S_{mj}) = - p_{ij}\log2(p_{ij}) \quad (2.13)$$

โดยที่

p_{ij} คือ ความน่าจะเป็นที่ใช้ในการตัดสินข้อมูล S_j ในกลุ่ม C_i

$p_{ij} = s_{ij}/S_j$

2.4.3 ข้อจำกัดของต้นไม้ตัดสินใจ

การแบ่งกลุ่มของต้นไม้ตัดสินใจในกรณีเป็นข้อมูลมีลักษณะเป็นค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคา ต้องทำการแปลงให้อยู่ในช่วงหรือแบ่งเป็นกลุ่มก่อน

2. เมื่ออัลกอริทึมเลือกว่าจะใช้ตัวแปรไหนเป็นตัวจำแนกประเภทแล้วก็จะไม่สนใจตัวแปรอื่นที่อาจมีความสำคัญเช่นเดียวกัน

3. ถึงแม้ต้นไม้ตัดสินใจจะมีความทนทานต่อความผิดพลาด แต่หากมีตัวแปรที่ไม่ทราบค่า ก็อาจส่งผลกระทบต่อผลลัพธ์ของต้นไม้ตัดสินใจได้

4. ต้นไม้ตัดสินใจที่มีระดับชั้นมากเกินไป จะทำให้ข้อมูลที่ผ่านโหนดแตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้น จะไม่มีประโยชน์ในการนำมาใช้ทำการวิเคราะห์

5. ปัญหาเรื่องความเหมาะพอดีมากเกินไป (overfitting) เกิดเนื่องจากแบบจำลองได้สร้างโครงสร้างต้นไม้ที่เฉพาะเจาะจงกับข้อมูลที่ใช้ในการเรียนรู้มากเกินไป

6. เทคนิคของต้นไม้ตัดสินใจจะจำกัดข้อมูลที่เป็นตัวแปรตาม (dependent variable) 1 ตัวต่อ 1 แบบจำลอง ถ้าต้องการทำนายตัวแปรตามหลาย ๆ ตัวจะต้องสร้างแบบจำลองสำหรับตัวแปรตามแต่ละตัว

2.4.4 ปัญหาความเหมาะพอดีมากเกินไป

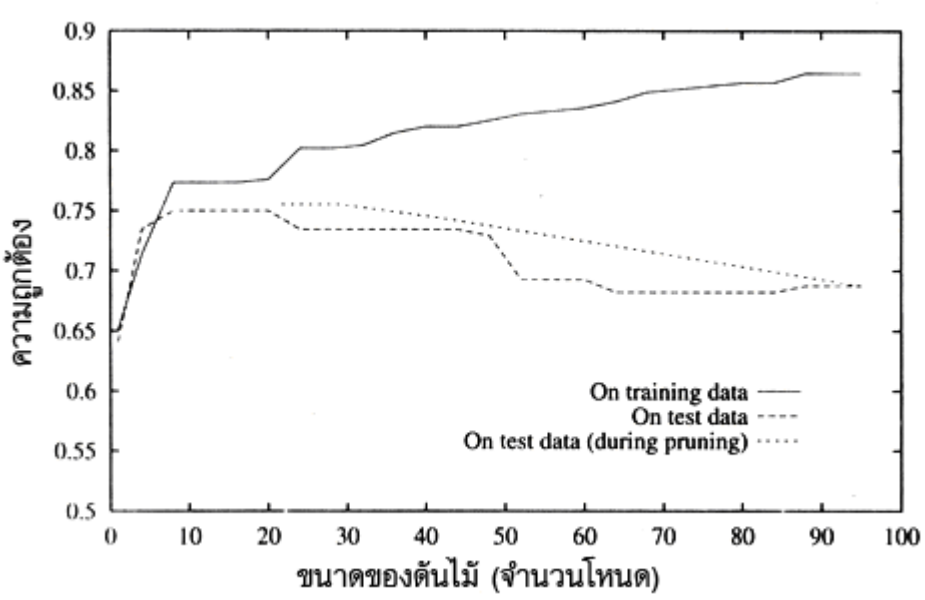
ปัญหาความเหมาะพอดีมากเกินไปของต้นไม้ตัดสินใจเกิดเนื่องจากแบบจำลองได้สร้างโครงสร้างต้นไม้ที่เฉพาะเจาะจงกับข้อมูลที่ใช้ในการเรียนรู้มากเกินไป ทำให้ขาดความ

ยืดหยุ่นในการนำไปปรับใช้กับข้อมูลอื่นที่ไม่เคยเห็นมาก่อน ซึ่งจะทำให้ความถูกต้องในการทำนายลดน้อยลง โดยปัญหาเรื่องความเหมาะพอดีมากเกินไปมักเกิดขึ้นเนื่องจากขนาดของต้นไม้มีการแตกโหนดและขยายขนาดมากเกินไป ภาพที่ 2.9 เป็นการเปรียบเทียบความถูกต้องของข้อมูลที่ใช้ในการเรียนรู้และข้อมูลที่ใช้ในการทดสอบ จะเห็นว่าเมื่อจำนวนโหนดของต้นไม้เพิ่มมากขึ้นความถูกต้องในการทำนายของข้อมูลชุดทดสอบจะเพิ่มขึ้นจนถึงจุดหนึ่งแล้วจะลดลง ในขณะที่ความถูกต้องในการทำนายของข้อมูลสำหรับเรียนรู้จะเพิ่มขึ้นเรื่อยๆ เมื่อจำนวนโหนดเพิ่มมากขึ้น

สำหรับวิธีแก้ปัญหาเรื่องความเหมาะพอดีมากเกินไปมี 2 วิธีได้แก่ การตัดกิ่งส่วนที่ไม่จำเป็นออก หรือการตัดเล็ม (pruning) และการหยุดต้นไม้ไม่ให้โตเกินไป

ภาพที่ 2.9

เปรียบเทียบขนาดของต้นไม้และความถูกต้องในการทำนาย

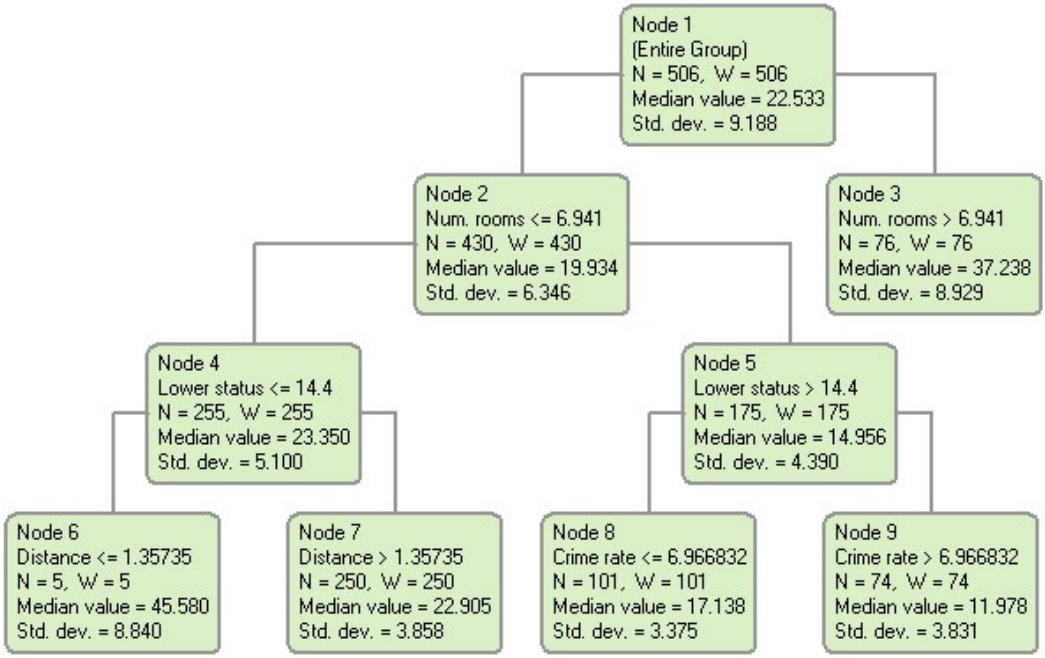


2.4.5 ต้นไม้ตัดสินใจแบบถดถอย (Regression Tree)

ในกรณีนำต้นไม้ตัดสินใจมาแก้ปัญหาแบบถดถอยซึ่งผลลัพธ์เป็นจำนวนจริง และเป็นค่าต่อเนื่องกันนั้น ในการทำนายจะต้องใช้ต้นไม้ตัดสินใจแบบถดถอย จะค่าที่ได้จากการทำนายจะนำมาจากค่าเฉลี่ยของค่าเอาต์พุตที่ตกอยู่ในโหนดนั้นๆ ตัวอย่างของต้นไม้ตัดสินใจแบบถดถอย

แสดงดังภาพที่ 2.10 โดยเอาต์พุตก็คือค่า “Median Value” จากตัวอย่างถ้าค่า “Num. Rooms” มีค่ามากกว่า 6.941 แล้ว ค่าเอาต์พุตที่ได้จากการเฉลี่ยจะมีค่าเท่ากับ 37.238 ในขณะที่ถ้าค่า “Num. Rooms” มีค่าน้อยกว่าหรือเท่ากับ 6.941 แล้ว ค่าเอาต์พุตจะเท่ากับ 19.934 เป็นต้น

ภาพที่ 2.10
ต้นไม้ตัดสินใจแบบถดถอย



ที่มา : Sherrod (2007)

2.4.6 การนำต้นไม้ตัดสินใจมาประยุกต์ใช้กับปัญหาแบบถดถอย

Li, Wang และ Gao (2007) ได้นำต้นไม้ตัดสินใจแบบถดถอยมาใช้ในการประมาณค่าลยาปูนอฟเอกซ์โพเนนต์มากที่สุด (Largest Lyapunov Exponent) ซึ่งเป็นค่าที่ใช้ในการวัดเสถียรภาพของระบบแบบไดนามิกและความใช้ได้ของวิธีการทดสอบสำหรับความโกลาหล (chaos) โดยการทดลองนี้ได้ทำการเปรียบเทียบผลการทดลองกับการใช้นิวรอลเน็ตเวิร์ก ซึ่งจากผลการทดลองสรุปได้ว่าต้นไม้ตัดสินใจแบบถดถอยให้ผลการทดลองที่พ้องกันและเป็นไปในทิศทางเดียวกันกับการใช้นิวรอลเน็ตเวิร์ก

Scaringella (2007) ได้นำต้นไม้ตัดสินใจแบบถดถอยมาใช้ในการประมาณราคาที่พักอาศัย โดยใช้ข้อมูลจากเซตตัวอย่างที่พักอาศัยในบอสตันซึ่งมีตัวแปรอินพุตทั้งหมด 14 ตัว เช่น อัตราการเกิดอาชญากรรม ปริมาณความเข้มข้นของไนตริกออกไซค์ จำนวนห้องเฉลี่ยต่อที่อยู่อาศัย อัตราส่วนระหว่างนักเรียนและอาจารย์ สัดส่วนของพื้นที่พักอาศัยที่มีพื้นที่มากกว่า 25,000 ตารางฟุต จำนวนโป๊ะของแม่น้ำ เป็นต้น และงานวิจัยนี้ยังได้ทำการทดลองปรับลดค่าปริมาณความเข้มข้นของไนตริกออกไซค์ลงในช่วง 0.25 ถึง 0.4 เพื่อหาความสัมพันธ์ระหว่างค่าปริมาณความเข้มข้นของไนตริกออกไซด์กับราคาที่พักอาศัย

2.5 ป่าของต้นไม้ตัดสินใจ

ป่าของต้นไม้ตัดสินใจ (decision tree forest หรือ random forest) เป็นอัลกอริทึมที่เสนอโดย Breiman (2001) ป่าของต้นไม้ตัดสินใจประกอบไปด้วยต้นไม้ตัดสินใจจำนวนหลายต้น แต่ละต้นเป็นอิสระต่อกัน ในการทำนายเอาต์พุตจะเริ่มต้นจากนำอินพุตใส่ลงไปในต้นไม้ตัดสินใจแต่ละต้น โดยต้นไม้แต่ละต้นจะทำการจำแนกประเภทหรือทำนายเอาต์พุตออกมา หลังจากนั้นเอาต์พุตสุดท้ายจะได้มาจากการโหวตของต้นไม้ตัดสินใจแต่ละต้น โดยเลือกค่าที่ได้รับการโหวตมากที่สุด

2.5.1 อัลกอริทึมในการสร้างป่าของต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจแต่ละต้นในป่าของต้นไม้ตัดสินใจจะถูกสร้างด้วยอัลกอริทึมต่อไปนี้

1. จำนวนแถวข้อมูลที่ใช้ในการเรียนรู้จำนวน n แถว และจำนวนตัวแปรอินพุตคือ M
2. เลือกแถวข้อมูลสำหรับเรียนรู้ โดยใช้วิธีสุ่มแบบใส่คืน (sampling with replacement) จำนวน 2 ใน 3 จากแถวข้อมูลทั้งหมด เพื่อนำไปใช้ในการสร้างต้นไม้ตัดสินใจ ส่วนแถวข้อมูลที่เหลืออีก 1 ใน 3 (ข้อมูล out-of-bag) จะใช้ในการประมาณความผิดพลาดของต้นไม้โดยนำมาใช้ทำนายเพื่อดูความถูกต้อง
3. m_{try} คือจำนวนตัวแปรอินพุตที่ถูกเลือกแบบสุ่มมาจาก M และนำมาใช้ในการสร้างโหนดของต้นไม้ตัดสินใจ โดยแต่ละโหนดของต้นไม้ตัดสินใจได้มาจากการหาตัวแปรที่สามารถแยกกลุ่มได้ดีที่สุด และค่า m_{try} นี้จะคงที่ตลอดระหว่างการสร้างป่าของต้นไม้ตัดสินใจ ยกตัวอย่างเช่น ถ้ามีตัวแปรอินพุตทั้งหมด 10 ตัว ถ้าค่า m คือ 5 ให้ทำการ

สุ่มเลือกตัวแปรอินพุตมา 5 ตัว และนำตัวแปร 5 ตัวนี้มาทำการหาว่าตัวแปรตัวใดสามารถทำการจำแนกประเภทได้ดีที่สุดก็จะใช้โหนดนั้นเป็นโหนดของต้นไม้ตัดสินใจ จากนั้นในการหาโหนดถัดไปก็จะทำการสุ่มตัวแปรขึ้นมาอีก 5 ตัวจากตัวแปรที่เหลือและหาตัวแปรที่สามารถทำการจำแนกได้ดีที่สุด ทำจนกระทั่งครบตัวแปรทุกตัว

4. ปล่อยให้ต้นไม้แต่ละต้นโตเต็มที่เท่าที่จะเป็นไปได้ โดยไม่ทำการตัดเล็ม

2.5.2 อัตราความผิดพลาด (Error Rate) ของป่าของต้นไม้ตัดสินใจ

อัตราความผิดพลาดของป่าของต้นไม้ตัดสินใจขึ้นอยู่กับ 2 สิ่ง ได้แก่

1. สหสัมพันธ์ (correlation) ระหว่างต้นไม้ตัดสินใจสองต้นใดๆ ในป่าของต้นไม้ตัดสินใจ โดยการเพิ่มขึ้นของสหสัมพันธ์จะทำให้อัตราความผิดพลาดของป่าของต้นไม้ตัดสินใจเพิ่มขึ้นด้วย
2. ความแข็งแรง (strength) ของต้นไม้ตัดสินใจแต่ละต้นในป่า
   ต้นไม้ตัดสินใจที่มีอัตราความผิดพลาดน้อยถือเป็นตัวจำแนกประเภทที่เข้มแข็ง การเพิ่มขึ้นของความเข้มแข็งของต้นไม้ตัดสินใจแต่ละต้นจะช่วยลดอัตราความผิดพลาดของป่าของต้นไม้ตัดสินใจ

การลดค่า m_{try} จะลดทั้งสหสัมพันธ์และความเข้มแข็งของต้นไม้ตัดสินใจ ในขณะที่การเพิ่มค่า m ก็จะเพิ่มค่าทั้งสองเช่นเดียวกัน ดังนั้นเราจึงจำเป็นต้องหาค่า m ที่เหมาะสม ซึ่งการใช้ทฤษฎีอัตราความผิดพลาดของข้อมูลที่เหลือ (out-of-bag error rate) ช่วยให้สามารถหาค่า m_{try} ที่เหมาะสมได้รวดเร็วขึ้น

2.5.3 การทำงานของป่าของต้นไม้ตัดสินใจ

เมื่อข้อมูลสำหรับเรียนรู้ถูกเลือกด้วยวิธีสุ่มแบบใส่คืนจากชุดข้อมูลทั้งหมด ต้นไม้ตัดสินใจจะถูกสร้างขึ้นด้วยข้อมูลชุดนี้ ส่วนข้อมูลที่เหลืออีกประมาณ 1 ใน 3 จะถูกนำมาใช้ในการประมาณความผิดพลาดของต้นไม้ตัดสินใจที่ถูกสร้างขึ้นมา นอกจากนี้ยังถูกใช้เพื่อหาความสำคัญของตัวแปรอีกด้วย

1. การประมาณความผิดพลาดจากข้อมูลที่เหลือจากการสุ่ม

ต้นไม้แต่ละต้นจะถูกสร้างขึ้นมาจากข้อมูลที่มาจาการสุ่มแบบใส่คืน ข้อมูลที่เหลืออีกประมาณ 1 ใน 3 จะไม่นำมาใช้ในการสร้างต้นไม้ดังกล่าว แต่จะถูกนำมาใช้ในการประมาณความผิดพลาด โดยจะถูกใช้เป็นชุดทดสอบความถูกต้อง โดยค่าประมาณความผิดพลาดจะได้มาจากค่าเฉลี่ยของอัตราส่วนในการทำนายผิดของต้นไม้ตัดสินใจทุกต้นในป่าของต้นไม้ตัดสินใจ

2. ความสำคัญของตัวแปร

สำหรับต้นไม้ตัดสินใจทุกๆ ต้นในป่าของต้นไม้ตัดสินใจ ให้ทำการนำข้อมูลที่เหลือจากการสุ่ม ไปทำการทำนายและนับจำนวนครั้งที่ทำนายถูกต้อง จากนั้นทำการสับเปลี่ยนตัวแปร m แบบสุ่มและนำชุดข้อมูลใหม่นี้ไปทำนายอีกครั้งหนึ่ง แล้วนำจำนวนครั้งที่ทายถูกของข้อมูลที่มีการสับเปลี่ยนตัวแปร m ไปลบออกจากจำนวนครั้งที่ทายถูกของข้อมูลที่ไม่ได้สับเปลี่ยนตัวแปร เมื่อนำค่าที่ได้นี้จากต้นไม้ทุกต้นมาหาค่าเฉลี่ยก็จะได้ค่าคะแนนดิบที่แสดงถึงความสำคัญของตัวแปร m

2.5.4 ความเหมาะพอดีมากเกินไปและการตัดเล็ม

ปัญหาเรื่องความเหมาะพอดีมากเกินไปเป็นปัญหาใหญ่สำหรับต้นไม้ตัดสินใจซึ่งเกิดเนื่องจากแบบจำลองได้สร้างโครงสร้างต้นไม้ที่เฉพาะเจาะจงกับข้อมูลที่ใช้ในการเรียนรู้มากเกินไป ทำให้ขาดความยืดหยุ่นในการนำไปปรับใช้กับข้อมูลอื่นที่ไม่เคยเห็นมาก่อน ซึ่งจะทำให้ความถูกต้องในการทำนายลดน้อยลง ซึ่งวิธีในการแก้ไขปัญหาก็คือการตัดเล็มซึ่งจะทำให้ได้ขนาดของต้นไม้ที่เหมาะสม แต่สำหรับป่าของต้นไม้ตัดสินใจนั้นไม่มีปัญหาเรื่องความเหมาะพอดีมากเกินไป เนื่องจากอัลกอริทึมในการสร้างต้นไม้ตัดสินใจในป่าเองซึ่งเหมือนเป็นการทดสอบความถูกต้องของข้อมูลอยู่แล้ว ดังนั้นจึงไม่มีความจำเป็นที่จะต้องทำการตัดกิ่งออกแต่อย่างใด

2.5.5 ป่าของต้นไม้ตัดสินใจสำหรับแก้ปัญหาแบบถดถอย

ใช้อัลกอริทึมในการสร้างต้นไม้ตัดสินใจเช่นเดียวกับการจำแนกประเภท แต่ต้นไม้ในป่าจะเป็นต้นไม้ตัดสินใจแบบถดถอยแทน และในการหาเอาต์พุตสุดท้ายต่างกันตรงที่สำหรับปัญหาแบบจำแนกประเภทจะใช้การโหวต ในขณะที่ค่าเอาต์พุตสุดท้ายของป่าของต้นไม้ตัดสินใจสำหรับแก้ปัญหาแบบถดถอยจะได้มาจากค่าเฉลี่ยของเอาต์พุตของต้นไม้ทุกต้น

2.5.6 การนำป่าของต้นไม้ตัดสินใจมาประยุกต์ใช้กับปัญหาแบบถดถอย

Arun และ Langmead (2005) ทำนาย chemical shifts โดยใช้วิธีป่าของต้นไม้ตัดสินใจเปรียบเทียบกับวิธีการทำนายที่นิยมใช้ในปัจจุบัน ได้แก่ SHIFTS, SHIFTX และ PROSHIFT ผลการทดลองแสดงให้เห็นว่าป่าของต้นไม้ตัดสินใจเป็นวิธีที่สามารถทำนายได้แม่นยำมากที่สุดเมื่อเปรียบเทียบกับวิธีทั้ง 3 ดังกล่าว โดยค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของนิวเคลียสชนิด H^N มีค่าเท่ากับ 0.49, H^{α} มีค่าเท่ากับ 0.28, ^{15}N มีค่าเท่ากับ 2.93, $^{13}C^{\alpha}$ มีค่าเท่ากับ 1.51, $^{13}C^{\beta}$ มีค่าเท่ากับ 2.93 และ $^{13}C'$ มีค่าเท่ากับ 1.19 ซึ่งค่าเฉลี่ยดังกล่าวหาได้จากวิธีทดสอบแบบไขว้ข้ามสิบชุด

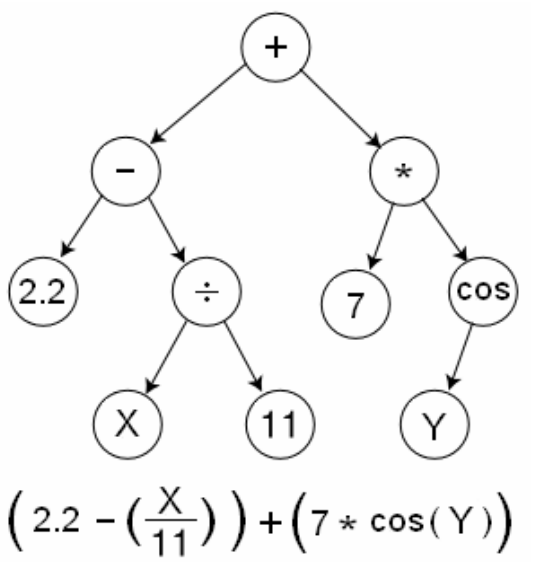
Jiang, Sun และ Lu (2007) นำป่าของต้นไม้ตัดสินใจมาใช้ในการทำนายค่า siRNA efficacy เปรียบเทียบกับวิธีซัพพอร์ตเวกเตอร์แบบถดถอย การทดลองใช้การหาค่าเฉลี่ยความผิดพลาดด้วยวิธีทดสอบความถูกต้องแบบข้าม 3 ชุดข้อมูลย่อย สำหรับป่าของต้นไม้ตัดสินใจที่ให้ผลการทำนายดีที่สุด ใช้จำนวนต้นไม้เท่ากับ 1,000 ต้น ตัวแปรอินพุต 15 ตัว กำหนดค้า m_{try} เท่ากับ 10 สำหรับซัพพอร์ตเวกเตอร์แบบถดถอยที่ใช้ในการทดลองใช้ฟังก์ชันเคอร์เนลแบบเรเดียลเบซิสค่าพารามิเตอร์ที่ให้ผลการทำนายดีที่สุดคือ ค่า C เท่ากับ 200 และ γ เท่ากับ 0.001 จากผลการทดลองพบว่าวิธีป่าของต้นไม้ตัดสินใจให้ผลการทำนายที่แม่นยำกว่าซัพพอร์ตเวกเตอร์แบบถดถอย โดยค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ได้จากวิธีป่าของต้นไม้ตัดสินใจและซัพพอร์ตเวกเตอร์แบบถดถอยเท่ากับ 8.924 และ 9.414 ตามลำดับ

2.6 โปรแกรมเชิงพันธุกรรม

โปรแกรมเชิงพันธุกรรมเป็นเทคนิครูปแบบหนึ่งของการคำนวณเชิงวิวัฒนาการ (Evolutionary Computation, EC) ซึ่งการพัฒนาโปรแกรมโดยเลียนแบบสิ่งมีชีวิต โดยอาศัยกลไกการเลือกสรรตามธรรมชาติเพื่อให้ได้สายพันธุ์ของสิ่งมีชีวิตที่ดีที่สุดด้วยวิธีการต่างๆ เช่น การสร้างประชากรใหม่ (reproduction) การเปลี่ยนถ่ายพันธุกรรม (crossover) และ การกลายพันธุ์ (mutation) เป็นต้น โปรแกรมเชิงพันธุกรรมถูกนำมาใช้ครั้งแรกโดย Stephen F. Smith และ Nichael L. Cramer ซึ่งต่อมา John R. Koza ได้เป็นผู้ที่นำโปรแกรมเชิงพันธุกรรมมาประยุกต์และพัฒนาใช้กับงานด้านต่างๆ โดยมีจุดประสงค์ให้คอมพิวเตอร์สามารถเขียนหรือสร้างโปรแกรมคอมพิวเตอร์ขึ้นมาได้ด้วยตัวเอง หลักการทำงานของโปรแกรมเชิงพันธุกรรมคือเริ่มแรกจะมีการ

สร้างโปรแกรมคอมพิวเตอร์ขึ้นมาจำนวนหนึ่ง (เรียกว่า individual หรือ ประชากรแต่ละตัว) มีลักษณะคล้ายโครงสร้างต้นไม้ ดังภาพที่ 2.11 โดยประชากรแต่ละตัวก็จะมีค่าฟิตเนสกำหนดอยู่เป็นค่าที่ใช้บอกว่าประชากรตัวนั้นมีความเหมาะสมในการแก้ปัญหามากน้อยเพียงใด

ภาพที่ 2.11
ตัวอย่างโครงสร้างต้นไม้ที่สร้างขึ้นโดยโปรแกรมเชิงพันธุกรรม



ขั้นตอนการทำงานของโปรแกรมเชิงพันธุกรรมสามารถอธิบายได้ดังนี้

1. เริ่มต้นที่รุ่น (generation) ที่ 0 (G=0) จะมีการสร้างประชากรแต่ละตัวแบบสุ่มจากเทอร์มินอลเซต (T) และฟังก์ชันเซต (F) ตามจำนวนของขนาดประชากรที่ต้องกำหนดไว้ตอนแรก
2. หาค่าฟิตเนสในการแก้ปัญหาของประชากรแต่ละตัว
3. ตรวจสอบเงื่อนไขในการหยุดทำงานดังนี้
   (1) โปรแกรมเชิงพันธุกรรมได้สร้างประชากรแต่ละตัวที่มีค่าฟิตเนสตามที่กำหนด
   (2) ทำงานครบจำนวนรอบหรือครบตามจำนวนรุ่นที่กำหนดไว้
4. ถ้าเงื่อนไขใดเงื่อนไขหนึ่งในข้อ 3 เป็นจริง โปรแกรมเชิงพันธุกรรมจะหยุดทำงานและได้ผลลัพธ์เป็นประชากรที่มีค่าฟิตเนสที่ดีที่สุด
5. ถ้าเงื่อนไขในข้อ 3 ยังเป็นเท็จทั้งคู่ ให้โปรแกรมเชิงพันธุกรรมทำงานต่อไป

6. โปรแกรมเชิงพันธุกรรมจะเลือกวิธีในการสร้างประชากรลูก หรือ ประชากรในรุ่นต่อไป จากวิธีการเปลี่ยนถ่ายพันธุกรรม และการสร้างประชากรใหม่
7. หลังจากนั้นจะมีการคัดเลือกประชากรหนึ่งหรือสองตัวเป็นประชากรผู้ให้กำเนิด ซึ่งจะใช้ในการสร้างประชากรลูก การพิจารณาประชากรผู้ให้กำเนิดจะพิจารณาจากค่าฟิตเนสโดยประชากรที่มีค่าฟิตเนสดีกว่าจะมีโอกาสถูกเลือกมากกว่า
8. โปรแกรมเชิงพันธุกรรมจะสร้างประชากรลูก ตามวิธีในข้อ 6
   (1) การสร้างประชากรใหม่ คือการสร้างประชากรลูกหนึ่งตัวที่เหมือนกับประชากรผู้ให้กำเนิด ทุกอย่าง
   (2) การเปลี่ยนถ่ายพันธุกรรม คือการสร้างประชากรลูก สองตัว ด้วยการแลกเปลี่ยนโครงสร้างของประชากรผู้ให้กำเนิดสองตัว
9. หลังจากสร้างประชากรลูกแล้ว ประชากรลูกอาจจะถูกเปลี่ยนแปลงโครงสร้างบางส่วนด้วยวิธีการกลายพันธุ์
10. เมื่อสร้างประชากรลูก ในรุ่นที่ 0 ครบตามจำนวนที่ต้องการประชากรลูกที่ถูกสร้างนี้จะถูกจัดให้อยู่ในรุ่นที่ 1 (G=1) โดยที่ในแต่ละรุ่นจะมีขนาดประชากรคงที่เสมอ
11. หาค่าฟิตเนสของประชากรในรุ่นที่ 1
12. โปรแกรมเชิงพันธุกรรมจะทำเช่นนี้ไปเรื่อยๆ จนกว่าเงื่อนไขในข้อ 3 ข้อใดข้อหนึ่งเป็นจริง

2.7 งานวิจัยที่เกี่ยวข้องกับการทำนายปริมาณสารโคแอกกูแลนต์

Mirsepaassi, Calthers และ Dharmappa (1995) นำนิวรอลเน็ตเวิร์กมาใช้ในการทำนายปริมาณสารส้ม และพอลิเมอร์ที่ใช้ในกระบวนการผลิตน้ำประปา สำหรับซอฟต์แวร์ที่ใช้ในการสร้างนิวรอลเน็ตเวิร์กชื่อ NeuralWorks Professional II/PLUS (1994) โครงสร้างของนิวรอลเน็ตเวิร์กใช้เป็นแบบส่งผ่านข้อมูลไปข้างหน้า มีตัวแปรอินพุต 47 ตัว ได้แก่ ค่าพีเอช ความขุ่น สีปรากฏ สีจริง อุณหภูมิ ปริมาณพอลิเมอร์ และปริมาณสารส้ม โดยค่าพีเอช ความขุ่น สีปรากฏ สีจริง และอุณหภูมิใช้ข้อมูลของวันที่จะคำนวณและข้อมูลย้อนหลังไป 6 วัน ส่วนปริมาณพอลิเมอร์และปริมาณสารส้มใช้ข้อมูลย้อนหลังไป 6 วัน รวมเป็นอินพุตทั้งหมด 47 ค่า รูปแบบนิวรอลเน็ตเวิร์ก แบบ 1 ชั้นแฝง และ 2 ชั้นแฝง ถูกนำมาทดสอบเพื่อหาโครงสร้างที่ให้การทำนายแม่นยำที่สุด การเรียนรู้ใช้อัลกอริทึมแบบแบ็กพรอพาเกชัน ทำการบันทึกผลการทดลองทุกๆ 1,000 รอบ

ผลการทดลองพบว่าจำนวนรอบการสอนที่ให้ความผิดพลาดในการทำนายน้อยที่สุดอยู่ที่ 67,000 รอบ และค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยสำหรับการทำนายปริมาณพอลิเมอร์เท่ากับ 0.0067 และสำหรับปริมาณสารส้มเท่ากับ 1.53

Nahm, Lee S. B. , Woo, Lee B. K. และ Shin (1996) พัฒนาโปรแกรมควบคุมเหมาะที่สุดสำหรับกระบวนการเติมโคแอกกูแลนต์ชนิดพอลิอะลูมิเนียมคลอไรด์ เพื่อหาอัตราการเติมโคแอกกูแลนต์ที่ดีที่สุดโดยใช้นิวรอลเน็ตเวิร์ก 2 รูปแบบ แบบแรกสำหรับน้ำที่มีสภาพปกติ (น้ำที่มีความขุ่นน้อยกว่า 30 NTU) อีกแบบหนึ่งสำหรับน้ำที่มีสภาพผิดปกติ (น้ำที่มีความขุ่นตั้งแต่ 30 NTU ขึ้นไป) โดยนิวรอลเน็ตเวิร์กแต่ละอันประกอบด้วยอินพุตจำนวน 5 โหนด ได้แก่ ความขุ่น อุณหภูมิ พีเอช ความเป็นด่าง และอัตราการเติมโคแอกกูแลนต์ ในส่วนของชั้นแฝงประกอบด้วยโหนดจำนวน 12 โหนด และชั้นเอาต์พุตประกอบด้วยโหนดจำนวน 1 โหนด ซึ่งเป็นค่าความขุ่นของน้ำที่ผ่านการบำบัดแล้ว ในการหาอัตราการเติมโคแอกกูแลนต์ที่ดีที่สุดสามารถหาได้จากการใช้อินพุตเป็นค่าอัตราการเติมโคแอกกูแลนต์ที่น้อยที่สุดจนถึงมากที่สุดที่เป็นไปได้ และค่าใดให้ค่าความขุ่นของน้ำที่ผ่านการบำบัดแล้วน้อยที่สุด ถือว่าค่านั้นเป็นอัตราการเติมโคแอกกูแลนต์ที่ดีที่สุดและจะทำการหาซ้ำทุกๆ 10 นาที สำหรับค่าความคลาดเคลื่อนจากการทดลองโดยใช้นิวรอลเน็ตเวิร์กพบว่าในกรณีสภาพน้ำปกติมีรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.138 และในกรณีสภาพน้ำผิดปกติเท่ากับ 1.193

Han, Nahm, Woo, Kim และ Ryu (1997) ใช้โมเดลคลุมเครือ (fuzzy model) และนิวรอลเน็ตเวิร์กในการหาอัตราการเติมโคแอกกูแลนต์ชนิดพอลิอะลูมิเนียมคลอไรด์ โดยใช้โมเดลคลุมเครือในกรณีน้ำมีสภาพปกติ (น้ำที่มีความขุ่นน้อยกว่า 30 NTU) และใช้นิวรอลเน็ตเวิร์กสำหรับน้ำที่มีสภาพผิดปกติ (น้ำที่มีความขุ่นตั้งแต่ 30 NTU ขึ้นไป) โดยทั้งโมเดลคลุมเครือและนิวรอลเน็ตเวิร์กมีอินพุตทั้งหมด 5 ตัวได้แก่ ความขุ่น อุณหภูมิ พีเอช ความเป็นด่าง และ Δพีเอช ซึ่งเป็นค่าที่ใช้แทนปริมาณสาหร่ายในน้ำ สามารถคำนวณได้จาก พีเอช$_{\text{flash mixer}}$ – พีเอช$_{\text{sedimentation basin}}$ สำหรับเอาต์พุตของนิวรอลเน็ตเวิร์กเป็นอัตราการเติมโคแอกกูแลนต์ และใช้อัลกอริทึมแบ็กพรอพาเกชันในการสอนนิวรอลเน็ตเวิร์ก สำหรับผลการทดลองพบว่าวิธีการนี้ให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 0.8 ในขณะที่วิธีการโมเดลแบบถดถอยให้ค่าความคลาดเคลื่อน 5.2

Chun, Kwak และ Ryu (1999) ใช้ระบบอนุมานแบบคลุมเครือโดยใช้เน็ตเวิร์กปรับตัวได้ (Adaptive Network-Based Fuzzy Inference System, ANFIS) เปรียบเทียบกับนิวรอลเน็ตเวิร์ก และการวิเคราะห์ถดถอยเชิงเส้น โดยที่ตัว ANFIS นี้เป็นการรวมข้อดีของนิวรอลเน็ตเวิร์ก

และฟัซซีโมเดลไว้ด้วยกัน สำหรับอินพุตของระบบได้แก่ ความขุ่น อุณหภูมิ พีเอช และความเป็นด่าง เพื่อใช้ในการหาความสัมพันธ์กับปริมาณการเติมโคแอกกูแลนต์ จากผลการทดลองพบว่า ANFIS ให้ผลการทำนายที่แม่นยำกว่านิวรอลเน็ตเวิร์กและการวิเคราะห์ถดถอยเชิงเส้น

Valentin, Denoeux และ Fotoohi (1999) ทำนายปริมาณการเติมโคแอกกูแลนต์ในน้ำ โดยระบบถูกแบ่งออกเป็น 3 โมดูล ได้แก่ 1. การตรวจสอบความถูกต้องของข้อมูลแบบพารามิเตอร์เดียว 2. การตรวจสอบความถูกต้องของข้อมูลแบบพารามิเตอร์หลายตัวและการสร้างขึ้นมาใหม่ 3. การสร้างโมเดลของปริมาณการเติมโคแอกกูแลนต์ โดยการตรวจสอบความถูกต้องของข้อมูลแบบพารามิเตอร์เดียวมีหน้าที่ในการตรวจสอบความถูกต้องของข้อมูลดิบที่เก็บมาจากกระบวนการ โดยวัดเป็นระดับความน่าเชื่อถือตั้งแต่ 0-1 โดยถ้ามีค่า 0 แปรว่าข้อมูลที่ได้มานั้นไม่มีความน่าเชื่อถือ และถ้ามีค่าเป็น 1 แปรว่าข้อมูลนั้นมีความน่าเชื่อถือสูงมาก ส่วนการตรวจสอบความถูกต้องของข้อมูลแบบพารามิเตอร์หลายตัวและการสร้างขึ้นมาใหม่จะใช้แผนผังการจัดระเบียบตัวเอง (Self-Organizing Map, SOM) ในการจับคู่ข้อมูลคุณภาพของน้ำดิบไปเป็นอาร์เรย์ 2 มิติ ในส่วนข้อมูลที่หายไปหรือมีไม่ครบถ้วนจะถูกประมาณค่าขึ้นมาใหม่ สำหรับโมดูลที่ 3 การสร้างโมเดลของปริมาณการเติมโคแอกกูแลนต์ เป็นการนำนิวรอลเน็ตเวิร์กมาเรียนรู้โดยใช้อัลกอริทึมแบ็กพรอพาเกชัน ใช้ตัวแปรอินพุตทั้งหมด 6 ตัว ได้แก่ ความขุ่น การนำไฟฟ้า พีเอช อุณหภูมิ ปริมาณออกซิเจนละลายน้ำ และ การดูดซับแสงยูวี ส่วนเอาต์พุตได้แก่ปริมาณการเติมสารโคแอกกูแลนต์ ข้อมูลที่นำมาใช้เป็นข้อมูลระยะเวลา 1 ปี ที่เก็บในช่วงพฤศจิกายน 1997 ถึง พฤศจิกายน 1998 จากการเทียบผลในการทำนายระหว่างนิวรอลเน็ตเวิร์กและโมเดลถดถอยเชิงเส้น พบว่านิวรอลเน็ตเวิร์กนั้นให้ผลในการทำนายที่ดีกว่า

Barker, Nickels และ Mayfield (2003) ได้ศึกษาถึงความเป็นไปได้ในการนำนิวรอลเน็ตเวิร์กแบบหลายชั้นมาประยุกต์ใช้ในการนำมาทำนายปริมาณความขุ่นของน้ำออก (effluent turbidity) โดยในเบื้องต้นตัวแปรอินพุตได้แก่ ปริมาณโคแอกกูแลนต์ (ชนิดของโคแอกกูแลนต์ที่ใช้มี 2 ชนิด ได้แก่ เฟอร์ริกซัลเฟต (ferric sulfate) และ Clar+ion) อุณหภูมิ ความขุ่นของน้ำเข้า (influent turbidity) พีเอช ความเป็นด่าง ความกระด้าง และปริมาณน้ำที่บำบัดแต่ละวัน ซึ่งหลังจากทำการทดลองซ้ำหลายๆ ครั้งพบว่าตัวแปรอินพุตที่ถือว่ามีนัยสำคัญต่อปริมาณความขุ่นของน้ำออกมากที่สุดได้แก่ ปริมาณโคแอกกูแลนต์ อุณหภูมิ และความขุ่นของน้ำเข้า อัลกอริทึมที่ใช้ในการเรียนรู้ได้แก่แบ็กพรอพาเกชัน ข้อมูลที่ใช้ในการเรียนรู้เป็นข้อมูลจากโรงงานผลิตน้ำประปาทางเหนือของรัฐเคนทักกี ในช่วงระยะเวลา 3 ปี ตั้งแต่ปี 1997-1999 และใช้นิวรอลเน็ตเวิร์กในการทำนายข้อมูลช่วงระยะเวลา 6 เดือนแรกของปี 2000 แต่ยกเว้นข้อมูลในช่วงฤดู

ใบไม้ผลิเนื่องจากข้อมูลในช่วงนี้พบว่ามีค่าความขุ่นสูงมากกว่าข้อมูลที่นำมาใช้เรียนรู้ สำหรับผลการทำนายพบว่าการทำนายร้อยละ 84 ของนิวรอลเน็ตเวิร์ก และร้อยละ 88 ของโมเดลแบบถดถอยสามารถทำนายได้ถูกต้องอยู่ในช่วง ± 0.5 NTU และเมื่อนิวรอลเน็ตเวิร์กและโมเดลแบบถดถอยมาทำนายข้อมูลในช่วงฤดูใบไม้ผลิของปี 2000 ที่ความขุ่นมีข้อมูลสูงกว่าข้อมูลที่นำมาเรียนรู้ พบว่าการทำนายร้อยละ 56 ของนิวรอลเน็ตเวิร์ก และร้อยละ 63.8 ของโมเดลแบบถดถอยสามารถทำนายได้ถูกต้องอยู่ในช่วง ± 0.5 NTU

Bae, Choi, Lee, Y. Kim และ S. Kim (2004) ได้นำต้นไม้ตัดสินใจและนิวรอลเน็ตเวิร์กมาใช้ในการทำนายชนิดของสารโคแอกกูแลนต์และปริมาณสารที่ใช้ ของโรงงานผลิตน้ำประปา Duksan เมือง Busan โดยนำต้นไม้ตัดสินใจมาใช้ในการเลือกประเภทสารโคแอกกูแลนต์ซึ่งมี 3 ชนิด ได้แก่ พอลิอะลูมิเนียมคลอไรด์ พอลิอะลูมิเนียมซัลเฟตซิลิเกต และพอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟต ให้เหมาะสมกับสภาพของน้ำดิบ หลังจากนั้นจะใช้นิวรอลเน็ตเวิร์กในการทำนายปริมาณสารโคแอกกูแลนต์ที่ต้องเติมเพื่อให้เกิดการตกตะกอนดีที่สุด สำหรับข้อมูลที่นำมาใช้เป็นข้อมูลที่ได้จากการทำจาร์เทสต์ซึ่งเก็บในช่วงปี 2003 โดยอินพุตของต้นไม้ตัดสินใจและนิวรอลเน็ตเวิร์กมี 5 ตัว ได้แก่ ความเป็นด่าง พีเอช คลอโรฟิลล์-เอ อุณหภูมิ และความขุ่น ซึ่งอินพุตสำหรับนิวรอลเน็ตเวิร์กจะถูกนำมาทำนอร์มอลไลซ์ด้วยค่า 0.1-0.9 โครงสร้างของนิวรอลเน็ตเวิร์กที่ใช้ทำนายปริมาณพอลิอะลูมิเนียมคลอไรด์ พอลิอะลูมิเนียมซัลเฟตซิลิเกต และพอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟต ประกอบด้วย 4 ชั้น ได้แก่ ชั้นอินพุต ชั้นแฝงที่ 1 ชั้นแฝงที่ 2 และชั้นเอาต์พุต สำหรับจำนวนโหนดของชั้นอินพุตมีจำนวน 5 โหนด ชั้นเอาต์พุตมีจำนวน 1 โหนด ส่วนชั้นแฝงที่ 1 และ 2 ของนิวรอลเน็ตเวิร์กที่ใช้ทำนายปริมาณโคแอกกูแลนต์ของทั้ง 3 ชนิดนั้นไม่เท่ากัน โดยชั้นแฝงที่ 1 ของพอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟต มีจำนวน 3 โหนด ชั้นแฝงที่ 2 มีจำนวน 4 โหนด ส่วนพอลิอะลูมิเนียมซัลเฟตซิลิเกต ชั้นแฝงที่ 1 มีจำนวน 2 โหนด ชั้นแฝงที่ 2 มีจำนวน 3 และสุดท้าย พอลิอะลูมิเนียมคลอไรด์ ชั้นแฝงที่ 1 จำนวน 1 โหนด ชั้นแฝงที่ 2 จำนวน 2 โหนด จากผลการทดลองพบว่าการทำนายปริมาณพอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟต โดยแบ่งเป็นข้อมูลชุดเรียนรู้ร้อยละ 50 และข้อมูลชุดทดสอบร้อยละ 50 ให้ผลการทำนายแม่นยำมากที่สุด คือมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยที่ 0.0058 ซึ่งสาเหตุหนึ่งเนื่องมาจากจำนวนข้อมูลที่นำมาใช้ในการเรียนรู้ของพอลิออแกนิกอะลูมิเนียมแมกนีเซียมซัลเฟตที่มีมากกว่าพอลิอะลูมิเนียมซัลเฟตซิลิเกตและพอลิอะลูมิเนียมคลอไรด์

บทที่ 3

แนวทางการทดลอง

ในบทนี้จะกล่าวถึง ที่มาและลักษณะของข้อมูลที่นำมาใช้ในการทดลอง การเตรียมการทดลอง วิธีการจัดเตรียมข้อมูลอินพุตสำหรับใช้ในการทดลอง รูปแบบโครงสร้างและวิธีการปรับค่าพารามิเตอร์ของนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ แนวทางการทดลองของวิธีการต่างๆ วิธีการสร้างตัวแปรอินพุตเพิ่มเติมด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม การปรับค่าพารามิเตอร์ของโปรแกรมเชิงพันธุกรรมและการนำมาใช้ การปรับลดตัวแปรโดยเลือกเฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อการเติมสารส้มมาทำการทดลอง และวิธีการเปรียบเทียบความแม่นยำของแต่ละวิธีการทำนาย

3.1 การเตรียมข้อมูล

ข้อมูลที่นำมาใช้ในการทดลองเป็นข้อมูลของโรงงานผลิตน้ำประปาบางเขน ซึ่งเก็บในช่วงระยะเวลา 1 ปี ตั้งแต่วันที่ 1 มกราคม 2549 ถึง 31 ธันวาคม 2549 โดยมีอินพุตจำนวน 7 ตัว ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน ส่วนเอาต์พุต คือ ปริมาณสารส้ม และตัวแปรแต่ละตัวมีหน่วยดังต่อไปนี้ ความขุ่นมีหน่วยเป็น NTU (Nephelometric Turbidity Units) ความเป็นด่างมีหน่วยเป็นมิลลิกรัมต่อลิตร (mg/l) พีเอชไม่มีหน่วย การนำไฟฟ้ามีหน่วยเป็นไมโครซีเมนส์ต่อเซนติเมตร (μ S/cm) สีมีหน่วยเป็น Pt-Co ปริมาณสารแขวนลอยมีหน่วยเป็นมิลลิกรัมต่อลิตร แอมโมเนียไนโตรเจนมีหน่วยเป็นมิลลิกรัมต่อลิตร และ ปริมาณสารส้มที่เติมมีหน่วยเป็นมิลลิกรัมต่อลิตร มีจำนวนข้อมูลทั้งสิ้น 2,014 เรคอร์ด ตัวอย่างข้อมูลบางส่วนแสดงดังตารางที่ 3.1 ในการเตรียมข้อมูลสำหรับทดลองลำดับการจดบันทึกข้อมูลจะถูกนำมาสลับแบบสุ่ม จากนั้นข้อมูลจะถูกเตรียมออกเป็น 10 ชุด สำหรับใช้ในการทดสอบแบบไขว้ข้ามสิบชุด (10 fold cross validation) แต่ละชุดข้อมูลจะถูกแบ่งออกเป็น 10 ชุดข้อมูลย่อย 1 ชุดข้อมูลย่อยจะใช้สำหรับเป็นชุดทดสอบความถูกต้อง (validation set) อีก 1 ชุดข้อมูลย่อยจะใช้เป็นชุดทดสอบ (test set) ส่วนที่เหลืออีก 8 ชุดใช้สำหรับเป็นชุดเรียนรู้ (training set) โดยชุดทดสอบความถูกต้อง ชุดทดสอบ และชุดเรียนรู้จะถูกสลับเปลี่ยนไปเรื่อยๆ ในแต่ละชุดข้อมูลซึ่งสามารถแสดงได้ดังภาพที่ 3.1

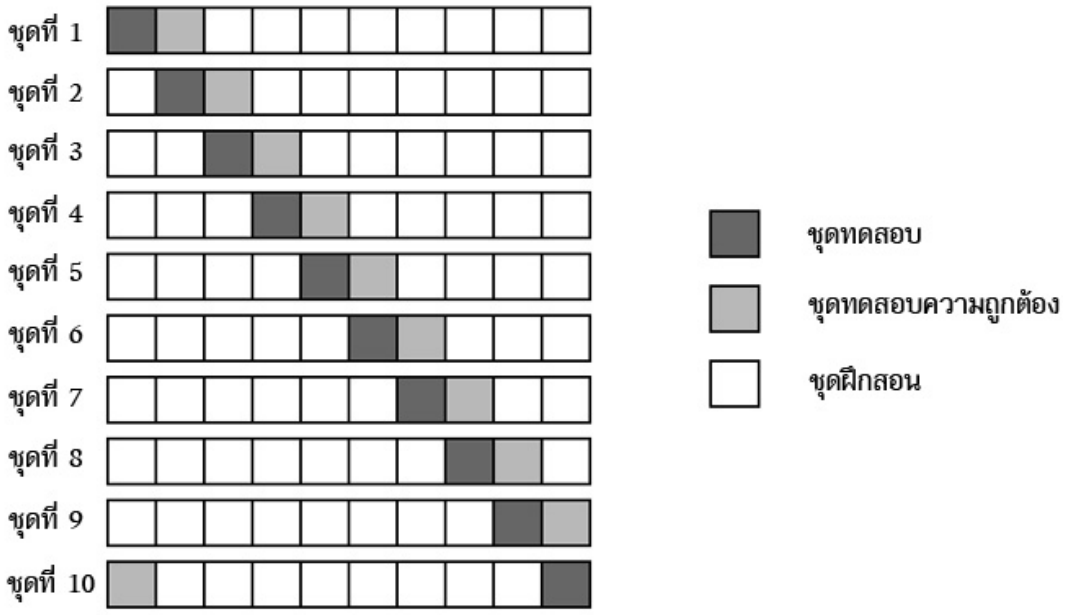
ตารางที่ 3.1

ตัวอย่างข้อมูลจากโรงงานผลิตน้ำประปาบางเขน

อินพุต							เอาต์พุต
ความขุ่น (NTU)	ความเป็นด่าง (มิลลิกรัมต่อลิตร)	พีเอช	การนำไฟฟ้า (ไมโครซีเมนส์ต่อเซนติเมตร)	สี (Pt-Co)	สารแขวนลอย (มิลลิกรัมต่อลิตร)	แอมโมเนียไนโตรเจน (มิลลิกรัมต่อลิตร)	สารส้ม (มิลลิกรัมต่อลิตร)
36	94	7.72	304	30	36	0.072	18
44	90	7.56	284	38	31	0.332	24
80	80	7.51	193	122	64	0.158	31
49	100	7.51	297	40	33	0.207	18
106	71	7.76	191	166	71	0.287	40
48	103	7.66	294	49	32	0.211	19
44	97	7.71	294	74	35	0.141	16
101	80	7.61	216	180	72	0.184	48
81	68	7.49	180	166	47	0.212	35

ภาพที่ 3.1

การแบ่งชุดข้อมูลเพื่อนำมาทำการทดสอบแบบไขว้ข้ามสิบชุด

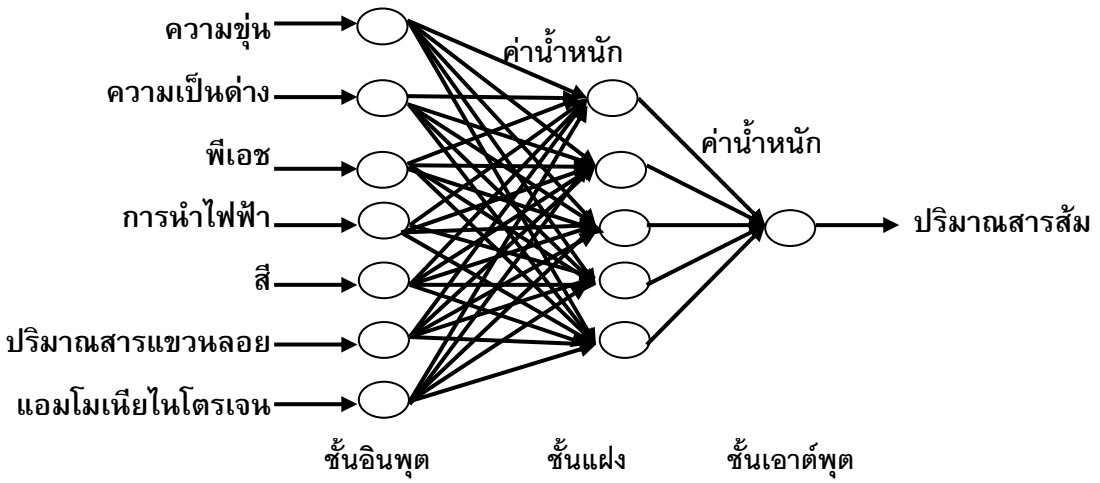


3.2 การเตรียมนิวรอลเน็ตเวิร์ก

3.2.1 โครงสร้างนิวรอลเน็ตเวิร์ก

โครงสร้างนิวรอลเน็ตเวิร์กที่นำมาใช้เป็นนิวรอลเน็ตเวิร์กแบบหลายชั้นดังภาพที่ 3.2 ซึ่งโครงสร้างประกอบไปด้วยชั้นอินพุตจำนวน 1 ชั้น ชั้นแฝงที่ 1 จำนวน 1 ชั้น ชั้นแฝงที่ 2 จำนวน 1 ชั้น ชั้นเอาต์พุตจำนวน 1 ชั้น การทดลองจะทำการปรับเปลี่ยนจำนวนโหนดในชั้นแฝงที่ 1 และ 2 ซึ่งจะได้รูปแบบของนิวรอลเน็ตเวิร์กที่จะทดสอบจำนวน 6 รูปแบบ ดังตารางที่ 3.2 สำหรับอัลกอริทึมที่ใช้ในการเรียนรู้ของนิวรอลเน็ตเวิร์กใช้อัลกอริทึมแบบแบ็กพรอพาเกชัน

ภาพที่ 3.2
ตัวอย่างโครงสร้างของนิวรอลเน็ตเวิร์กแบบหลายชั้นที่ใช้ในการทดลอง



ตารางที่ 3.2
รูปแบบการปรับเปลี่ยนจำนวนโหนดของนิวรอลเน็ตเวิร์ก

รูปแบบที่	จำนวนโหนดในชั้นอินพุต	จำนวนโหนดในชั้นแฝงที่ 1	จำนวนโหนดในชั้นแฝงที่ 2	จำนวนโหนดในชั้นเอาต์พุต
1	7	10	ไม่ใช้	1
2	7	15	ไม่ใช้	1
3	7	20	ไม่ใช้	1

ตารางที่ 3.2 (ต่อ)

รูปแบบที่	จำนวนโหนดในชั้นอินพุต	จำนวนโหนดในชั้นแฝงที่ 1	จำนวนโหนดในชั้นแฝงที่ 2	จำนวนโหนดในชั้นเอาต์พุต
4	7	5	5	1
5	7	10	5	1
6	7	10	10	1

3.2.2 การปรับค่าพารามิเตอร์

ทดลองปรับค่าพารามิเตอร์เพื่อหาค่าที่เหมาะสมที่สุดดังนี้

ค่าอัตราการเรียนรู้ (μ) 3 ค่า ได้แก่ 0.05 และ 0.1

ค่าโมเมนตัม (n) 3 ค่า ได้แก่ 0.25, 0.5 และ 0.75

3.2.3 ขั้นตอนการทดลอง

1.เพื่อหาโครงสร้างของนิวรอลเน็ตเวิร์กที่เหมาะสมที่สุด การทดลองนี้ได้ปรับเปลี่ยนค่า μ , n และจำนวนโหนดในชั้นแฝง โดยทดสอบทั้งหมด 36 กรณี (1 โครงสร้างนิวรอลเน็ตเวิร์ก x 6 รูปแบบของจำนวนโหนดในชั้นแฝง x 2 ค่า μ x 3 ค่า n)

2.หาจำนวนรอบการสอนที่เหมาะสม โดยใช้ข้อมูลสำหรับทดสอบความถูกต้อง กำหนดจำนวนรอบสูงสุดที่ 100,000 รอบ

3. ทดสอบแบบไขว้ข้ามสิบชุดเพื่อหาโครงสร้างของนิวรอลเน็ตเวิร์กที่ให้ผลการทำนายผิดพลาดน้อยที่สุด

3.3 การเตรียมซัพพอร์ตเวกเตอร์แมชชีน

3.3.1 ชนิดของซัพพอร์ตเวกเตอร์แมชชีน

ชนิดของซัพพอร์ตเวกเตอร์แมชชีนที่นำมาใช้ในการทดลองได้แก่ ε-SVR (epsilon-Support Vector Regression) และใช้ฟังก์ชันเคอร์เนลชนิดเรเดียลเบซิส (RBF) ดังนั้นจึงมีพารามิเตอร์ที่ต้องปรับจำนวน 3 ตัวด้วยกันได้แก่ γ (Gamma), C และ epsilon (ε)

3.3.2 การปรับค่าพารามิเตอร์

ทดลองปรับค่าพารามิเตอร์เพื่อหาค่าที่เหมาะสมที่สุดดังนี้

ค่า γ จำนวน 3 ค่า ได้แก่ 5 และ 10

ค่า C จำนวน 3 ค่า ได้แก่ 1,000, 2,000 และ 3,000

ค่า ε จำนวน 3 ค่า ได้แก่ 0.1 และ 1

ดังนั้นจะได้รูปแบบที่จะต้องทดสอบทั้งหมด 12 กรณี (2 ค่า γ x 3 ค่า C X 2 ค่า ε) โดยทุกกรณีจะทำการทดสอบแบบไขว้ข้ามสิบชุด

3.4 การเตรียมต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจมีพารามิเตอร์ที่ต้องปรับเปลี่ยนค่าจำนวน 2 ตัว ได้แก่

1. ค่าจำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อ (minimum size node to split) ซึ่งมีผลคือถ้าโหนดนั้นมีจำนวนโหนดที่จะแตกต่อน้อยกว่าค่านี้โหนดนั้นจะหยุดแตกโหนด

2. ค่าจำนวนระดับชั้นสูงสุดของต้นไม้ตัดสินใจ (maximum tree level)

ทดลองปรับค่าพารามิเตอร์จำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อจำนวน 5 ค่าได้แก่ 2, 3, 5, 10 และ 15 ส่วนจำนวนระดับชั้นสูงสุดของต้นไม้ตัดสินใจทดลองปรับจำนวน 4 ค่า ได้แก่ 5, 10, 50, 100

ดังนั้นจะได้รูปแบบของต้นไม้ตัดสินใจสำหรับทดสอบจำนวน 20 กรณี (จำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อ 5 ค่า x ระดับชั้นสูงสุดของต้นไม้ตัดสินใจ 4 ค่า) โดยทุกรูปแบบทำการทดสอบแบบไขว้ข้ามสิบชุดเพื่อหาค่าพารามิเตอร์ที่เหมาะสมที่สุด

3.5 การเตรียมป่าของต้นไม้ตัดสินใจ

ป่าของต้นไม้ตัดสินใจที่นำมาใช้ในการทดลองกำหนดค่า m_{try} เท่ากับ 5 มีค่าพารามิเตอร์ที่ทำการปรับเปลี่ยนได้แก่ จำนวนของต้นไม้ตัดสินใจในป่า (number of trees in forest) โดยทดลองปรับเปลี่ยนค่าจำนวน 5 ค่า ได้แก่ 50, 100, 200, 500 และ 1,000 โดยทุกค่า

จะทำการทดสอบแบบไขว้ข้ามสิบชุด เพื่อหาจำนวนของต้นไม้ตัดสินใจในป่าที่ให้ผลการทำนายที่ดีที่สุด

3.6 การสร้างข้อมูลอินพุตเพิ่มเติม

เพื่อช่วยให้การทำนายปริมาณสารส้มด้วยวิธีต่างๆ สามารถทำนายผลได้แม่นยำมากขึ้น งานวิจัยนี้ได้ใช้วิธีสร้างตัวแปรอินพุตเพิ่มเติมเพื่อช่วยในการทำนาย โดยวิธีการสร้างตัวแปรอินพุตเพิ่มเติมในงานวิจัยนี้ได้เสนอเทคนิคในการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมในการสร้างตัวแปรอินพุตเพิ่มเติมเพื่อทำงานร่วมกับวิธีในการทำนายปริมาณสารส้มแบบต่างๆ โดยตัวแปรอินพุตที่สร้างขึ้นมาใหม่เป็นค่าที่ได้จากการทำนายค่าเอาต์พุตโดยใช้โปรแกรมเชิงพันธุกรรม ซึ่งโปรแกรมเชิงพันธุกรรมจะทำการหาฟังก์ชันความสัมพันธ์ระหว่างอินพุตและเอาต์พุตของชุดข้อมูลสำหรับเรียนรู้ และแต่ละชุดข้อมูลจะทำการหาซ้ำ 5 ครั้ง แล้วจึงเลือกฟังก์ชันที่มีค่าฟิตเนส (fitness) ดีที่สุด (ฟังก์ชันที่สามารถคำนวณเอาต์พุตได้ใกล้เคียงกับเอาต์พุตจริงมากที่สุด) ฟังก์ชันดังกล่าวจะนำมาคำนวณหาค่าที่จะใช้เป็นตัวแปรอินพุตเพิ่มเติมสำหรับข้อมูลอินพุตชุดใหม่

เมื่อได้ข้อมูลอินพุตชุดใหม่แล้ว จึงนำวิธีป่าของต้นไม้ตัดสินใจและรูปแบบของนิวรอลเน็ตเวิร์กที่ให้ผลการทำนายแม่นยำที่สุดสำหรับตัวแปรอินพุตแบบ 7 ตัว มาใช้ทำนายข้อมูลอินพุตชุดใหม่ เพื่อดูแนวโน้มว่าผลของการสร้างข้อมูลอินพุตเพิ่มเติมช่วยให้วิธีการป่าของต้นไม้ตัดสินใจและนิวรอลเน็ตเวิร์กทำนายผลได้แม่นยำขึ้นหรือไม่

การปรับค่าพารามิเตอร์ของโปรแกรมเชิงพันธุกรรม มีพารามิเตอร์ที่ต้องทำการปรับค่าดังนี้

1. ขนาดประชากร (population-size) คือ จำนวนประชากรหรือจำนวนฟังก์ชันที่โปรแกรมเชิงพันธุกรรมสร้างขึ้นมาในแต่ละรุ่น
2. ขนาดของประชากรที่นำมาเปรียบเทียบ (tournament-size) คือ จำนวนประชากรในแต่ละกลุ่มที่จะนำมาเปรียบเทียบกันเพื่อหาประชากรที่มีค่าฟิตเนสดีที่สุด
3. จำนวนรุ่น (generations) คือ จำนวนรุ่นหรือ จำนวนรอบการทำงานของโปรแกรมเชิงพันธุกรรม
4. ความน่าจะเป็นที่จะทำการการเปลี่ยนถ่ายพันธุกรรม (crossover-probability)
5. ความน่าจะเป็นที่จะทำการการกลายพันธุ์ (node-mutation-probability)

เพื่อหาฟังก์ชันความสัมพันธ์ระหว่างอินพุตและเอาต์พุตที่มีค่าฟิตเนสดีที่สุด ในงานวิจัยนี้ได้ทำการปรับค่าพารามิเตอร์สำหรับการทดลองดังนี้ ขนาดประชากรเท่ากับ 5,000 ขนาดของประชากรที่นำมาเปรียบเทียบเท่ากับ 20 จำนวนรุ่นเท่ากับ 200 ความน่าจะเป็นที่จะทำการการเปลี่ยนถ่ายพันธุกรรมเท่ากับ 0.9 และความน่าจะเป็นที่จะทำการการกลายพันธุ์เท่ากับ 0.01

3.7 การทำนายปริมาณสารส้มโดยใช้เฉพาะตัวแปรอินพุตหลัก

เพื่อศึกษาถึงผลของตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม จากงานวิจัยของ Mirsepaassi (1995) Han (1997) Chun (1999) Valentin (1999) และ Bae (2004) พบว่าตัวแปรอินพุตหลักที่นิยมใช้ในงานวิจัยส่วนใหญ่ในการทำนายปริมาณสารโคแอกกูแลนต์ ได้แก่ ความขุ่น อุณหภูมิ พีเอช และความเป็นด่าง ดังนั้นในงานวิจัยนี้จึงทดลองใช้อินพุตดังกล่าวเพื่อศึกษาถึงผลของตัวแปรอินพุตหลักที่มีต่อความแม่นยำในทำนายปริมาณการเติมสารส้ม แต่เนื่องจากโรงผลิตน้ำประปาบางเขนไม่มีการจดบันทึกค่าอุณหภูมิของน้ำไว้ งานวิจัยนี้จึงเลือกใช้เฉพาะค่าตัวแปรอินพุต 3 ตัว ได้แก่ ความขุ่น พีเอช และความเป็นด่าง

3.8 วิธีการคำนวณหาความผิดพลาด

ในการหาว่าวิธีการทำนายวิธีใดสามารถทำนายผลได้แม่นยำที่สุด ในงานวิจัยนี้ได้ทำการเปรียบเทียบผลความแม่นยำของแต่ละวิธีด้วยการหาค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (Root Mean Square Error, RMSE) ซึ่งแสดงได้ดังสมการที่ 3.1 โดยวิธีการทำนายที่มีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยน้อยที่สุดแปลว่าวิธีการนั้นสามารถทำนายผลได้ใกล้เคียงความถูกต้องหรือมีความแม่นยำมากที่สุด

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(D_i - Y_i\right)^2} \qquad (3.1)$$

โดยที่

D_i คือค่าปริมาณการเติมสารส้มที่ต้องการ ลำดับที่ i

Y_i คือค่าปริมาณการเติมสารส้มได้จากการทำนาย ลำดับที่ i

N คือจำนวนข้อมูล

บทที่ 4

ผลการทดลอง

ในบทนี้จะกล่าวถึง ผลการทดลองของการปรับโครงสร้าง และค่าพารามิเตอร์ ของวิธีการทำนายต่างๆ ได้แก่ นิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ โดยผลการทดลองจะแบ่งออกเป็นสามส่วน ส่วนที่หนึ่งจะเป็นผลการทดลองกรณีการใช้อินพุตแบบ 7 ตัวแปร ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน ส่วนที่สองเป็นผลการทดลองกรณีตัวแปรอินพุตแบบ 8 ตัว ซึ่งตัวแปรอินพุตที่เพิ่มขึ้นได้มาจากการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม ส่วนที่สามเป็นผลการทดลองกรณีตัวแปรอินพุตแบบ 3 ตัว ซึ่งเลือกเฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม ได้แก่ ความขุ่น พีเอช และความเป็นด่าง

4.1 กรณีอินพุตแบบ 7 ตัว

4.1.1 ผลการทดลองของนิวรอลเน็ตเวิร์ก

ผลการทดลองปรับค่าตัวแปรของนิวรอลเน็ตเวิร์กแสดงได้ดังตารางที่ 4.1 ซึ่งพบว่านิวรอลเน็ตเวิร์กที่สามารถทำนายได้แม่นยำที่สุดคือนิวรอลเน็ตเวิร์กที่มีจำนวนโหนดในชั้นแฝงที่ 1 เท่ากับ 10 และจำนวนโหนดในชั้นแฝงที่ 2 เท่ากับ 5 ค่าอัตราการเรียนรู้เท่ากับ 0.1 และค่าโมเมนตัมเท่ากับ 0.25 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.391

ตารางที่ 4.1
ผลการทำนายด้วยนิวรอลเน็ตเวิร์ก

จำนวนโหนดในชั้นแฝงที่ 1	จำนวนโหนดในชั้นแฝงที่ 2	ค่าอัตราการเรียนรู้	ค่าโมเมนตัม	RMSE
10	ไม่ใช้	0.05	0.25	4.438
10	ไม่ใช้	0.05	0.5	4.776
10	ไม่ใช้	0.05	0.75	4.250

ตารางที่ 4.1 (ต่อ)

จำนวนโหนดในชั้นแฝงที่ 1	จำนวนโหนดในชั้นแฝงที่ 2	ค่าอัตราการเรียนรู้	ค่าโมเมนตัม	RMSE
10	ไม่ใช้	0.1	0.25	4.206
10	ไม่ใช้	0.1	0.5	4.199
10	ไม่ใช้	0.1	0.75	4.783
15	ไม่ใช้	0.05	0.25	4.251
15	ไม่ใช้	0.05	0.5	4.401
15	ไม่ใช้	0.05	0.75	4.402
15	ไม่ใช้	0.1	0.25	4.129
15	ไม่ใช้	0.1	0.5	4.317
15	ไม่ใช้	0.1	0.75	5.219
20	ไม่ใช้	0.05	0.25	4.666
20	ไม่ใช้	0.05	0.5	4.620
20	ไม่ใช้	0.05	0.75	4.208
20	ไม่ใช้	0.1	0.25	4.553
20	ไม่ใช้	0.1	0.5	4.142
20	ไม่ใช้	0.1	0.75	5.670
5	5	0.05	0.25	2.568
5	5	0.05	0.5	2.572
5	5	0.05	0.75	2.664
5	5	0.1	0.25	2.559
5	5	0.1	0.5	2.553
5	5	0.1	0.75	2.682
10	5	0.05	0.25	2.469
10	5	0.05	0.5	2.503
10	5	0.05	0.75	2.477
10	5	0.1	0.25	2.391
10	5	0.1	0.5	2.409
10	5	0.1	0.75	2.575
10	10	0.05	0.25	2.460
10	10	0.05	0.5	2.427
10	10	0.05	0.75	2.483
10	10	0.1	0.25	2.4299
10	10	0.1	0.5	2.429
10	10	0.1	0.75	2.430

4.1.2 ผลการทดลองของซัพพอร์ตเวกเตอร์แมชชีน

ผลการทดลองปรับค่าตัวแปรของซัพพอร์ตเวกเตอร์แมชชีนแสดงได้ดังตารางที่ 4.2 จากผลการทดลองพบว่าค่าของตัวแปรที่ทำให้ซัพพอร์ตเวกเตอร์แมชชีนทำนายได้แม่นยำที่สุดคือ ค่า C เท่ากับ 1,000 ค่า γ เท่ากับ 5 ค่า ε เท่ากับ 0.1 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.662

ตารางที่ 4.2
ผลการทำนายด้วยซัพพอร์ตเวกเตอร์แมชชีน

ค่า C	ค่า γ	ค่า ε	RMSE
1,000	5	0.1	3.662
1,000	5	1	3.679
1,000	10	0.1	3.727
1,000	10	1	3.754
2,000	5	0.1	3.830
2,000	5	1	3.766
2,000	10	0.1	3.844
2,000	10	1	3.844
3,000	5	0.1	3.914
3,000	5	1	3.800
3,000	10	0.1	3.950
3,000	10	1	3.944

4.1.3 ผลการทดลองของต้นไม้ตัดสินใจ

ผลการทดลองปรับค่าตัวแปรของต้นไม้ตัดสินใจแสดงได้ดังตารางที่ 4.3 จากผลการทดลองพบว่าค่าตัวแปรที่ให้ผลการทำนายแม่นยำที่สุดคือ ค่าจำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อเท่ากับ 10 จำนวนระดับชั้นสูงสุดของต้นไม้ตัดสินใจเท่ากับ 50 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.763

ตารางที่ 4.3
ผลการทำนายด้วยต้นไม้ตัดสินใจ

จำนวนโหนดต่ำที่สุด ที่ต้นไม้จะแตกต่อ	จำนวนระดับชั้นสูงสุด ของต้นไม้ตัดสินใจ	RMSE
2	5	5.316
2	10	3.885
2	50	3.844
2	100	3.844
3	5	5.316
3	10	3.883
3	50	3.829
3	100	3.829
5	5	5.316
5	10	3.867
5	50	3.766
5	100	3.766
10	5	5.316
10	10	3.865
10	50	3.763
10	100	3.763
15	5	5.316
15	10	3.868
15	50	3.797
15	100	3.797

4.1.4 ผลการทดลองของป่าของต้นไม้ตัดสินใจ

ผลการทดลองปรับค่าตัวแปรของป่าของต้นไม้ตัดสินใจแสดงได้ดังตารางที่ 4.4 จากผลการทดลองพบว่าค่าตัวแปรที่ให้ผลการทำนายแม่นยำที่สุดคือ จำนวนป่าเท่ากับ 1,000 ต้น โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.486

ตารางที่ 4.4

ผลการทำนายด้วยป่าของต้นไม้ตัดสินใจ

จำนวนต้นไม้ตัดสินใจในป่า	RMSE
50	2.535
100	2.518
200	2.518
500	2.500
1,000	2.486

4.2 กรณีอินพุตแบบ 8 ตัว

ในกรณีตัวแปรอินพุต 8 ตัว ซึ่งตัวแปรที่สร้างเพิ่มขึ้นมาได้จากการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม ตัวอย่างของสมการความสัมพันธ์ระหว่างอินพุตตัวใหม่ซึ่งเป็นค่าการทำนายปริมาณสารส้ม และตัวแปรอินพุตทั้ง 7 ตัว สามารถแสดงได้ดังนี้

ข้อมูลอินพุตตัวใหม่ = ((((พีเอช+((สี+สี)/ความขุ่น))/((ปริมาณสารแขวนลอย* 0.13)-(ปริมาณสารแขวนลอย/(((((พีเอช/((ปริมาณสารแขวนลอย* 0.13)- 0.69))+ปริมาณสารแขวนลอย)/พีเอช)+(พีเอช/(ปริมาณสารแขวนลอย/ความขุ่น))))))+((((พีเอช/การนำไฟฟ้า)+(((ปริมาณสารแขวนลอย+สี)/(0.13+พีเอช))+(พีเอช/(ปริมาณสารแขวนลอย/ความขุ่น))))- 0.76)-(((สี-ความเป็นด่าง)/ 0.13)/การนำไฟฟ้า)))-((((การนำไฟฟ้า* 0.14)-(ปริมาณสารแขวนลอย/(((ความเป็นด่าง+ปริมาณสารแขวนลอย)/ปริมาณสารแขวนลอย)+ 0.29)))+(0.14/(0.13-(การนำไฟฟ้า/(((สี+ปริมาณสารแขวนลอย)/ 0.13)+(ปริมาณสารแขวนลอย/(พีเอช/ความขุ่น)))))))/(พีเอช+((ความขุ่น+ความขุ่น)/การนำไฟฟ้า))))

ตัวอย่างของชุดข้อมูลชุดใหม่สามารถแสดงได้ดังตารางที่ 4.5 โดยคอลัมน์จีพี (GP) คือตัวแปรอินพุตตัวใหม่ที่ได้จากการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม สำหรับวิธีที่นำมาใช้ทำนายปริมาณสารส้มเลือกมาจากโครงสร้างและค่าตัวแปรของแต่ละวิธีที่ทำนายได้แม่นยำที่สุดในกรณีอินพุต 7 ตัว มาทำการทดลอง โดยผลการทดลองสามารถสรุปได้ ดังต่อไปนี้

ตารางที่ 4.5

ตัวอย่างข้อมูลที่ได้จากการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม

ความขุ่น (NTU)	ความเป็นด่าง (มิลลิกรัมต่อลิตร)	พีเอช	การนำไฟฟ้า (ไมโครซีเมนส์ต่อเซนติเมตร)	สี (Pt-Co)	สารแขวนลอย (มิลลิกรัมต่อลิตร)	แอมโมเนียไนโตรเจน (มิลลิกรัมต่อลิตร)	จีพี	ปริมาณสารส้ม (มิลลิกรัมต่อลิตร)
36	94	7.72	304	30	36	0.072	17.9	18
44	90	7.56	284	38	31	0.332	20.65	24
80	80	7.51	193	122	64	0.158	32.54	31
49	100	7.51	297	40	33	0.207	21.1	18
106	71	7.76	191	166	71	0.287	38.89	40
48	103	7.66	294	49	32	0.211	22.37	19
44	97	7.71	294	74	35	0.141	23.68	16
101	80	7.61	216	180	72	0.184	40.2	48
81	68	7.49	180	166	47	0.212	36.98	35
41	102	7.63	307	26	39	0.201	17.64	18
79	86	7.72	211	122	64	0.175	31.87	29
108	63	7.35	154	178	69	0.196	40.56	48
31	100	7.53	318	37	28	0.151	19.03	18
232	60	7.39	141	340	152	0.274	68.07	44
61	76	7.4	184	99	41	0.31	33.12	39
84	70	8.06	203	142	71	0.206	33.85	50
38	80	7.25	184	79	30	0.128	26.15	28
30	112	7.32	302	59	18	0.101	24.93	40
15	108	7.56	274	29	7	0.107	31.54	28
35	92	7.66	296	29	37	0.091	17.78	22
37	93	7.62	303	38	31	0.332	19.36	24
110	76	7.32	215	196	70	0.23	43.35	50
131	70	7.26	187	164	91	0.186	43.38	43
46	109	7.45	298	65	27	0.116	24.97	43
13	101	7.32	266	36	15	0.135	36.78	28
38	88	7.52	301	62	34	0.115	21.48	18
43	102	7.68	301	37	29	0.102	20.83	18
42	102	7.41	304	57	26	0.145	23.33	20
49	99	7.74	304	31	33	0.188	20.23	26
50	90	7.58	280	42	45	0.125	20.42	16

ตารางที่ 4.5 (ต่อ)

ความขุ่น (NTU)	ความเป็นด่าง (มิลลิกรัมต่อลิตร)	พีเอช	การนำไฟฟ้า (ไมโครซีเมนส์ต่อเซนติเมตร)	สี (Pt-Co)	สารแขวนลอย (มิลลิกรัมต่อลิตร)	แอมโมเนียไนโตรเจน (มิลลิกรัมต่อลิตร)	จีพี	ปริมาณสารส้ม (มิลลิกรัมต่อลิตร)
146	64	7.4	160	186	82	0.277	45.13	40
117	72	7.38	169	168	57	0.27	41.36	37
108	79	7.45	202	167	71	0.131	40.05	38
19	103	7.41	246	38	13	0.145	26	28
87	69	7.47	171	196	63	0.281	39.9	35
45	97	7.48	276	47	34	0.15	21.46	16

4.2.1 ผลการทดลองของนิวรอลเน็ตเวิร์ก

ในกรณีอินพุตแบบ 7 ตัว นิวรอลเน็ตเวิร์กที่สามารถทำนายได้แม่นยำที่สุดมีจำนวนโหนดในชั้นแฝงที่ 1 เท่ากับ 10 และจำนวนโหนดในชั้นแฝงที่ 2 เท่ากับ 5 ค่าอัตราการเรียนรู้เท่ากับ 0.1 และค่าโมเมนตัมเท่ากับ 0.25 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.391 เมื่อนำนิวรอลเน็ตเวิร์กดังกล่าวมาทำนายอินพุตแบบ 8 ตัว พบว่าให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.396

4.2.2 ผลการทดลองของซัพพอร์ตเวกเตอร์แมชชีน

ค่าตัวแปรของซัพพอร์ตเวกเตอร์แมชชีนที่ให้ค่าการทำนายอินพุตแบบ 7 ตัวแม่นยำที่สุดได้แก่ ค่า C เท่ากับ 1,000 ค่า γ เท่ากับ 5 ค่า ε เท่ากับ 0.1 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.662 เมื่อนำค่าตัวแปรดังกล่าวมาทำนายตัวแปรอินพุตแบบ 8 ตัวแปร พบว่าซัพพอร์ตเวกเตอร์แมชชีนให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.678

4.2.3 ผลการทดลองของต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจที่สามารถทำนายได้แม่นยำที่สุดในกรณีอินพุตแบบ 7 ตัว มีค่าตัวแปรจำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อ และจำนวนระดับชั้นสูงสุดของต้นไม้ตัดสินใจ เท่ากับ 10

และ 50 โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.763 เมื่อนำมาทำนายอินพุตแบบ 8 ตัว พบว่าต้นไม้ตัดสินใจดังกล่าวให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.579

4.2.4 ผลการทดลองของป่าของต้นไม้ตัดสินใจ

ป่าของต้นไม้ตัดสินใจที่ทำนายกรณีอินพุต 7 ตัวแม่นยำที่สุด มีจำนวนต้นไม้ในป่าเท่ากับ 1,000 ต้น โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.486 เมื่อนำมาทำนายกรณีอินพุต 8 ตัว พบว่าให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.335

4.3 กรณีอินพุตแบบ 3 ตัว

อินพุต 3 ตัวที่เลือกมาใช้ในการทดลอง เป็นอินพุตหลักที่นิยมใช้ในงานวิจัยส่วนใหญ่ ได้แก่ ความขุ่น พีเอช และความเป็นด่าง การทดลองจะเลือกรูปแบบของโครงสร้างและค่าตัวแปรของแต่ละวิธีที่สามารถทำนายได้แม่นยำที่สุดในกรณีอินพุตแบบ 7 ตัว มาใช้ในการทำนายกรณีอินพุตแบบ 3 ตัว เพื่อดูแนวโน้มความแม่นยำในการทำนาย

4.3.1 ผลการทดลองของนิวรอลเน็ตเวิร์ก

เมื่อนำนิวรอลเน็ตเวิร์กมาทำนายอินพุตแบบ 3 ตัว พบว่าให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.731

4.3.2 ผลการทดลองของซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนเมื่อนำมาทำนายอินพุตแบบ 3 ตัวให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 6.427

4.3.3 ผลการทดลองของต้นไม้ตัดสินใจ

ต้นไม้ตัดสินใจเมื่อนำมาทำนายอินพุตแบบ 3 ตัวให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 5.565

4.3.4 ผลการทดลองของป่าของต้นไม้ตัดสินใจ

วิธีการป่าของต้นไม้ตัดสินใจเมื่อนำมาทำนายอินพุตแบบ 3 ตัวให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 5.008

บทที่ 5

สรุปผลการศึกษา วิเคราะห์ และข้อเสนอแนะ

5.1 สรุปผลการศึกษา

งานวิจัยนี้มีวัตถุประสงค์เพื่อหาวิธีการที่เหมาะสมในการทำนายปริมาณสารส้มที่ใช้ในกระบวนการผลิตน้ำประปา โดยเปรียบเทียบวิธีต่างๆ ในการทำนาย ได้แก่ นิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ เพื่อหาวิธีในการทำนายปริมาณสารส้มที่แม่นยำที่สุด โดยผลการทดลองแบ่งออกเป็นสามส่วนด้วยกันตามจำนวนตัวแปรอินพุตที่ใช้ ส่วนที่หนึ่งเป็นการใช้ตัวแปรอินพุตจำนวน 7 ตัว ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน ส่วนที่สองเป็นการทดลองโดยใช้ตัวแปรอินพุตจำนวน 8 ตัว โดยทำการสร้างตัวแปรอินพุตเพิ่มเติม โดยในงานวิจัยนี้ได้เสนอเทคนิคในการสร้างตัวแปรอินพุตเพิ่มเติมด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม ส่วนที่สามเลือกใช้ตัวแปรอินพุตเฉพาะตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม จำนวน 3 ตัว ได้แก่ ความขุ่น พีเอช และความเป็นด่าง

ผลการทดลองส่วนที่หนึ่ง ในกรณีตัวแปรอินพุตแบบ 7 ตัว ซึ่งเลือกใช้ตัวแปรอินพุตหลักๆ และตัวแปรอินพุตที่คาดว่าจะมีผลต่อปริมาณการเติมสารส้ม ได้แก่ ความขุ่น ความเป็นด่าง พีเอช การนำไฟฟ้า สี ปริมาณสารแขวนลอย และ แอมโมเนียไนโตรเจน ผลการทดลองพบว่ารูปแบบของนิวรอลเน็ตเวิร์กที่ให้ผลการทำนายแม่นยำที่สุดได้แก่ นิวรอลเน็ตเวิร์กที่มีจำนวนโหนดในชั้นแฝงชั้นที่ 1 เท่ากับ 10 และจำนวนโหนดในชั้นแฝงชั้นที่ 2 เท่ากับ 5 และมีค่าอัตราการเรียนรู้และค่าโมเมนตัมเท่ากับ 0.1 และ 0.25 ตามลำดับ โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.391 ซัพพอร์ตเวกเตอร์แมชชีนที่ทำนายผลได้แม่นยำที่สุดมีค่า C เท่ากับ 1,000 ค่า γ เท่ากับ 5 ค่า ε เท่ากับ 0.1 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.662 ต้นไม้ตัดสินใจที่ทำนายผลได้แม่นยำที่สุดมีค่าจำนวนโหนดต่ำที่สุดที่ต้นไม้จะแตกต่อเท่ากับ 10 จำนวนระดับชั้นสูงสุดของต้นไม้ตัดสินใจเท่ากับ 50 โดยให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.763 ป่าของต้นไม้ตัดสินใจพบว่าจำนวนต้นไม้ในป่าที่ให้ผลการทำนายแม่นยำที่สุดมีค่าเท่ากับ 1,000 ต้น โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.486 ดังนั้นในกรณีตัวแปรอินพุตแบบ 7 ตัว นิวรอลเน็ตเวิร์กเป็นวิธีที่สามารถทำนาย

ผลได้แม่นยำที่สุด รองลงมาได้แก่วิธีการป่าของต้นไม้ตัดสินใจ ซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้ตัดสินใจ ตามลำดับ

ผลการทดลองส่วนที่สอง ในกรณีตัวแปรอินพุตแบบ 8 ตัว ซึ่งตัวแปรตัวใหม่ที่เพิ่มขึ้นมาได้มาจากการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมนั้น เมื่อนำนิวรอลเน็ตเวิร์ก ซัพพอร์ต-เวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ ที่ทำนายตัวแปรอินพุตแบบ 7 ตัวได้แม่นยำที่สุดมาทำนายตัวแปรอินพุตแบบ 8 ตัว พบว่านิวรอลเน็ตเวิร์กสามารถทำนายผลได้แม่นยำน้อยกว่ากรณีตัวแปรอินพุตแบบ 7 ตัวเล็กน้อย โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.396 ผลการทำนายของซัพพอร์ตเวกเตอร์แมชชีนพบว่ามีความแม่นยำน้อยกว่ากรณีอินพุตแบบ 7 ตัว โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.678 สำหรับต้นไม้ตัดสินใจพบว่าเมื่อนำมาทำนายอินพุตแบบ 8 ตัว ต้นไม้ตัดสินใจสามารถทำนายได้แม่นยำขึ้นกว่ากรณีอินพุต 7 ตัว โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.579 เช่นเดียวกับป่าของต้นไม้ตัดสินใจซึ่งสามารถทำนายผลได้แม่นยำขึ้นมากกว่ากรณีอินพุตแบบ 7 ตัว โดยมีค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 2.335 โดยจะเห็นว่าวิธีการที่เสนอในงานวิจัยนี้ซึ่งใช้การประยุกต์ใช้โปรแกรมเชิงพันธุกรรมในการสร้างตัวแปรอินพุตเพิ่มเติมช่วยให้ต้นไม้ตัดสินใจและป่าของต้นไม้ตัดสินใจทำนายผลได้แม่นยำมากขึ้นจากกรณีอินพุตแบบ 7 ตัว และเมื่อเปรียบเทียบผลความแม่นยำของทุกวิธีและทุกรูปแบบของตัวแปรอินพุตแล้วพบว่าวิธีการป่าของต้นไม้ตัดสินใจที่ใช้ตัวแปรอินพุตแบบ 8 ตัว เป็นวิธีที่มีความแม่นยำมากที่สุด

ผลการทดลองส่วนที่สาม ในกรณีตัวแปรอินพุตแบบ 3 ตัว ได้แก่ ความขุ่น พีเอช และความเป็นด่าง ซึ่งเป็นตัวแปรอินพุตหลักๆ ที่มีผลต่อปริมาณการเติมสารส้ม ผลการทดลองพบว่าเมื่อนำนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจที่ทำนายตัวแปรอินพุตแบบ 7 ตัวได้แม่นยำที่สุดมาทำนายตัวแปรอินพุตแบบ 3 ตัว พบว่าทุกวิธีให้ผลการทำนายที่แม่นยำน้อยกว่ากรณีอินพุต 7 ตัว และ 8 ตัว โดยนิวรอลเน็ตเวิร์กให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 3.731 ซัพพอร์ตเวกเตอร์แมชชีนให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 6.427 ต้นไม้ตัดสินใจให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 5.565 ป่าของต้นไม้ตัดสินใจให้ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยเท่ากับ 5.008 ซึ่งจะเห็นได้ว่านิวรอลเน็ตเวิร์กเป็นวิธีที่สามารถทำนายตัวแปรอินพุตแบบ 3 ตัวได้ดีกว่าวิธีอื่นๆ อย่างชัดเจน

ผลการทดลองของนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ ในการทำนายปริมาณสารส้มกรณีอินพุต 7 ตัว 8 ตัว และ 3 ตัว สามารถสรุปได้ดังตารางที่ 5.1

ตารางที่ 5.1

สรุปผลการทดลองกรณีอินพุต 7 ตัว 8 ตัว และ 3 ตัว

วิธีการทำนาย	RMSE กรณีอินพุต 7 ตัว	RMSE กรณีอินพุต 8 ตัว	RMSE กรณีอินพุต 3 ตัว
นิวรอลเน็ตเวิร์ก	2.391	2.396	3.731
ซัพพอร์ตเวกเตอร์แมชชีน	3.661	3.678	6.427
ต้นไม้ตัดสินใจ	3.763	3.579	5.565
ป่าของต้นไม้ตัดสินใจ	2.486	2.335	5.008

และเพื่อศึกษาว่าผลลัพธ์ของวิธีการทำนายแต่ละวิธีมีความแตกต่างกันอย่างมีนัยสำคัญทางสถิติหรือไม่ งานวิจัยนี้ได้เลือกทำการทดสอบแบบทีจับคู่ (Paired T-Test) เพื่อเปรียบเทียบผลลัพธ์ของแต่ละวิธี โดยนำผลการทำนายของข้อมูลชุดที่ 1 ถึง 10 ของวิธีที่ทำนายได้แม่นยำที่สุดในแต่ละกรณีมาเปรียบเทียบกับวิธีการที่เหลือ สำหรับค่าพี (P-value) ที่ได้จากการทดสอบแบบทีจับคู่ถ้ามีค่าน้อยกว่า 0.05 แปลว่ากลุ่มของค่า 2 กลุ่มที่นำมาเปรียบเทียบกันมีความแตกต่างอย่างมีนัยสำคัญทางสถิติ ในขณะที่ถ้าค่าพีมีค่ามากกว่า 0.05 แปลว่ากลุ่มของค่า 2 กลุ่มที่นำมาเปรียบเทียบกันนั้นไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติที่ระดับความเชื่อมั่นร้อยละ 95

ผลการทำการทดสอบแบบทีจับคู่ในกรณีตัวแปรอินพุต 7 ตัว พบว่าผลการทำนายของนิวรอลเน็ตเวิร์กแตกต่างกับผลการทำนายของซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้ตัดสินใจอย่างมีนัยสำคัญ โดยมีค่าพี < 0.05 ส่วนผลการทำนายของนิวรอลเน็ตเวิร์กและผลการทำนายของป่าของต้นไม้ตัดสินใจแตกต่างกันอย่างไม่มีนัยสำคัญ โดยมีค่าพี > 0.05 ในกรณีอินพุตแบบ 8 ตัว พบว่าผลการทำนายของป่าของต้นไม้ตัดสินใจกับผลการทำนายของนิวรอลเน็ตเวิร์กมีความแตกต่างกันอย่างไม่มีนัยสำคัญโดยมีค่าพี > 0.05 ในขณะที่ผลการทำนายของซัพพอร์ตเวกเตอร์แมชชีนและต้นไม้ตัดสินใจ พบว่ามีความแตกต่างกันอย่างมีนัยสำคัญเมื่อเปรียบเทียบกับป่าของต้นไม้

ตัดสินใจ โดยมีค่าพี < 0.05 ในกรณีอินพุต 3 ตัว ผลการทำนายของนิวรอลเน็ตเวิร์กแตกต่างอย่างมีนัยสำคัญกับผลการทำนายของซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจ โดยมีค่าพี < 0.05 อย่างไรก็ตามวิธีที่เหมาะสมในการนำไปใช้ในการทำนายปริมาณการเติมสารส้มมากที่สุด ได้แก่ วิธีการป่าของต้นไม้ตัดสินใจที่ใช้ตัวแปรอินพุต 8 ตัว ถึงแม้ผลจากการทำการทดสอบแบบทีจับคู่จะพบว่าผลการทำนายของป่าของต้นไม้ตัดสินใจและนิวรอลเน็ตเวิร์กไม่มีความแตกต่างอย่างมีนัยสำคัญทางสถิติ แต่เนื่องจากค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ยของป่าของต้นไม้ตัดสินใจมีค่าน้อยกว่านิวรอลเน็ตเวิร์ก

เพื่อศึกษาว่าค่าปริมาณสารส้มที่ได้จากการทำนายด้วยป่าของต้นไม้ตัดสินใจมีความแตกต่างจากค่าปริมาณสารส้มที่ใช้จริงในกระบวนการผลิตน้ำประปาอย่างมีนัยสำคัญทางสถิติหรือไม่เพียงใด งานวิจัยนี้จึงได้ทำการทดสอบแบบทีจับคู่ระหว่างค่าปริมาณสารส้มที่ทำนายได้จากป่าของต้นไม้ตัดสินใจกับค่าปริมาณสารส้มที่เติมจริง ในแต่ละชุดข้อมูลจำนวน 10 ชุด ค่าพีของข้อมูลแต่ละชุดสามารถแสดงได้ดังตารางที่ 5.2 ซึ่งจะเห็นได้ว่ามีข้อมูลชุดที่ 5 เพียงชุดเดียวที่มีค่าพี < 0.05 และข้อมูลชุดที่ 10 มีค่าพี = 0.05 นอกนั้นอีก 8 ชุดมีค่าพี > 0.05 ซึ่งแปลว่าค่าที่ได้จากการทำนายด้วยป่าของต้นไม้ตัดสินใจส่วนใหญ่มีความแตกต่างกันอย่างไม่มีนัยสำคัญกับค่าจริง ซึ่งในทางสถิติถือว่าค่าที่ได้จากการทำนายสามารถนำไปใช้ได้

ตารางที่ 5.2
ค่าพีจากผลของการทดสอบแบบทีจับคู่ระหว่างค่าปริมาณสารส้มที่เติมจริง
และค่าที่ได้จากการทำนายด้วยป่าของต้นไม้ตัดสินใจ

ข้อมูลชุดที่	1	2	3	4	5	6	7	8	9	10
ค่าพี	0.056	0.106	0.490	0.663	0.017	0.889	0.269	0.244	0.183	0.050

5.2 วิเคราะห์ผลการทดลอง

จากการเปรียบเทียบวิธีในการทำนายกรณีอินพุตแบบ 7 ตัว กับอินพุตแบบ 8 ตัว พบว่าต้นไม้ตัดสินใจและป่าของต้นไม้ตัดสินใจสามารถทำนายผลได้แม่นยำมากขึ้นเมื่อใช้ตัวแปรอินพุตแบบ 8 ตัว ในขณะที่นิวรอลเน็ตเวิร์กที่ใช้ตัวแปรอินพุต 8 ตัว ให้ผลการทำนายใกล้เคียงกับการทำนายอินพุตแบบ 7 ตัว และซัพพอร์ตเวกเตอร์แมชชีนเมื่อใช้ตัวแปรอินพุต 8 ตัว พบว่าทำนายผลได้แย่ลงกว่าเดิม โดยสาเหตุที่ต้นไม้ตัดสินใจและป่าของต้นไม้ตัดสินใจสามารถทำนายผลได้

แม่นยำมากขึ้นก็เนื่องมาจากตัวแปรที่สร้างขึ้นมาใหม่ด้วยการประยุกต์ใช้โปรแกรมเชิงพันธุกรรม นั้นเมื่อนำไปใช้กับต้นไม้ตัดสินใจและป่าของต้นไม้ตัดสินใจ พบว่าเมื่อคำนวณหาค่าความสำคัญของตัวแปรแต่ละตัวแล้ว ตัวแปรตัวใหม่ที่สร้างขึ้นมาด้วยโปรแกรมเชิงพันธุกรรมเป็นตัวแปรที่มีความสำคัญสูงสุดเมื่อเทียบกับตัวแปรอื่นๆ ดังนั้นตัวแปรนี้จึงช่วยให้ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจสามารถทำนายผลได้แม่นยำมากขึ้น ในขณะที่เมื่อนำตัวแปรตัวใหม่นี้ไปใช้กับนิวรอลเน็ตเวิร์กซึ่งใช้วิธีการปรับค่าน้ำหนักด้วยอัลกอริทึมแบ็กพรอพาเกชัน พบว่าตัวแปรดังกล่าวมีแนวโน้มไม่ได้ช่วยให้นิวรอลเน็ตเวิร์กทำนายผลได้ดีขึ้น เช่นเดียวกับซัพพอร์ตเวกเตอร์แมชชีนซึ่งผลการทำนายในกรณีอินพุต 7 ตัว และ 8 ตัว มีผลที่ใกล้เคียงกัน ส่วนในกรณีตัวแปรอินพุตแบบ 3 ตัวแปร พบว่านิวรอลเน็ตเวิร์กสามารถทำนายผลได้ดีถึงแม้จะมีตัวแปรอินพุตที่น้อย โดยสามารถทำนายผลได้ดีกว่าซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจอย่างเห็นได้ชัด ดังจะเห็นได้จากผลของการทดสอบแบบทีจับคู่ซึ่งเมื่อเปรียบเทียบกับวิธีอื่นๆ ให้ค่าพี < 0.05 ซึ่งแปลว่าให้ผลที่แตกต่างกันอย่างมีนัยสำคัญ ผลของงานวิจัยนี้สามารถสรุปได้ว่าเทคนิคที่งานวิจัยนี้เสนอคือการใช้ป่าของต้นไม้ตัดสินใจร่วมกับการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมในการสร้างตัวแปรอินพุตเพิ่มเติมเป็นวิธีที่สามารถทำนายผลได้แม่นยำมากที่สุดเมื่อเปรียบเทียบกับวิธีการทำนายด้วยนิวรอลเน็ตเวิร์ก ซัพพอร์ตเวกเตอร์แมชชีน และต้นไม้ตัดสินใจ

อย่างไรก็ตาม เนื่องจากการทดลองนี้เป็นการใช้ข้อมูลจากโรงงานผลิตน้ำประปาบางเขนมาใช้ในการฝึกสอนเครื่องเพื่อให้เกิดการเรียนรู้ ดังนั้นโครงสร้างและค่าตัวแปรที่ได้จากงานวิจัยนี้จึงอาจจะไม่เหมาะที่จะนำไปใช้กับโรงงานผลิตน้ำประปาแห่งอื่น และการทดลองนี้อาจจะไม่สามารถเลือกโครงสร้าง และค่าพารามิเตอร์ได้อย่างครอบคลุมทั่วถึงทั้งหมด ด้วยสาเหตุในเรื่องของข้อจำกัดทางเวลา และขอบเขตของค่าตัวแปรที่ค่อนข้างกว้างมาก รวมถึงประสิทธิภาพของคอมพิวเตอร์ที่นำมาใช้ในการทดลองซึ่งเป็นคอมพิวเตอร์ส่วนบุคคล ทำให้ไม่สามารถทำการทดลองที่มีการคำนวณสูงมากๆ ได้ เช่น การกำหนดจำนวนต้นไม้ในป่าตัดสินใจที่มีจำนวนสูงมากๆ เช่น 2,000 ต้น เป็นต้น ดังนั้นผลลัพทธ์ที่ได้จากการทดลองครั้งนี้อาจไม่ใช่ค่าของผลลัพธ์ที่ดีที่สุดของคำตอบทั้งหมด

5.3 ข้อเสนอแนะ

จากผลการทดลองเพื่อหาปริมาณการเติมสารส้มในน้ำดิบพบว่าวิธีการที่ทำนายได้แม่นยำที่สุดคือการใช้ป่าของต้นไม้ตัดสินใจร่วมกับการประยุกต์ใช้โปรแกรมเชิงพันธุกรรมในการสร้างตัว

แปรอินพุตเพิ่มเติม ซึ่งโมเดลที่ได้จากการทดลองนี้ควรนำไปใช้ในการหาปริมาณการเติมสารส้มของโรงงานผลิตน้ำประปาบางเขน ซึ่งจะช่วยให้สามารถตอบสนองของลักษณะของน้ำดิบที่เข้ามาได้รวดเร็วยิ่งขึ้น สามารถช่วยประหยัดเวลาและงบประมาณได้เป็นจำนวนมาก

สำหรับโรงงานผลิตน้ำประปาอื่นๆ ที่อาจจะมีการตรวจวัดและจดบันทึกค่าของตัวแปรอินพุตไม่มากเท่ากับโรงงานผลิตน้ำประปาบางเขน โดยมีการจดบันทึกเฉพาะอินพุตหลักๆ เช่น ความขุ่น อุณหภูมิ พีเอช และความเป็นด่าง ในกรณีนี้การนำนิวรอลเน็ตเวิร์กไปใช้ในการหาปริมาณสารส้มจะมีความเหมาะสมมากกว่าวิธีการอื่นๆ เนื่องจากให้ผลการทำนายที่แม่นยำกว่าซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และป่าของต้นไม้ตัดสินใจอย่างเห็นได้ชัด

งานวิจัยอื่นๆ ที่มีการนำนิวรอลเน็ตเวิร์กไปใช้ในการทดลอง ควรจะวางแนวทางการทดลองโดยกำหนดโครงสร้างของนิวรอลเน็ตเวิร์กให้ครอบคลุมโครงสร้างของนิวรอลเน็ตเวิร์กที่มีจำนวนชั้นแฝงเท่ากับ 2 ชั้นด้วย เนื่องจากผลของการทดลองนี้แสดงให้เห็นได้ว่านิวรอลเน็ตเวิร์กที่มีจำนวนชั้นแฝงเท่ากับ 1 ชั้น และนิวรอลเน็ตเวิร์กที่มีจำนวนชั้นแฝงเท่ากับ 2 ชั้น ให้ผลการทำนายที่แตกต่างกันอย่างชัดเจน

สำหรับงานวิจัยในอนาคตจะนำเทคนิคใหม่ๆ มาผสมผสานกับวิธีการป่าของต้นไม้ตัดสินใจเพื่อช่วยให้วิธีการนี้สามารถใช้ในการทำนายปริมาณสารส้มได้อย่างถูกต้องแม่นยำเพิ่มมากขึ้น