การพัฒนาเครื่องมือสำหรับแก้ปัญหาการจำแนกประเภทของดาต้าไมน์นิ่ง

THE DEVELOPMENT OF DATA MINING TOOL FOR SOLVING CLASSIFICATION PROBLEMS

โดย

ณัฐวรรณ รุจิราณิชาภัทร

NATTAWAN RUJIRANICHAPAT

อาจารย์ที่ปรึกษา

รศ.ดร. อาริต ธรรมโน

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2552

THE DEVELOPMENT OF DATA MINING TOOL FOR SOLVING CLASSIFICATION PROBLEMS

NATTAWAN RUJIRANICHAPAT

A REPORT SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS OF THE COURSE

SYSTEM DEVELOPMENT PROJECT

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2009

ใบรับรองโครงการพัฒนาระบบงาน (SYSTEM DEVELOPMENT PROJECT)

เรื่อง

การพัฒนาเครื่องมือสำหรับแก้ปัญหาการจำแนกประเภทของดาต้าไมน์นิ่ง

THE DEVELOPMENT OF DATA MINING TOOL FOR SOLVING CLASSIFICATION PROBLEMS

นางสาวณัฐวรรณ รุจิราณิชาภัทร

รหัสประจำตัว 49066720

ขอรับรองว่ารายงานฉบับนี้ ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด

รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ

การศึกษาวิชาโครงการพัฒนาระบบงาน หลักสูตรวิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศ)

ภาคเรียนที่ 2 ปีการศึกษา 2552

....อาจารย์ที่ปรึกษา

(รศ.ดร. อาริต ธรรมโน)

....กรรมการสอบ

(รศ.ดร. วรพจน์ กรีสุระเดช)

....กรรมการสอบ

(ผศ.ดร. พรฤดี เนติโสภากุล)

หัวข้อ	การพัฒนาเครื่องมือสำหรับการแก้ปัญหาการจำแนกประเภทของดาต้าไมน์นิ่ง
นักศึกษา	นางสาวณัฐวรรณ รุจิราณิชาภัทร
รหัสนักศึกษา	49066720
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2552
อาจารย์ที่ปรึกษา	รศ.ดร.อาริต ธรรมโน

บทคัดย่อ

เนื่องจากปัจจุบันมีอัตราการแข่งขันที่สูงขึ้น ส่งผลให้องค์กรต้องหาวิธีการเพิ่มประสิทธิภาพในการวิเคราะห์ วางแผน และตัดสินใจ ดังนั้นการทำดาต้าไมนิ่งจึงมีความสำคัญมากในการนำข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่ มากลั่นกรองและวิเคราะห์ด้วยกระบวนการวิธีต่างๆ เพื่อนำข้อมูลที่ซ่อนอยู่นั้นมาใช้ให้เกิดประโยชน์และมีประสิทธิภาพสูงสุด โดยโครงงานนี้พัฒนาขึ้นสำหรับการทำดาต้าไมน์นิ่งด้วยวิธีจำแนกประเภทของข้อมูล โดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียมด้วยอัลกอริธึม Backpropagation มาช่วยในการแบ่งประเภทของข้อมูลที่เราสนใจว่าอยู่ในกลุ่มใด เพื่อนำผลลัพธ์ที่ได้มาช่วยในการตัดสินใจและนำไปใช้ประโยชน์ต่อไป

Title	The development of data mining tool for solving classification problems
Student	Ms. Nattawan Rujiranichapat
Student ID.	49066720
Degree	Master of Science
Program	Information Technology
Major	Information Science
Academic Year	2009
Advisor	Assoc. Prof. Dr. Arit Thammano

ABSTRACT

In the present, the competitive rate is grows up. Resulting organization must find the ways to increase efficiency in analysis, planning and decision. So, Data Mining is still very important in bringing the information contained in large databases filtered by a filtration and analytical process in different ways to expose the hidden information of data for maximum benefit and efficiency. This project is developed for Classification in Data Mining. By using Decision Tree techniques like ID3 algorithm and Neural Network techniques like Backpropagation algorithm for Helping in the classification of information that we are interested in any group and using the results for helping in decision making and used to advantage.

กิตติกรรมประกาศ

โครงการศึกษาพัฒนาเครื่องมือสำหรับการแก้ปัญหาการจำแนกประเภทของดาต้าไมน์นิ่ง ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี ด้วยคำแนะนำและคำปรึกษาจาก รศ.ดร. อาริต ธรรมโน ซึ่งเป็นอาจารย์ผู้ควบคุมโครงการพัฒนาระบบนี้ โดยอาจารย์ได้สละเวลาให้คำปรึกษา ชี้แนะข้าพเจ้าเป็นอย่างดี ข้าพเจ้ารู้สึกซาบซึ้งในความอนุเคราะห์ของอาจารย์เป็นอย่างมาก และขอขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณคณาจารย์ ภาควิชาวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกๆท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ ในภาควิชาวิทยาการสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกคนที่ให้คำแนะนำต่างๆ คอยช่วยเหลือและคอยให้กำลังใจเสมอมา

ขอขอบคุณพี่ๆ น้องๆ ชาวโมทีฟและทีมบาสทุกๆ คน ที่สละเวลายื่นมือให้ความช่วยเหลือในยามที่ต้องการ ขอบคุณหัวหน้าทีมที่เข้าใจและให้การสนับสนุนอย่างเต็มที่

ขอขอบคุณคณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง รวมไปถึงคณาจารย์ทุกๆท่าน ที่ได้ปูพื้นฐานชีวิตในรั้วมหาวิทยาลัยตั้งแต่ในเรื่องของความรู้ สังคม และหล่อหลอมจริยธรรมให้แก่ข้าพเจ้า

และสุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา และทุกคนในครอบครัวของข้าพเจ้าที่ช่วยผลักดัน สนับสนุนในทุกๆ เรื่อง และคอยเป็นกำลังใจอยู่ตลอดเวลา ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

คุณค่าและประโยชน์อันพึงมาจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน หากผิดพลาดประการใด ขออภัยไว้ ณ ที่นี้ด้วยค่ะ

ณัฐวรรณ รุจิราณิชาภัทร

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VI

สารบัญรูป VII

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา 1

1.3 ขอบเขตของการศึกษา 2

1.4 องค์ประกอบของการพัฒนาโครงการ 2

1.5 ขั้นตอนของการศึกษา 2

1.6 ประโยชน์ที่คาดว่าจะได้รับ 3

บทที่ 2 เทคนิคการทำดาต้าไมน์นิ่งด้วยวิธีจำแนกประเภทข้อมูล 4

2.1 Classification Concept 4

2.1.1 การทำ Data Classification 4

2.2 เทคนิคในการจำแนกประเภทข้อมูลด้วย Decision Tree 5

2.2.1 วิธีการสร้าง Decision Tree 6

2.2.2 ข้อจำกัดของ Decision Tree 7

2.2.3 ไอดีทรี (ID3) 7

2.2.4 อัลกอริธึมสำหรับการสร้าง Decision Tree 8

2.2.5 ตัวอย่างการหา Information gain เพื่อสร้าง Classification 9

2.3 โครงข่ายประสาทเทียม (Artificial Neural Network) 13

2.3.1 หลักการและการทำงานของโครงข่ายประสาทเทียม 14

2.3.2 การเรียนรู้ของโครงข่ายประสาทเทียม 14

2.3.3 สถาปัตยกรรมของโครงข่ายประสาทเทียม 15

2.3.4 การประยุกต์ใช้งาน Neural Network 17

สารบัญ (ต่อ)

หน้า

2.3.5 Classification by Backpropagation 18

2.3.5.1 ขั้นตอนการทำ Backpropagation 18

2.3.5.2 อัลกอริธึมสำหรับการสร้าง Backpropagation 19

2.3.5.3 ตัวอย่างการสร้างการเรียนรู้ด้วยวิธี Backpropagation 20

บทที่ 3 วิธีดำเนินการศึกษา 24

3.1 ข้อมูลที่ใช้กับระบบ 24

3.2 ขั้นตอนและวิธีการดำเนินงาน 25

3.2.1 การทำงานภาพรวมของระบบ 27

3.2.2 ขั้นตอนการสร้าง Decision Tree ด้วยอัลกอริธึม ID3 29

3.2.3 ขั้นตอนการสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation 31

3.3 การพัฒนาโปรแกรมประยุกต์ 33

3.3.1 เครื่องมือที่ใช้ในการพัฒนาระบบ 31

3.3.2 ส่วนติดต่อผู้ใช้ระบบ 33

บทที่ 4 ผลการทดลอง 39

4.1 การประยุกต์ใช้งานระบบ 39

4.1.1 การจำแนกประเภทข้อมูลด้วยวิธี Neural Network ด้วยอัลกอริธึม Backpropagation 40

4.1.2 การจำแนกประเภทข้อมูลด้วยวิธี Decision Tree ด้วยอัลกอริธึม ID3 46

4.2 ข้อมูลที่ใช้กับโปรแกรมประยุกต์ 50

4.3 ผลการทดลองและวิเคราะห์ 53

4.3.1 ข้อมูลดอกไอริส (Iris) 53

4.3.2 ข้อมูล Wine 54

4.3.3 ข้อมูล Image 55

4.3.4 วิเคราะห์ผลการทดลอง 56

บทที่ 5 สรุปผลการศึกษาและข้อเสนอแนะ 57

5.1 สรุปผลการศึกษา 57

สารบัญ (ต่อ)

หน้า

5.2 ประโยชน์ที่ได้รับจากการศึกษาและพัฒนาระบบ 58

5.3 ข้อเสนอแนะ 58

บรรณานุกรม 59

ประวัติผู้เขียน 60

สารบัญตาราง

ตารางที่ หน้า

2.1 แสดง Training Data ที่ใช้สร้าง Decision Tree 9
4.1 แสดงผลลัพธ์การทดลอง Iris ด้วยข้อมูลเทรน (Training Data) 90 ตัวอย่าง 53
4.2 แสดงผลลัพธ์การทดลอง Wine ด้วยข้อมูลเทรน (Training Data) 106 ตัวอย่าง....... 54
4.3 แสดงผลลัพธ์การทดลอง Image ด้วยข้อมูลเทรน (Training Data) 210 ตัวอย่าง 55

สารบัญรูป

รูปที่ หน้า

2.1 กระบวนการเรียนรู้ข้อมูล 4
2.2 กระบวนการจำแนกประเภทข้อมูล 5
2.3 แสดงการทำงานของ Decision tree 6
2.4 แสดงตัวอย่างของ Decision Tree 6
2.5 การแบ่งค่าของแอททริบิวต์ age 10
2.6 แสดงรูปแบบของ decision tree จากอัลกอริธึมที่ได้ 12
2.7 แบบจำลองของ Neuron ในสมองมนุษย์ 14
2.8 การเรียนรู้แบบ Supervised Learning 15
2.9 การเรียนรู้แบบ Unsupervised Learning 15
2.10 สถาปัตยกรรมของ Feedforward Network 16
2.11 สถาปัตยกรรมของ Feedback Network 16
2.12 โครงสร้างของโครงข่ายประสาทเทียม 16
2.13 แสดงขั้นตอนในการทำ Backpropagation 18
2.14 ตัวอย่างที่ใช้สำหรับการเรียนรู้ 20
3.1 ตัวอย่างข้อมูลที่นำมาใช้ 24
3.2 ตัวอย่างรูปแบบของข้อมูล Input 26
3.3 การทำงานภาพรวมของระบบ 27
3.4 การสร้าง Decision Tree ด้วยอัลกอริธึม ID3 29
3.5 การสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation 31
3.6 หน้าจอหลักของโปรแกรม 33
3.7 หน้าจอในส่วนของ Decision Tree (ID3) 34
3.8 หน้าจอในส่วนของการเทรนข้อมูล 35
3.9 หน้าจอในส่วนของการเทสข้อมูล 35
3.10 หน้าจอในส่วนของ Neural Network (Backpropagation) 36
3.11 หน้าจอในส่วนของ Training and Configuration 37
3.12 หน้าจอในส่วนของการ Testing 38
4.1 หน้าจอหลักของโปรแกรม 39
4.2 หน้าจอในส่วนของ Neural Network (Backpropagation) 40

สารบัญรูป (ต่อ)

รูปที่ หน้า

4.3 หน้าจอในส่วนของการ Training and Configuration 41
4.4 หน้าจอในการเทรนข้อมูลของ Backpropagation 42
4.5 หน้าจอแสดงข้อความเมื่อระบบทำการเทรนข้อมูลเสร็จ 43
4.6 หน้าจอแสดงไฟล์ที่ถูกสร้างขึ้นมา 43
4.7 หน้าจอในส่วนของการ Testing ของ Backpropagation 44
4.8 หน้าจอสำหรับเลือกไฟล์ที่ต้องการใช้ทดสอบ 45
4.9 หน้าจอแสดงผลลัพธ์ที่ได้จากการจำแนกประเภทข้อมูลด้วยวิธี Backpropagation... 46
4.10 หน้าจอในส่วนของ Decision Tree (ID3) 46
4.11 หน้าจอในส่วนของการ Training 47
4.12 หน้าจอในการเทรนข้อมูลของ ID3 47
4.13 หน้าจอแสดงข้อความเมื่อระบบทำการทดสอบข้อมูลเสร็จ 48
4.14 หน้าจอแสดงผลลัพธ์ที่ได้จากการจำแนกประเภทข้อมูลด้วยวิธี ID3 49
4.15 ข้อมูล Iris ที่ใช้สำหรับการ Train 50
4.16 ข้อมูล Wine ที่ใช้สำหรับการ Train 51
4.17 ข้อมูล Image ที่ใช้สำหรับการ Train 52

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบันซึ่งเป็นยุคของข้อมูลข่าวสาร แต่ละองค์กรได้นำเอาระบบสารสนเทศมาใช้งานในองค์กรเพื่อให้การทำงานเป็นไปอย่างมีประสิทธิภาพ จึงทำให้มีข้อมูลต่างๆ ที่ต้องจัดเก็บไว้เป็นจำนวนมาก โดยในปัจจุบันนิยมที่จะจัดเก็บลงในระบบฐานข้อมูลซึ่งอาจจะมีลักษณะที่แตกต่างกันขึ้นอยู่กับลักษณะและวัตถุประสงค์ที่นำไปใช้งานขององค์กร เช่น Microsoft SQL Server , Oracle เป็นต้น โดยข้อมูลทรานแซคชันต่างๆ ที่ถูกจัดเก็บในแต่ละวันเหล่านี้ล้วนแล้วแต่มีประโยชน์ในการทำงานสำหรับองค์กรแต่กลับไม่เพียงพอต่อการนำไปใช้ในการตัดสินใจทางธุรกิจเท่าที่ควร ดังนั้นจึงมีการนำเอาแนวคิดเกี่ยวกับเทคโนโลยีทางด้านดาต้าไมน์นิ่งซึ่งเป็นกระบวนการในการเพิ่มประสิทธิภาพของระบบข้อมูลข่าวสารที่มีอยู่ให้มีประโยชน์สูงสุด เข้ามาใช้ในการกลั่นกรองและวิเคราะห์ข้อมูลเพื่อจะสกัดข้อมูลสารสนเทศที่มีความสำคัญและน่าสนใจเพื่อนำไปใช้ในกระบวนการตัดสินใจทางธุรกิจ เทคนิคทางด้านดาต้าไมน์นิ่งจำนวนมากถูกนำมาใช้ในองค์กรเพื่อสร้างความได้เปรียบในด้านการแข่งขัน ตัวอย่างของการนำเอาดาต้าไมน์นิ่งมาใช้ เช่น การจัดกลุ่มของลูกค้า (Customer Segmentation) เมื่อเราสามารถจัดกลุ่มของลูกค้าได้แล้วก็สามารถนำกระบวนการทางซีอาร์เอ็ม (Customer Relationship Management) เข้ามาประยุกต์ใช้เพื่อสร้างความได้เปรียบทางธุรกิจ การสร้างโมเดลเพื่อช่วยในการตัดสินใจเรื่องการอนุมัติบัตรเครดิตหรือการวิเคราะห์ความสัมพันธ์ของข้อมูล (Link Analysis) เป็นต้น ซึ่งประโยชน์ในการทำดาต้าไมน์นิ่งนั้น นอกจากจะช่วยในกระบวนการตัดสินใจแล้ว ยังอาจจะได้รับข้อมูลใหม่ๆ ที่เป็นประโยชน์ทำให้การวิเคราะห์ข้อมูลมีประสิทธิภาพมากขึ้น

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับขั้นตอนกระบวนการทำดาต้าไมน์นิ่ง
2. เพื่อศึกษาการนำหลักการทางด้านดาต้าไมน์นิ่ง โดยใช้เทคนิควิธีในการแบ่งประเภทข้อมูล (Classification)
3. เพื่อนำความรู้ ความเข้าใจที่ได้ศึกษา ค้นคว้ามาประยุกต์ใช้หรือช่วยในการแก้ปัญหา
4. เพื่อสร้างซอฟต์แวร์ที่สามารถใช้ในการแบ่งประเภทและหาความสัมพันธ์ของข้อมูล
5. เพื่อนำข้อมูลที่มีอยู่ในคลังข้อมูลมาทำให้เกิดประโยชน์สูงสุด
6. เพื่อให้ได้ข้อมูลที่สามารถนำมาวิเคราะห์ วางแผนและช่วยในการตัดสินใจได้อย่างรวดเร็ว

1.3 ขอบเขตของการศึกษา

โครงการนี้เป็นการศึกษาและพัฒนาระบบงานเกี่ยวกับดาต้าไมน์นิ่ง โดยนำหลักการและเทคนิคในการแบ่งประเภทของข้อมูล (Classification) มาใช้ในพัฒนาระบบ โดยใช้เทคนิควิธีโครงสร้างต้นไม้ (Decision Tree) ด้วยอัลกอริธึม ID3 และโครงข่ายประสาทเทียม (Neural network) ด้วยอัลกอริธึม Backpropagation มาใช้ในการประมวลผลข้อมูลและแสดงผลลัพธ์ที่เกิดขึ้นหลังจากข้อมูลนั้นผ่านกระบวนการทางดาต้าไมน์นิ่งมาแล้ว

โดยขอบเขตของระบบจะเป็นการดำเนินการตามกระบวนการทางดาต้าไมน์นิ่ง โดยจะเริ่มตั้งแต่กระบวนการจัดเตรียมข้อมูล (Data Preparation) ซึ่งจะมีการกำจัดข้อมูลส่วนที่ไม่จำเป็นหรือไม่มีความเกี่ยวข้องออกไป ต่อมาก็จะทำการเลือกข้อมูลให้ตรงกับความต้องการ ผ่านทางกระบวนการทางด้านดาต้าไมน์นิ่งจนกระทั่งได้ผลลัพธ์ที่ต้องการจากข้อมูลที่มีอยู่

1.4 องค์ประกอบของการพัฒนาโครงการ

1) คอมพิวเตอร์ Laptop ยี่ห้อ Dell รุ่น Latitude E6400 : Intel® Core™ 2 Duo CPU P8400 2.26GHz, RAM 4 GB.
2) ระบบปฏิบัติการ Microsoft Windows XP Professional Ver. 2002 Service Pack 3
3) Microsoft Visual Studio.Net Team System 2008
4) Microsoft Office System 2007 Professional Edition

1.5 ขั้นตอนของการศึกษา

1) ศึกษากระบวนการทางด้านดาต้าไมน์นิ่ง
2) เลือกใช้อัลกอริธึมทางด้านการแบ่งประเภทข้อมูลที่เหมาะสมกับโครงการ
3) ศึกษาอัลกอริธึมที่เลือกใช้ และการนำไปใช้งานให้เหมาะสมกับโครงการ
4) ศึกษาการทำงานและการใช้งาน Microsoft Visual Studio.Net Team System 2008
5) วิเคราะห์และออกแบบระบบ
6) พัฒนาระบบ
7) ทดสอบและปรับปรุงแก้ไขระบบ
8) จัดทำเอกสารประกอบการใช้งานระบบ

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1) เพิ่มประโยชน์ของข้อมูลที่มีอยู่ ทำให้เกิดประโยชน์สูงสุดในการนำไปใช้งาน และช่วยในการประกอบการตัดสินใจทางธุรกิจ
2) เพิ่มความรวดเร็วในการช่วยตัดสินใจทางธุรกิจ
3) นำความรู้และเทคนิคที่ได้จากการศึกษาทางด้านดาต้าไมน์นิ่ง ไปประยุกต์ใช้ในการวิเคราะห์ข้อมูลและพัฒนาระบบเพื่อนำไปใช้งานจริง

บทที่ 2

เทคนิคการจำแนกประเภทข้อมูล

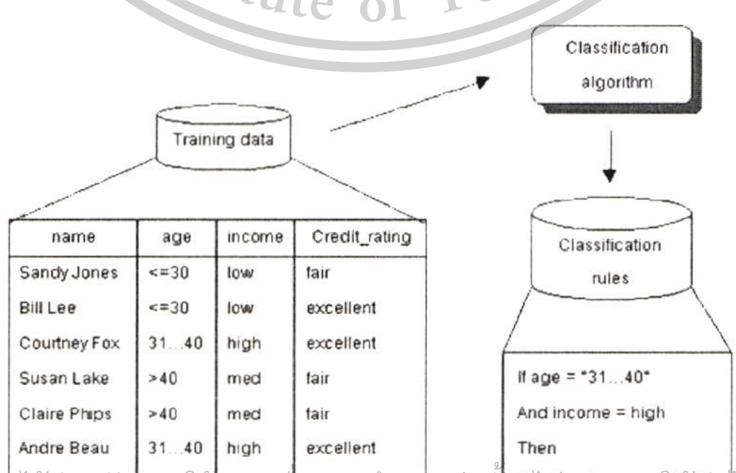
2.1 Classification Concept

การจำแนกประเภทของข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เพื่อแสดงให้เห็นความแตกต่างระหว่างคลาสหรือกลุ่มของข้อมูลได้ และเพื่อทำนายว่าข้อมูลนี้ควรจัดอยู่ในคลาสใด ซึ่งโมเดลที่ใช้ในการจำแนกประเภทของข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ จะขึ้นอยู่กับการวิเคราะห์เซตของข้อมูลทดลอง (Training Data) โดยนำ Training Data มาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในคลาสเดียวกันบ้าง โดยผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลจำแนกประเภทของข้อมูล (Classifier) โมเดลนี้สามารถแทนได้ในหลายรูปแบบ เช่น Classification Rules, Decision Tree, Mathematical Formula หรือ Neural Networks และจะนำข้อมูลส่วนที่เหลือจากการ Training Data มาเป็นข้อมูลที่ใช้ในการทดสอบ (Testing Data) ซึ่งเป็นกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบ โดยข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง โดยจะทำการปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา เราจะนำข้อมูลผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลนี้ได้

2.1.1 การทำ Data Classification

ขั้นตอนการทำ Data Classification นั้นแบ่งออกเป็น 2 ขั้นตอนคือ

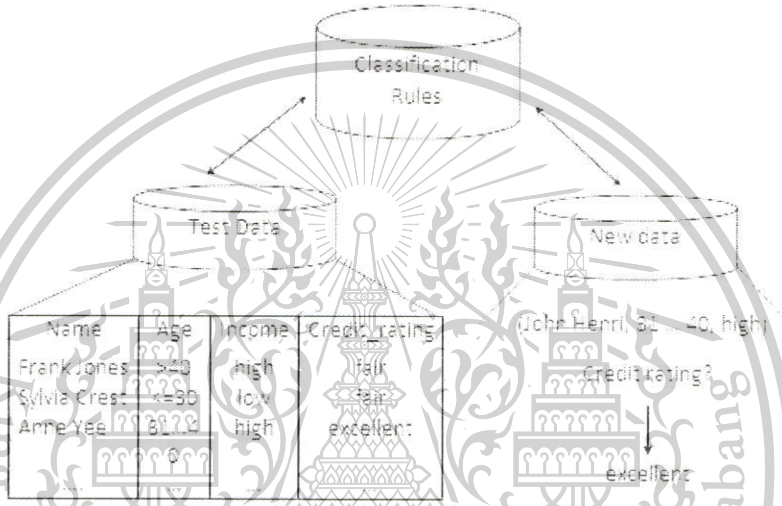
1) Learning โดยข้อมูลทดลอง (Training Data) จะถูกวิเคราะห์โดยอัลกอริธึมของ Classification และ Learning Model ถูกแทนในรูปของ Classification Rules



รูปที่ 2.1 กระบวนการเรียนรู้ข้อมูล

จากรูปอธิบายได้ว่า Classification Rules ที่ได้จะสามารถเรียนรู้ว่าลูกค้าคนใดที่มี credit_rating ที่ excellent หรือ fair ซึ่งกฎนี้สามารถใช้ในการจำแนกประเภทของข้อมูลเข้ามาใหม่ได้

2) Classification เมื่อได้ Classification Rules จะมีการตรวจสอบว่ากฎที่ได้สามารถทำนายได้ถูกต้องแม่นยำหรือไม่ โดยการนำเอา Test Data ที่ทราบแล้วว่าอยู่ในกลุ่มใดไปเปรียบเทียบกับ Learning Model ถ้าหากว่าผลที่ได้มีความถูกต้อง ก็จะสามารถนำโมเดลหรือกฎที่ได้ไปทำนาย credit_rating ของลูกค้าที่เข้ามาใหม่ได้

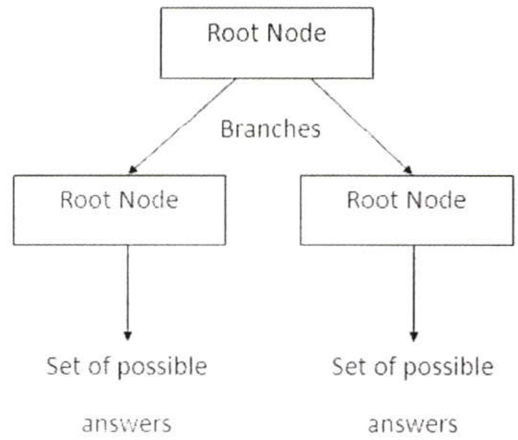


รูปที่ 2.2 กระบวนการจำแนกประเภทข้อมูล

2.2 เทคนิคในการจำแนกประเภทข้อมูลด้วย Decision Tree

Decision tree เป็นหนึ่งวิธีที่สำคัญใน Classification โดย Decision Tree จะมีลักษณะเป็น Flow-Chart เหมือนโครงสร้างต้นไม้ ที่แต่ละโหนดแสดงคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล แต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนด (Leaf Node) แสดงกลุ่มหรือ Class ที่กำหนดไว้ ซึ่ง Decision Tree นี้ง่ายต่อการปรับเปลี่ยนเป็น Classification Rule

เมื่อมีข้อมูลที่ต้องการที่จะจำแนกประเภทก็จะนำแอททริบิวต์ ต่างๆของข้อมูลนั้นไปเปรียบเทียบกับ Decision Tree นี้ตามเส้นทางใน Tree จนกระทั่งได้คลาสปลายทาง ซึ่งก็คือกลุ่มของข้อมูลที่เหมือนกัน



รูปที่ 2.3 แสดงการทำงานของ Decision tree

รูปแบบของ Tree จะประกอบด้วยโหนดแรกสุดที่เรียกว่า Root Node จาก Root Node จะแตกออกเป็นโหนดลูกและที่โหนดลูกจะมีลูกของตัวเองซึ่งโหนดที่ระดับสุดท้ายเรียกว่า Leaf Node

รูปที่ 2.4 แสดงตัวอย่างของ Decision Tree

จะเห็นว่าจาก Root Node จนถึง Leaf Node จะมีเพียงเส้นทางเดียวเท่านั้น ซึ่งเส้นทางนี้จะอธิบายถึงกฎที่ใช้สำหรับการจัดหมวดหมู่ของแต่ละกลุ่ม ซึ่งในแต่ละ Leaf Node นั้นอาจเป็นกลุ่มเดียวกัน ซึ่งเกิดจากเหตุผลที่แตกต่างกันได้

2.2.1 วิธีการสร้าง Decision Tree

การนำข้อมูลมาสร้าง Tree มีขั้นตอนดังนี้

1) หาแอททริบิวต์ที่สำคัญที่สุดมาจำแนกประเภทข้อมูลโดย แอททริบิวต์ นี้จะถูกนำมาสร้างเป็น Root Node โดยจะมี Target Attribute เป็นผลลัพธ์ซึ่งเป็น Leaf Node ที่ถูกกำหนดไว้ก่อน

2) นำข้อมูลทั้งหมดมาเปรียบเทียบกับเงื่อนไขที่ใช้ในการจำแนกประเภทข้อมูลเพื่อดูว่าข้อมูลใดจะถูกจัดไปอยู่ในกลุ่มใด

3) วนกลับไปทำที่ขั้นตอนแรก คือ หาแอททริบิวต์ ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวจำแนกประเภทต่อไป

2.2.2 ข้อจำกัดของ Decision Tree

1) การจำแนกประเภทข้อมูลแบบ Decision Tree กรณีเป็นข้อมูลที่มีค่าต่อเนื่อง เช่น ข้อมูลรายได้ ข้อมูลราคา ต้องทำการแปลงให้อยู่ในช่วงหรือตัดเป็นกลุ่มก่อน

2) เมื่อเลือกว่าจะใช้ค่าใดเป็นตัวจำแนกประเภทแล้วก็จะไม่สนใจค่าอื่นที่อาจมีความสำคัญเช่นเดียวกัน

3) การจัดการกับข้อมูลที่ไม่ทราบค่า อาจมีผลกระทบกับผลลัพธ์ของ Decision Tree

4) Tree ที่มีระดับชั้นมากเกินไป จะทำให้ข้อมูลที่ผ่านโหนดแตกออกเป็นชิ้นเล็กชิ้นน้อย ซึ่งข้อมูลเหล่านั้นจะไม่มีประโยชน์ในการนำมาวิเคราะห์

5) ปัญหาเรื่อง Overfitting เกิดจากการที่โมเดลเข้าไปถึงรายละเอียดของข้อมูลมากเกินไป ทำให้เกิดโหนดที่เป็นส่วนเฉพาะเจาะจงกับกลุ่มข้อมูล ซึ่งจะต้องหาวิธีในการตัดกิ่งนี้ออกไป โดยสามารถใช้วิธี Tree Pruning เพื่อนำมาใช้แก้ปัญหานี้

2.2.3 ไอดีทรี (ID3)

จากการศึกษาการสร้าง Decision Tree ในการทำ Classification มีทฤษฎีที่ใช้ในการสร้าง Decision Tree คือ Information Gain เพื่อใช้ในการเลือกแอททริบิวต์ ในแต่ละโหนดของ tree ซึ่งแอททริบิวต์ตัวใดที่มีค่า Information Gain สูงสุดหรือมีค่า Entropy น้อยจะถูกเลือกให้เป็นแอททริบิวต์ของโหนดนั้น และแอททริบิวต์ตัวนี้จะลดจำนวนข้อมูลที่จะใช้ในการสร้าง Tree

ทฤษฎีนี้มีสมการ การคำนวณหา Information gain โดยมีการกำหนดตัวแปรดังนี้

ให้ S เป็นเซตของข้อมูล

m แทนจำนวนคลาสที่แตกต่างกันของข้อมูล ซึ่งแต่ละค่าแทนด้วย C_i (for i = 1,…, m)

s_i เป็นกลุ่มข้อมูลย่อยของ S ในคลาส C_i

สมการที่ใช้ในการจำแนกประเภทของข้อมูลคือ สมการ (2.1)

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A) \tag{2.1}$$

เมื่อ A คือแอททริบิวต์ ที่ใช้หา Information gain

Gain (A) ที่ได้จะเป็นตัวบอกว่าแอททริบิวต์ ตัวนี้เป็นตัวจำแนกประเภทแยกข้อมูลหรือไม่ (Gain ที่มีค่ามากที่สุด) เมื่อเปรียบเทียบกับ Gain ของแอททริบิวต์ ตัวอื่นๆจากสมการข้างต้น โดย $I(s_1, s_2, \ldots, s_m)$ หาได้จากสมการ (2.2)

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2.2}$$

โดยที่ p_i คือ ความน่าจะเป็นที่ใช้ในการตัดสินข้อมูล S ในคลาส C_i

$p_i = s_i/S$

ให้แอททริบิวต์ A ที่มีค่าของข้อมูลที่แตกต่างกัน v ค่าแทนด้วย $a_1, a_2, \ldots, a_v$ ซึ่งแอททริบิวต์ A จะแบ่งข้อมูล S ออกเป็น v subset แทนด้วย $s_1, s_2, \ldots, s_v$

ให้ s_i แทนจำนวนข้อมูลในแต่ละค่าของแอททริบิวต์ A

s_{ij} แทนจำนวนข้อมูลในแต่ละคลาสของแต่ละค่าแอททริบิวต์ A แทนด้วย $s_{1j}, s_{2j}, \ldots, s_{mj}$

หา Entropy ของกลุ่มข้อมูลย่อย ดังสมการ (2.3)

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \tag{2.3}$$

จากสมการ (3) สามารถหาค่า $I(s_{1j},\ldots,s_{mj})$ ได้จากสมการ (2.4)

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} \log_2(p_{ij}) \tag{2.4}$$

โดยที่

p_{ij} คือ ความน่าจะเป็นที่ใช้ในการตัดสินข้อมูล s_i ลงในคลาส C_i

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

2.2.4 อัลกอริธึมสำหรับการสร้าง Decision Tree

Algorithm: Generate_decision_tree

Input: Training sample, *sample*;

The set of candidate attributes, *attribute-list*;

Output: A decision tree.

Method:

(1) create a node N;

(2) **if** *sample* are all of the same class, C **then**

(3) return N as a leaf node labeled with the class C;

(4) **if** *attribute-list* is empty **then**

(5) return *N* as a leaf node labeled with the most common class in *samples*; // majority voting

(6) select *test-attribute*, the attribute among *attribute-list* with the highest information gain;

(7) label node *N* with *test-attribute*;

(8) **for each** known value a_i **of** *test-attribute* // partition the samples

(9) grow a branch from node N for the condition *test-attribute* = a_i;

(10) let s_i be the set of sample in samples for which *test-attribute* = a_i; // a partition

(11) **if** s_i is empty **then**

(12) attach a leaf labeled with the most common class in *sample;*

(13) **else** attach the node returned by Generate_decision_tree (s_i, *attribute-list-test-attribute*);

2.2.5 ตัวอย่างการหา Information gain เพื่อสร้าง Classification

RID	age	income	student	credit_rating	Class: buys_computer
1	<=30	high	no	fair	no
2	<=30	high	no	excellent	no
3	31...40	high	no	fair	yes
4	>40	medium	no	fair	yes
5	>40	low	yes	fair	yes
6	>40	low	yes	excellent	no
7	31...40	low	yes	excellent	yes
8	<=30	medium	no	fair	no
9	<=30	low	yes	fair	yes
10	>40	medium	yes	fair	yes
11	<=30	medium	yes	excellent	yes
12	31...40	medium	no	excellent	yes
13	31...40	high	yes	fair	yes
14	>40	medium	no	excellent	no

ตารางที่ 2.1 แสดง Training Data ที่ใช้สร้าง Decision tree

จากตารางได้ข้อมูลดังนี้

- มีคลาสชื่อว่า buys_computer ซึ่งมีค่าที่แบ่งออกได้ 2 ค่า คือ yes และ no ดังนั้น m = 2
- ให้ C_1 แทน yes และ C_2 แทน no
- ข้อมูลทั้งหมดมี 14 แถว ดังนั้น S = 14
- มีข้อมูลที่อยู่ใน C_1 9 แถว
  (ข้อมูลที่มีค่าในคลาส buys_computer เป็น yes) ดังนั้น $s_1 = 9$
- มีข้อมูลที่อยู่ใน C_2 5 แถว
  (ข้อมูลที่มีค่าในคลาส buys_computer เป็น no) ดังนั้น $s_2 = 5$

ใช้สมการ (2) ในการจำแนกประเภทของข้อมูลได้ดังนี้

$$I(s_1,s_2) = I(9,5)$$

$$= -\left(\frac{9}{14}\log_2\frac{9}{14}\right) - \left(\frac{5}{14}\log_2\frac{5}{14}\right)$$

$$= 0.940$$

จากนั้นทำการเลือกแอททริบิวต์ ที่จะทดสอบ Gain เพื่อดูการกระจายของข้อมูลของคลาส yes และ no โดยแอททริบิวต์ที่จะนำมาพิจารณา คือ age, income, student และ credit_rating โดย

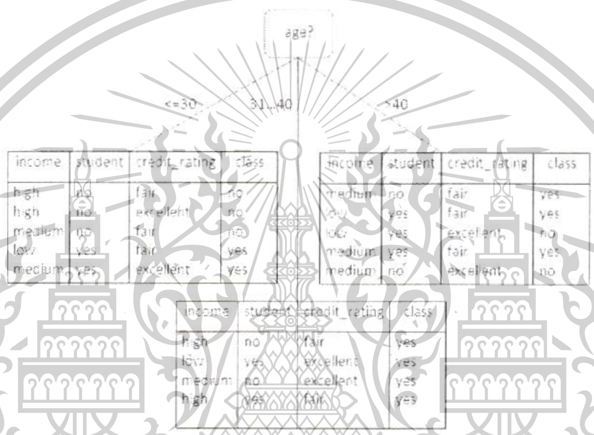
age แบ่งออกไว้ 3 ค่า คือ "<= 30", "31…40" และ ">40"

income แบ่งออกไว้ 3 ค่า คือ high, medium และ low

student แบ่งออกไว้ 2 ค่า คือ yes และ no

credit_rating แบ่งออกไว้ 2 ค่า คือ fair และ excellent

เริ่มทดสอบแอททริบิวต์ age ซึ่งแอททริบิวต์ นี้มีค่าที่ต่างกัน 3 ค่า สามารถแยกค่าที่แตกต่างกันของแอททริบิวต์ age ออกเป็นตารางได้ดังนี้



รูปที่ 2.5 การแบ่งค่าของแอททริบิวต์ age

ค่าของแอททริบิวต์ age = "<= 30"

ได้จำนวนข้อมูลของคลาส C_1 คือ $s_{11} = 2$

ได้จำนวนข้อมูลของคลาส C_2 คือ $s_{21} = 3$

จากสมการที่ 4 หา $I(s_{11}, s_{21})$ จะได้

$$I(s_{11}, s_{21}) = I(2,3)$$

$$= -\left(\frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right)$$

$$= 0.971$$

ค่าของแอททริบิวต์ age = "31…40"

ได้จำนวนข้อมูลของคลาส C_1 คือ $s_{12} = 4$

ได้จำนวนข้อมูลของคลาส C_2 คือ $s_{22} = 0$

จากสมการที่ 4 หา $I(s_{11}, s_{21})$ จะได้

$$I(s_{12}, s_{22}) = I(4,0)$$
$$= -\left(\frac{4}{4}\log_2\frac{4}{4}\right) - (0\log_2 0)$$
$$= 0$$

ค่าของแอททริบิวต์ age = “>40”

ได้จำนวนข้อมูลของคลาส C_1 คือ $s_{13} = 3$

ได้จำนวนข้อมูลของคลาส C_2 คือ $s_{23} = 2$

จากสมการที่ 4 หา $I(s_{13}, s_{23})$ จะได้

$$I(s_{13}, s_{23}) = I(3,2)$$
$$= -\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right)$$
$$= 0.971$$

จากนั้นคำนวณหา Entropy ของแอททริบิวต์ age ได้ดังนี้

$$E(age) = \frac{5}{14}I(s_{11}, s_{21}) + \frac{4}{14}I(s_{12}, s_{22}) + \frac{5}{14}I(s_{13}, s_{23})$$
$$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$$
$$= 0.694$$

จากนั้นคำนวณหา Gain ของแอททริบิวต์ age ได้ดังนี้

$$Gain(age) = I(s_1, s_2) - E(age)$$
$$= 0.940 - 0.694$$
$$= 0.246$$

จากวิธีการนี้เมื่อทำการคำนวณหา Information gain ของแอททริบิวต์ ทั้งหมด จะได้

Gain (income) = 0.029

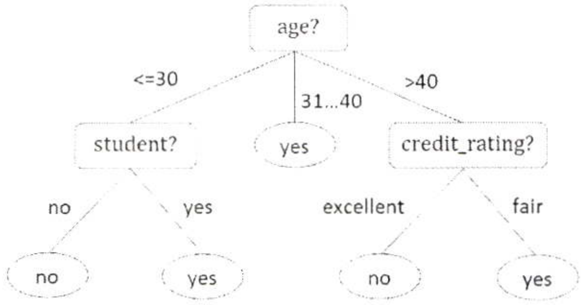
Gain (student) = 0.151

Gain (credit_rating) = 0.048

จาก Gain ของแอททริบิวต์ ต่างๆที่ได้ จะเห็นว่า ค่า Gain(age) มีค่ามากที่สุด ดังนั้นแอททริบิวต์ age จึงถูกเลือกให้เป็น root node ของ decision tree branch ของ decision tree ก็จะแตกออกจากค่าแต่ละค่าของแอททริบิวต์ age

วิธีการที่จะแตกโหนดต่อไปจนได้คลาส yes และ no นั้น จะใช้วิธีการเดียวกันนี้ กระทำแบบ recursive เมื่อได้แอททริบิวต์ ที่เป็นตัวจำแนกประเภทแล้วจะตัดแอททริบิวต์ นั้นออกไปไม่นำมาพิจารณาอีก สังเกตได้ว่าการแยกตารางของค่า age = “31...40” ของแอททริบิวต์ age นั้น

คลาสที่ได้จะเป็นโหนดที่ไม่สามารถแตกได้อีก อัลกอริธึมที่ได้สามารถเขียนเป็น decision tree ได้ดังนี้



รูปที่ 2.6 แสดงรูปแบบของ decision tree จากอัลกอริธึมที่ได้

จากรูปสรุปได้ว่า

- กลุ่ม "<=30"
  ต้องพิจารณาว่าเป็นนักเรียนหรือไม่
  ถ้า ไม่ใช่ ผลคือ จะไม่ซื้อสินค้า
  ถ้า ใช่ ผลคือ จะซื้อสินค้า
- กลุ่ม "31…40" ทุกคนจะซื้อสินค้า
- กลุ่ม ">40"
  ต้องพิจารณาเครดิต
  ถ้า excellent ผลคือ จะไม่ซื้อสินค้า
  ถ้า fair ผลคือ จะซื้อสินค้า

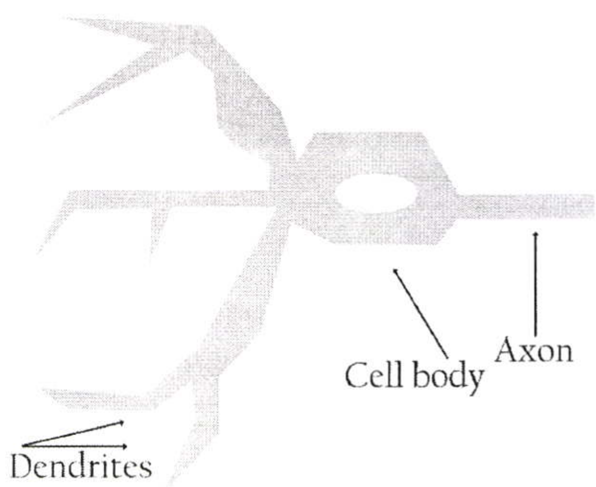
จากอัลกอริธึมที่ได้สามารถนำมาแปลงตามกฎ IT_THEN ได้ดังนี้

```
IF age = "<=30" AND student = "no"
        THEN buys_computer = "no"
IF age = "<=30" AND student = "yes"
        THEN buys_computer = "yes"
IF age = "31…40"
        THEN buys_computer = "yes"
IF age = ">40" AND student = "excellent"
        THEN buys_computer = "no"
IF age = ">40" AND student = "fair"
        THEN buys_computer = "yes"
```

2.3 โครงข่ายประสาทเทียม (Artificial Neural Network)

โครงข่ายประสาทเทียม (Artificial Neural Network) คือแบบจำลองทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศด้วยการคำนวณแบบคอนเนคชันนิสต์ (Connectionist) เพื่อจำลองการทำงานของเครือข่ายประสาทในสมองมนุษย์ ด้วยวัตถุประสงค์ที่จะสร้างเครื่องมือซึ่งมีความสามารถในการเรียนรู้การจดจำรูปแบบ (Pattern Recognition) และการสร้างความรู้ใหม่ (Knowledge Extraction) เช่นเดียวกับความสามารถที่มีในสมองมนุษย์ แนวความคิดเริ่มต้นของเทคนิคนี้ได้มาจากการศึกษาข่ายงานไฟฟ้าชีวภาพ (Bioelectric Network) ในสมอง ซึ่งประกอบด้วย เซลล์ประสาท หรือ นิวรอน (Neurons) และ จุดประสานประสาท (Synapses) แต่ละเซลล์ประสาทประกอบด้วยปลายในการรับกระแสประสาทเรียกว่า เดนไดรท์ (Dendrite) ซึ่งเป็น input และปลายในการส่งกระแสประสาทเรียกว่า แอคซอน (Axon) ซึ่งเป็นเหมือน output ของเซลล์ เซลล์เหล่านี้ทำงานด้วยปฏิกริยาไฟฟ้าเคมี เมื่อมีการกระตุ้นด้วยสิ่งเร้าภายนอกหรือกระตุ้นด้วยเซลล์ด้วยกัน กระแสประสาทจะวิ่งผ่านเดนไดรท์เข้าสู่นิวเคลียสซึ่งจะเป็นตัวตัดสินว่าต้องกระตุ้นเซลล์อื่นๆต่อหรือไม่ ถ้ากระแสประสาทแรงพอ นิวเคลียสก็จะกระตุ้นเซลล์อื่นๆต่อไปผ่านทางแอคซอนของมัน

ตามแบบจำลองนี้โครงข่ายประสาทเกิดจากการเชื่อมต่อระหว่างเซลล์ประสาทจนเป็นเครือข่ายที่ทำงานร่วมกัน



รูปที่ 2.7 แบบจำลองของ Neuron ในสมองมนุษย์

2.3.1 หลักการและการทำงานของโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม เป็นแบบจำลองระบบการประมวลผลที่เหมือนเครือข่ายเส้นใยประสาทของมนุษย์ที่เชื่อมโยงกัน ที่เรียกว่า เส้นประสาท (Neural) ที่ประกอบด้วยสถาปัตยกรรมพื้นฐานจำนวนมากด้วยลักษณะที่คล้ายคลึงกับสมองที่มีกระบวนการเชื่อมต่อในลักษณะคู่ขนาน (Parallel) และมีการเคลื่อนที่สื่อสารระหว่างกัน ทำให้เครือข่ายเกิดความสามารถในการเรียนรู้ (Learn) จากข้อมูลที่ได้รับและมีการนำข้อมูลที่ได้รับนั้นไปประมวลผลและศึกษารูปแบบหรือความสัมพันธ์ของข้อมูลนั้นและมีการตอบสนองเพื่อเปลี่ยนแปลงรูปแบบ (Pattern) ของข้อมูลที่ได้รับและผลลัพธ์ที่เกิดขึ้น

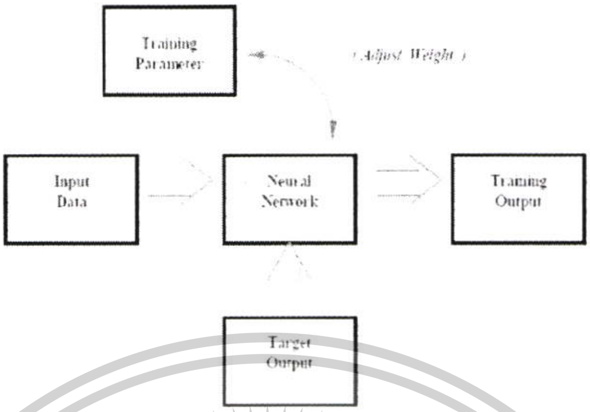
แม้ว่าโครงข่ายประสาทเทียมจะมีโครงสร้างแตกต่างจากโครงข่ายในสมอง แต่ก็ยังเหมือนสมองในแง่ที่ว่าโครงข่ายประสาทเทียม คือการรวมกลุ่มแบบขนานของหน่วยประมวลผลย่อยๆ และการเชื่อมต่อนี้เป็นส่วนสำคัญที่ทำให้เกิดสติปัญญาของโครงข่าย เมื่อพิจารณาขนาดแล้วสมองมีขนาดใหญ่กว่าโครงข่ายประสาทเทียมอย่างมาก รวมทั้งเซลล์ประสาทยังมีความซับซ้อนกว่าหน่วยย่อยของโครงข่าย อย่างไรก็ตามหน้าที่สำคัญของสมอง เช่น การเรียนรู้ยังคงสามารถถูกจำลองขึ้นอย่างง่ายด้วยโครงข่ายประสาทนี้

2.3.2 การเรียนรู้ของโครงข่ายประสาทเทียม

1. Supervised Learning

การเรียนแบบมีการสอนเป็นการเรียนแบบที่มีการตรวจคำตอบเพื่อให้โครงข่ายประสาทเทียมปรับตัว ชุดข้อมูลที่ใช้สอนโครงข่ายประสาทเทียมจะมีคำตอบไว้คอยตรวจดูว่าโครงข่ายประสาท

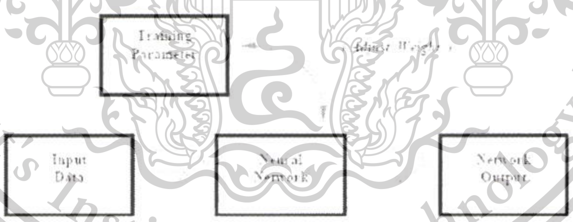
เทียมให้คำตอบที่ถูกหรือไม่ ถ้าตอบไม่ถูก โครงข่ายประสาทเทียมก็จะปรับตัวเองเพื่อให้ได้คำตอบที่ดีขึ้น



รูปที่ 2.8 การเรียนรู้แบบ Supervised Learning

2. Unsupervised Learning

การเรียนแบบไม่มีการสอนเป็นการเรียนแบบไม่มีผู้แนะนำ ไม่มีการตรวจคำตอบว่าถูกหรือผิด โครงข่ายประสาทเทียมจะพยายามจัดเรียงโครงสร้างด้วยตัวเองตามลักษณะของข้อมูลผลลัพธ์ที่ได้ โครงข่ายประสาทเทียมจะสามารถจัดหมวดหมู่ของข้อมูลได้

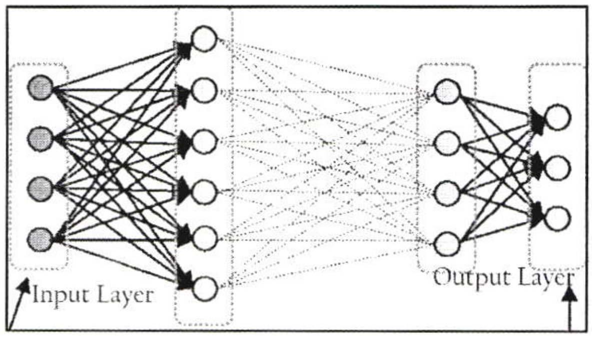


รูปที่ 2.9 การเรียนรู้แบบ Unsupervised Learning

2.3.3 สถาปัตยกรรมของโครงข่ายประสาทเทียม

1. Feedforward Network

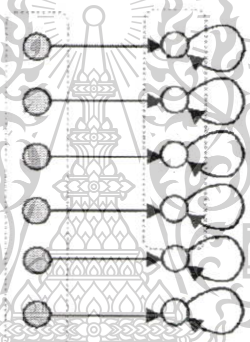
ข้อมูลที่ประมวลผลในโครงข่ายประสาทเทียมจะถูกส่งไปในทิศทางเดียวจาก Input Nodes ส่งต่อมาเรื่อยๆ จนถึง Output Nodes โดยไม่มีการย้อนกลับของข้อมูล หรือแม้แต่ Nodes ในชั้นเดียวกันก็ไม่มีการเชื่อมต่อ



รูปที่ 2.10 สถาปัตยกรรมของ Feedforward Network

2. Feedback Network

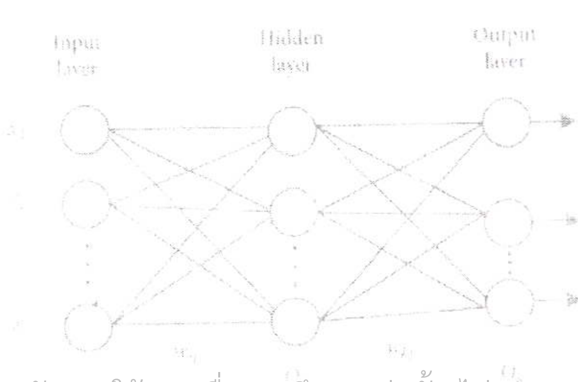
ข้อมูลที่ประมวลผลในโครงข่ายประสาทเทียม จะมีการป้อนกลับเข้าไปยังโครงข่ายประสาทเทียมหลายๆครั้ง จนกระทั่งได้คำตอบออกมา



รูปที่ 2.11 สถาปัตยกรรมของ Feedback Network

3. Network Layer

โดยทั่วไปโครงข่ายประสาทเทียมมีการแบ่งโครงสร้างออกเป็นชั้นๆโดยจะแบ่งเป็น 3 ชั้นหลักๆ ได้แก่ Input Layer ที่เชื่อมต่อเข้ากับชั้นของ Hidden Layer ซึ่งเชื่อมต่อเข้ากับชั้นของ Output Layer



รูปที่ 2.12 โครงสร้างของโครงข่ายประสาทเทียม

Input Layer มีหน้าที่รับข้อมูลจากภายนอกเข้ามาโดยไม่ได้มีกระบวนการจัดการกับข้อมูลก่อนจะกระจายข้อมูลไปในเครือข่าย

Hidden Layer มีหน้าที่ประมวลผลข้อมูลตามค่าน้ำหนัก (Weight) บนความสัมพันธ์ระหว่าง Input Layer และ Hidden Layer ที่ได้กำหนดไว้

Output Layer มีหน้าที่ส่งผลลัพธ์ที่ได้ออกนอกโครงข่าย โดยการทำงานของชั้นนี้ขึ้นอยู่กับการทำงานของชั้น Hidden Layer และ ค่าน้ำหนัก (Weight) ระหว่างชั้น Hidden Layer และชั้น Output

Architecture of Layer สามารถแบ่งสถาปัตยกรรมของชั้น ออกเป็น 2 ประเภทคือ

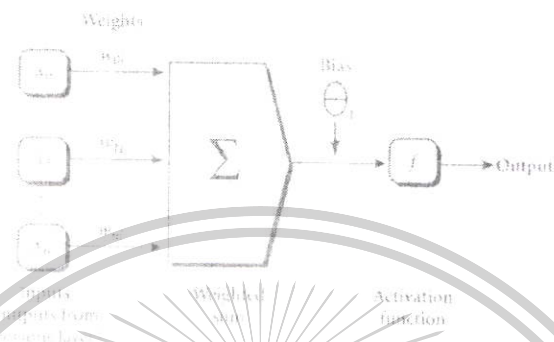
1. **Single-layer perceptron** โครงข่ายประสาทเทียมที่ประกอบด้วยชั้นเพียงชั้นเดียว จำนวน Input Nodes ขึ้นอยู่กับจำนวน Component ของ Input data และ Activation Function ขึ้นอยู่กับลักษณะข้อมูลของ Output
2. **Multi-layer perceptron** โครงข่ายประสาทเทียมจะประกอบด้วยชั้นหลายชั้นโดยในแต่ละชั้นจะประกอบด้วยโหนด (Nodes) หรือเปรียบได้กับตัวเซลล์ประสาท (Neurons) ค่าน้ำหนักของเส้นที่เชื่อมต่อระหว่างโหนดของแต่ละชั้น ค่า Bias Vector และค่า Output Vector

2.3.4 การประยุกต์ใช้งาน Neural Network

โครงข่ายประสาทเทียม (Neural Network) มีความสามารถในการจำลองพฤติกรรมทางกายภาพของระบบที่มีความซับซ้อนจากข้อมูลที่ป้อนให้เรียนรู้ การประยุกต์ใช้โครงข่ายประสาทเทียมนั้น สามารถนำมาประยุกต์ใช้งานหลายประเภท เช่น งานจดจำรูปแบบที่มีความไม่แน่นอน การประมาณค่าฟังก์ชันหรือการประมาณความสัมพันธ์ งานที่สิ่งแวดล้อมเปลี่ยนแปลงอยู่เสมอเนื่องจากโครงข่ายประสาทเทียมสามารถปรับตัวเองได้ งานจัดหมวดหมู่และแยกแยะสิ่งของ การประยุกต์ใช้โครงข่ายประสาทเทียมควบคุมกระบวนการทางเคมีโดยวิธีพยากรณ์แบบจำลอง (Model Predictive Control) การประยุกต์ใช้งานโครงข่ายประสาทเทียมแบบแพร่กระจายกลับในการทำนายพลังงานความร้อนที่สะสมอยู่ในตัวอาคาร และนอกจากนี้ยังสามารถนำไปประยุกต์ในงานต่างๆอีกหลายงาน

2.3.5 Classification by Backpropagation

Backpropagation เป็นอัลกอริธึมที่ใช้ในการเรียนรู้ของโครงข่ายประสาทเทียมวิธีหนึ่งที่นิยมใช้ Multilayer Perceptron เพื่อปรับค่าน้ำหนักในเส้นเชื่อมต่อระหว่างโหนดให้เหมาะสม โดยการปรับค่านี้จะขึ้นกับความแตกต่างของค่า Output ที่คำนวณได้กับค่า Output ที่ต้องการ



รูปที่ 2.13 แสดงขั้นตอนในการทำ Backpropagation

โดยกำหนดให้

i, j และ k คือ โหนดในชั้นรับข้อมูล ชั้นซ่อน และชั้นผลลัพธ์

W_{ij} คือ ค่าน้ำหนักสำหรับการเชื่อมต่อระหว่างโหนด i ในชั้นรับข้อมูล กับ โหนด j ในชั้นซ่อน

W_{jk} คือ ค่าน้ำหนักสำหรับการเชื่อมต่อระหว่างโหนด j ในชั้นซ่อน กับ โหนด k ในชั้นผลลัพธ์

2.3.5.1 ขั้นตอนการทำ Backpropagation

ขั้นตอนที่ 1. กำหนดค่าเริ่มต้นของค่า Weight และ Bias โดยค่า Weight จะถูกกำหนดค่าเริ่มต้นด้วยตัวเลขสุ่ม (random) ที่มีค่าอยู่ในช่วง -1.0 ถึง 1.0 หรือ -0.5 ถึง 0.5

ขั้นตอนที่ 2. ขั้นตอนการผ่าน Input เข้าสู่โครงข่าย (Propagate the input forward) ในขั้นตอนนี้จะทำการคำนวณค่า net input และ output ของแต่ละโหนดในชั้น Hidden layer และ Output layer โดยคำนวณค่า net input ได้จาก

$$I_j = \sum_i w_{ij} O_i + \theta_j \tag{2.5}$$

จากนั้นคำนวณหา Output ของแต่ละโหนด j ได้จาก

$$O_j = \frac{1}{1+e^{-I_j}} \tag{2.6}$$

ขั้นตอนที่ 3. คำนวณค่าความคลาดเคลื่อน (Error) ของแต่ละโหนด j ในชั้น Output layer

$$Err_j = O_j(1 - O_j)(T_j - O_j) \tag{2.7}$$

คำนวณค่าความคลาดเคลื่อน (Error) ของแต่ละโหนด j ในชั้น Hidden layer

$$Err_j = O_j(1 - O_j)\sum_k Err_k w_{jk} \tag{2.8}$$

ขั้นตอนที่ 4. คำนวณหาค่า Weight และค่า Bias ที่ได้เปลี่ยนแปลงไป

$$\Delta w_{ij} = (l)Err_j O_i \tag{2.9}$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \tag{2.10}$$

และ

$$\Delta\theta_j = (l)Err_j \tag{2.11}$$

$$\theta_j = \theta_j + \Delta\theta_j \tag{2.12}$$

ขั้นตอนที่ 5. วนกลับไปทำซ้ำในขั้นตอนแรกจนกว่าจะได้ผลลัพธ์ที่ใกล้เคียงกับค่าที่เราคาดหวังมากที่สุด

2.3.5.2 อัลกอริธึมสำหรับการสร้าง Backpropagation

Algorithm: Backpropagation

Input: The training samples, *sample*;

The learning rate, l;

A multilayer feed-forward network, *network*;

Output: A neural network trained to classify the samples.

Method:

Initialize all weights and biases in *network*;

while termination condition is not satisfied {

  for each training sample X in *samples* {

    //Propagate the inputs forward:

    for each hidden or output layer unit j {

      $I_j = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit j with respect to the previous layer, i

      $O_j = \frac{1}{1+e^{-I_j}}$; //compute the output of each unit j

```
//backpropagate the errors:
for each unit j in the output layer
    Err_j = O_j(1 - O_j)(T_j - O_j); } //compute the error
for each unit j in the hidden layers, from the last to the first hidden layer
    Err_j = O_j(1 - O_j) Σ_k Err_k w_jk ; } //compute the error with respect to the
    next higher layer, k
for each weight w_ij in network {
    Δw_ij = (l)Err_j O_i ; } //weight increment
    w_ij = w_ij + Δw_ij ; } //weight update
for each bias θ_j in network {
    Δθ_j = (l)Err_j ; } // bias increment
    θ_j = θ_j + Δθ_j ; } // bias update
}}
```

2.3.5.3 ตัวอย่างการสร้างการเรียนรู้ด้วยวิธี Backpropagation

รูปที่ 2.14 ตัวอย่างที่ใช้สำหรับการเรียนรู้

<u>ค่าเริ่มต้นกำหนดให้</u>

$x_1 = 1$, $x_2 = 0$, $x_3 = 1$

$w_{14} = 0.2$, $w_{15} = -0.3$, $w_{24} = 0.4$, $w_{25} = 0.1$

$w_{34} = -0.5$, $w_{35} = 0.2$, $w_{46} = -0.3$, $w_{56} = -0.2$

$\theta_4 = -0.4$, $\theta_5 = 0.2$, $\theta_6 = 0.1$ และ ค่า Learning rate = 0.9

1) คำนวณหา Net input และ Output

จากสูตร $I_j = \sum_i w_{ij} O_i + \theta_j$ และ $O_j = \dfrac{1}{1+e^{-I_j}}$

<u>พิจารณาโหนดที่ 4</u>

Net input, I_j $= x_1\ w_{14} + x_2\ w_{24} + x_3\ w_{34} + \theta_4$

= 0.2+0-0.5-0.4

= -0.7

Output, O_j $= \dfrac{1}{1+e^{0.7}}$

= 0.332

<u>พิจารณาโหนดที่ 5</u>

Net input, I_j $= x_1\ w_{15} + x_2\ w_{25} + x_3\ w_{35} + \theta_5$

= -0.3+0+0.2+0.2

= 0.1

Output, O_j $= \dfrac{1}{1+e^{-0.1}}$

= 0.525

<u>พิจารณาโหนดที่ 6</u>

Net input, I_k $= w_{46}\ o_4 + w_{56}\ o_5 + \theta_6$

= (-0.3)(0.332)-(0.2)(0.525)+0.1

= -0.105

Output, O_k $= \dfrac{1}{1+e^{0.105}}$

= 0.474

2) คำนวณหาค่า error ของแต่ละโหนด

จากสูตร $Err_j = O_j(1-O_j)(T_j - O_j)$ ซึ่งใช้สำหรับหาค่า Error บนชั้น Output Layer

<u>พิจารณาโหนดที่ 6</u>

Err_j $= O_6(1-O_6)(1-O_6)$

= (0.474)(1-0.474)(1-0.474)

= 0.1311

จากสูตร $Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$ ซึ่งใช้สำหรับหา Error บนชั้น Hidden Layer

<u>พิจารณาโหนดที่ 5</u>

$Err_j = O_5(1-O_5)(Err_6 w_{56})$

= (0.525)(1-0.525)(0.1311-0.2)

= -0.0065

<u>พิจารณาโหนดที่ 4</u>

$Err_j = O_4(1-O_4)(Err_6 w_{46})$

= (0.332)(1-0.332)(0.1311)(-0.3)

= -0.0087

3) คำนวณหาค่า Weight

จากสูตร $\Delta w_{ij} = (l) Err_j O_i$ //weight increment

$w_{ij} = w_{ij} + \Delta w_{ij}$ //weight update

จะได้ว่า $w_{ij} = w_{ij} + (l) Err_j O_i$

ดังนั้น $w_{46} = w_{46} + (l) Err_6 O_4$

= -0.3+(0.9)(0.1311)(0.332)

= -0.261

$w_{56} = w_{56} + (l) Err_6 O_5$

= -0.2+(0.9)(0.1311)(0.525)

= -0.138

$w_{14} = w_{14} + (l) Err_4 O_1$

= 0.2+(0.9)(-0.0087)(1)

= 0.192

$w_{15} = w_{15} + (l) Err_5 O_1$

= -0.3+(0.9)(-0.0065)(1)

= -0.306

$w_{24} = w_{24} + (l) Err_4 O_2$

$= 0.4+(0.9)(-0.0087)(0)$

$= 0.4$

$w_{25} \quad = w_{25} + (l)\, Err_5 O_2$

$= 0.1+(0.9)(-0.0065)(0)$

$= 0.1$

$w_{34} \quad = w_{34} + (l)\, Err_4 O_3$

$= -0.5+(0.9)(-0.0087)(1)$

$= -0.508$

$w_{35} \quad = w_{35} + (l)\, Err_5 O_3$

$= 0.2+(0.9)(-0.0065)(1)$

$= 0.194$

4) คำนวณหาค่า Bias

จากสูตร $\Delta\theta_j = (l) Err_j$ //bias increment

$\theta_j = \theta_j + \Delta\theta_j$ //bias update

จะได้ว่า $\theta_j = \theta_j + (l)\, Err_j$

ดังนั้น $\theta_6 \quad = \theta_6 + (l)\, Err_6$

$= 0.1+(0.9)(0.1311)$

$= 0.218$

$\theta_5 \quad = \theta_5 + (l)\, Err_5$

$= 0.2+(0.9)(-0.0065)$

$= 0.194$

$\theta_4 \quad = \theta_4 + (l)\, Err_4$

$= -0.4+(0.9)(-0.0087)$

$= -0.408$

หลังจากที่ทำการปรับค่า Weight และ Bias แล้ว เมื่อทำการใส่ค่า Input ตัวเดิมในรอบถัดไปโดยใช้ค่า Weight และค่า Bias ที่ได้ทำการปรับค่า ไปทำการคำนวณในรอบต่อไปจะพบว่า ค่าความผิดพลาดที่เกิดขึ้นในแต่ละโหนดนั้นมีค่าลดลง

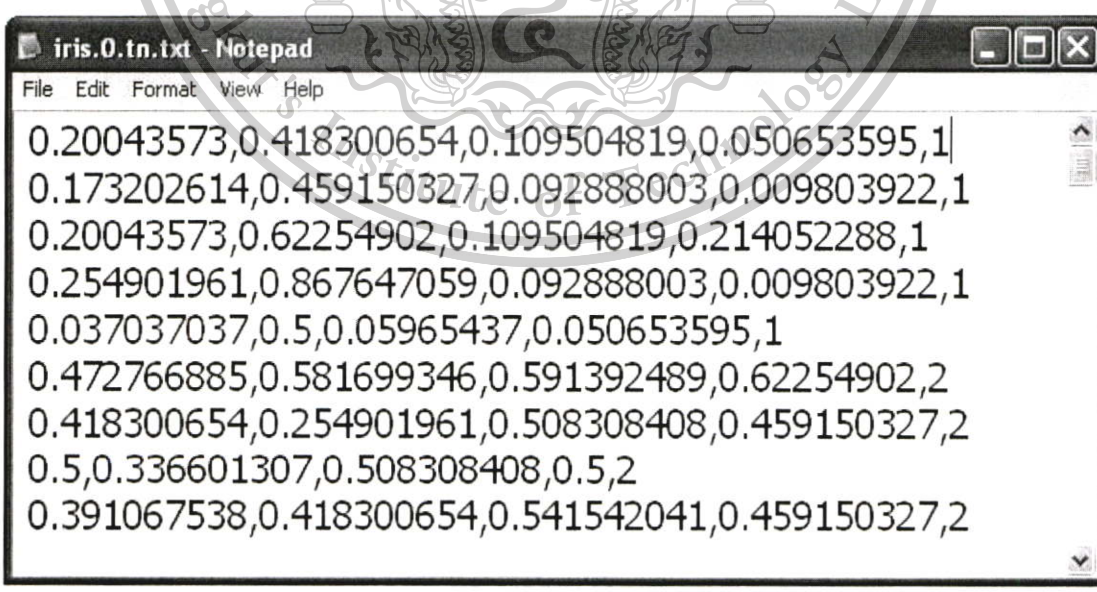
บทที่ 3
วิธีดำเนินการศึกษา

ในการศึกษาโครงงานนี้ เป็นการนำทฤษฎีการทำดาต้าไมน์นิ่งด้วยวิธีการจำแนกประเภทของข้อมูล โดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation มาพัฒนาเป็นโปรแกรมคอมพิวเตอร์เพื่อประยุกต์ใช้ในการจำแนกประเภทของข้อมูล โดยแบ่งการดำเนินศึกษาออกเป็น 3 ส่วน ดังนี้

1. ข้อมูลที่ใช้กับระบบ
2. ขั้นตอนและวิธีการดำเนินงาน
3. การพัฒนาโปรแกรม

3.1 ข้อมูลที่ใช้กับระบบ

โปรแกรมประยุกต์ที่พัฒนาขึ้นในโครงการนี้จะรับข้อมูลเข้าในลักษณะของ Text File (.txt) ซึ่งลักษณะของข้อมูลแต่ละชุดจะประกอบด้วยคุณลักษณะต่างๆ ของชุดข้อมูลที่นำมาใช้ โดยที่ค่าสุดท้ายของข้อมูลแต่ละชุดนั้นจะบอกถึงกลุ่มที่ชุดของข้อมูลนั้นๆ ถูกจัดให้อยู่ โดยแต่ละคุณลักษณะจะเป็นข้อมูลชนิดตัวเลขที่ถูกคั่นด้วยเครื่องหมายจุลภาค (,) ดังรูปที่ 3.1



รูปที่ 3.1 ตัวอย่างข้อมูลที่นำมาใช้

โดยลักษณะของข้อมูลที่นำมาใช้สามารถสรุปได้ ดังนี้ คือ

- ข้อมูลแต่ละบรรทัดจะบอกถึงคุณลักษณะของข้อมูลนั้นๆ 1 ชุด
- ข้อมูลแต่ละชุดจะประกอบไปด้วยคุณลักษณะของข้อมูล ซึ่งข้อมูลแต่ละชุดนั้นจะต้องประกอบไปด้วยจำนวนของคุณลักษณะที่เท่ากัน
- ค่าสุดท้ายของข้อมูลแต่ละชุดนั้นจะบอกถึงกลุ่มที่ชุดข้อมูลนั้นๆ จัดให้อยู่
- คุณลักษณะของข้อมูลจะต้องเป็นข้อมูลชนิดตัวเลขเท่านั้น
- คุณลักษณะของข้อมูลแต่ละตัวจะถูกคั่นด้วยเครื่องหมายจุลภาค (,)

ข้อมูลที่นำมาใช้จะแบ่งเป็น 2 ชุดคือ

1. ข้อมูลฝึกฝน (Training Data) ซึ่งจะนำมาใช้เพื่อสอนการเรียนรู้ของระบบ
2. ข้อมูลทดสอบ (Testing Data) เป็นข้อมูลที่นำมาใช้ทดสอบการทำงานของระบบ

โดยข้อมูลฝึกฝน (Training Data) จะถูกนำเข้าสู่ระบบการแบ่งข้อมูล เมื่อระบบทำการเรียนรู้จากข้อมูลที่ได้รับแล้ว จึงนำข้อมูลทดสอบ (Testing Data) เข้าสู่ระบบเพื่อทดสอบประสิทธิภาพและความถูกต้องของระบบที่ได้จากการเรียนรู้

3.2 ขั้นตอนและวิธีการดำเนินงาน

ในการศึกษาโครงงานนี้จะทำการศึกษาการทำดาต้าไมน์นิ่งด้วยวิธีการจำแนกประเภทของข้อมูล โดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation โดยทำการพัฒนาระบบเป็นโปรแกรมคอมพิวเตอร์ เพื่อประยุกต์ใช้ในการศึกษาเพื่อใช้ในการทำนายการจำแนกประเภทของข้อมูลโดยมีขั้นตอนการดำเนินงานดังนี้

1. การเตรียมข้อมูล (Data Preparation)

ในขั้นตอนนี้เป็นการเตรียมข้อมูลเพื่อเข้าสู่กระบวนการทำดาต้าไมน์นิ่ง เมื่อคัดเลือกข้อมูลที่ต้องการ กำจัดข้อมูลที่ไม่สมบูรณ์ ข้อมูลที่ผิดปกติรวมไปถึงข้อมูลที่ไม่ได้ใช้และข้อมูลที่ไม่มีความสำคัญออกไป เพื่อจัดเตรียมข้อมูล Input และ Output ที่เหมาะสมเพื่อป้อนเข้าสู่ระบบที่เราสร้างขึ้น โดยทำการแปลงข้อมูลให้อยู่ในรูปของ Text File (.txt) ดังนี้

iris.0.tn.txt - Notepad

File Edit Format View Help

```
0.20043573,0.418300654,0.109504819,0.050653595,1
0.173202614,0.459150327,0.092888003,0.009803922,1
0.20043573,0.62254902,0.109504819,0.214052288,1
0.254901961,0.867647059,0.092888003,0.009803922,1
0.037037037,0.5,0.05965437,0.050653595,1
0.472766885,0.581699346,0.591392489,0.62254902,2
0.418300654,0.254901961,0.508308408,0.459150327,2
0.5,0.336601307,0.508308408,0.5,2
0.391067538,0.418300654,0.541542041,0.459150327,2
0.445533769,0.5,0.641242938,0.704248366,2
0.990196078,0.745098039,0.907111997,0.785947712,3
0.363834423,0.336601307,0.657859754,0.785947712,3
0.608932462,0.418300654,0.757560651,0.704248366,3
0.826797386,0.37745098,0.890495181,0.704248366,3
0.445533769,0.418300654,0.691093387,0.704248366,3
0.254901961,0.62254902,0.092888003,0.050653595,1
0.145969499,0.581699346,0.109504819,0.050653595,1
0.145969499,0.581699346,0.159355268,0.050653595,1
```

รูปที่ 3.2 ตัวอย่างรูปแบบของข้อมูล Input

2. การเลือกข้อมูล

ในการเลือกข้อมูลนั้น ข้อมูลจะประกอบไปด้วยข้อมูลฝึกฝนและข้อมูลทดสอบ โดยการเลือกข้อมูลฝึกฝนนั้นจะต้องเลือกข้อมูลให้ครอบคลุมทุกคุณลักษณะของผลลัพธ์ (Output) และข้อมูลทดสอบนั้นจะต้องไม่เป็นข้อมูลชุดเดียวกันกับข้อมูลฝึกฝน

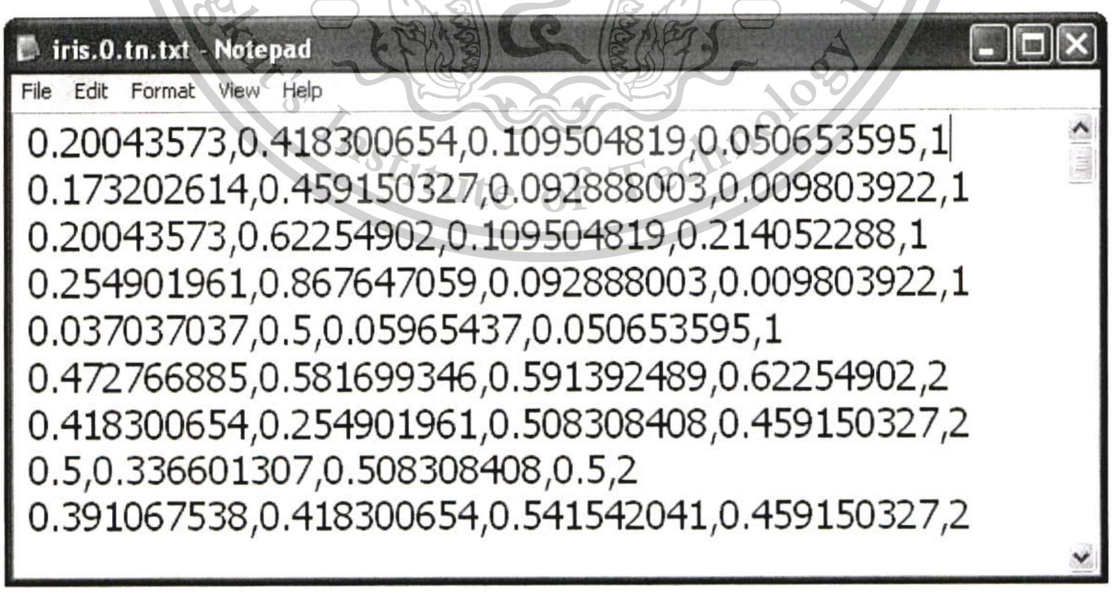
บทที่ 3
วิธีดำเนินการศึกษา

ในการศึกษาโครงงานนี้ เป็นการนำทฤษฎีการทำดาต้าไมน์นิ่งด้วยวิธีการจำแนกประเภทของข้อมูล โดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation มาพัฒนาเป็นโปรแกรมคอมพิวเตอร์เพื่อประยุกต์ใช้ในการจำแนกประเภทของข้อมูล โดยแบ่งการดำเนินศึกษาออกเป็น 3 ส่วน ดังนี้

1. ข้อมูลที่ใช้กับระบบ
2. ขั้นตอนและวิธีการดำเนินงาน
3. การพัฒนาโปรแกรม

3.1 ข้อมูลที่ใช้กับระบบ

โปรแกรมประยุกต์ที่พัฒนาขึ้นในโครงการนี้จะรับข้อมูลเข้าในลักษณะของ Text File (.txt) ซึ่งลักษณะของข้อมูลแต่ละชุดจะประกอบด้วยคุณลักษณะต่างๆ ของชุดข้อมูลที่นำมาใช้ โดยที่ค่าสุดท้ายของข้อมูลแต่ละชุดนั้นจะบอกถึงกลุ่มที่ชุดของข้อมูลนั้นๆ ถูกจัดให้อยู่ โดยแต่ละคุณลักษณะจะเป็นข้อมูลชนิดตัวเลขที่ถูกคั่นด้วยเครื่องหมายจุลภาค (,) ดังรูปที่ 3.1


iris.0.tn.txt - Notepad

File Edit Format View Help

0.20043573,0.418300654,0.109504819,0.050653595,1
0.173202614,0.459150327,0.092888003,0.009803922,1
0.20043573,0.62254902,0.109504819,0.214052288,1
0.254901961,0.867647059,0.092888003,0.009803922,1
0.037037037,0.5,0.05965437,0.050653595,1
0.472766885,0.581699346,0.591392489,0.62254902,2
0.418300654,0.254901961,0.508308408,0.459150327,2
0.5,0.336601307,0.508308408,0.5,2
0.391067538,0.418300654,0.541542041,0.459150327,2

รูปที่ 3.1 ตัวอย่างข้อมูลที่นำมาใช้

โดยลักษณะของข้อมูลที่นำมาใช้สามารถสรุปได้ ดังนี้ คือ

- ข้อมูลแต่ละบรรทัดจะบอกถึงคุณลักษณะของข้อมูลนั้นๆ 1 ชุด
- ข้อมูลแต่ละชุดจะประกอบไปด้วยคุณลักษณะของข้อมูล ซึ่งข้อมูลแต่ละชุดนั้นจะต้องประกอบไปด้วยจำนวนของคุณลักษณะที่เท่ากัน
- ค่าสุดท้ายของข้อมูลแต่ละชุดนั้นจะบอกถึงกลุ่มที่ชุดข้อมูลนั้นๆ จัดให้อยู่
- คุณลักษณะของข้อมูลจะต้องเป็นข้อมูลชนิดตัวเลขเท่านั้น
- คุณลักษณะของข้อมูลแต่ละตัวจะถูกคั่นด้วยเครื่องหมายจุลภาค (,)

ข้อมูลที่นำมาใช้จะแบ่งเป็น 2 ชุดคือ

1. ข้อมูลฝึกฝน (Training Data) ซึ่งจะนำมาใช้เพื่อสอนการเรียนรู้ของระบบ
2. ข้อมูลทดสอบ (Testing Data) เป็นข้อมูลที่นำมาใช้ทดสอบการทำงานของระบบ

โดยข้อมูลฝึกฝน (Training Data) จะถูกนำเข้าสู่ระบบการแบ่งข้อมูล เมื่อระบบทำการเรียนรู้จากข้อมูลที่ได้รับแล้ว จึงนำข้อมูลทดสอบ (Testing Data) เข้าสู่ระบบเพื่อทดสอบประสิทธิภาพและความถูกต้องของระบบที่ได้จากการเรียนรู้

3.2 ขั้นตอนและวิธีการดำเนินงาน

ในการศึกษาโครงงานนี้จะทำการศึกษาการทำดาต้าไมน์นิ่งด้วยวิธีการจำแนกประเภทของข้อมูล โดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation โดยทำการพัฒนาระบบเป็นโปรแกรมคอมพิวเตอร์ เพื่อประยุกต์ใช้ในการศึกษาเพื่อใช้ในการทำนายการจำแนกประเภทของข้อมูลโดยมีขั้นตอนการดำเนินงานดังนี้

1. การเตรียมข้อมูล (Data Preparation)

ในขั้นตอนนี้เป็นการเตรียมข้อมูลเพื่อเข้าสู่กระบวนการทำดาต้าไมน์นิ่ง เมื่อคัดเลือกข้อมูลที่ต้องการ กำจัดข้อมูลที่ไม่สมบูรณ์ ข้อมูลที่ผิดปกติรวมไปถึงข้อมูลที่ไม่ได้ใช้และข้อมูลที่ไม่มีความสำคัญออกไป เพื่อจัดเตรียมข้อมูล Input และ Output ที่เหมาะสมเพื่อป้อนเข้าสู่ระบบที่เราสร้างขึ้น โดยทำการแปลงข้อมูลให้อยู่ในรูปของ Text File (.txt) ดังนี้

iris.0.tn.txt - Notepad

File Edit Format View Help

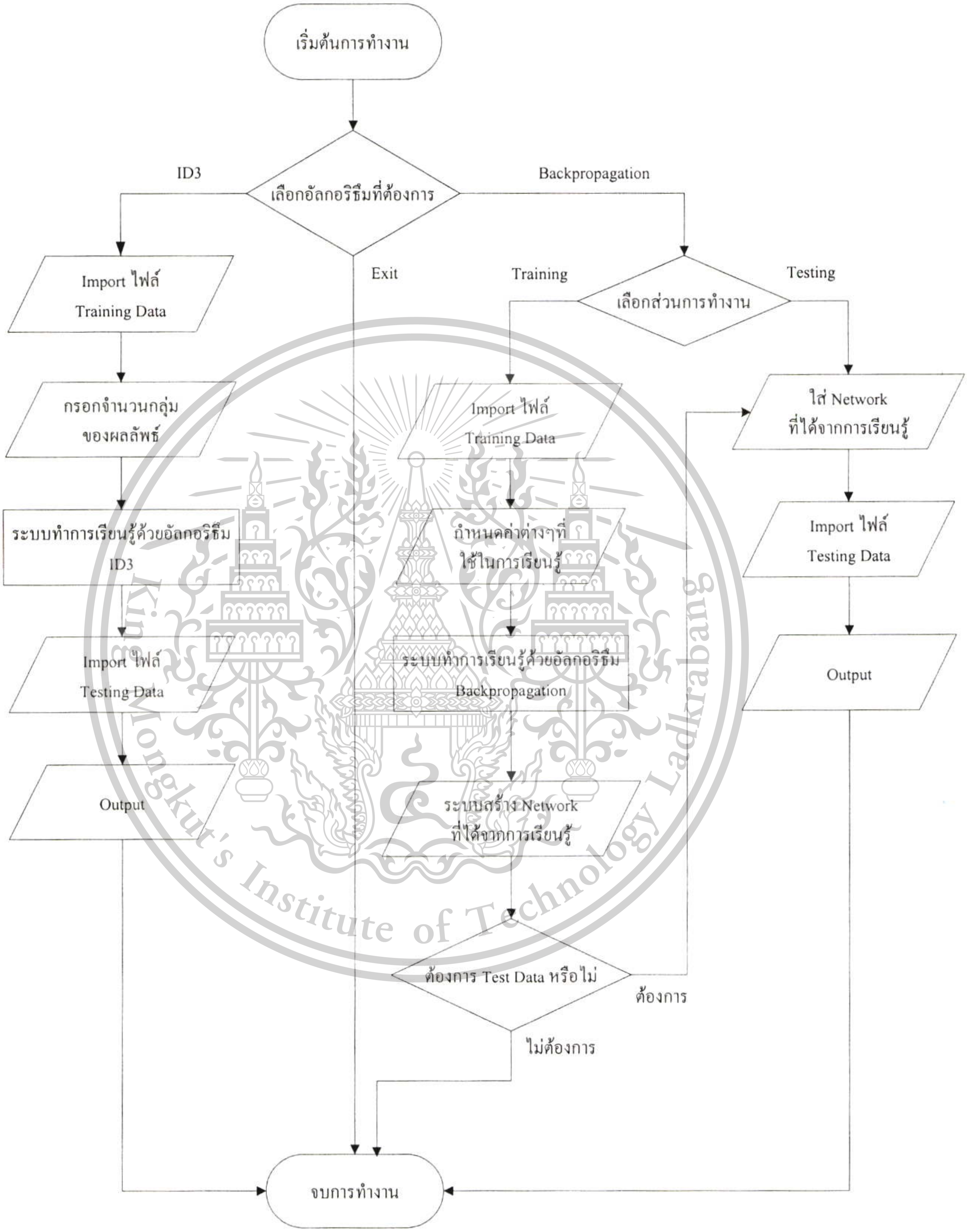
```
0.20043573,0.418300654,0.109504819,0.050653595,1
0.173202614,0.459150327,0.092888003,0.009803922,1
0.20043573,0.62254902,0.109504819,0.214052288,1
0.254901961,0.867647059,0.092888003,0.009803922,1
0.037037037,0.5,0.05965437,0.050653595,1
0.472766885,0.581699346,0.591392489,0.62254902,2
0.418300654,0.254901961,0.508308408,0.459150327,2
0.5,0.336601307,0.508308408,0.5,2
0.391067538,0.418300654,0.541542041,0.459150327,2
0.445533769,0.5,0.641242938,0.704248366,2
0.990196078,0.745098039,0.907111997,0.785947712,3
0.363834423,0.336601307,0.657859754,0.785947712,3
0.608932462,0.418300654,0.757560651,0.704248366,3
0.826797386,0.37745098,0.890495181,0.704248366,3
0.445533769,0.418300654,0.691093387,0.704248366,3
0.254901961,0.62254902,0.092888003,0.050653595,1
0.145969499,0.581699346,0.109504819,0.050653595,1
0.145969499,0.581699346,0.159355268,0.050653595,1
```

รูปที่ 3.2 ตัวอย่างรูปแบบของข้อมูล Input

2. การเลือกข้อมูล

ในการเลือกข้อมูลนั้น ข้อมูลจะประกอบไปด้วยข้อมูลฝึกฝนและข้อมูลทดสอบ โดยการเลือกข้อมูลฝึกฝนนั้นจะต้องเลือกข้อมูลให้ครอบคลุมทุกคุณลักษณะของผลลัพธ์ (Output) และข้อมูลทดสอบนั้นจะต้องไม่เป็นข้อมูลชุดเดียวกันกับข้อมูลฝึกฝน

3.2.1 การทำงานภาพรวมของระบบ



รูปที่ 3.3 การทำงานภาพรวมของระบบ

คำอธิบาย Flow Chart การทำงานภาพรวมของระบบ

จากภาพรวมการทำงานของระบบอธิบายการทำงานได้ดังนี้ คือ เมื่อระบบเริ่มต้นทำงาน ผู้ใช้ทำการเลือกอัลกอริธึมที่ต้องการใช้สำหรับการจำแนกประเภทของข้อมูล

หากผู้ใช้เลือกอัลกอริธึม ID3

1. ผู้ใช้ทำการ Import ไฟล์ Training Data เข้าสู่ระบบ ไฟล์ที่ใช้จะอยู่ในรูปแบบของ Text File (.txt) โดยลักษณะของข้อมูลเข้าสามารถดูคำอธิบายได้ในหัวข้อที่ 3.1
2. ผู้ใช้กรอกจำนวนกลุ่มของผลลัพธ์ (Target)
3. ระบบทำการเรียนรู้ด้วยอัลกอริธึม ID3 จาก Training Data ที่ถูกป้อนเข้าไป (โดยสามารถดูรายละเอียดขั้นตอนการสร้าง Decision Tree ด้วยอัลกอริธึม ID3 ได้ในหัวข้อที่ 3.2.1)
4. เมื่อผู้ใช้ต้องการทดสอบการจำแนกประเภทของข้อมูล ให้ผู้ใช้ Import ไฟล์ Testing Data
5. ระบบทำการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม ID3 และให้ Output ออกมาในรูปของ Text File โดยกลุ่มที่ถูกจำแนกได้จะอยู่ใน Text File เดียวกันกับที่ผู้ใช้ป้อนเข้าไป

หากผู้ใช้เลือกอัลกอริธึม Backpropagation

1. ผู้ใช้ทำการ Import ไฟล์ Training Data เข้าสู่ระบบ ไฟล์ที่ใช้จะอยู่ในรูปแบบของ Text File (.txt) โดยลักษณะของข้อมูลเข้าสามารถดูคำอธิบายได้ในหัวข้อที่ 3.1
2. ผู้ใช้กำหนดค่าต่างๆ ที่ใช้ในการเรียนรู้ของข้อมูล
3. ผู้ใช้กรอกจำนวนกลุ่มของผลลัพธ์ (Target)
4. ระบบทำการเรียนรู้ด้วยอัลกอริธึม Backpropagation จาก Training Data ที่ถูกป้อนเข้าไป (โดยสามารถดูรายละเอียดขั้นตอนการสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation ได้ในหัวข้อที่ 3.2.2)
5. ระบบทำการเรียนรู้และสร้าง Network จากข้อมูล Training Data ที่ป้อนเข้าไป โดยระบบจะสร้างไฟล์ .nn ขึ้นมา และออกมาเป็น Output
6. เมื่อผู้ใช้ต้องการทดสอบการจำแนกประเภทของข้อมูล ให้ผู้ใช้ Import ไฟล์ Testing Data
7. ระบบทำการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation และให้ Output ออกมาในรูปของ Text File โดยกลุ่มที่ถูกจำแนกได้จะอยู่ใน Text File เดียวกันกับที่ผู้ใช้ป้อนเข้าไป

3.2.2 ขั้นตอนการสร้าง Decision Tree ด้วยอัลกอริธึม ID3

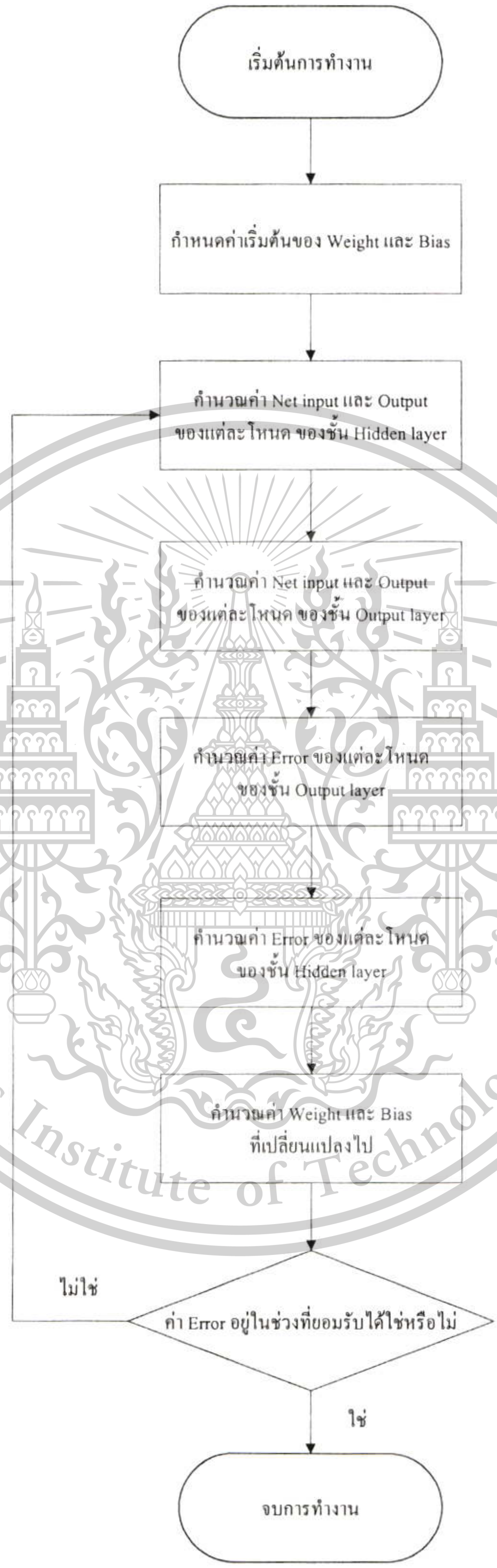
รูปที่ 3.4 การสร้าง Decision Tree ด้วยอัลกอริธึม ID3

คำอธิบาย Flow Chart การสร้าง Decision Tree ด้วยอัลกอริธึม ID3

จากภาพการทำงานของระบบอธิบายการทำงานได้ดังนี้ คือ

1. หาแอททริบิวต์ที่สำคัญที่สุดมาจำแนกประเภทข้อมูล โดยในขั้นตอนที่ 1 จะคำนวณหาค่า Information Gain ของแต่ละแอททริบิวต์
2. คำนวณหาค่า Entropy ของแต่ละแอททริบิวต์
3. คำนวณหาค่า Gain ของทุกแอททริบิวต์
4. เลือกแอททริบิวต์ที่มีค่า Gain สูงที่สุดมาใช้ในการสร้างเป็น Root node โดยมี Target Attribute เป็นผลลัพธ์ ซึ่งเป็น Leaf Node ที่ถูกกำหนดไว้ก่อน
5. แตกกิ่งจากโหนดที่ถูกสร้างขึ้นมา โดยแตกกิ่งตามค่าที่มีในคลาส
6. วนกลับไปทำในขั้นตอนที่แรก คือ หาแอททริบิวต์ที่สำคัญที่สุดจากข้อมูลที่เข้ามาเพื่อหาตัวจำแนกประเภทต่อไป จนกว่าจะไม่มีแอททริบิวต์เป้าหมาย (Target Attribute) เหลืออยู่

3.2.3 ขั้นตอนการสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation



รูปที่ 3.5 การสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation

คำอธิบาย Flow Chart การสร้าง Neural Network ด้วยอัลกอริธึม Backpropagation

จากภาพการทำงานของระบบอธิบายการทำงานได้ดังนี้ คือ

1. กำหนดค่าเริ่มต้นของค่า Weight และ Bias โดยค่า Weight จะถูกกำหนดค่าเริ่มต้นด้วยตัวเลขสุ่ม (random) ที่มีค่าอยู่ในช่วง -1.0 ถึง 1.0
2. ขั้นตอนการผ่าน Input เข้าสู่โครงข่าย โดยขั้นตอนนี้จะทำการคำนวณหาค่า Net input และ Output ของแต่ละโหนดในชั้น Hidden Layer
3. คำนวณหาค่า Net input และ Output ของแต่ละโหนดในชั้น Output Layer
4. คำนวณหาค่าความคลาดเคลื่อน (Error) ของแต่ละโหนดในชั้น Output Layer
5. คำนวณหาค่าความคลาดเคลื่อน (Error) ของแต่ละโหนดในชั้น Hidden Layer
6. คำนวณหาค่า Weight และค่า Bias ที่ได้เปลี่ยนแปลงไป
7. วนกลับไปทำซ้ำในขั้นตอนแรกจนกว่าค่า Error อยู่ในช่วงที่ยอมรับได้

3.3 การพัฒนาโปรแกรมประยุกต์

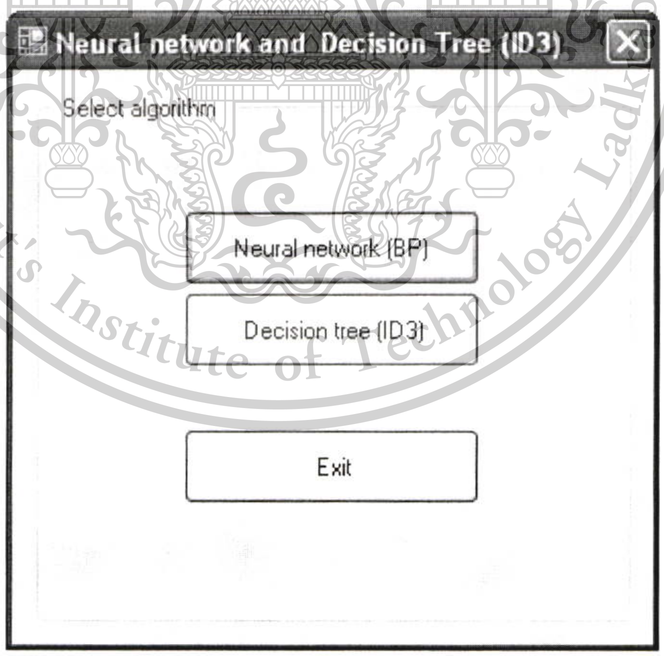
3.3.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

การออกแบบและพัฒนาระบบงาน มีเครื่องมือที่ใช้ในการพัฒนา ดังนี้

- คอมพิวเตอร์ Laptop ยี่ห้อ Dell รุ่น Latitude E6400 : Intel® Core™ 2 Duo CPU P8400 2.26GHz. RAM 4 GB.
- ระบบปฏิบัติการ Microsoft Windows XP Professional Ver. 2002 Service Pack 3 โดยเลือกใช้ ในส่วนของ VB.NET ในการพัฒนาระบบงาน
- Microsoft Visual Studio.Net Team System 2008
- Microsoft Office System 2007 Professional Edition

3.3.2 ส่วนติดต่อผู้ใช้ระบบ

ในการออกแบบโปรแกรมประยุกต์นั้น เราคำนึงถึงความง่ายในการใช้งานเป็นหลัก โดยมีการแบ่งหน้าจอการทำงานหลักๆ ดังนี้

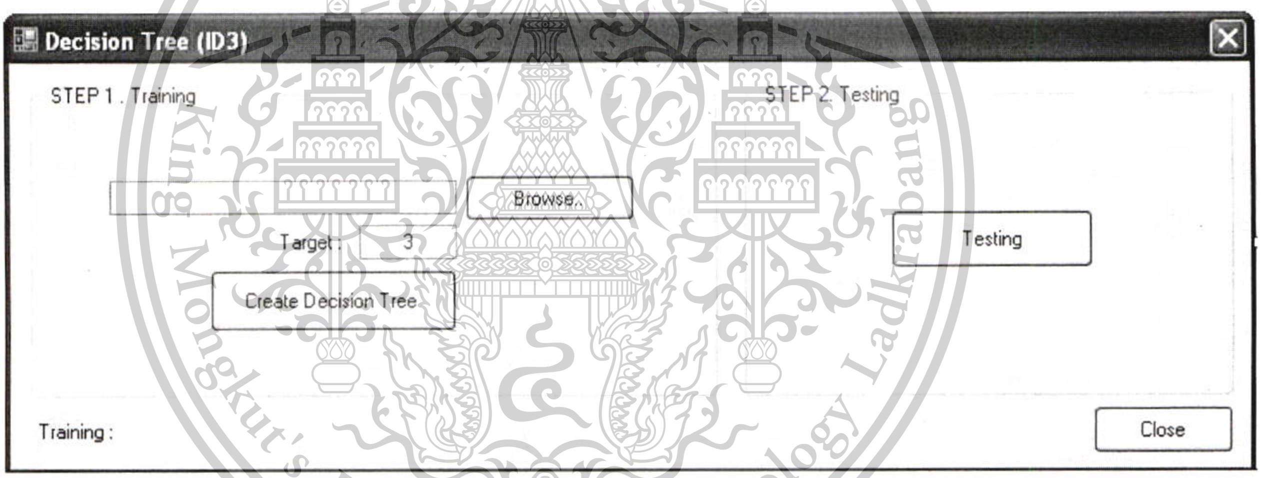


รูปที่ 3.6 หน้าจอหลักของโปรแกรม

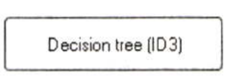
ในส่วนแรกคือหน้าจอหลัก (Main Menu) เป็นหน้าจอสำหรับการเลือกอัลกอริธึมที่ต้องการ ประกอบด้วย

- **Neural Network (BP)**
  คือส่วนที่นำเข้าสู่โปรแกรมการแบ่งข้อมูลด้วยเทคนิควิธีโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation
- **Decision Tree (ID3)**
  คือส่วนที่นำเข้าสู่โปรแกรมการแบ่งข้อมูลด้วยเทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3
- **Exit**
  คือส่วนที่ออกจากโปรแกรม

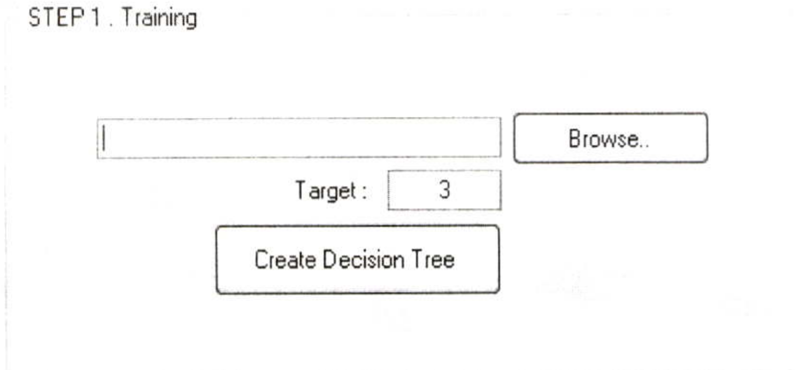
หน้าจอในส่วนของ Decision Tree (ID3)



รูปที่ 3.7 หน้าจอในส่วนของ Decision Tree (ID3)

เมื่อกดปุ่ม 

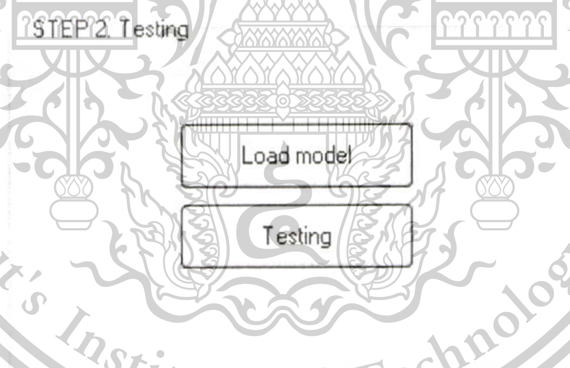
จากหน้าจอหลักแล้วโปรแกรมจะนำเข้าสู่หน้าจอในส่วนของ Decision Tree (ID3) ดังรูปที่ 3.7 โดยจะแบ่งการทำงานออกเป็น 2 ส่วนคือ



รูปที่ 3.8 หน้าจอในส่วนของการเทรนข้อมูล

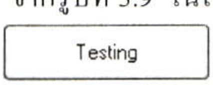
ส่วนที่ 1 : Training

จากรูปที่ 3.8 ในส่วนนี้ใช้สำหรับการนำเข้าข้อมูลฝึกฝน (Training Data) เพื่อให้ระบบทำการเรียนรู้จากข้อมูลในลักษณะของ Text File ที่ป้อนเข้าไป โดยในส่วนของ Target ผู้ใช้ต้องกรอกจำนวนกลุ่มของเป้าหมายหรือจำนวนกลุ่มของผลลัพธ์ที่จะเป็นไปได้ในการจำแนกประเภทของข้อมูล



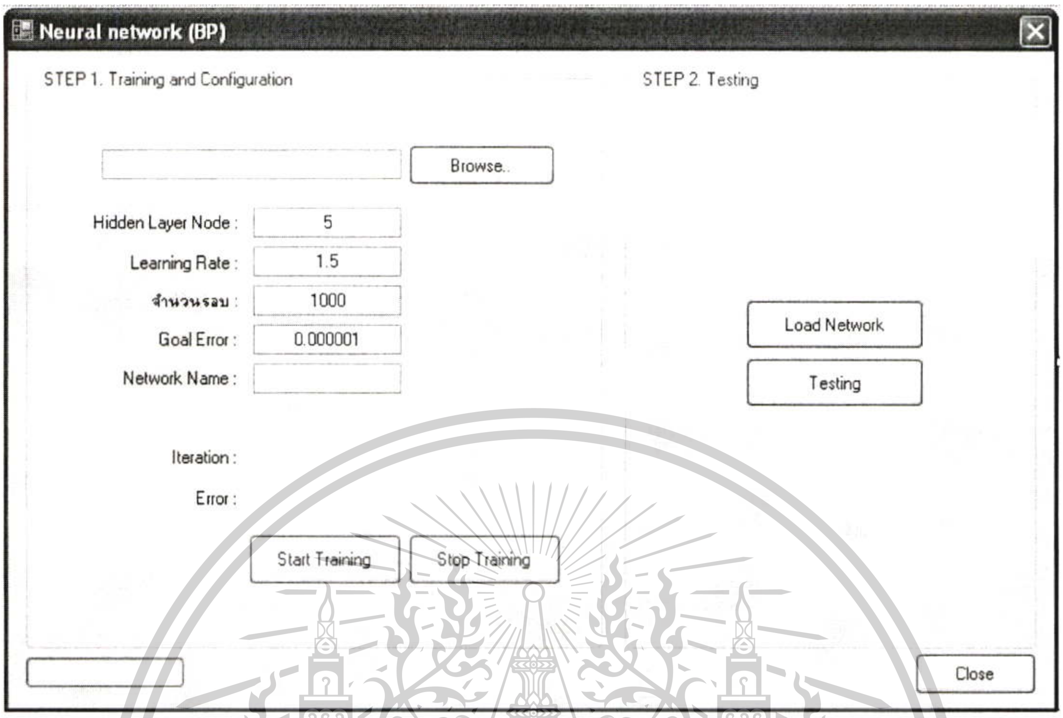
รูปที่ 3.9 หน้าจอในส่วนของการเทสข้อมูล

ส่วนที่ 2 : Testing

จากรูปที่ 3.9 ในส่วนนี้ใช้สำหรับการทดสอบการจำแนกประเภทของข้อมูล โดยให้ผู้ใช้กดปุ่ม 

ป้อนข้อมูลทดสอบ (Testing Data)ในรูปแบบของ Text File ที่ต้องการทดสอบเข้าไป

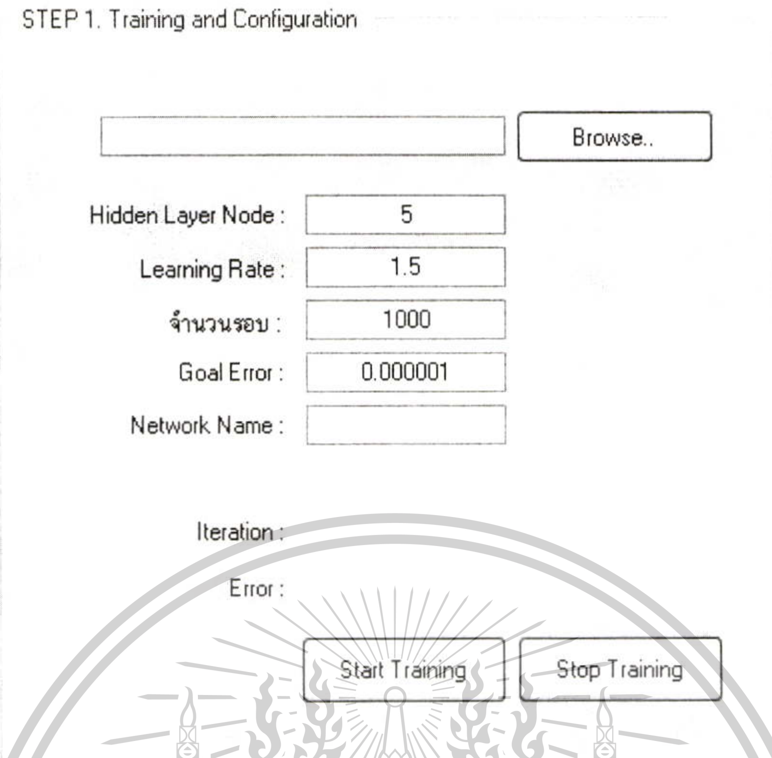
หน้าจอในส่วนของ Neural Network (Backpropagation)



รูปที่ 3.10 หน้าจอในส่วนของ Neural Network (Backpropagation)

เมื่อกดปุ่ม 

จากหน้าจอหลักแล้วโปรแกรมจะนำเข้าสู่หน้าจอในส่วนของ Neural Network (Backpropagation) ดังรูปที่ 3.10 โดยจะแบ่งการทำงานออกเป็น 2 ส่วนคือ



รูปที่ 3.11 หน้าจอในส่วนของ Training and Configuration

ส่วนที่ 1 : Training and Configuration

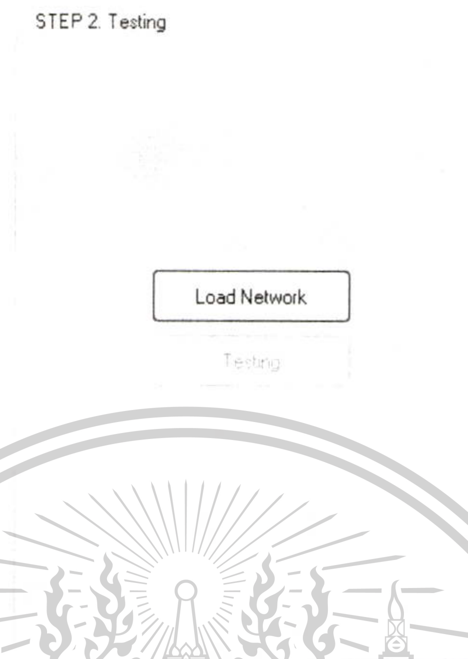
จากรูปที่ 3.11 ในส่วนนี้ใช้สำหรับการนำเข้าข้อมูลฝึกฝน (Training Data) เพื่อให้ระบบทำการเรียนรู้จากข้อมูลในลักษณะของText File ที่ป้อนเข้าไป ในส่วนนี้ผู้ใช้จะต้องกำหนดข้อมูลต่างๆ คือ

- Hidden Layer Node คือ จำนวนของ Hidden Layer Node
- Learning Rate คือ อัตราการเรียนรู้
- จำนวนรอบ คือ จำนวนรอบของการเทรนข้อมูล ซึ่งระบบจะหยุดการเทรนข้อมูลหากมีการวนครบจำนวนรอบที่กำหนด
- Goal Error คือ ค่า Error ที่ยอมรับได้ ซึ่งระบบจะหยุดการเทรนข้อมูลหากค่า Error อยู่ในเกณฑ์ที่ยอมรับได้ตามที่ผู้ใช้กำหนด
- Network Name คือ ชื่อของ Network ที่ได้จากการ Train

นอกจากนี้ยังมีส่วนที่แสดงข้อมูล อีก 2 ชนิด คือ

- Iteration คือ จำนวนรอบของการเทรนข้อมูล ซึ่งจะแสดงจำนวนรอบขณะที่ทำการเทรนข้อมูล

- Error คือ ค่า Error ขณะที่ทำการเทรนข้อมูล



รูปที่ 3.12 หน้าจอในส่วนของการ Testing

ส่วนที่ 2 : **Testing**

จากรูปที่ 3.12 ในส่วนนี้ใช้สำหรับการทดสอบการจำแนกประเภทของข้อมูล โดยให้ผู้ใช้กดปุ่ม Load network เพื่อทำการป้อนไฟล์ที่ได้จากการเรียนรู้ (Training) เข้าไป จากนั้นกดปุ่ม Testing เพื่อทำการป้อนข้อมูลทดสอบ (Testing Data) ในรูปแบบของText File ที่ต้องการทดสอบเข้าไป

บทที่ 4

ผลการทดลอง

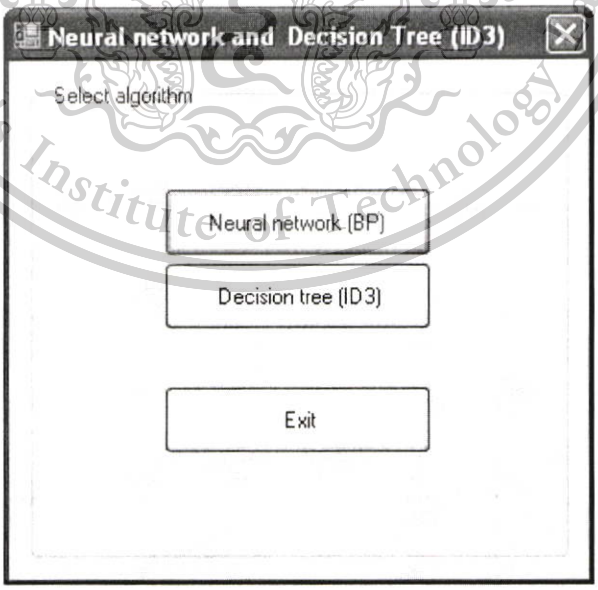
ในบทนี้จะกล่าวถึงขั้นตอนและวิธีการใช้งานโปรแกรมประยุกต์ โดยแบ่งออกเป็นส่วนๆ ดังนี้

1. การประยุกต์ใช้งานระบบ
2. ข้อมูลที่ใช้กับโปรแกรมประยุกต์
3. ผลการทดลองและการวิเคราะห์

4.1 การประยุกต์ใช้งานระบบ

ในการทำงานโปรแกรมประยุกต์สำหรับการแก้ปัญหาการจำแนกประเภทของข้อมูลนั้น ได้ถูกออกแบบมาเพื่อความสะดวกและมีหน้าจอการทำงานที่ใช้งานได้โดยง่าย แต่สามารถที่จะประมวลผลและจำแนกประเภทของข้อมูลให้ได้ผลลัพธ์ที่น่าพอใจ โดยมีขั้นตอนวิธีการทำงานดังนี้

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอหลัก (Main Menu) ดังรูปที่ 5.1



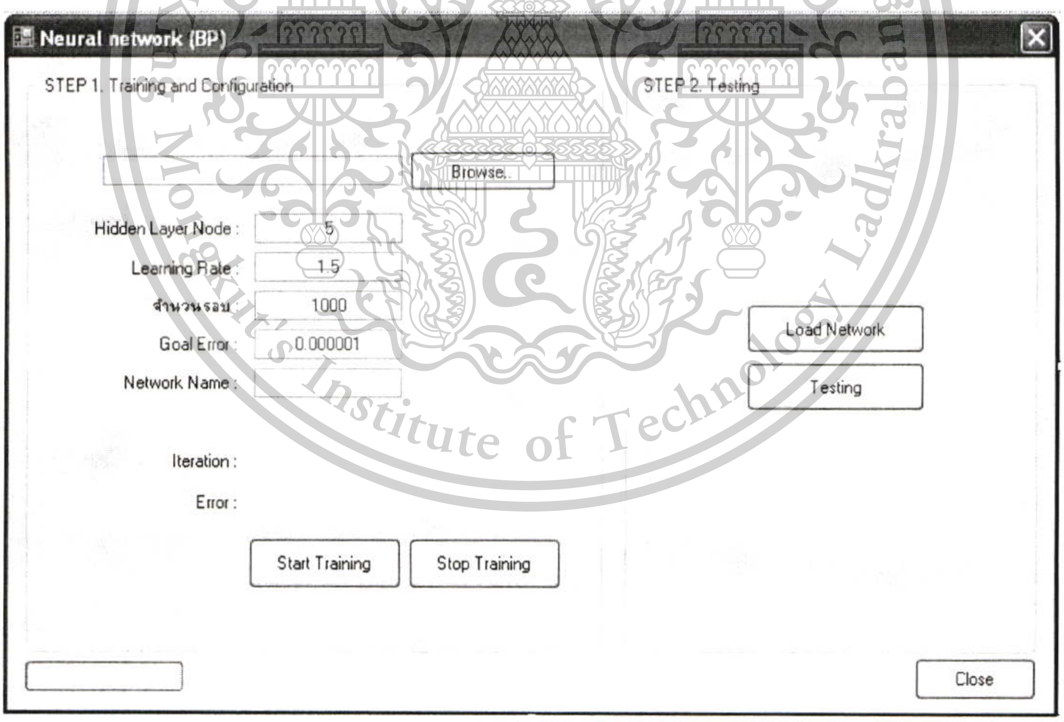
รูปที่ 4.1 หน้าจอหลักของโปรแกรม

เมื่อปรากฏหน้าจอนี้ ผู้ใช้สามารถทำการเลือกอัลกอริธึมที่ต้องการใช้ในการทำนายข้อมูลได้ โดยสามารถใช้งานได้ ดังนี้

1. ผู้ใช้กดปุ่ม Neural network (BP) หากต้องการเลือกใช้วิธี Neural Network ด้วยอัลกอริธึม Backpropagation
2. ผู้ใช้กดปุ่ม Decision tree (ID3) หากต้องการเลือกใช้วิธี Decision Tree ด้วยอัลกอริธึม ID3
3. ผู้ใช้กดปุ่ม Exit หากต้องการจบการทำงาน

4.1.1 การจำแนกประเภทข้อมูลด้วยวิธี Neural Network ด้วยอัลกอริธึม Backpropagation

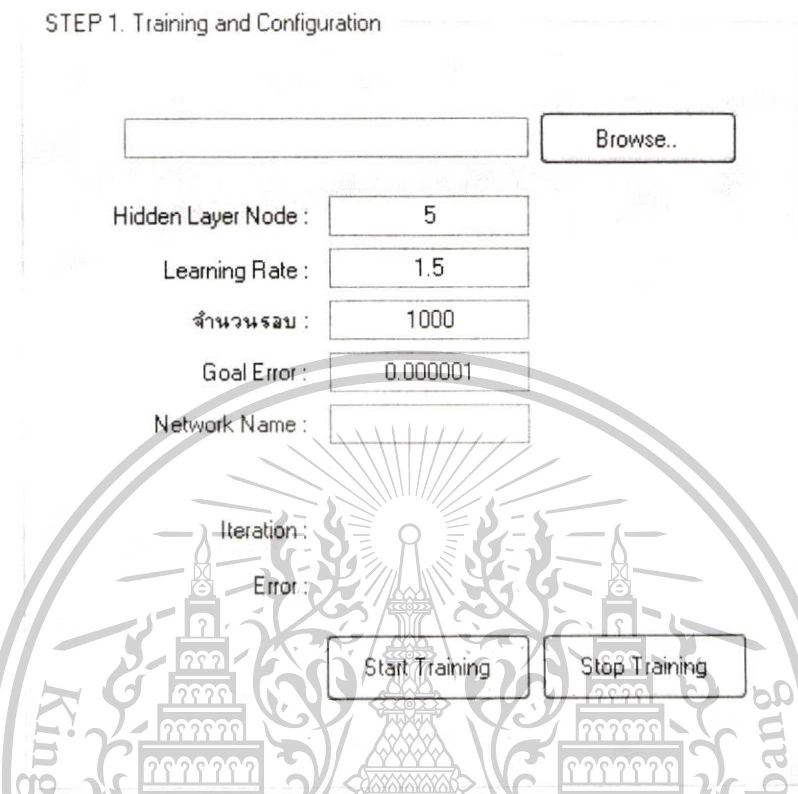
เมื่อผู้ใช้กดปุ่ม Neural network (BP) เพื่อเลือกใช้วิธี Neural Network ด้วยอัลกอริธึม Backpropagation แล้วจะปรากฏหน้าจอ ดังรูปที่ 4.2



รูปที่ 4.2 หน้าจอในส่วนของ Neural Network (Backpropagation)

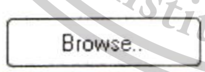
ในหน้าจอการทำงานนี้จะแบ่งออกเป็น 2 ส่วนคือ

ส่วนที่ 1 : Training and Configuration



รูปที่ 4.3 หน้าจอในส่วนของการ Training and Configuration

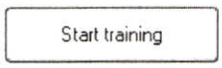
ในส่วนนี้คือส่วนที่ใช้ในการเรียนรู้ข้อมูลจากข้อมูลฝึกฝน(Training Data) ที่ป้อนเข้าสู่ระบบในรูปแบบของ Text File โดยจะถูกวิเคราะห์ด้วยอัลกอริธึม Backpropagation ซึ่งมีขั้นตอนการทำงาน ดังนี้

1. กดปุ่ม  เพื่อทำการเลือก Text File ที่ต้องการป้อนเข้าสู่ระบบ

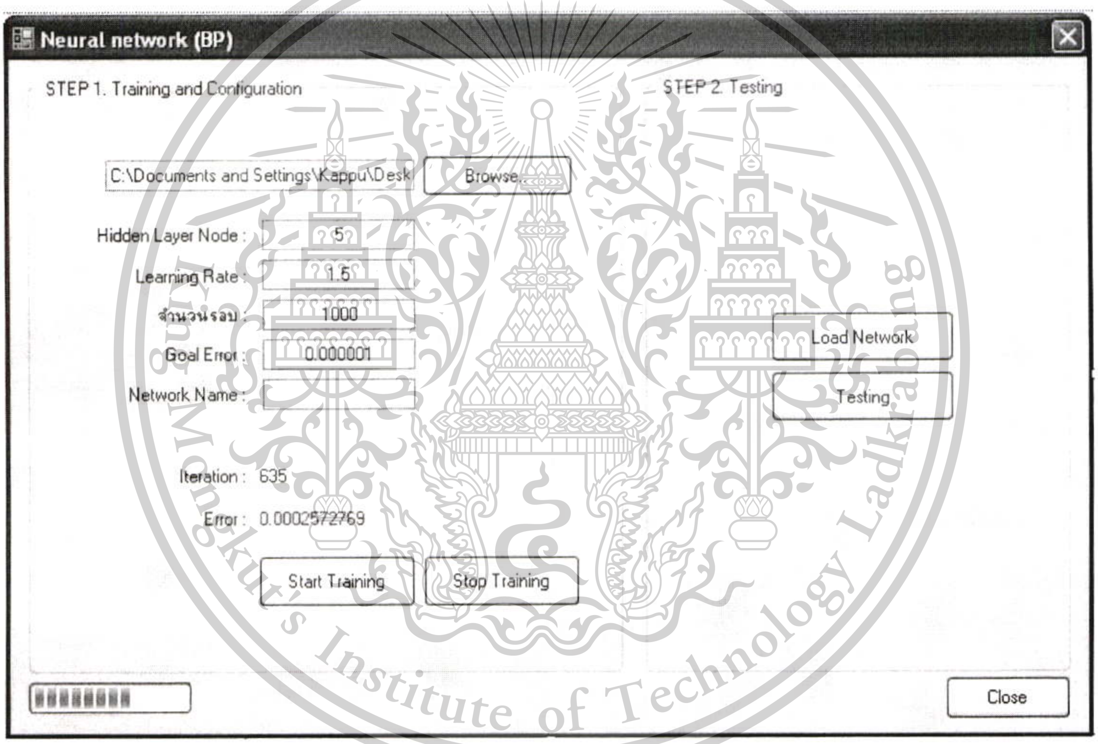
2. กำหนดค่าต่างๆ ดังนี้
    - Hidden Layer Node คือ จำนวนของ Hidden Layer Node
    - Learning Rate คือ อัตราการเรียนรู้
    - จำนวนรอบ คือ จำนวนรอบของการเทรนข้อมูล ซึ่งระบบจะหยุดการเทรนข้อมูลหากมีการวนครบจำนวนรอบที่กำหนด
    - Goal Error คือ ค่า Error ที่ยอมรับได้ ซึ่งระบบจะหยุดการเทรนข้อมูลหากค่า Error อยู่ในเกณฑ์ที่ยอมรับได้ตามที่ผู้ใช้กำหนด
    - Network Name คือ ชื่อของ Network ที่ได้จากการ Train

นอกจากนี้ยังมีส่วนที่แสดงข้อมูล อีก 2 ชนิด คือ

- Iteration คือ จำนวนรอบของการเทรนข้อมูล ซึ่งจะแสดงจำนวนรอบขณะที่ทำการเทรนข้อมูล
- Error คือ ค่า Error ขณะที่ทำการเทรนข้อมูล

3. จากนั้นกดปุ่ม 

เพื่อเริ่มต้นในการเทรนข้อมูล เมื่อระบบเริ่มทำการเทรนข้อมูล จำนวนรอบ Iteration จะเพิ่มขึ้นตามจำนวนรอบ และค่า Error จะลดลงเรื่อยๆ เมื่อจำนวนรอบในการเทรนเพิ่มขึ้น

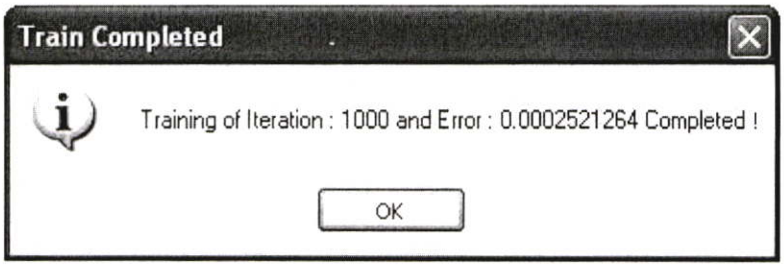


รูปที่ 4.4 หน้าจอในการเทรนข้อมูลของ Backpropagation

4. ผู้ใช้สามารถกดปุ่ม 

หากต้องการหยุดการทำงาน

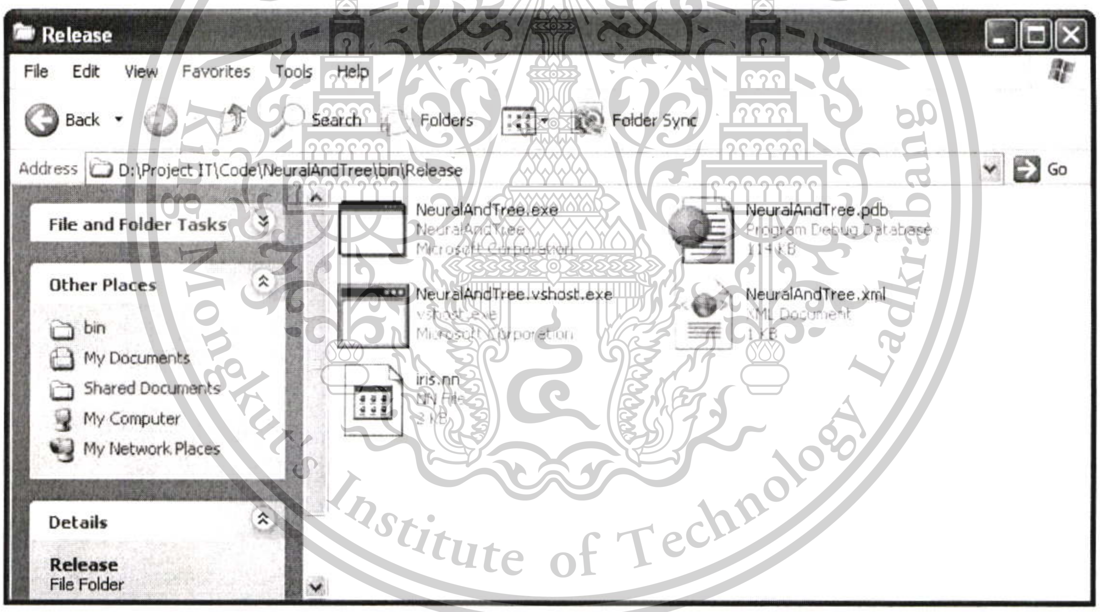
5. ระบบจะทำการหยุดเทรนเมื่อครบจำนวนรอบที่กำหนดไว้ หรือค่า Error อยู่ในช่วงที่ยอมรับได้ซึ่งค่านี้ผู้ใช้เป็นผู้กำหนดไว้เช่นเดียวกัน เมื่อระบบทำการเทรนข้อมูลเสร็จเรียบร้อยแล้วจะปรากฎหน้าจอ ดังรูปที่ 4.5



รูปที่ 4.5 หน้าจอแสดงข้อความเมื่อระบบทำการเทรนข้อมูลเสร็จ

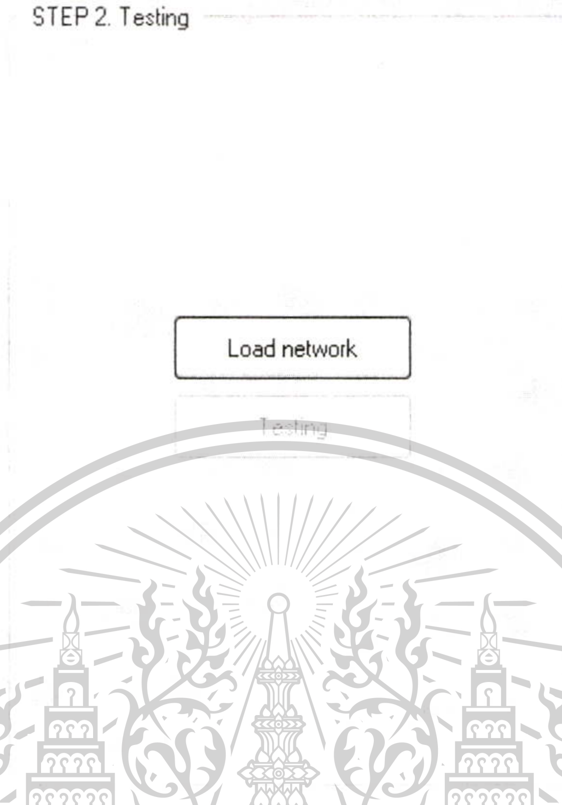
6. เมื่อทำการเทรนข้อมูลเสร็จเรียบร้อย ระบบจะสร้างไฟล์ .nn ขึ้นมา 

โดยชื่อของไฟล์จะเป็นไปตามที่ผู้ใช้ได้ตั้งไว้ในขั้นตอนที่ 2 ในส่วนของ Network Name ซึ่งไฟล์นี้จะทำการเก็บค่า Weight ของแต่ละกิ่งของ Network เพื่อนำไปใช้สำหรับการทดสอบ (Test) การจำแนกประเภทของข้อมูลในภายหลังได้



รูปที่ 4.6 หน้าจอแสดงไฟล์ที่ถูกสร้างขึ้นมา

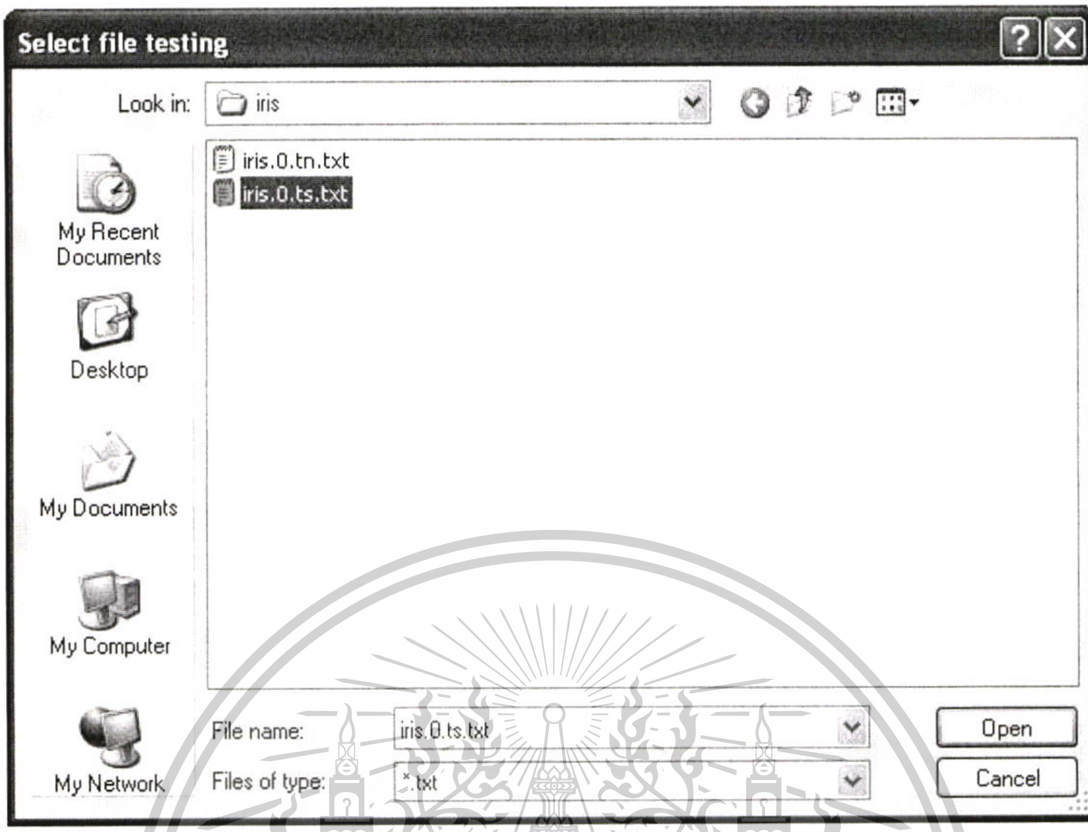
ส่วนที่ 2 : Testing



รูปที่ 4.7 หน้าจอในส่วนของการ Testing ของ Backpropagation

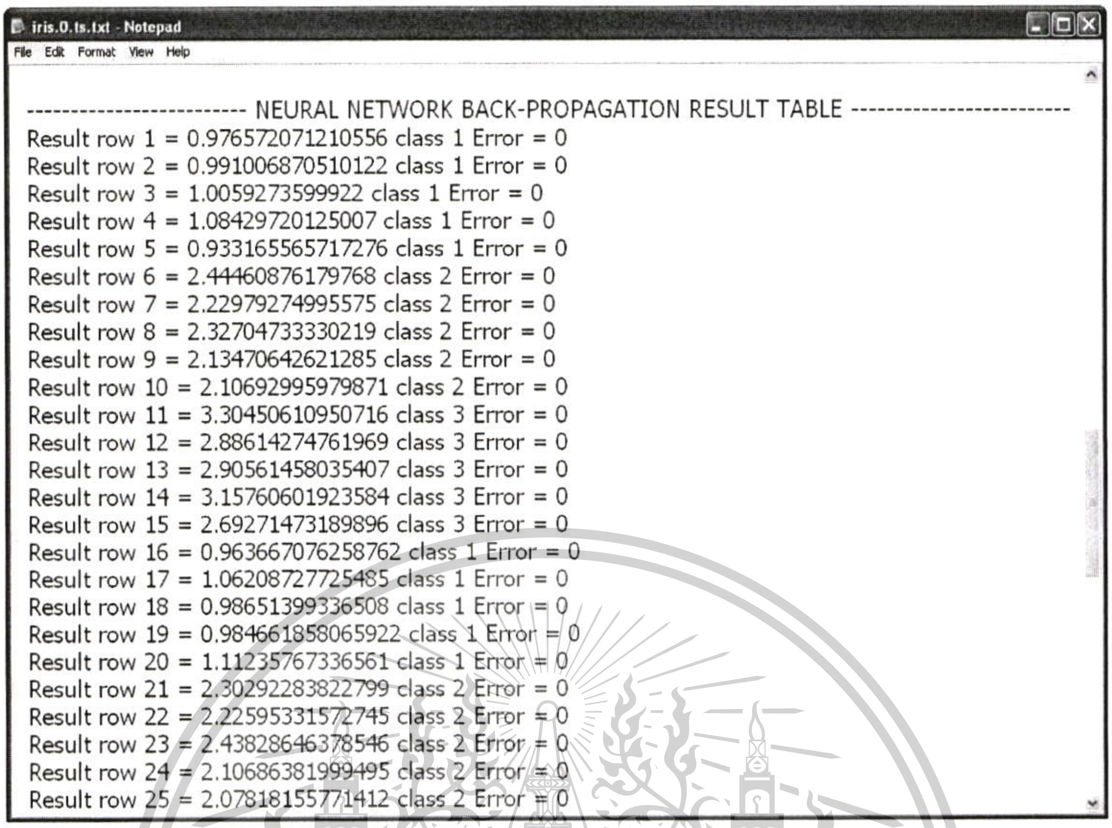
ในส่วนนี้คือส่วนที่นำเอา Test Data ที่ทราบผลอยู่แล้วว่าอยู่ในกลุ่มใดไปเปรียบเทียบกับ Learning Model ที่ตัวระบบเรียนรู้ไว้ ว่าผลที่ได้มีความถูกต้องมากน้อยเพียงใด ซึ่งมีขั้นตอนการทำงานดังนี้

1. กดปุ่ม [Load network] เพื่อทำการป้อนไฟล์ .nn ที่ถูกสร้างไว้ในขั้นตอนของการ Training Data
2. กดปุ่ม [Testing] เมื่อต้องการทดสอบข้อมูล (Testing Data)
3. เมื่อกดปุ่ม [Testing] จะปรากฏหน้าจอสำหรับเลือก Text File ที่ต้องการใช้ในการทดสอบ ดังรูปที่ 4.8



รูปที่ 4.8 หน้าจอสำหรับเลือกไฟล์ที่ต้องการใช้ทดสอบ

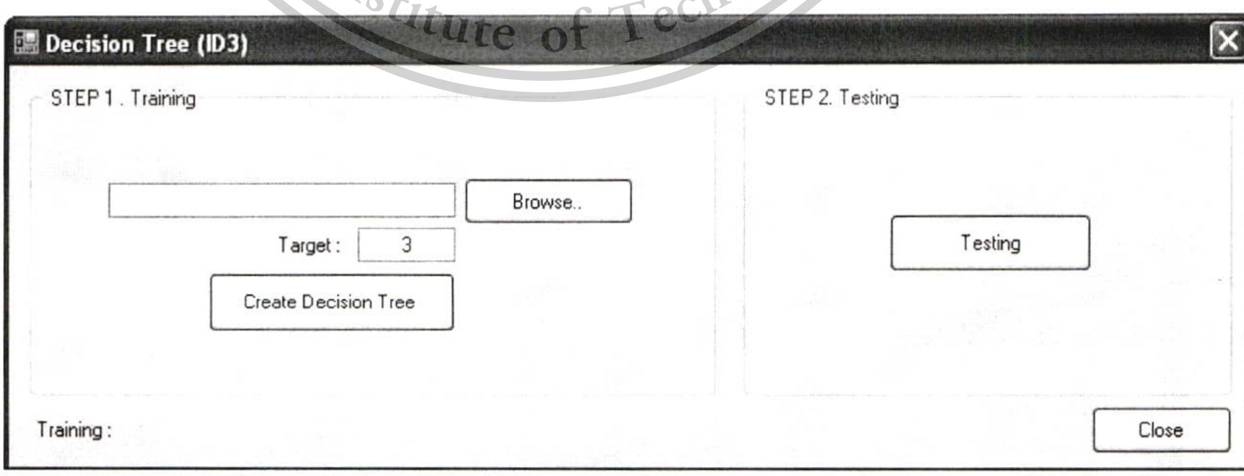
4. ระบบจะทำการจำแนกประเภทของข้อมูล โดยผลของการจำแนกประเภทของข้อมูลนั้น จะถูกเขียนต่อไว้จากข้อมูลทดสอบใน Text File เดิมที่ป้อนเข้าไปในขั้นตอนที่ 3



รูปที่ 4.9 หน้าจอแสดงผลลัพธ์ที่ได้จากการจำแนกประเภทข้อมูลด้วยวิธี Backpropagation

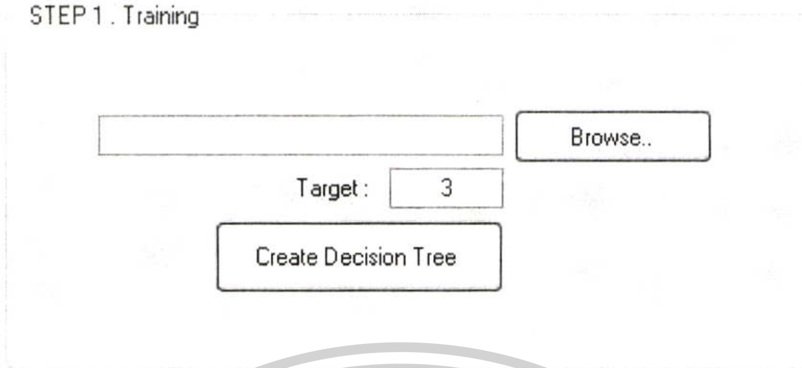
4.1.2 การจำแนกประเภทข้อมูลด้วยวิธี Decision Tree ด้วยอัลกอริธึม ID3

เมื่อกดปุ่ม Decision tree (ID3) แล้วโปรแกรมจะนำเข้าสู่หน้าจอในส่วนของ Decision Tree (ID3) ดังรูปที่ 4.10



รูปที่ 4.10 หน้าจอในส่วนของ Decision Tree (ID3)

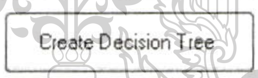
ในหน้าจอการทำงานนี้จะแบ่งออกเป็น 2 ส่วนคือ

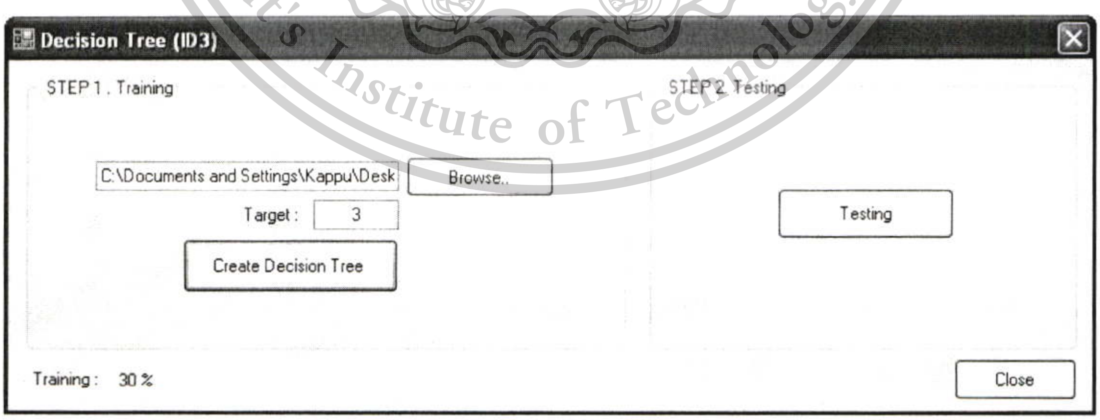


รูปที่ 4.11 หน้าจอในส่วนของการ Training

ส่วนที่ 1 : Training

ในส่วนนี้ใช้สำหรับการนำเข้าข้อมูลฝึกฝน (Training Data) เพื่อให้ระบบทำการเรียนรู้จากข้อมูลในลักษณะของ Text File ที่ป้อนเข้าไป ซึ่งมีขั้นตอนการทำงาน ดังนี้

1. กดปุ่ม  เพื่อทำการเลือก Text File ที่ต้องการป้อนเข้าสู่ระบบ
2. ในช่อง "Target" ให้กรอกจำนวนกลุ่มของเป้าหมายหรือจำนวนกลุ่มของผลลัพธ์ที่จะเป็นไปได้ในการจำแนกประเภทของข้อมูล
3. จากนั้นกดปุ่ม  เพื่อเริ่มต้นในการเทรนข้อมูล



รูปที่ 4.12 หน้าจอในการเทรนข้อมูลของ ID3

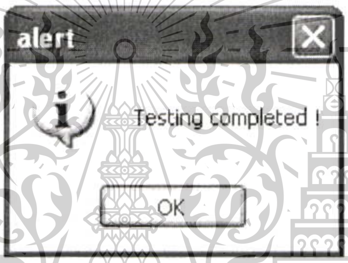
ส่วนที่ 2 : Testing

ในส่วนนี้ใช้สำหรับการทดสอบการจำแนกประเภทข้อมูล ซึ่งมีขั้นตอนการทำงาน ดังนี้

1. โดยให้ผู้ใช้กดปุ่ม Testing ป้อนข้อมูลทดสอบ (Testing Data)ในรูปแบบของ Text File ที่ต้องการทดสอบเข้าไป

2. ระบบจะทำการจำแนกประเภทของข้อมูล โดยผลของการจำแนกประเภทของข้อมูลนั้น จะถูกเขียนต่อไว้จากข้อมูลทดสอบใน Text File เดิมที่ป้อนเข้าไป

3. เมื่อระบบทำการเทรนข้อมูลเสร็จเรียบร้อยแล้วจะปรากฏหน้าจอ ดังรูปที่ 4.13 โดยระบบจะทำการจำแนกประเภทของข้อมูล โดยผลของการจำแนกประเภทของข้อมูลนั้น จะถูกเขียนต่อไว้จากข้อมูลทดสอบใน Text File เดิมที่ป้อนเข้าไปในขั้นตอนที่ 1



รูปที่ 4.13 หน้าจอแสดงข้อความเมื่อระบบทำการทดสอบข้อมูลเสร็จ

```
iris.0.ts.txt - Notepad
File Edit Format View Help
------------------------- DECESSION TREE RESULT TABLE -------------------------
Result row 1 = class 1 Error = 0
Result row 2 = class 1 Error = 0
Result row 3 = class 1 Error = 0
Result row 4 = class 1 Error = 0
Result row 5 = class 1 Error = 0
Result row 6 = class 2 Error = 0
Result row 7 = class 2 Error = 0
Result row 8 = class 2 Error = 0
Result row 9 = class 2 Error = 0
Result row 10 = class 2 Error = 0
Result row 11 = class 4 Error = 1
Result row 12 = class 3 Error = 0
Result row 13 = class 3 Error = 0
Result row 14 = class 3 Error = 0
Result row 15 = class 3 Error = 0
Result row 16 = class 1 Error = 0
Result row 17 = class 1 Error = 0
Result row 18 = class 1 Error = 0
Result row 19 = class 1 Error = 0
Result row 20 = class 1 Error = 0
Result row 21 = class 2 Error = 0
Result row 22 = class 2 Error = 0
Result row 23 = class 2 Error = 0
Result row 24 = class 2 Error = 0
Result row 25 = class 2 Error = 0
Result row 26 = class 3 Error = 0
```

รูปที่ 4.14 หน้าจอแสดงผลลัพธ์ที่ได้จากการจำแนกประเภทข้อมูลด้วยวิธี ID3

4.2 ข้อมูลที่ใช้กับโปรแกรมประยุกต์

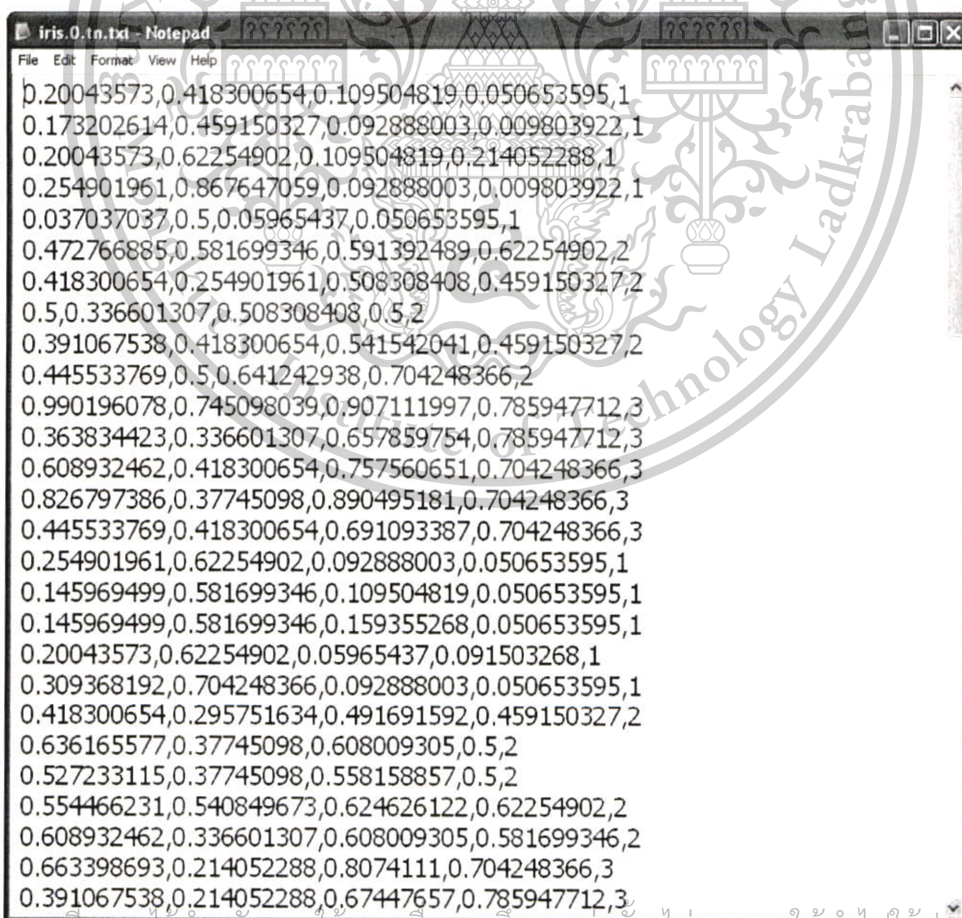
ข้อมูลที่นำมาใช้ทดสอบกับโปรแกรมประยุกต์ จะถูกจัดเก็บอยู่ในรูปของ Text File ซึ่งมีข้อมูลตัวอย่าง ดังนี้

ข้อมูลดอกไอริส (Iris)

ข้อมูลดอกไม้ Iris มีลักษณะข้อมูลทั้งหมด 4 คุณลักษณะที่แตกต่างกัน คือ

- ความยาวกลีบเลี้ยงดอกไม้
- ค่าความกว้างกลีบเลี้ยงดอกไม้
- ค่าความยาวของกลีบดอกไม้
- ค่าความกว้างของกลีบดอกไม้

โดยชนิดของดอกไม้ Iris จะมีทั้งหมด 3 ชนิด โดยมีจำนวนข้อมูลทั้งหมด 150 ตัวอย่าง โดยแบ่งเป็นชนิดละ 50 ตัวอย่าง โดยจะทำการแบ่งข้อมูลเป็นข้อมูลเทรน (Training Data) 90 ตัวอย่าง และข้อมูลสำหรับทดสอบ (Testing Data) 60 ตัวอย่าง


iris.0.tn.txt - Notepad

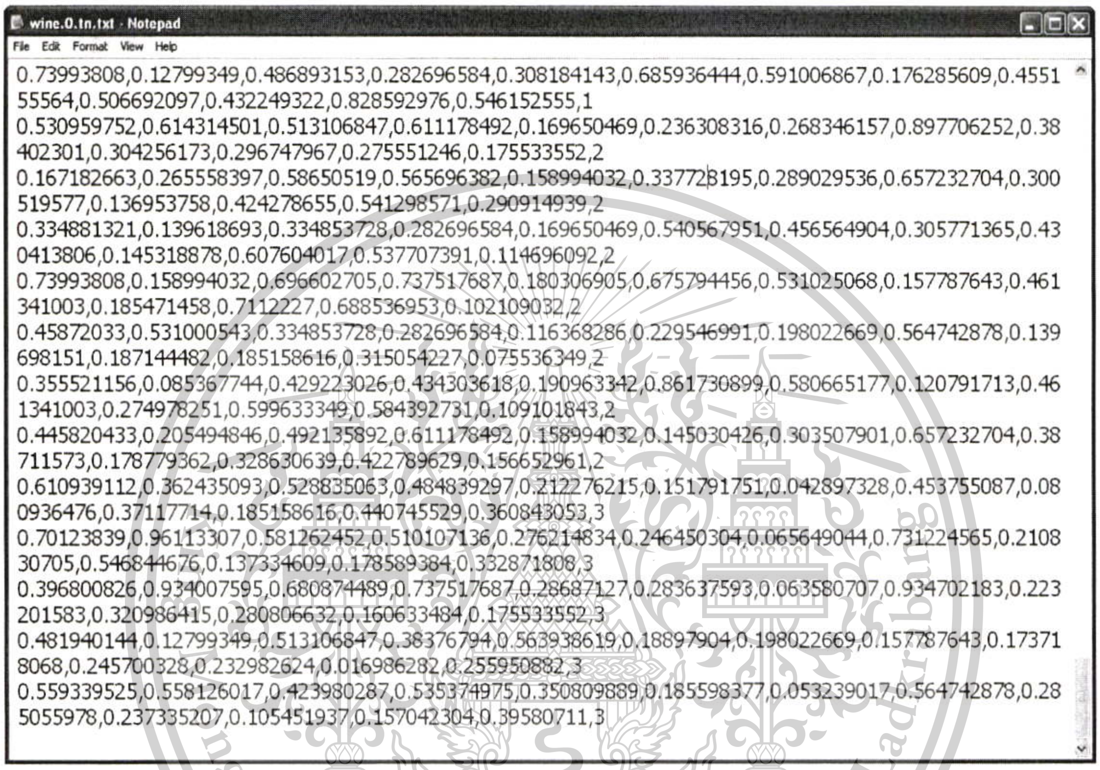
File Edit Format View Help

```
0.20043573,0.418300654,0.109504819,0.050653595,1
0.173202614,0.459150327,0.092888003,0.009803922,1
0.20043573,0.62254902,0.109504819,0.214052288,1
0.254901961,0.867647059,0.092888003,0.009803922,1
0.037037037,0.5,0.05965437,0.050653595,1
0.472766885,0.581699346,0.591392489,0.62254902,2
0.418300654,0.254901961,0.508308408,0.459150327,2
0.5,0.336601307,0.508308408,0.5,2
0.391067538,0.418300654,0.541542041,0.459150327,2
0.445533769,0.5,0.641242938,0.704248366,2
0.990196078,0.745098039,0.907111997,0.785947712,3
0.363834423,0.336601307,0.657859754,0.785947712,3
0.608932462,0.418300654,0.757560651,0.704248366,3
0.826797386,0.37745098,0.890495181,0.704248366,3
0.445533769,0.418300654,0.691093387,0.704248366,3
0.254901961,0.62254902,0.092888003,0.050653595,1
0.145969499,0.581699346,0.109504819,0.050653595,1
0.145969499,0.581699346,0.159355268,0.050653595,1
0.20043573,0.62254902,0.05965437,0.091503268,1
0.309368192,0.704248366,0.092888003,0.050653595,1
0.418300654,0.295751634,0.491691592,0.459150327,2
0.636165577,0.37745098,0.608009305,0.5,2
0.527233115,0.37745098,0.558158857,0.5,2
0.554466231,0.540849673,0.624626122,0.62254902,2
0.608932462,0.336601307,0.608009305,0.581699346,2
0.663398693,0.214052288,0.8074111,0.704248366,3
0.391067538,0.214052288,0.67447657,0.785947712,3
```

รูปที่ 4.15 ข้อมูล Iris ที่ใช้สำหรับการ Train

ข้อมูล Wine

ข้อมูลของ Wine มีลักษณะข้อมูลทั้งหมด 13 คุณลักษณะที่แตกต่างกัน โดยแบ่งข้อมูลออกเป็น 3 กลุ่ม คือ กลุ่มที่ 1 จำนวน 59 ตัวอย่าง กลุ่มที่ 2 จำนวน 71 ตัวอย่าง กลุ่มที่ 3 จำนวน 48 ตัวอย่าง รวมทั้งหมด 178 ตัวอย่าง โดยจะแบ่งเป็นข้อมูลเทรน (Training Data) จำนวน 106 ตัวอย่าง และข้อมูลสำหรับทดสอบ (Testing Data) จำนวน 72 ตัวอย่าง


wine.0.tn.txt - Notepad

File Edit Format View Help

```
0.73993808,0.12799349,0.486893153,0.282696584,0.308184143,0.685936444,0.591006867,0.176285609,0.455155564,0.506692097,0.432249322,0.828592976,0.546152555,1
0.530959752,0.614314501,0.513106847,0.611178492,0.169650469,0.236308316,0.268346157,0.897706252,0.38402301,0.304256173,0.296747967,0.275551246,0.175533552,2
0.167182663,0.265558397,0.58650519,0.565696382,0.158994032,0.337728195,0.289029536,0.657232704,0.300519577,0.136953758,0.424278655,0.541298571,0.290914939,2
0.334881321,0.139618693,0.334853728,0.282696584,0.169650469,0.540567951,0.456564904,0.305771365,0.430413806,0.145318878,0.607604017,0.537707391,0.114696092,2
0.73993808,0.158994032,0.696602705,0.737517687,0.180306905,0.675794456,0.531025068,0.157787643,0.461341003,0.185471458,0.7112227,0.688536953,0.102109032,2
0.45872033,0.531000543,0.334853728,0.282696584,0.116368286,0.229546991,0.198022669,0.564742878,0.139698151,0.187144482,0.185158616,0.315054227,0.075536349,2
0.355521156,0.085367744,0.429223026,0.434303618,0.190963342,0.861730899,0.580665177,0.120791713,0.461341003,0.274978251,0.599633349,0.584392731,0.109101843,2
0.445820433,0.205494846,0.492135892,0.611178492,0.158994032,0.145030426,0.303507901,0.657232704,0.38711573,0.178779362,0.328630639,0.422789629,0.156652961,2
0.610939112,0.362435093,0.528835063,0.484839297,0.212276215,0.151791751,0.042897328,0.453755087,0.080936476,0.37117714,0.185158616,0.440745529,0.360843053,3
0.70123839,0.96113307,0.581262452,0.510107136,0.276214834,0.246450304,0.065649044,0.731224565,0.210830705,0.546844676,0.137334609,0.178589384,0.332871808,3
0.396800826,0.934007595,0.680874489,0.737517687,0.28687127,0.283637593,0.063580707,0.934702183,0.223201583,0.320986415,0.280806632,0.160633484,0.175533552,3
0.481940144,0.12799349,0.513106847,0.38376794,0.563938619,0.18897904,0.198022669,0.157787643,0.173718068,0.245700328,0.232982624,0.016986282,0.255950882,3
0.559339525,0.558126017,0.423980287,0.535374975,0.350809889,0.185598377,0.053239017,0.564742878,0.285055978,0.237335207,0.105451937,0.157042304,0.39580711,3
```

รูปที่ 4.16 ข้อมูล Wine ที่ใช้สำหรับการ Train

ข้อมูล Image

ข้อมูลของ Image มีลักษณะข้อมูลทั้งหมด 19 คุณลักษณะที่แตกต่างกัน โดยแบ่งข้อมูลออกเป็น 7 กลุ่ม คือ กลุ่มที่ 1 จำนวน 330 ตัวอย่าง กลุ่มที่ 2 จำนวน 330 ตัวอย่าง กลุ่มที่ 3 จำนวน 330 ตัวอย่าง กลุ่มที่ 4 จำนวน 330 ตัวอย่าง กลุ่มที่ 5 จำนวน 330 ตัวอย่าง กลุ่มที่ 6 จำนวน 330 ตัวอย่าง กลุ่มที่ 7 จำนวน 330 ตัวอย่าง รวมทั้งหมด 2,310 ตัวอย่าง โดยจะแบ่งเป็นข้อมูลเทรน (Training Data) จำนวน 210 ตัวอย่าง และข้อมูลสำหรับทดสอบ (Testing Data) จำนวน 2,200 ตัวอย่าง


wine.0.tn.txt - Notepad

File Edit Format View Help

0.73993808,0.12799349,0.486893153,0.282696584,0.308184143,0.685936444,0.591006867,0.176285609,0.455155564,0.506692097,0.432249322,0.828592976,0.546152555,1
0.530959752,0.614314501,0.513106847,0.611178492,0.169650469,0.236308316,0.268346157,0.897706252,0.38402301,0.304256173,0.296747967,0.275551246,0.175533552,2
0.167182663,0.265558397,0.58650519,0.565696382,0.158994032,0.337728195,0.289029536,0.657232704,0.300519577,0.136953758,0.424278655,0.541298571,0.290914939,2
0.334881321,0.139618693,0.334853728,0.282696584,0.169650469,0.540567951,0.456564904,0.305771365,0.430413806,0.145318878,0.607604017,0.537707391,0.114696092,2
0.73993808,0.158994032,0.696602705,0.737517687,0.180306905,0.675794456,0.531025068,0.157787643,0.461341003,0.185471458,0.7112227,0.688536953,0.102109032,2
0.45872033,0.531000543,0.334853728,0.282696584,0.116368286,0.229546991,0.198022669,0.564742878,0.139698151,0.187144482,0.185158616,0.315054227,0.075536349,2
0.355521156,0.085367744,0.429223026,0.434303618,0.190963342,0.861730899,0.580665177,0.120791713,0.461341003,0.274978251,0.599633349,0.584392731,0.109101843,2
0.445820433,0.205494846,0.492135892,0.611178492,0.158994032,0.145030426,0.303507901,0.657232704,0.38711573,0.178779362,0.328630639,0.422789629,0.156652961,2
0.610939112,0.362435093,0.528835063,0.484839297,0.212276215,0.151791751,0.042897328,0.453755087,0.080936476,0.37117714,0.185158616,0.440745529,0.360843053,3
0.70123839,0.96113307,0.581262452,0.510107136,0.276214834,0.246450304,0.065649044,0.731224565,0.210830705,0.546844676,0.137334609,0.178589384,0.332871808,3
0.396800826,0.934007595,0.680874489,0.737517687,0.28687127,0.283637593,0.063580707,0.934702183,0.223201583,0.320986415,0.280806632,0.160633484,0.175533552,3
0.481940144,0.12799349,0.513106847,0.38376794,0.563938619,0.18897904,0.198022669,0.157787643,0.173718068,0.245700328,0.232982624,0.016986282,0.255950882,3
0.559339525,0.558126017,0.423980287,0.535374975,0.350809889,0.185598377,0.053239017,0.564742878,0.285055978,0.237335207,0.105451937,0.157042304,0.39580711,3

รูปที่ 4.17 ข้อมูล Image ที่ใช้สำหรับการ Train

4.3 ผลการทดลองและวิเคราะห์

4.3.1 ข้อมูลดอกไอริส (Iris)

ทำการเปรียบเทียบความถูกต้องของข้อมูลดอกไอริสโดยมีจำนวนข้อมูลทั้งหมด 150 ตัวอย่าง มีทั้งหมด 3 ชนิด โดยแบ่งเป็นชนิดละ 50 ตัวอย่าง

โดยกำหนด

- จำนวนข้อมูลเทรน (Training Data) 90 ตัวอย่าง
- จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 60 ตัวอย่าง
- Hidden Layar Node = 6
- Learning Rate = 1.5

ผลที่ได้จากการจำแนกข้อมูล Iris คือ

ตารางที่ 4.1 แสดงผลลัพธ์การทดลอง Iris ด้วยข้อมูลเทรน (Training Data) 90 ตัวอย่าง

อัลกอริทึม	Backpropagation (10 รอบ)	Backpropagation (100 รอบ)	Backpropagation (1,000 รอบ)	Backpropagation (5,000 รอบ)	ID3
จำนวนตัวอย่างในการทดสอบ	60	60	60	60	60
จำแนกถูก	57	60	60	60	60
จำแนกผิด	3	0	0	0	0
เปอร์เซ็นต์ความถูกต้อง	94%	100%	100%	100%	100%

4.3.2 ข้อมูล Wine

ทำการเปรียบเทียบความถูกต้องของข้อมูล Wine โดยมีลักษณะข้อมูลทั้งหมด 13 คุณลักษณะที่แตกต่างกัน โดยแบ่งข้อมูลออกเป็น 3 กลุ่ม คือ กลุ่มที่ 1 จำนวน 59 ตัวอย่าง กลุ่มที่ 2 จำนวน 71 ตัวอย่าง กลุ่มที่ 3 จำนวน 48 ตัวอย่าง จำนวนข้อมูลรวมทั้งหมด 178 ตัวอย่าง

โดยกำหนด

- จำนวนข้อมูลเทรน (Training Data) 106 ตัวอย่าง
- จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 72 ตัวอย่าง
- Hidden Layar Node = 5
- Learning Rate = 1.5

ผลที่ได้จากการจำแนกข้อมูล Wine คือ

ตารางที่ 4.2 แสดงผลลัพธ์การทดลอง Wine ด้วยข้อมูลเทรน (Training Data) 106 ตัวอย่าง

อัลกอริทึม	Backpropagation (10 รอบ)	Backpropagation (100 รอบ)	Backpropagation (1,000 รอบ)	Backpropagation (5,000 รอบ)	ID3
จำนวนตัวอย่างในการทดสอบ	72	72	72	72	72
จำแนกถูก	44	54	69	69	65
จำแนกผิด	28	18	3	3	7
เปอร์เซ็นต์ความถูกต้อง	61.11%	75%	95.83%	95.83%	90.28%

4.3.3 ข้อมูล Image

ทำการเปรียบเทียบความถูกต้องของข้อมูล Image โดยมีลักษณะข้อมูลทั้งหมด 19 คุณลักษณะที่แตกต่างกัน โดยแบ่งข้อมูลออกเป็น 7 กลุ่ม คือ กลุ่มที่ 1 จำนวน 330 ตัวอย่าง กลุ่มที่ 2 จำนวน 330 ตัวอย่าง กลุ่มที่ 3 จำนวน 330 ตัวอย่าง กลุ่มที่ 4 จำนวน 330 ตัวอย่าง กลุ่มที่ 5 จำนวน 330 ตัวอย่าง กลุ่มที่ 6 จำนวน 330 ตัวอย่าง กลุ่มที่ 7 จำนวน 330 ตัวอย่าง จำนวนข้อมูลรวมทั้งหมด 2,310 ตัวอย่าง

โดยกำหนด

- จำนวนข้อมูลเทรน (Training Data) 210 ตัวอย่าง
- จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 2,200 ตัวอย่าง
- Hidden Layar Node = 9
- Learning Rate = 1.5

ผลที่ได้จากการจำแนกข้อมูล Image คือ

ตารางที่ 4.3 แสดงผลลัพธ์การทดลอง Image ด้วยข้อมูลเทรน (Training Data) 210 ตัวอย่าง

อัลกอริทึม	Backpropagation (10 รอบ)	Backpropagation (100 รอบ)	Backpropagation (1,000 รอบ)	Backpropagation (5,000 รอบ)	ID3
จำนวนตัวอย่างในการทดสอบ	2,200	2,200	2,200	2,200	2,200
จำแนกถูก	213	1,231	1,617	1,716	300
จำแนกผิด	1,987	969	583	484	1,900
เปอร์เซ็นต์ความถูกต้อง	9.68%	55.95%	73.50%	78%	13.64%

4.3.5 วิเคราะห์ผลการทดลอง

จากผลการทดลองการจำแนกประเภทของข้อมูล จากตัวอย่างข้อมูล 4 ชุด คือ ข้อมูลดอกไอริส, ข้อมูล Wine และ Image ด้วยเทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และเทคนิควิธีโครงข่ายประสาทเทียม ด้วยอัลกอริธึม Backpropagation สามารถสรุปและวิคราะห์ผลการทดลองได้ ดังนี้

ข้อมูลดอกไอริส จะเห็นว่าการทดลองการจำแนกประเภทของข้อมูลดอกไอริสนั้นได้ใช้จำนวนข้อมูลเทรน (Training Data) 90 ตัวอย่าง จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 60 ตัวอย่าง จะเห็นว่าผลลัพธ์จากการทดลองจำแนกประเภทของข้อมูลด้วยอัลกอริธึม ID3 นั้นได้ค่าเปอร์เซ็นต์ความถูกต้องแม่นยำ 100% แต่การจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation นั้นค่าเปอร์เซ็นต์ความถูกต้องแม่นยำของผลลัพธ์ที่ได้จากการเทรนด้วยจำนวนรอบที่สูงกว่าจะให้ค่าความถูกต้องที่มากกว่า

ข้อมูล Wine จะเห็นว่าการทดลองการจำแนกประเภทของข้อมูล Wine นั้นได้ใช้ จำนวนข้อมูลเทรน (Training Data) 106 ตัวอย่าง จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 72 ตัวอย่าง จะเห็นว่าผลลัพธ์จากการทดลองจำแนกประเภทของข้อมูลด้วยอัลกอริธึม ID3 นั้นได้ค่าเปอร์เซ็นต์ความถูกต้องแม่นยำน้อยกว่าการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation ซึ่งการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation นั้นค่าเปอร์เซ็นต์ความถูกต้องแม่นยำของผลลัพธ์ที่ได้จากการเทรนด้วยจำนวนรอบที่สูงกว่าจะให้ค่าความถูกต้องที่มากกว่า

ข้อมูล Image จะเห็นว่าการทดลองการจำแนกประเภทของข้อมูล Image นั้นได้ใช้ จำนวนข้อมูลเทรน (Training Data) 210 ตัวอย่าง จำนวนข้อมูลสำหรับทดสอบ (Testing Data) 2,200 ตัวอย่าง จะเห็นว่าผลลัพธ์จากการทดลองจำแนกประเภทของข้อมูลด้วยอัลกอริธึม ID3 นั้นได้ค่าเปอร์เซ็นต์ความถูกต้องแม่นยำน้อยมากเมื่อเทียบกับการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation ซึ่งเกิดจากจำนวนของข้อมูลเทรน (Training Data) ที่นำมาใช้เทรนนั้นมีจำนวนน้อย และมีจำนวนของกลุ่มของข้อมูลเป็นจำนวนมาก แต่เมื่อนำมาข้อมูลมาทำการจำแนกประเภทของข้อมูลด้วยอัลกอริธึม Backpropagation นั้นค่าเปอร์เซ็นต์ความถูกต้องแม่นยำของผลลัพธ์ที่ได้จากการเทรนนั้นสูงกว่าเนื่องจากวิธีนี้สามารถกำหนดจำนวนรอบของการเทรนได้ ดังนั้นเมื่อมีจำนวนรอบในการเทรนที่สูงกว่าทำให้ค่าความถูกต้องที่ได้มากกว่า

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

จากการศึกษาการจำแนกประเภทของข้อมูลโดยใช้เทคนิควิธี Decision Tree ด้วยอัลกอริธึม ID3 และโครงข่ายประสาทเทียม (Neural Network) ด้วยอัลกอริธึม Backpropagation ในโครงการนี้ สามารถสรุปผลการดำเนินงานและสรุปผลการทดลองรวมถึงข้อเสนอแนะดังต่อไปนี้

5.1 สรุปผลการศึกษา

จากการศึกษาในโครงการนี้สามารถสรุปผลดำเนินงาน ได้ดังนี้

ในการทำ Data Mining ซึ่งเป็นกระบวนการที่ใช้วิเคราะห์และค้นหาข้อมูลที่ซ่อนเร้นอยู่ภายในฐานข้อมูล ซึ่งข้อมูลที่ได้จากการทำ Data Mining นั้นมีประโยชน์อย่างมากในการนำไปใช้ ซึ่งในการเตรียมข้อมูลในการทำการจำแนกประเภทของข้อมูลนั้น ข้อมูลที่นำไปฝึกสอนโครงข่ายนั้นมีความสำคัญมากเพราะว่าถ้าข้อมูลที่นำไปฝึกสอนโครงข่ายมีลักษณะครบถ้วนในทุกลักษณะของข้อมูลจะทำให้ความสามารถในการจำแนกประเภทของข้อมูลของโครงข่ายมีประสิทธิภาพดีขึ้น

ในการเรียนรู้แต่ละครั้งของอัลกอริธึมแบบ Backpropagation ผลจากการทดลองจะให้ค่าที่ต่างกัน เนื่องจากผลลัพธ์จากการจำแนกประเภทนั้นจะขึ้นอยู่กับการกำหนดค่าโครงสร้างของโครงข่าย ต่าง ๆ ที่ใช้ในการ Train เช่น ค่าอัตราการเรียนรู้ และค่าจำนวนรอบในการ Train เป็นต้น ดังนั้นการเรียนรู้ต้องอาศัยการเรียนรู้หลาย ๆ ครั้ง เพื่อสังเกตว่าการกำหนดค่าต่าง ๆ นั้น ควรจะกำหนดค่าเหล่านั้นอย่างไร เพื่อให้ได้ประสิทธิภาพในการจำแนกประเภทของข้อมูลได้ถูกต้องมากที่สุด

ในการจำแนกประเภทของข้อมูลด้วยอัลกอริธึมแบบ ID3 นั้น จำนวนข้อมูลฝึกฝน (Training Data) และจำนวนกลุ่มของข้อมูล ที่นำมาใช้เทรนนั้นมีผลต่อประสิทธิภาพในการแบ่งกลุ่มของข้อมูล

5.2 ประโยชน์ที่ได้รับจากการศึกษาและพัฒนาระบบ

1. ทำให้เข้าใจทฤษฎี และหลักการทำงานของอัลกอริธึมต่างๆในการทำ Classification ข้อมูลมากขึ้น
2. ทำให้ทราบว่าข้อมูลชุดนั้นเหมาะสมกับอัลกอริธึมแบบใด ทำให้เราสามารถเลือกใช้อัลกอริธึมได้ตามวัตถุประสงค์
3. ทำให้ทราบถึงปัจจัยต่างๆ ที่มีผลต่อการแบ่งกลุ่มของข้อมูล ซึ่งทำให้เราสามารถนำความรู้ที่ได้รับไปปรับปรุงและใช้ในอนาคตได้

5.3 ข้อเสนอแนะ

โครงงานนี้สามารถนำไปปรับให้สามารถทำงานได้ดียิ่งขึ้น เพื่อเพิ่มความเร็วให้กับการเรียนรู้แบบแบคพรอพพาเกชันเน็ตเวิร์ค (Back-propagation) ควรค้นหาวิธีที่เหมาะสมเพื่อเพิ่มประสิทธิภาพในการสุ่มค่าน้ำหนัก หรือการปรับค่าน้ำหนักให้มีประสิทธิภาพมากขึ้น เนื่องจากค่าน้ำหนักเริ่มต้นมีผลต่อการทำงานของโปรแกรม นอกจากนี้ยังสามารถนำเอาเทคนิควิธีอื่นๆ หรืออัลกอริธึมอื่นๆ ที่ใช้ในการแบ่งกลุ่มของข้อมูลมาพัฒนาเพื่อเพิ่มประสิทธิภาพของโปรแกรมประยุกต์ให้มากขึ้น

บรรณานุกรม

dol. 2007 **DataMining Techniques For Classification Problems.** [online] Available: http://www.dol.go.th/chaiwut/ism7/downloadism/decision_tree.pdf

gotoknow. 2007 **Customer Segmentation and Data Mining.** [online] Available: http://gotoknow.org/blog/business-intelligence/56616

Han, J. and Kamber, M. 2001. **Data Mining : Concepts and Techniques.** San Francisco: Morgan Kanfmann

Larose, D. T. 2005. **Discovering Knowledge in Data : An Introduction to Data Mining.** New Jersey: John Wiley and Sons

Markov, Z. and Larose, D. T. 2007. **Data Mining the Web: uncovering patterns in Web content, structure, and usage.** New Jersey: John Wiley and Sons

Nectec. 2008 **NaiveBayesModel.** [online] Available: http://wiki.nectec.or.th/ntl/Project/NaiveBayesModel

Swu. 2008 **Engineering Statistics (Probability).** [online] Available: http://mylesson.swu.ac.th/ine201/detail/2prob_p6.htm

Thaiall. 2010 **Flowchart.** [online]. Available: http://www.thaiall.com/flowchart/

Wikipedia. 2007 **Data Mining.** [online] Available: http://th.wikipedia.org/wiki/Data_mining

Wikipedia. 2008 **Bayes' Theorem.** [online] Available: http://en.wikipedia.org/wiki/Bayes'_theorem

ประวัติผู้เขียน

ชื่อ-นามสกุล	ณัฐวรรณ รุจิราณิชาภัทร
วัน เดือน ปีเกิด	13 มกราคม 2527
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
สถานที่สำเร็จการศึกษา	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีที่สำเร็จการศึกษา	2549