การพัฒนาเครื่องมือสำหรับการจัดหมวดหมู่โดยใช้ดิซิชันทรี

DEVELOPING TOOL FOR CLASSIFICATION USING DECISION TREE

โดย

กรวิภา เกตุเรืองโรจน์

KONVIPA KETRUENGROJ

อาจารย์ที่ปรึกษา

ผศ.ดร.พรฤดี เนติโสภากุล

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2549

DEVELOPING TOOL FOR CLASSIFICATION

USING DECISION TREE

KONVIPA KETRUENGROJ

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2006

หัวข้อ	การพัฒนาเครื่องมือสำหรับการจัดหมวดหมู่โดยใช้ดิซิชันทรี
นักศึกษา	นางสาวกรวิภา เกตุเรืองโรจน์
รหัสนักศึกษา	47066129
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	ผศ.ดร.พรฤดี เนติโสภากุล

บทคัดย่อ

โครงการพัฒนาระบบงานนี้เกี่ยวกับการจัดหมวดหมู่ของข้อมูลโดยใช้เทคนิคดิซิชันทรี ประกอบกับใช้อัลกอริธึม ไอดีสาม (ID3) ในการสร้างต้นไม้ประกอบการตัดสินใจ ซึ่งโครงการนี้สามารถวิเคราะห์ฐานข้อมูลและแบ่งกลุ่มข้อมูลออกตามคลาสที่กำหนดไว้ เพื่อยังประโยชน์ในการทำนายความเป็นไปได้หรือหาแนวโน้มของข้อมูล ซึ่งเทคนิคในการจัดหมวดหมู่สามารถแสดงให้เห็นได้ด้วยต้นไม้ประกอบการตัดสินใจและสามารถเลือกตัวแบ่งกลุ่มข้อมูลที่ดีที่สุดได้ รวมทั้งบอกได้ว่าปัจจัยใดมีอิทธิพลต่อการทำนายมากที่สุด ผลการทดสอบของโครงการยังสามารถบอกค่าความน่าเชื่อถือของแต่ละกิ่งตัดสินใจเป็นเปอร์เซ็นต์ เพื่อให้ผู้ใช้สามารถตรวจสอบผลที่ได้จากการใช้งานและนำไปประกอบการตัดสินใจได้ต่อไป

Title	Developing Tool for Classification Using Decision Tree
Student	Ms. Konvipa Ketruengroj
Student ID.	47066129
Degree	Master of Science
Programme	Information Technology Management
Academic Year	2006
Advisor	Asst. Prof. Dr. Ponrudee Netisopakul

ABSTRACT

This system development project is about classification data by using Decision Tree technique together with ID3 algorithm to create the decision tree model. This project will be able to analyze and classify each database into specific class in order to predict or search for any possible trends. The Classification technique willing to select the best data classify and willing to tell the most important factor for the predicted result. The result of each decision tree is including with reliability values (as %) of each node in the decision tree model. This reliability values will help user check all results and bring it forward to make any decision for further more.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้เป็นอย่างดี ด้วยคำแนะนำ และคำปรึกษาจาก ผศ.ดร.พรฤดี เนติโสภากุล ซึ่งเป็นอาจารย์ผู้ควบคุมการพัฒนาระบบงานของข้าพเจ้า และสละเวลาช่วยข้าพเจ้าตรวจทานแก้ไข ข้อบกพร่องในการทำงานของข้าพเจ้ามาโดยตลอด และขอขอบคุณ รศ.ดร.วรพจน์ กรีสุระเดช ที่ให้คำแนะนำและอนุญาตให้ข้าพเจ้ารับฟังการบรรยายวิชา Data Mining ที่เพิ่มพูนความเข้าใจแก่ข้าพเจ้าเป็นอย่างดี ข้าพเจ้ารู้สึกซาบซึ้งในความอนุเคราะห์จากท่านอาจารย์ทั้งสองท่าน และขอขอบพระคุณอาจารย์เป็นอย่างสูง

ขอกราบพระคุณคณาจารย์ภาควิชาวิทยาการสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุก ๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้าตลอดมา ให้ข้าพเจ้าได้รู้แจ้งในสิ่งที่ข้าพเจ้ายังไม่เคยรู้มาก่อน ขอขอบพระคุณค่ะ

ขอขอบคุณบัณฑิตศึกษาและบัณฑิตวิทยาลัยรวมไปถึงพี่ ๆ ทุกคนที่ห้องบริการนักศึกษา ที่เอื้ออำนวยความสะดวกทุกอย่างแก่พวกนักศึกษาตาดำ ๆ ด้วยรอยยิ้มเสมอมา คอยให้ความช่วยเหลือและ แนะนำในเรื่องต่าง ๆ อย่างอดทน น่ารักจริง ๆ ค่ะ ขอบคุณมาก ๆ เลยนะคะ

ขอขอบคุณ สองแม่ครัวที่ยืนยงเคียงข้างกันมาตลอด, ขอบคุณโป้งกับนุชที่ยื่นมือมาช่วยพี่ในยามท้อแท้จริง ๆ ขอบคุณมากจ้า, ขอบคุณนิวมาก ๆ จ้ะที่ให้เรามากกว่าคำว่าช่วยเหลือและทุกคนที่บ้านนิวสำหรับความน่ารักเสมอมา, ขอบคุณบีมที่ไม่เคยหันหลังให้เพื่อนคนนี้เลยจริง ๆ, ขอบคุณพี่เถี่ยวที่แวะเวียนมาให้ความช่วยเหลือเสมอมานะคะ, ขอบคุณ เอิง ที่สละเวลาวันหยุดมาดูงานให้พี่นะ, ขอบคุณจ๊อบที่ให้ไอเดียและช่วยเหลือเราด้วยความเต็มใจ, ขอบคุณ พี่อั๋น, พี่เสือ, พี่โก้, พี่เน็ก, แอ้, หลิว, แหม่ม, ตุลย์, วุฒิ, น้องบีม, น้องอร ที่ไม่เคยเหนื่อยกับการหยิบยื่นกำลังใจและคำว่าสู้ ๆ, ขอบคุณเพื่อน ๆ พี่ๆ น้องๆ ทุกคนในรุ่น 17.1 ที่เป็นกำลังใจ และให้คำแนะนำต่างๆ ตลอดมา ขอบคุณนะคุงที่ให้เข้าใจ, ให้โอกาสและกำลังใจแถมเป็นห่วงกันเสมอจริง ๆ

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา และทุกคนในครอบครัวของข้าพเจ้าที่เป็นกำลังใจ และให้การสนับสนุนในทุกเรื่องๆ ขอบคุณพี่อัฐที่มอบกำลังใจดี ๆ ผนวกความรักและความคิดถึงส่งให้กันเสมอมา ทำให้กำลังกายและใจจากศูนย์ขึ้นเป็นร้อยอย่างอัศจรรย์ ขอบคุณพี่แอ้มที่อยู่เคียงข้างข้าพเจ้าทำให้อุ่นใจทุกครั้งและทำให้รู้ว่า Someone watching over me จริง ๆ ขอขอบคุณทุก ๆ คนที่ทำให้ข้าพเจ้าสามารถทำวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์อันพึงมาจากวิทยานิพนธ์ฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน หากผิดพลาดประการใด ขออภัยไว้ ณ ที่นี้ด้วยค่ะ

กรวิภา เกตุเรืองโรจน์

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....I

บทคัดย่อภาษาอังกฤษ.....II

กิตติกรรมประกาศ.....III

สารบัญ.....IV

สารบัญรูป.....VII

บทที่ 1 บทนำ.....1

1.1 ความเป็นมาและความสำคัญของปัญหา.....1

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา.....2

1.3 ขอบเขตของโครงการ.....2

1.4 ขั้นตอนการดำเนินงาน.....3

1.5 ข้อจำกัดของการศึกษา.....3

1.6 ประโยชน์ที่คาดว่าน่าจะได้รับ.....4

บทที่ 2 แนวคิดและทฤษฎีที่เกี่ยวข้อง.....5

2.1 ความเป็นมาของดาต้าไมนิ่ง.....4

2.2 ดาต้าไมนิ่ง.....5

2.2.1 Predictive Modeling.....7

2.2.2 Forensic Analysis.....7

2.2.3 Discovery.....7

2.3 การทำงานของดาต้าไมนิ่ง.....7

2.3.1 กระบวนการจัดเตรียมข้อมูล.....7

2.3.2 กระบวนการวิเคราะห์และจัดหมวดหมู่.....7

2.3.3 กระบวนการเรียนรู้ความรู้.....8

2.3.4 กระบวนการคาดคะเน.....8

2.4 รูปแบบการเก็บข้อมูลที่สามารถทำดาต้าไมนิ่ง.....10

2.5 รูปแบบข้อมูลตัวแปรในการทำดาต้าไมนิ่ง (Type of Attributes).....11

2.5.1 ตัวแปรแบบ Categorical.....11

2.5.2 ตัวแปรแบบ Quantitative.....11

สารบัญ (ต่อ)

หน้า

2.6 ขั้นตอนในการทำดาต้าไมนิ่ง....11

2.7 ประเภทงานทางด้านดาต้าไมนิ่ง....12

2.7.1 Descriptive mining....12

2.7.2 Predictive mining....12

2.8 ประเภทของแบบจำลองสำหรับการทำดาต้าไมนิ่ง....12

2.8.1 Classification Model....12

2.8.2 Clustering Model....12

2.8.3 Association Model....12

2.8.4 Deviation Detection....13

2.8.5 Sequential Analysis....13

บทที่ 3 การจัดกลุ่มข้อมูลแบบ Classification ด้วย Decision Tree....14

3.1 Classification Model....14

3.1.1 Tree induction (Decision Tree)....15

3.1.2 Mathematical....16

3.1.3 Neural induction....17

3.2 ต้นไม้การตัดสินใจ (Decision Tree)....17

3.2.1 เหตุผลที่ Decision นั้นเป็นที่นิยมในการนำไปใช้....18

3.2.2 ข้อเสียของดิซิชันทรี....19

3.2.3 วิธีการต้นไม้การตัดสินใจ....19

3.3 เอนโทรปี....20

3.4 การวัดค่าอินฟอเมชันเกน....21

3.5 อัลกอริธึม ID3....21

3.5.1 ID3....22

3.6 การหลีกเลี่ยงการเกิดโอเวอร์ฟิต....25

3.6.1 พรี-พรุนทรี....25

3.6.2 โพส-พรุนทรี....25

สารบัญ (ต่อ)

หน้า

บทที่ 4 การวิเคราะห์ระบบงาน.....26

4.1 องค์ประกอบของระบบงาน.....26

4.1.1 Data Selection.....26

4.1.2 Data Preparation.....26

4.1.3 Data Mining.....26

4.2 Flow การทำงานของระบบ.....26

4.3 การออกแบบส่วนต่อประสานระบบกับผู้ใช้.....32

บทที่ 5 การประยุกต์การใช้งานของระบบ.....39

5.1 การประยุกต์ใช้งานของระบบ.....39

5.1.1 ต้อนรับเข้าสู่การใช้งานโปรแกรม.....39

5.1.2 การล็อกอินเข้าสู่ระบบเพื่อเชื่อมต่อกับเซิฟเวอร์.....40

5.1.3 การเลือกตารางจากฐานข้อมูล.....41

5.1.4 การเลือกใช้แอททริบิวที่ใช้ในไมนิ่ง.....42

5.1.5 แสดงข้อมูลที่ทำการเลือกมา.....43

5.1.6 การเลือกแอททริบิวเป้าหมาย.....44

5.1.7 การแสดงต้นไม้ประกอบการตัดสินใจ.....45

บทที่ 6 สรุปผลการวิจัยและข้อเสนอแนะ.....46

6.1 สรุปผลการวิจัย.....46

6.2 ข้อเสนอแนะ.....46

บรรณานุกรม.....47

ประวัติผู้เขียน.....48

สารบัญรูป

รูปที่ หน้า

2.1 เป้าหมายการทำดาต้าไมนิ่งในรูปแบบต่าง ๆ.....6

2.2 แสดงเฟสการทำงานของดาต้าไมนิ่ง.....9

2.3 แสดงการนำความรู้ออกจากข้อมูล.....10

3.1 แสดงการจัดกลุ่มของวัตถุ.....14

3.2 แสดง Tree Representation.....15

3.3 แสดง Binary Decision Tree.....16

3.4 แสดงตัวอย่าง Neural Network.....17

3.5 ภาพการทำงานของต้นไม้การตัดสินใจ.....18

3.6 ความสัมพันธ์ของการสุ่มข้อมูลและค่าเอนโทรปี.....20

3.7 แสดงถึงแนวคิดอัลกอริธึม ID3.....22

4.1 แสดงภาพรวมการทำงานของระบบทั้งหมด.....27

4.2 แสดงขั้นตอนการคลีนข้อมูล.....29

4.3 แสดงขั้นตอนการสร้างต้นไม้การตัดสินใจด้วยอัลกอริธึม ID3.....30

4.4 แสดงขั้นตอนการสร้างต้นไม้การตัดสินใจด้วยอัลกอริธึม ID3 (ต่อ).....31

4.5 แสดงหน้าจอต้อนรับเข้าสู่การใช้งานระบบ.....32

4.6 แสดงหน้าจอล็อกอินเชื่อมต่อดาต้าเบสเซิฟเวอร์.....33

4.7 แสดงหน้าจอเลือกตารางฐานข้อมูล.....34

4.8 แสดงหน้าจอการเลือกแอททริบิวที่สำหรับการทำไมนิ่ง.....35

4.9 หน้าจอแสดงข้อมูลทั้งหมดที่เลือกมาก่อนเข้าไมนิ่งเพื่อยืนยัน.....36

4.10 แสดงหน้าจอการเลือกแอททริบิวเป้าหมายเพื่อสร้างต้นไม้.....37

4.11 แสดงหน้าจอสร้างต้นไม้การตัดสินใจพร้อมแสดงค่าความถูกต้องของแต่ละกิ่งที่แตก..... 38

5.1 หน้าจอต้อนรับเข้าสู่โปรแกรมการทำงาน.....39

5.2 หน้าจอล็อกอินเข้าใช้ฐานข้อมูล.....40

5.3 หน้าจอให้ผู้ใช้เลือกตารางฐานข้อมูล..... 41

5.4 หน้าจอเลือกแอททริบิวที่ใช้ในไมนิ่ง.....42

5.5 หน้าจอแสดงข้อมูลที่ทำการเลือกมา.....43

5.6 หน้าจอแสดงข้อมูลที่ทำการเลือกมา.....44

5.7 หน้าจอแสดงต้นไม้การตัดสินใจและค่าความถูกต้อง..... 45

บทที่1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันดาต้าไมนิ่ง (Data Mining) เป็นเทคโนโลยีที่กำลังได้รับความสนใจและถูกประยุกต์ใช้ประโยชน์ในหลาย ๆ ธุรกิจ เนื่องจากสามารถค้นหาองค์ความรู้ (Knowledge) ที่ซ่อนอยู่ในข้อมูลที่รวบรวมไว้จำนวนมากมาประกอบการดำเนินการทางธุรกิจ ทั้งนี้เนื่องจากเทคโนโลยีดาต้าไมนิ่งเป็นเทคโนโลยีที่ไม่เพียงแต่ประมวลผลข้อมูลในอดีตจนถึงปัจจุบันเท่านั้น แต่ยังสามารถพยากรณ์ (Prediction) ไปสู่ข้อมูลในอนาคต โดยการสร้างแบบจำลอง (Model) ข้อมูลขนาดใหญ่ สำหรับค้นหารูปแบบ (Pattern) การเกิดข้อมูลและความสัมพันธ์ของข้อมูลเพื่อนำผลที่ได้จากการวิเคราะห์ข้อมูลไปช่วยสนับสนุนการตัดสินใจซึ่งในการดำเนินธุรกิจองค์กรที่สามารถคาดการณ์ล่วงหน้าเพื่อใช้ในการวางแผนงานต่าง ๆ ในอนาคต จะเป็นผู้ที่ได้เปรียบและสร้างโอกาสทางธุรกิจเหนือคู่แข่ง

เนื่องจากจำนวนของข้อมูลในปัจจุบันที่เก็บในระบบฐานข้อมูลมีขนาดใหญ่จนเกินความสามารถที่จะทำการวิเคราะห์ข้อมูลที่เก็บไว้โดยปราศจากเครื่องมือที่ใช้ในการวิเคราะห์โดยอัตโนมัติ ดังจะเห็นได้จากการที่ระบบฐานข้อมูลทางด้านวิทยาศาสตร์ และระบบทางธุรกิจ มีอัตราการเติบโตอย่างต่อเนื่อง ดังนั้น Knowledge Discovery in Database (KDD) จึงเกี่ยวกับการหาวิถีทางในการวิเคราะห์ข้อมูลขนาดใหญ่โดยอัตโนมัติ เพื่อหาหรือค้นพบข้อมูลใหม่ที่มี เนื่องจากข้อมูลที่มีอยู่เป็นจำนวนมากนั้น อาจประกอบด้วยข้อมูลที่ไม่มีประโยชน์แฝงอยู่ ซึ่งหากสามารถนำเอาเทคนิคและวิธีการในการที่จะวิเคราะห์แล้วดึงเอาแต่สิ่งที่ต้องการที่อาจซ่อนอยู่ในระบบฐานข้อมูลที่มีขนาดใหญ่ ๆ เพื่อนำผลลัพธ์ที่ได้ไปใช้ประโยชน์ เช่นนำไปใช้ในการประกอบการตัดสินใจทางด้านธุรกิจ (Decision Support System) หรือทำให้องค์กรทางธุรกิจเข้าใจพฤติกรรมของลูกค้า หรือผู้บริโภคได้ดีขึ้น ทำให้สามารถรักษาลูกค้าเก่าไว้ได้ และ อาจจะหาลูกค้าใหม่ได้มากขึ้น เมื่อระบบธุรกิจนั้นๆ มีการแข่งขันสูง เป็นต้น

ดังจะเห็นได้ว่าหากต้องการดึงเอาองค์ความรู้ที่มีอยู่ภายในฐานข้อมูลขนาดใหญ่ออกมานั้น มีความจำเป็นที่จะต้องทราบถึงกระบวนการทั้งหมดในการดึงเอาองค์ความรู้ออกมาด้วยเพื่อให้ทราบถึงกระบวนการต่าง ๆ ไปตามลำดับ โดยที่ดาต้าไมนิ่งเป็นขั้นตอนหนึ่งใน KDD Process ซึ่งใช้ในการหารูปแบบ (Pattern) และ แนวโน้ม (Trend) จากข้อมูล ซึ่งการทำงานของ KDD จะเป็นการนำเอาผลลัพธ์ที่ได้ออกมาจากการทำดาต้าไมนิ่งมาแปลงอย่างระมัดระวังและละเอียดลออ ให้เป็น

ข้อมูลที่มีประโยชน์ มีคุณค่า และสามารถเข้าใจได้ง่ายและมีประโยชน์อย่างมากเมื่อใช้ประกอบการตัดสินใจในเรื่องต่าง ๆ ต่อไป

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

โครงการพัฒนาระบบงานนี้มีวัตถุประสงค์ดังต่อไปนี้

1. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับแนวคิดและขั้นตอนการทำดาต้าไมนิ่ง
2. เพื่อให้สามารถประยุกต์ความรู้ที่ได้เรียนนำมาพัฒนา ศึกษาค้นคว้าและแก้ปัญหาในโครงการพัฒนาระบบงานได้
3. เพื่อศึกษาและทำความเข้าใจเกี่ยวกับการใช้แนวคิด Classification โดยนำหลักการดิซิชันทรีเข้ามาใช้
4. เพื่อศึกษาถึงแนวทางและความเป็นไปได้ในการนำแนวคิดดิซิชันทรีมาใช้ในการวิเคราะห์ข้อมูลว่าสามารถนำมาวิเคราะห์ข้อมูลได้จริงและมีประสิทธิภาพหรือไม่
5. เพื่อเป็นแนวทางในการประยุกต์ใช้ดาต้าไมนิ่งในการสนับสนุนการตัดสินใจ
6. เพื่อให้เป็นเครื่องมือในการวิเคราะห์ความรู้จากฐานข้อมูลขนาดใหญ่ในปัจจุบันแบบอัตโนมัติที่สามารถวิเคราะห์ได้เร็วและมีประสิทธิภาพมากยิ่งขึ้น

1.3 ขอบเขตของโครงการ

1. โครงการพัฒนาระบบงานนี้ เป็นการศึกษาหลักการในแนวทางของกระบวนการทำดาต้าไมนิ่งแบบ Classification แสดงผลออกมาในรูปแบบของต้นไม้การตัดสินใจ
2. ทำการจัดหมวดหมู่ของฐานข้อมูลโดยใช้ อัลกอริธึมID3
3. ข้อมูลที่สามารถทำการวิเคราะห์ในการจัดกลุ่มข้อมูลต้องเป็นฐานข้อมูล Microsoft SQL Server เท่านั้น
4. สามารถแสดงผลการจัดหมวดหมู่ออกเป็นต้นไม้การตัดสินใจทางทรีวิวพร้อมแสดงผลค่าความถูกต้องของแต่ละกิ่งของต้นไม้ได้
5. ข้อมูลที่รับเข้ามาต้องเป็นฐานข้อมูลที่พร้อมต่อการเข้าสู่กระบวนการดาต้าไมนิ่งเรียบร้อยแล้ว กล่าวคือข้อมูลในทุกคอลัมน์ต้องเป็นตัวอักษรแล้วเท่านั้น
6. ฐานข้อมูลที่เรียกใช้หากมีค่าของแถวที่เป็นค่าว่างจะทำการตัดแถวนั้นทิ้งทันที
7. การแตกกิ่งของต้นไม้การตัดสินใจจะแตกกิ่งจนกว่าจะเหลือค่าของแถวภายในคอลัมน์เดียวกันไม่ต่ำกว่า 3 แถวเท่านั้น จากนั้นจะทำการหยุดแตกกิ่งการตัดสินใจ (Pre-Prune Tree)

1.4 ขั้นตอนของการดำเนินงาน

การพัฒนาระบบงานในการศึกษานี้สามารถจำแนกขั้นตอนการทำงานออกเป็นข้อ ๆ ได้ดังนี้

1. ศึกษาหลักการและกระบวนการทำงานของดาต้าไมนิ่งและกระบวนการในการค้นพบความรู้ KDD Process
2. ศึกษาการนำดาต้าไมนิ่งมาประยุกต์ใช้ในการพัฒนาโครงการ โดยใช้เทคนิค Classification ด้วยอัลกอริธึม ID3
3. ศึกษาอัลกอริธึม ID3 เพื่อนำมาประยุกต์ใช้กับโครงการ
4. ศึกษาเครื่องมือต่าง ๆ ที่นำมาใช้ในการพัฒนาระบบงาน
5. รวบรวมและเตรียมข้อมูล รวมทั้งกำหนดเครื่องมือที่จะนำมาใช้ในการพัฒนาโครงการ
   5.1 การจัดเตรียมข้อมูล : จัดเตรียมข้อมูลให้เหมาะสมกับเป้าหมายการดำเนินงาน ทั้งปริมาณของข้อมูลก็ควรให้มีความพอเพียง
   5.2 เลือกเครื่องมือที่ใช้ในการพัฒนาโครงการ : ปัจจุบันการพัฒนาระบบงานหนึ่ง ๆ มีวิธีการและเครื่องมือมากมายให้เลือกใช้ตามแต่ความเหมาะสม สำหรับโครงการพัฒนาระบบงานนี้ได้เลือกใช้เครื่องมือดังต่อไปนี้
      - Application ที่ใช้ในการพัฒนาระบบคือ Microsoft Visual Studio.Net 2003
      - ระบบฐานข้อมูล Microsoft SQL Server
      - เครื่องคอมพิวเตอร์ที่นำมาใช้ในการพัฒนา คือ Microsoft Window XP Professional 2002, Pentium 3 Processor, CPU 666 Mhz, Ram 320 MB
6. ออกแบบส่วนติดต่อกับผู้ใช้ (User Interface) และออกแบบฐานข้อมูล
7. ออกแบบกระบวนการทำงานและพัฒนาระบบงาน
8. ทดสอบผลลัพธ์ของระบบงาน
9. สรุปผลการทำโครงการ

1.5 ข้อจำกัดของการศึกษา

การพัฒนาโครงการนี้ได้มีข้อจำกัดต่าง ๆ ของระบบงานดังนี้

1. ฐานข้อมูลที่ใช้ต้อง Microsoft SQL Server เท่านั้น
2. ข้อมูลที่รับเข้ามาต้องพร้อมต่อการเข้าสู่กระบวนการดาต้าไมนิ่งเรียบร้อยแล้ว กล่าวคือ ข้อมูลในทุกคอลัมน์ต้องเป็นตัวอักษรแล้วเท่านั้น
3. ฐานข้อมูลที่เรียกใช้หากมีค่าของแถวที่เป็นค่าว่างจะทำการตัดแถวนั้นทิ้งทันที
4. การแตกกิ่งของต้นไม้การตัดสินใจจะแตกกิ่งจนกว่าจะเหลือค่าของแถวภายในคอลัมน์เดียวกันไม่ต่ำกว่า 3 แถวเท่านั้น จากนั้นจะทำการหยุดแตกกิ่งการตัดสินใจ

1.6 ประโยชน์ที่คาดว่าจะได้รับ

การพัฒนาโครงการนี้ขึ้นก็เพื่อประโยชน์ต่าง ๆ ที่คาดว่าจะได้รับดังต่อไปนี้

1. สามารถประยุกต์ความรู้ที่ได้ศึกษามาเพื่อแก้ปัญหาในการพัฒนาและสามารถจัดการโครงการพัฒนาระบบงานด้านเทคโนโลยีสารสนเทศที่เกิดขึ้นในสภาพที่เป็นจริงได้
2. สามารถเข้าใจถึงขั้นตอนการค้นพบองค์ความรู้และขั้นตอนการทำดาต้าไมนิ่งในรูปแบบของ Classification โดยการใช้ดิซิชันทรี
3. เพื่อให้เกิดทักษะและความรู้ในด้านการวิเคราะห์และออกแบบระบบ รวมทั้งการพัฒนาระบบที่ใช้เทคโนโลยีดาต้าไมนิ่งในส่วนระบบฐานข้อมูล
4. ระบบที่พัฒนาขึ้นสามารถช่วยในการตัดสินใจให้กับผู้ใช้งานได้มีประสิทธิภาพยิ่งขึ้น
5. ระบบที่พัฒนาขึ้นนี้จะสามารถช่วยจัดหมวดหมู่และวิเคราะห์ข้อมูลที่อยู่ภายในฐานข้อมูล
6. การศึกษาและพัฒนานี้สามารถเป็นแนวทางในการประยุกต์ใช้ดาต้าไมนิ่งกับงานด้านอื่น ได้ต่อไป

บทที่2
แนวคิดและทฤษฎีที่เกี่ยวข้อง

ดาต้าไมนิ่งเป็นเทคโนโลยีที่ขยายมาจากเทคนิคทางสถิติร่วมกับเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence) และ Machine Learning เพื่อสร้างแบบจำลองสำหรับช่วยตัดสินใจปัญหาทางธุรกิจ การทำดาต้าไมนน์นิ่งเป็นขั้นตอนหนึ่งในกระบวนการค้นพบความรู้ ที่ช่วยในการค้นหารูปแบบและแนวโน้มที่ซ่อนอยู่ในข้อมูล เพื่อแปลงออกมาเป็นผลลัพธ์ที่สามารถทำความเข้าใจได้ง่ายกับผู้ใช้งานในทุก ๆ ระดับ จึงสามารถกล่าวได้ว่าการค้นพบความรู้เป็นกระบวนการที่มีการโต้ตอบและเกี่ยวข้องกับมนุษย์ (Human-Centered) ซึ่งจำเป็นต้องอาศัยเทคนิคที่เหมาะสมเช่นกัน

การทำงานในลักษณะดังกล่าวต้องอาศัยข้อมูลจำนวนมากและพิจารณาความสัมพันธ์ระหว่างข้อมูลเหล่านั้นโดยใช้แนวความคิดและเทคนิคต่าง ๆ ได้แก่ การนำหลักสถิติมาประยุกต์ใช้กับปัญญาประดิษฐ์, ฐานข้อมูลหรือแนวความคิดอื่น เพื่อนำมาสร้างกฎและรูปแบบเพื่อนำไปวิเคราะห์ให้เกิดประโยชน์ต่อไป

2.1 ความเป็นมาของดาต้าไมนิ่ง [1]

ในอดีตเมื่อเริ่มมีการเก็บข้อมูลด้วยฐานข้อมูลในทศวรรษที่ 60 (1960s) จากนั้นเทคโนโลยีการเก็บข้อมูลด้วยฐานข้อมูลได้ถูกพัฒนาขึ้นมาอย่างต่อเนื่อง จนมาสู่ยุคข้อมูลข่าวสาร (1990s – 2000s) ข้อมูลมีจำนวนมาก แต่ไม่สามารถนำข้อมูลเหล่านั้นมาใช้ให้เกิดประโยชน์ได้ ซึ่งข้อมูลมากมายที่มีอยู่อาจมีข้อมูลที่มีประโยชน์เพียงบางส่วน แต่เป็นข้อมูลส่วนที่มีประโยชน์อย่างมาก ดังนั้นเทคนิคดาต้าไมนิ่งจึงได้ถูกพัฒนาขึ้นเพื่อช่วยจัดการปัญหาดังกล่าว และได้รับความสนใจเป็นอย่างมาก

ในปัจจุบันเทคโนโลยีทำให้กระบวนการดาต้าไมนิ่งเป็นไปอย่างอัตโนมัติ มีการรวมเข้ากับดาต้าแวร์เฮาส์ และนำเสนอผลลัพธ์ในหลาย ๆ ทางที่ผู้ใช้ต้องการ

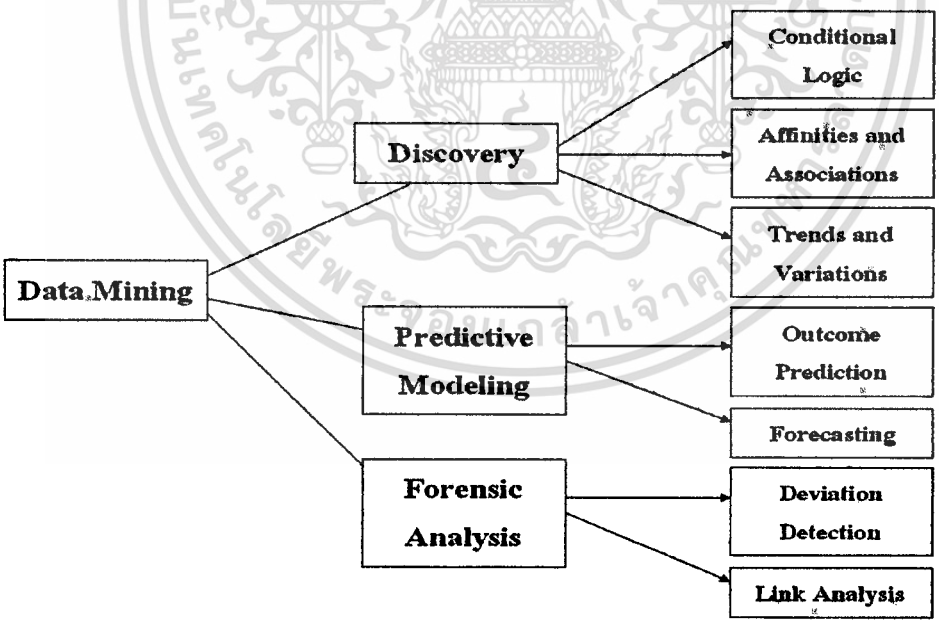
2.2 ดาต้าไมนิ่ง [1]

ดาต้าไมนิ่งเป็นเทคนิคที่ใช้ในการค้นหารูปแบบสารสนเทศที่ซ่อนเร้นจากข้อมูลขนาดใหญ่ เปรียบเสมือนกระบวนการสนับสนุนการตัดสินใจจากรูปแบบที่แอบแฝงอยู่ในข้อมูลเป็นกระบวนการของการตัดทอนข้อมูลจากฐานข้อมูลขนาดใหญ่ เช่น ความสัมพันธ์ แนวโน้มหรือรูปแบบเฉพาะ (Pattern) ซึ่งสารสนเทศเหล่านี้มีประโยชน์อย่างมากในการตัดสินใจเชิงธุรกิจ เพื่อให้องค์กรมีประสิทธิภาพในการแข่งขันมากยิ่งขึ้น ในปัจจุบันองค์กรส่วนใหญ่จะเผชิญกับปัญหาของข้อมูลดิบจำนวนมากแต่ข้อมูลที่ประยุกต์ใช้ได้นั้นมีน้อย

ในโลกของธุรกิจปัจจุบัน บริษัทต่าง ๆ จะพยายามหาเทคนิคต่าง ๆ ที่สามารถนำความสำเร็จมาสู่บริษัท เช่น ธุรกิจขนาดย่อมจะสร้างความสัมพันธ์กับลูกค้าโดยสังเกตจากความต้องการ ความชอบ และความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากข้อมูลในอดีตว่าจะทำอย่างไรให้การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต หรือบริษัทที่เป็นผู้ออกบัตรเครดิตและธนาคารต่าง ๆ จะมีขบวนการที่ใช้ดาต้าไมนิ่งให้เป็นประโยชน์ ในการช่วยตัดสินใจว่าลูกค้ากลุ่มใดเป็นกลุ่มที่ดี, ทำความเข้าใจลูกค้า, ช่วยในการแยกประเภทของลูกค้าและทำนายกลุ่มของประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต เป็นต้น

การนำดาต้าไมนิ่งมาประยุกต์ใช้กับงานนั้นมีหลายรูปแบบ เช่น ธุรกิจด้านการสื่อสาร ได้แก่การค้นหาลักษณะของกลุ่มลูกค้าที่มีแนวโน้มจะยกเลิกการใช้บริการ หรือในธุรกิจด้านการตลาดแบบการขายตรง (Direct marketing) ที่นำดาต้าไมนิ่งมาใช้ในการทำนายคุณลักษณะของลูกค้าที่มีแนวโน้มซื้อสินค้าประเภทต่าง ๆ ซึ่งจะทำให้สามารถวางแผนการทำตลาดของสินค้าไปที่กลุ่มลูกค้าที่ทำรายได้ให้สินค้ามาก ๆ ได้ดียิ่งขึ้น

เป้าหมายของการทำดาต้าไมนิ่งมีแตกต่างกันไปได้แก่ Discovery, Predictive Modeling และ Forensic Analysis แต่ละแบบมีข้อดีข้อด้อยและได้มาจากเทคนิคที่แตกต่างกัน จึงมีความจำเป็นที่จะต้องคัดเลือกเทคนิคที่จะนำมาใช้ในการวิเคราะห์ให้เหมาะสมกับข้อมูล



รูปภาพ 2.1 เป้าหมายการทำดาต้าไมนิ่งในรูปแบบต่าง ๆ

2.2.1. Predictive Modeling เป็นรูปแบบที่ค้นพบจากฐานข้อมูล เพื่อใช้คาดคะเนอนาคต โดยยอมให้ผู้ใช้สามารถกำหนดแถวที่ไม่ทราบค่าได้ และระบบจะคาดเดาค่าที่ไม่ทราบนั้น โดยใช้รูปแบบที่มีอยู่ก่อนในฐานข้อมูล

2.2.2. Forensic Analysis เป็นรูปแบบที่ใช้หาข้อมูลที่ผิดปกติ เพื่อค้นหาสิ่งที่ไม่คุ้นเคย โดยสิ่งแรกจะต้องค้นหาสิ่งที่เป็นปกติก่อน จึงจะตรวจพบสิ่งที่หักเหไปจากปกติ

2.2.3. Discovery ต่างจาก Predictive Modeling อย่างมากคือแบบ Predictive Modeling นั้นไม่สามารถช่วยให้เข้าใจข้อมูลของตนเอง เพียงแค่ใช้ในการคาดคะเนเท่านั้น อย่างไรก็ตาม นอกจากการคาดคะเน ไอดีไอเอส (IDIS-The Information Discovery) ยังสามารถค้นหาข้อมูลที่แน่นอน และบอกสิ่งที่ไม่เคยเปิดเผยมาก่อนของฐานข้อมูลด้วยภาษาอังกฤษที่ชัดเจน

2.3 การทำงานของดาต้าไมนิ่ง [1]

ดาต้าไมนิ่งเป็นกระบวนการในการค้นหาแนวโน้มและรูปแบบของข้อมูลที่ซ่อนอยู่ เพื่อสร้างความรู้ใหม่เกี่ยวกับข้อมูลนั้นๆ โดยใช้การวิเคราะห์ทางสถิติและเทคนิคในการสร้างแบบจำลอง (Model)

ดาต้าไมนิ่งนำเอาวิธีการสร้างแบบจำลองมาช่วยในการค้นหารูปแบบและความสัมพันธ์ของข้อมูล แบบจำลองเป็นเสมือนแบบจำลองของสถานการณ์จริง ซึ่งแบบจำลองที่ดีจะมีประโยชน์ในการทำความเข้าใจกับธุรกิจและบอกได้ถึงสิ่งที่ควรปฏิบัติเพื่อทำให้เกิดความสำเร็จในธุรกิจ กระบวนการทำงานของ Data Mining ประกอบไปด้วย 4 Phases ดังรูปที่ 2.2

2.3.1 กระบวนการจัดเตรียมข้อมูล (Data Preparation)

ชุดข้อมูลสำคัญที่จะจัดการโดยดาต้าไมนิ่งจะผ่านการแสดงตัว (Identified) และกลั่นกรองมาแล้ว เนื่องจากกระบวนการจัดเก็บข้อมูลในคลังข้อมูลได้ถูกจัดเก็บอย่างเป็นระบบมาแล้ว

2.3.2 กระบวนการวิเคราะห์และจัดหมวดหมู่ (Data Analysis and Classification)

เป็นการศึกษาคุณลักษณะของข้อมูลและรูปแบบของข้อมูล ขณะที่กำลังดำเนินกระบวนการนี้ เครื่องมือดาต้าไมนิ่งจะจัดการอัลกอริธึมเฉพาะ (Applies specific algorithms) ได้แก่

- จัดกลุ่มแบ่งประเภทและเรียงลำดับข้อมูล
- ทำการเชื่อมโยงความสัมพันธ์ของข้อมูล
- ชี้ให้เห็นถึงแนวโน้มของข้อมูล

2.3.3 กระบวนการเรียนรู้ความรู้ (Knowledge Acquisition)

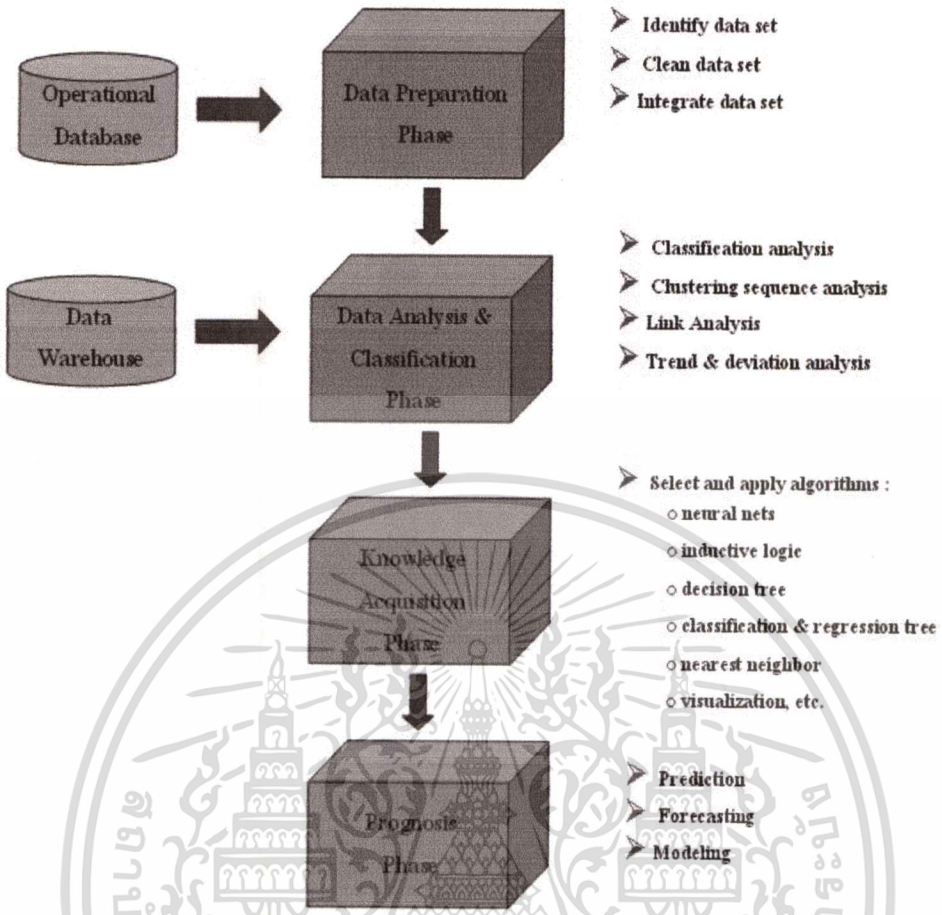
กระบวนการนี้จะใช้ผลของกระบวนการวิเคราะห์และจัดหมวดหมู่ของข้อมูล เครื่องมือดาต้าไมนิ่งจะเลือกแบบจำลองหรือ Knowledge Acquisition Algorithms ซึ่งอัลกอริธึมส่วนที่ใช้ในดาต้าไมนิ่งจะใช้พื้นฐานของอัลกอริทึมต่าง ๆ เช่น

- Neural networks
- Decision trees
- Rules induction
- Generic algorithms
- Classification and regression trees
- Memory-based reasoning

เครื่องมือดาต้าไมนิ่งจะใช้อัลกอริธึมเหล่านี้ในการรวบรวมเพื่อที่จะสร้างแบบจำลองคอมพิวเตอร์ (Computer model) ที่จะสะท้อนพฤติกรรมของข้อมูลที่ต้องการ ถึงแม้ว่าเครื่องมือดาต้าไมนิ่งบางตัวจะเสร็จสิ้นการทำงานที่กระบวนการเรียนรู้ความรู้ แต่ยังมีเครื่องมือดาต้าไมนิ่งอีกหลายตัวที่ยังคงทำงานต่อไปในกระบวนการคาดคะเน

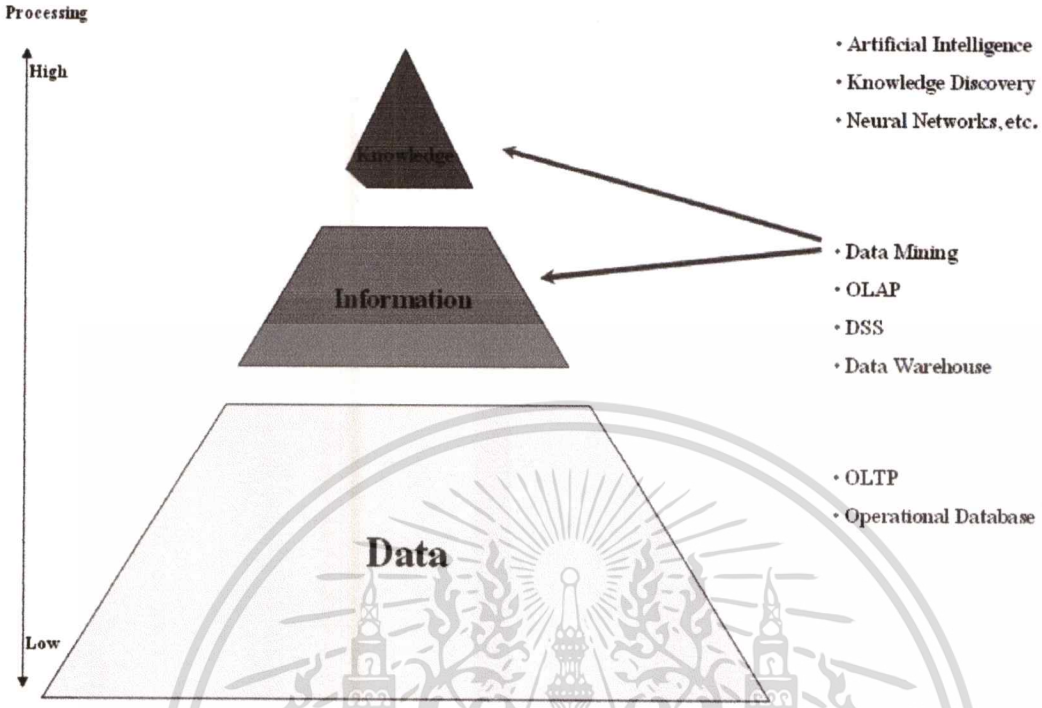
2.3.4 กระบวนการคาดคะเน (Prognosis)

กระบวนการนี้ดาต้าไมนิ่งจะทำการวินิจฉัยพฤติกรรมในอนาคตและพยากรณ์ธุรกิจที่กำลังจะเกิดขึ้น



รูปที่ 2.2 แสดงเฟสการทำงานของดาต้าไมนิ่ง

เครื่องมือดาต้าไมนิ่งใช้เทคนิคชั้นสูงจากการค้นหาความรู้ (Knowledge discovery), ปัญญาประดิษฐ์ (Artificial intelligence), และอื่น ๆ ที่จะบรรจุ "ความรู้" ประยุกต์ใช้กับความต้องการทางธุรกิจ เพื่อใช้ในการวินิจฉัยเหตุการณ์หรือพยากรณ์ค่าของข้อมูลต่าง ๆ โดยใช้เครื่องมือโอแอลเอพี (OLAP) ร่วมกับดาต้าไมนิ่ง ดังรูปที่ 2.3



รูปที่ 2.3 แสดงการนำความรู้ออกจากข้อมูล

2.4 รูปแบบการเก็บข้อมูลที่สามารถทำดาต้าไมนิ่ง

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced Database Systems and Advanced Database Applications
- Object-Oriented Databases
- Object-Relational Databases
- Spatial Databases
- Temporal Databases and Time-Series Databases
- Text Databases and Multimedia Databases
- Heterogeneous Databases and Legacy Databases
- The World Wild Web

2.5 รูปแบบข้อมูลตัวแปรในการทำดาต้าไมนิ่ง (Type of Attributes) [2]

2.5.1 ตัวแปรแบบ Categorical จำแนกออกเป็น

- Nominal เป็นตัวแปรที่กำหนดความเป็นไปได้อย่างชัดเจน เช่น yes, no เป็นต้น
- Ordinal เป็นตัวแปรที่มีการจัดลำดับ เช่น อุณหภูมิ hot, mild, cool โดยที่ hot > mild and cool หรือ hot < mild < cool
- Interval เป็นตัวแปรที่ไม่เพียงแต่สามารถจัดลำดับเท่านั้นแต่สามารถวัดได้ในหน่วยที่เหมือนกัน เช่น อุณหภูมิมักจะวัดกันในหน่วยขององศาเซลเซียสมากกว่า hot, mild และ cool จึงสามารถเปรียบเทียบได้ เช่น 20 องศาเซลเซียส ต่ำกว่า 22 องศาเซลเซียสอยู่ 2 องศา
- Ratio ตัวแปรที่อยู่ในรูปแบบของสัดส่วน

2.5.2. ตัวแปรแบบ Quantitative จำแนกออกเป็น

- Continuous ค่าที่เก็บเป็นตัวเลขจำนวนจริง (Real Number) หรือค่าที่ต่อเนื่อง เช่น รายได้ เป็นต้น
- Discrete ค่าที่เก็บเป็นตัวเลขจำนวนเต็ม (Integer) เช่น ข้อมูลจำนวนพนักงาน เป็นต้น

2.6 ขั้นตอนในการทำดาต้าไมนิ่ง

1. Data cleaning เป็นขั้นตอนของการทำความสะอาดข้อมูลโดยจะมีการกำจัด noisy และ inconsistency data
2. Data integration เป็นขั้นตอนของการรวมข้อมูลจากหลายแหล่งเข้าด้วยกัน จะทำเมื่อมีการรวมข้อมูลจากหลาย ๆ แหล่งข้อมูลมาไว้ที่เดียวกัน
3. Data selection เป็นขั้นตอนของการเลือกข้อมูล เมื่อมีข้อมูลที่เกี่ยวข้องกับสิ่งที่กำลังทำการวิเคราะห์จะดึงข้อมูลเหล่านั้นขึ้นมาจากฐานข้อมูล
4. Data transformation เป็นขั้นตอนของการแปลงข้อมูลที่ทำการจัดรูปแบบข้อมูลให้อยู่ในรูปแบบที่เหมาะสมในการทำดาต้าไมนิ่งตามอัลกอริทึมของดาต้าไมนิ่งที่เลือกใช้ เช่น การแปลงข้อมูลจากตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical
5. Data mining เป็นขั้นตอนที่ทำการไมน์นิ่งข้อมูลเพื่อให้ได้รูปแบบ ของข้อมูลที่ซ่อนอยู่
6. Pattern Evaluation เป็นขั้นตอนที่ทำการเลือกและวิเคราะห์รูปแบบที่สนใจ
7. Knowledge Representation เป็นขั้นตอนของเทคนิคของการนำเสนอความรู้ที่ได้จากการไมนิ่งให้ไปสู่ผู้ที่ใช้งานข้อมูลเหล่านั้น

2.7 ประเภทงานทางด้านดาต้าไมนิ่ง

เราสามารถจำแนกงานทางด้านดาต้าไมนิ่งออกเป็น 2 ประเภท คือ

2.7.1 Descriptive mining

เป็นการหาลักษณะคุณสมบัติของข้อมูลในฐานข้อมูล อธิบายเป็นรูปแบบ (Pattern) เพื่อใช้เป็นรูปแบบในการตัดสินใจหรือวางแผนในอนาคตได้

2.7.2 Predictive mining

เป็นการอ้างอิงจากข้อมูลเดิมที่มีอยู่ไปสู่การพยากรณ์ (Prediction) ข้อมูลที่แตกต่างออกไป เช่น การนำข้อมูลประวัติการชำระลูกหนี้ของลูกหนี้มาสร้างแบบจำลอง (Model) เพื่อระบุลักษณะของลูกหนี้ที่อาจมีปัญหา

ความแตกต่างโดยพื้นฐานของงานด้านดาต้าไมนิ่งทั้ง 2 ประเภทข้างต้น คือ Predictive mining แสดงผลการพยากรณ์อย่างชัดเจน ในขณะที่ Descriptive mining สามารถนำมาใช้ในการทำ Predictive mining ได้อีกทีหนึ่ง

2.8 ประเภทของแบบจำลองสำหรับการทำดาต้าไมนิ่ง (Data Mining Model)

2.8.1 Classification Model

Classification เป็นกระบวนการที่ใช้ในการค้นหาแบบจำลองที่จะใช้ในการพยากรณ์ (Prediction) โดยจะเป็นการจัดประเภทของสิ่งที่สนใจให้อยู่ในคลาส หรือกลุ่มที่ได้กำหนดไว้ล่วงหน้า ซึ่งคลาสจะต้องเป็นเซตของความเป็นไปได้ที่มีค่าแน่นอน

2.8.2 Clustering Model

Clustering เป็นวิธีการในการจัดกลุ่มข้อมูลที่มีลักษณะคล้ายกันเข้าไว้ด้วยกัน เพื่อลดขนาดของข้อมูลซึ่งสามารถเรียกอีกอย่างได้ว่าเป็นการทำ Segmentation โดยการจัดกลุ่มไม่ได้มีการกำหนดกลุ่มหรือคลาสไว้ล่วงหน้าแบบ Classification Model ช่วยให้สามารถค้นหาข้อมูลที่ถูกละเลยไปได้ เทคนิคนี้มักถูกใช้เป็นขั้นตอนเบื้องต้นในการทำดาต้าไมนิ่ง

2.8.3 Association Model

Association เป็นวิธีการที่ทำการวิเคราะห์หาความสัมพันธ์ของข้อมูลในรายการ (Transaction) เดียวกัน เป็นการค้นความสัมพันธ์ที่ถูกซ่อนอยู่ ไม่เฉพาะความสัมพันธ์ของแอททริบิวที่มีต่อคลาสเท่านั้น ยังหาความสัมพันธ์ระหว่างแอททริบิวด้วยกันเองอีกด้วย เช่น

การวิเคราะห์จากรายการการซื้อสินค้าของลูกค้าในซุปเปอร์มาร์เก็ต พบว่าส่วนใหญ่เมื่อลูกค้าซื้อผ้าอ้อม จะซื้อเบียร์ด้วย เป็นต้น ตัวอย่างรูปแบบของกฎ Classification จะเป็นดังนี้

" IF condition1 THEN condition2 " หรือ

" WHEN condition1 THEN condition2 "

โดยที่ condition1 และ condition2 เกิดขึ้นพร้อมกันในรายการเดียวกันเรียก condition1 ว่า Rule Body หรือ Left-hand หรือ Antecedent และเรียก condition2 ว่า Rule Head หรือ Right-hand หรือ Consequent ซึ่งสามารถเรียก condition1 ว่า "เหตุ" และ condition2 ว่า "ผล" อย่างไรก็ดี Association rules จะใช้กับแอททริบิวที่เป็น Non-numeric attribute

2.8.4 Deviation Detection

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) หรือเป็นความพยายามหาสิ่งแปลกปลอมออกจากกลุ่มของมัน ส่วนมากอาศัยการพล๊อตกราฟแล้วดูการกระจายของจุด สำหรับเทคนิคนี้ใช้ในการตรวจสอบลายเซ็นปลอม หรือบัตรเครดิตปลอม เป็นต้น

2.8.5 Sequential Analysis

ในการวิเคราะห์ลำดับเพื่อค้นพบรูปแบบของการปรากฏของข้อมูล ซึ่งปรากฏในรายการที่แยกออกมา เช่นถ้าผู้ซื้อของซื้อสินค้า A แล้วเขาจะต้องการซื้อสินค้า B ในภายหลัง

ในการศึกษาเพื่อการทำโครงการในครั้งนี้จะมุ่งประเด็นไปที่การสร้างแบบจำลองของ Classification Model โดยใช้เทคนิค Tree induction (Decision Tree) เป็นหลัก ซึ่งจะกล่าวถึงรายละเอียดในบทถัดไป

บทที่ 3

การจัดกลุ่มข้อมูลแบบ Classification ด้วย Decision Tree

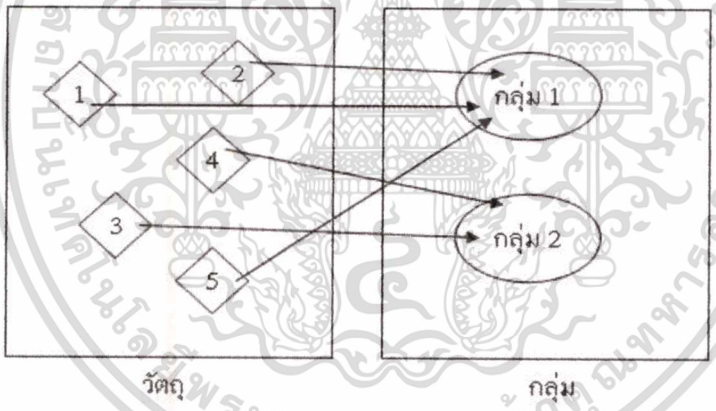
3.1 Classification Model

เป็นแบบจำลองที่สร้างขึ้นมาจากข้อมูลในอดีต โดยมีเฟสในการสร้างแบบจำลอง 2 เฟส คือ

- **Training Phase** เป็นการนำข้อมูลที่มีอยู่มาค้นหารูปแบบจำลองโดยผ่านอัลกอริทึมในเทคนิคของดาต้าไมนิ่งที่เลือกไว้ โดยข้อมูลที่ใช้เรียกว่า Training data
- **Test Phase** เป็นการตรวจสอบความถูกต้องของรูปแบบที่ได้จาก Training phase

กระบวนการของการทำการจัดกลุ่มข้อมูลนี้ถูกจัดให้เป็นการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) คือ มีการกำหนดรูปแบบข้อมูลนำเข้า (input)และข้อมูลส่งออก (Output) มาก่อน

Classification เป็นการสร้างขอบเขตของข้อมูลเบื้องต้นที่มีอยู่จริง เช่น



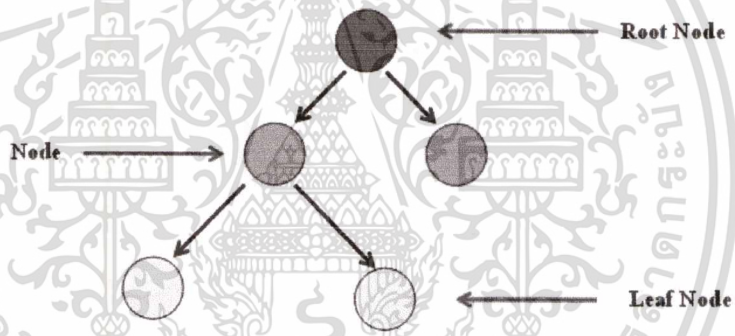
รูปที่3.1 แสดงการจัดกลุ่มของวัตถุ

จากรูปมีวัตถุเบื้องต้น 5 elements จะพยายามจัดกลุ่มนี้ออกได้เป็นกี่กลุ่ม ซึ่งขอบเขตนั้นรู้แล้วว่าแบ่งได้เป็น 2 กลุ่ม ถ้าปัญหาที่เราจัดกลุ่มของ elements ได้จะเป็นการจัดกลุ่มข้อมูลเช่น เราจะบอกว่าวัตถุชิ้นนี้ผ่านการทดสอบหรือไม่ ฉะนั้นเราสามารถบอกได้ว่าแบ่งเป็น 2 กลุ่ม คือกลุ่มที่ผ่านกับกลุ่มที่ไม่ผ่าน แล้วดูคุณสมบัติต่าง ๆ ว่าวัตถุชิ้นนี้มีคุณสมบัติตามที่ต้องการหรือไม่ ถ้าผ่านหมดก็ไปอยู่กับกลุ่ม 1 ถ้าไม่ผ่านก็ไปอยู่กับกลุ่มที่ 2 ซึ่งตัวอย่างแบบจำลองที่ใช้ในการทำการจัดกลุ่มข้อมูลมีอยู่หลายรูปแบบ แบบที่รู้จักกันทั่วไป ได้แก่

3.1.1 Tree Induction (Decision Tree)

เป็นลักษณะของ Flow-Chart ที่มีโครงสร้างเหมือนต้นไม้ แต่ละโหนด (Node) คือจุดที่ใช้กำหนดการตัดสินใจเลือกทางเลือกและเป็นตัวแทนของแอททริบิวในแต่ละโหนด โดยมี Root node เป็นปัจจัยสำคัญที่สุดที่มีผลต่อการตัดสินใจ และในแต่ละกิ่ง (Branch) จะเป็นตัวแทนผลที่ได้จากการทดสอบ ส่วนใบ (Leafs) ซึ่งเป็นโหนดในชั้นล่างสุด จะเป็นตัวแทนของคลาส ซึ่ง Decision Tree เปลี่ยนเป็นแบบ Classification rules ได้ง่าย แบบ Tree induction มีหลักการทำงาน ดังนี้

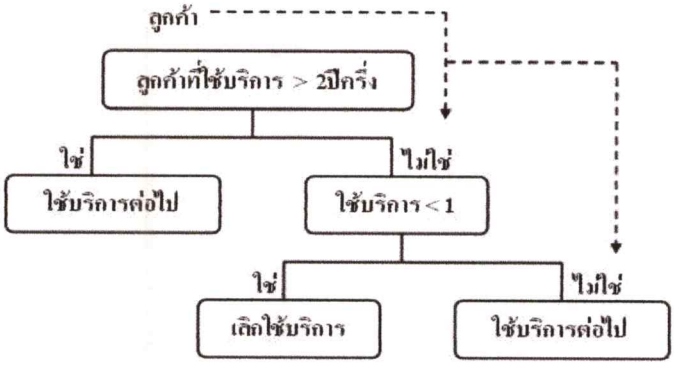
- กำหนดคอลัมน์ที่มีผลต่อการจัดกลุ่มที่สำคัญที่สุดในการทำนายเป็น Root node
- เมื่อเลือกคอลัมน์ที่สำคัญได้ก็การแตกกิ่งจากโหนดราก (Root node) จากการใช้ค่าของคอลัมน์ที่กำหนดนั้น โหนดที่ได้จากการแตกกิ่งเรียกว่าโหนดลูก (Child node)
- หลังจากนั้นกระบวนการทำงานจะวนซ้ำเดิม สำหรับแต่ละโหนดที่แตกออกมา



รูปที่ 3.2 แสดง Tree Representation

หลักการทำงานพื้นฐานของ Tree induction มีวิธีการทำงานโดยการกำหนดคอลัมน์ที่มีค่าความสำคัญที่สุดที่มีผลกระทบในการกำหนดการจัดกลุ่มเพื่อนำมาเป็นโหนดรากจากนั้นทางเลือกจากโหนดรากจะมีการกำหนดเป็นโหนดต่อ ๆ ไป สำหรับค่าที่เป็นไปได้ในการเลือกทางเลือกต่อไป กระบวนการทั้งหมดจะถูกกระทำซ้ำ ๆ บน Training data ที่คัดเลือกไว้ ปัญหาที่เกิดขึ้นคือจำนวนโหนดที่เหมาะสมควรเป็นกี่โหนด และการพิจารณาคัดเลือกคอลัมน์ที่ใช้กำหนดเป็นโหนดรากที่ต่างกันจะมีผลกระทบต่อแบบจำลองหรือไม่ อย่างไรก็ตาม Tree induction เป็นเทคนิคหนึ่งที่มีประสิทธิภาพในแง่ของการประหยัดเวลาที่ใช้ในการ Process และวิธีวิเคราะห์ผลลัพธ์

บ่อยครั้งที่มีการนำ Binary Tree มานำเสนอเพื่อให้เข้าใจง่าย ดังตัวอย่างต่อไปนี้



รูปที่ 3.3 แสดง Binary Decision Tree

จากตัวอย่างนำเสนอกลุ่มตัวอย่างของลูกค้าประกันภัยที่มีแนวโน้มที่จะใช้บริการของบริษัทต่อไปหรือเลิกใช้บริการของบริษัท อัลกอริธึมจะเริ่มที่ค่าที่มีความสำคัญที่สุด ค่านั้นถือว่ามีผลต่อการตัดสินใจในการจัดแบ่งกลุ่ม ซึ่งจากตัวอย่างนี้ค่าที่สำคัญที่สุดคือจำนวนปีที่เป็นลูกค้าของบริษัท ผลที่ได้จากการจัดกลุ่มจากฐานข้อมูลได้ 2 กลุ่ม คือกลุ่มที่เป็นลูกมากกว่า 2 ปีครึ่ง และกลุ่มที่เป็นลูกค้าน้อยกว่า 2 ปีครึ่ง ก็จะกำหนดให้เป็นค่าโหนดบนสุดของต้นไม้ ถ้าเป็นไปตามเงื่อนไขที่ตั้งไว้ก็จะนำไปใช้ทางโหนดซ้าย หากไม่ตรงตามเงื่อนไขก็จะอยู่ทางโหนดขวา อัลกอริธึมจะตัดสินใจในค่าที่สำคัญถัดไปซึ่งก็คือจำนวนบริการที่ลูกค้าใช้ และก็จะทำวนจนกระทั่งต้นไม้เสร็จสมบูรณ์

ตัวอย่างอัลกอริทึมของการสร้างต้นไม้ ได้แก่

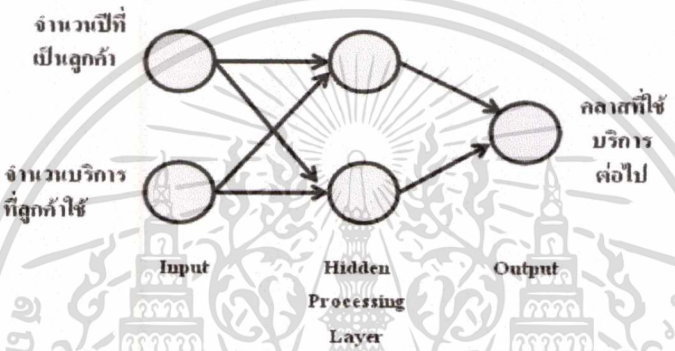
- CLS (1966) :เป็นหนึ่งในอัลกอริธึมเริ่มแรกของการทำ Decision tree
- CART (1984) : Classification And Regression Trees
- ID3 (1986) : Induction Decision Tree พัฒนาโดย Quilan
- C4.5 (1993) : Decision Tree Induction Algorithm ซึ่งพัฒนาต่อมาจาก ID3
- SLIQ (1996) : A Fast Scalable Classifier for Data Mining
- SPRINT (1996) : A Scalable Parallel Classifier for Data Mining
- PUBLIC (1998) :ใช้เทคนิค Tree splitting และ Tree pruning Integration
- RainForest (1998) : A Framework of Fast Decision Tree Construction of Large Dataset

3.1.2 Mathematical formula

เป็นการใช้รูปแบบทางสถิติ (Statistical Model) หรือกระบวนการเชิงเส้นทางคณิตศาสตร์ (Linear Model) มักใช้กับคลาสที่เป็นตัวเลข ตัวอย่างอัลกอริทึมประเภทนี้ เช่น Linear regression, Non-Linear regression, Logistic regression เป็นต้น

3.1.3 Neural Induction

เป็นรูปแบบที่เลียนแบบจากการทำงานของสมองมนุษย์ในการเรียนรู้และจัดการ เทคนิคนี้นำเสนอตัวอย่างที่เป็นโครงสร้างของโหนดและปรับค่าถ่วงน้ำหนัก (Weight) ของความสัมพันธ์ในการเชื่อมต่อนั้น ซึ่งเป็นพื้นฐานของเครือข่ายศูนย์กลางประสาท Neural networks ซึ่งเป็นการรวมกลุ่มของการเชื่อมต่อโหนดที่ได้ Input Output และการประมวลผลแต่ละโหนด ระหว่างค่า Input และ Output มีชั้นการประมวลผลที่ถูกซ่อนไว้ (Hidden processing layers) ดังรูปที่3.4 เป็นการนำเสนอ Neural network ที่ต่อเนื่องมาจากตัวอย่างก่อนหน้านี้ในเรื่องการจัดกลุ่มแนวโน้มของลูกค้าบริษัทประกันภัย



รูปที่ 3.4 แสดงตัวอย่าง Neural Network

จากรูปวงกลมแสดงถึงการประมวลผลหนึ่งยูนิต แต่ละยูนิตนั้นเชื่อมต่อกับแต่ละการประมวลผลในชั้นถัดไปโดยค่าน้ำหนักจะสนับสนุนความสัมพันธ์ ค่าน้ำหนักจะถูกแสดงที่เส้นที่เชื่อต่อแต่ละยูนิต และมีค่าเริ่มต้นที่ไม่ใช่เลขศูนย์ และมีการปรับค่าระหว่างการ เทรนนิ่งในเครือข่ายดังนั้นค่า Output ที่ได้จะต้องเป็นไปตามค่าที่ถูกคำนวณและนำไปใช้ประมวลผลในโหนดในเครือข่าย และจะทำเงื่อนไขนั้นตอนนี้จนกระทั่งผลที่จัดกลุ่มออกมาได้ถูกต้องในที่สุด

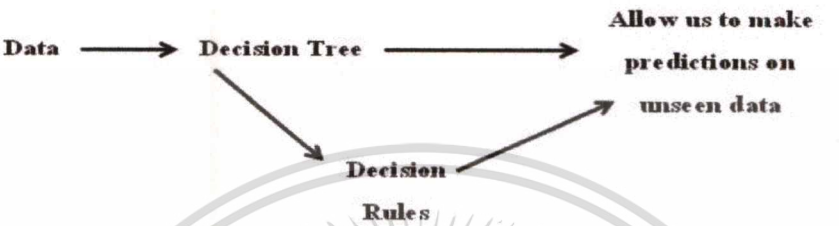
จากที่กล่าวมาข้างต้นเป็นการทำความเข้าใจเกี่ยวกับทฤษฎีดาต้าไมนิ่ง ทำให้สามารถเลือกเทคนิคและวิธีการที่เหมาะสมที่จะนำไปประยุกต์ใช้ในการพัฒนาระบบงาน คือ เทคนิคการจัดกลุ่มข้อมูลโดยใช้แนวทางดิซิชั่นทรี หรือที่เรียกว่าต้นไม้ตัดสินใจมาใช้ โดยเลือกการทำงานของอัลกอริธึม ID3 ดังที่จะได้กล่าวต่อไป

3.2 ต้นไม้การตัดสินใจ (Decision Tree)

ต้นไม้การตัดสินใจเป็นการทำนายในเทคนิค Tree induction ซึ่งเป็นแนวทางในเทคนิคการจัดกลุ่มข้อมูล เทคนิคการจัดกลุ่มข้อมูลเป็นการเรียนรู้ฟังก์ชันการจัดกลุ่มข้อมูลจากกลุ่มข้อมูลที่กำหนดให้ ในขั้นตอนการเรียนรู้ของการจัดกลุ่มข้อมูลจะนำไปสู่กฎที่สรุปออกมาได้ แต่ปัญหาหลัก

ของการเรียนรู้กฎของการจัดกลุ่มข้อมูลคือผลของต้นไม้การตัดสินใจที่เป็นไปได้จะมีขนาดใหญ่มาก และต้นไม้การตัดสินใจมีการประมวลผลที่ซับซ้อน

ต้นไม้การตัดสินใจ คือต้นไม้ที่ในแต่ละกิ่งหรือโหนดมีทางเลือกให้ตัดสินใจคั่นระหว่างตัวเลขของทางเลือกนั้น และแต่ละโหนดใบจะแสดงข้อมูลที่แบ่งแยกเป็นประเภทหรือข้อมูลที่ได้ตัดสินแล้ว เช่น เรามักจะใช้สถานการณ์ทางการเงินตัดสินใจในการให้กู้เงิน



รูปที่ 3.5 ภาพการทำงานของต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจใช้สำหรับการตัดสินใจด้วยการค้นหาง่าย ๆ จัดกลุ่มประเภทให้ถูกต้องสำหรับตำแหน่งหรือสถานการณ์ที่เจาะจง ต้นไม้การตัดสินใจทำงานโดยการสร้างกฎในรูปแบบโครงสร้างต้นไม้ โดยเงื่อนไขบนสุด (root node) จะถูกเปรียบเทียบก่อน เช่นการเปรียบเทียบอายุ อายุน้อยกว่า 20 อายุมากกว่าหรือเท่ากับ 20 และน้อยกว่า 40 หรืออายุมากกว่าหรือเท่ากับ 40 ตามคำตอบที่ได้จะอ้างถึงหนึ่งในหลายโหนดย่อย (Sub node) ของต้นไม้ แต่ละการทดสอบจะระบุเงื่อนไขออกมาได้ การประมวลผลของการทดสอบเงื่อนไขและการอ้างถึงซับโหนดถูกวนซ้ำจนกระทั่งเข้าถึงโหนดใบของต้นไม้ เพราะว่าการสร้างต้นไม้มีจุดมุ่งหมายคือโหนดใบที่จะแสดงถึงประเภทของคลาสที่จะแสดงถึงประเภทของคลาสที่แบ่งแยกได้ ซึ่งคือความรู้ที่ได้นั่นเอง

3.2.1 เหตุผลที่ Decision นั้นเป็นที่นิยมในการนำไปใช้มีดังต่อไปนี้

- ของเขตของการตัดสินใจที่ซับซ้อนและยุ่งยากระดับ Global (โดยเฉพาะใน High-dimensional space) นั้นสามารถประมาณออกมาได้ด้วยการนำขอบเขตการตัดสินใจที่ง่ายกว่าในระดับ Local ในหลาย ๆ Level ของ Tree มารวมกัน
- เมื่อเปรียบเทียบกับ Single-stage classifier ที่ต้องนำแต่ละข้อมูลตัวอย่างไปทดสอบเทียบกับ Class เป็นเหตุทำให้ประสิทธิภาพนั้นลดลง ซึ่งใน Tree classifier นั้น ตัวอย่างจะถูกทดสอบเทียบเพียงแค่ Subset ของ Class เท่านั้น ดังนั้นจึงเป็นการตัดการประมวลผลที่ไม่จำเป็นออกไป
- ใน Single-stage classifier นั้นจะใช้เพียง 1 Subset เท่านั้นสำหรับการจำแนกในจำนวน Class ทั้งหมด ซึ่ง Subset นี้มักจะถูกเลือกขึ้นมาโดยพิจารณาจาก Globally optimal

criterion เช่น พิจารณาจาก Maximum average interclass separability ส่วนในต้นไม้การตัดสินใจนั้นจะเป็นในทางกลับกันคือ มีความยืดหยุ่นในการเลือก Subset ที่แตกต่างกันของ Internal node ที่ต่างกัน ดังนั้น Feature subset ที่เลือกนั้นเป็น Subset ที่ดีที่สุดที่จะแบ่งแยกคลาสต่าง ๆ ในโหนดนั้น ๆ และความยืดหยุ่นที่เกิดขึ้นนั้นอาจเป็นตัวที่ทำให้ประสิทธิภาพของการตัดสินใจนั้นเหนือกว่า Single-stage classifier

- ใน Multivariate analysis โดยที่จำนวน Feature และ Class มีเป็นจำนวนมาก ๆ มักจะทำการประมาณค่าของ High-dimensional distributions (possibly multimodal) หรือไม่ก็ Certain parameters ของ Class distributions เช่น Priori probabilities จากเซตของ Training data ที่มีขนาดเล็ก ๆ ในทางปฏิบัติมักจะพบปัญหาในเรื่องของ High-dimensionality ซึ่งปัญหานี้อาจหลีกเลี่ยงในการตัดสินใจได้โดยการใช้จำนวนที่น้อยกว่าของลักษณะในแต่ละ Internal node โดยไม่คิดถึงเรื่องของประสิทธิภาพในการใช้งาน

3.2.2 ข้อเสียของ Decision Tree

- การซ้อนทับกัน (Overlap) โดยเฉพาะอย่างยิ่งเมื่อจำนวนของ Class มีเป็นจำนวนมาก ๆ เป็นสาเหตุให้จำนวนของ Class label มีมากกว่าจำนวนของคลาสที่มีอยู่จริง Actual classes ดังนั้นทำให้เวลาในการค้นหาเพิ่มขึ้น และความต้องการพื้นที่หน่วยความจำก็เพิ่มขึ้น
- ข้อผิดพลาดอาจมีการสะสมมาจากแต่ละระดับต้นไม้ ที่เพิ่มขึ้นเรื่อย ๆ ในต้นไม้ที่มีขนาดใหญ่ขึ้น ซึ่งไม่สามารถทำให้ความแม่นยำและประสิทธิภาพมีค่าที่ดีที่สุดพร้อมกัน
- การออกแบบการตัดสินใจให้ค่าที่ดีที่สุดค่อนข้างยาก ซึ่งประสิทธิภาพของการตัดสินใจจะขึ้นอยู่กับการออกแบบว่าออกแบบมาได้ดีเพียงใด

3.2.3 วิธีการต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจเป็นโครงสร้างที่สามารถมองได้ในรูปแบบแผนภูมิต้นไม้ โดยแต่ละกิ่งของโหนดแสดงให้เห็นถึงทางเลือกระหว่างจำนวนของทางเลือก และในส่วนปลายของต้นไม้หรือโหนดที่เป็นใบแสดงให้เห็นถึงการจำแนกพวกหรือการตัดสินใจ

ต้นไม้การตัดสินใจเป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคยและเข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดงแอททริบิว แต่ละกิ่งแสดงผลในการทดสอบ และโหนดใบแสดงคลาสที่กำหนดไว้

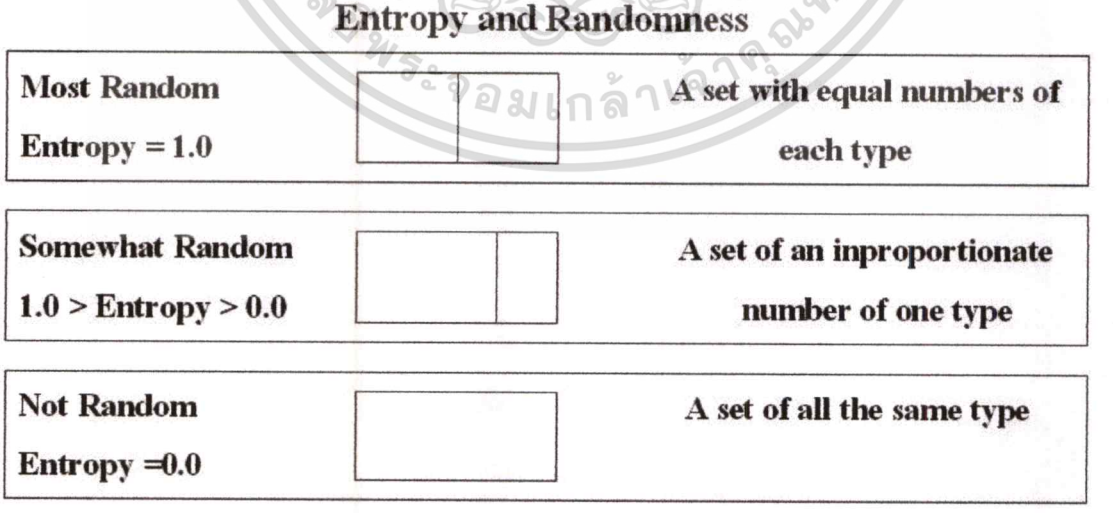
สมมติว่ามีบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์ มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะดูแลอาคารต่าง ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้เมื่อรวบรวมข้อมูลแบ่งเป็นตาราง

พื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา ข้อมูลพนักงาน ข้อมูลทรัพย์สิน และข้อมูลลูกค้า พร้อมทั้งกำหนดความสัมพันธ์ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า รายการให้เช่า รายการขายสินทรัพย์ เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า " 40% ของลูกค้าที่เช่าบ้านนานกว่าสองปี และมีอายุเกิน 25 ปี จะซื้อบ้านเป็นของตนเอง โดยกรณีเช่นนี้เกิดขึ้นกับ 35% ของลูกค้าผู้เช่าบ้านของบริษัท" เป็นต้น

โดยอัลกอริธึมที่ใช้ในโครงงานนี้คืออังกอริธึม ID3 โดย ID3 เป็นอัลกอริธึมพื้นฐานของต้นไม้การตัดสินใจ ซึ่งเป็นการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) เป็นการสร้างกฎในการแบ่งระดับ (Classification Rules) ในรูปแบบของต้นไม้การตัดสินใจ ซึ่งจะใช้เซตของการเทรนนิ่ง ซึ่งแอททริบิวจะถูกเลือกใช้เพื่อที่จะเป็นตัวแบ่งเซตของข้อมูล โดยอาศัยหลักของเอนโทรปี (Entropy) และอินฟอเมชันเกน (Information Gain) โดยต้นไม้จะถูกสร้างตามค่าย่อย ๆ ของแอททริบิวนั้น และทำเช่นนี้จนกระทั่งสมาชิกทุกตัวมีคลาสเดียวกัน

3.3 เอนโทรปี (Entropy)

เอนโทรปีคือกระบวนการที่ใช้ค้นหาระดับของความบริสุทธ์ ที่สร้างโดยตัวพรรณา (Descriptor) เพื่อสะท้อนการกระจายของตัวแบ่งแยก (Classifier) เอนโทรปีเป็นแนวคิดที่จะนำมาเพื่อใช้จัดเรียงตัวแบ่งแยกต่าง ๆ ตามระดับนัยสำคัญ (Significant) ของตัวแบ่งแยก หรืออธิบายได้อีกอย่างหนึ่งคือ เอนโทรปีจะเป็นตัวที่บ่งบอกถึงระดับของความไม่แน่นอนของตัวแบ่งแยกนั่นเอง



รูปที่ 3.6 ความสัมพันธ์ของการสุ่มข้อมูลและค่าเอนโทรปี

จากแนวความคิดของการสุ่มข้อมูล ค่าเอนโทรปีถูกนำมาอ้างอิงถึงโดยค่าเอนโทรปีจะแสดงถึงความเป็นชุดข้อมูลที่เหมือนกัน จากรูปกลุ่มตัวอย่างที่ประกอบด้วยชุดของ 2 ชนิดที่เป็นไปได้ ชุดแบบสุ่มมากที่สุดคือชุดที่มีความเป็นไปได้ของข้อมูลสองชนิดเท่ากันทั้งสองชนิด คือชุดข้อมูลที่มีค่าเอนโทรปี = 1 ในขณะที่ชุดข้อมูลอีกชุดหนึ่งที่มีความเป็นไปได้ของข้อมูลเพียงชนิดเดียว หรือไม่ว่าสุ่มอย่างไรก็จะได้ข้อมูลแบบเดียวกันเสมอ ชุดข้อมูลชุดนี้มีค่าเอนโทรปี = 0 ชุดข้อมูลนี้เรียดอีกอย่างหนึ่งว่าชุดข้อมูลที่มีโครงสร้าง

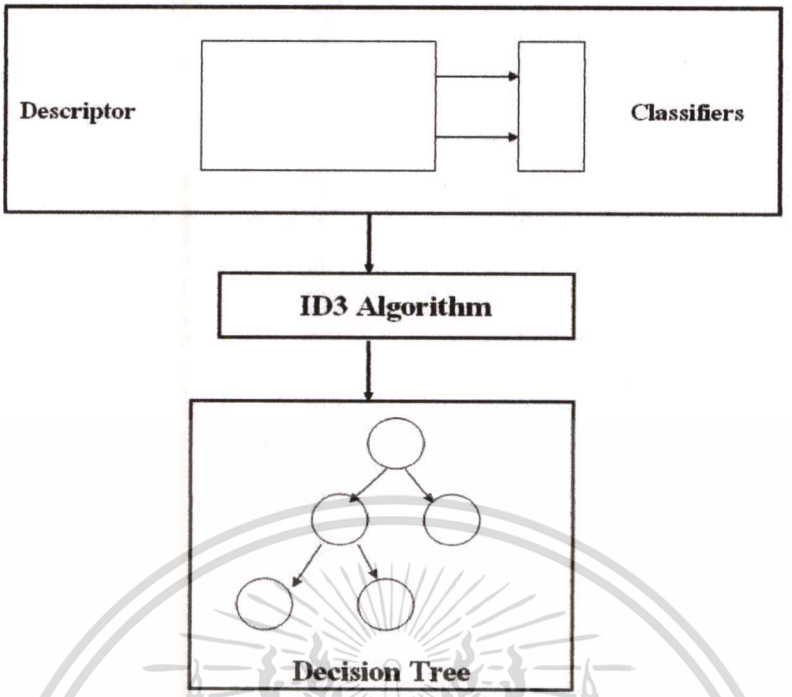
3.4 การวัดค่าอินฟอเมชันเกน (Information Gain Measure)

เป็นวิธีที่ใช้ในการวัดการปรับปรุงค่าเอนโทรปีหลังการแบ่งกับก่อนการแบ่ง เพื่อคำนวณหาค่าอินฟอร์เมชันเกนที่สูงที่สุดไปสร้างต้นไม้การตัดสินใจ

Information Gain = Entropy ก่อน Split – Entropy หลัง Split

3.5 อัลกอริธึม ID3

อัลกอริธึม ID3 เป็นอัลกอริธึมที่มีเป้าหมายในการสร้างต้นไม้การตัดสินใจจากข้อมูลที่ให้คำอธิบายที่มีประสิทธิภาพของการแยกประเภท (Classify) บนพื้นฐานของเอนโทรปี ซึ่งจะทำการสร้างต้นไม้การตัดสินใจที่สามารถนำไปใช้เพื่อทำนายหรือแยกประเภทของข้อมูล โดยอาศัยแนวความคิดของเอนโทรปีมาช่วยในการพิจารณาพฤติกรรมของตัวแบ่งแยก (Classifier)



รูปที่ 3.7 แสดงถึงแนวคิดอัลกอริธึม ID3

ID3 เป็นอัลกอริธึมพื้นฐานโดยที่ชุดตัวอย่าง (Examples) คือเซตของข้อมูลที่ใช้ในการเรียนรู้ (Training Examples) แอททริบิวเป้าหมาย (Target Attribute) คือแอททริบิวที่นำค่าไปใช้ในการทำนายผลในต้นไม้ และแอททริบิว(Attributes) คือแอททริบิวอื่น ๆ ที่ใช้ในการสร้างโหนดในต้นไม้ และไม่ใช่แอททรบิวเป้าหมาย ซึ่งลักษณะของอัลกอริธึมมีดังนี้

3.5.1 ID3 (Examples, Target_Attribute, Attributes)

- Create a root node for the tree
- If all examples are positive, return the Single-node tree root. With label = +
- If all examples are negative, return the single-node tree root, with label = -
- If number of predicting attributes is empty, then return the single node tree root, with label = most common value target of the target attribute in the examples
- Otherwise Begin
  * A ← The Attribute that best classifies examples
  * Decision tree attribute for root ← A
  * For each positive value, vi, of A,
    + Add a new tree branch below root, corresponding to the test A = vi
    + Let Examples(vi), be the subset of examples that have the value vi for A

+ If examples(vi) is empty

- Then below this new branch add aleaf node with label = most common target value in the example
- Else below this new branch add the subtree ID3 (Examples(vi), Target_Attribute, Attribute – {A})

- End

- Return Root

ซึ่ง ID3 เป็นอัลกอริธึมที่จะสร้างต้นไม้การตัดสินใจโดยจะมีการนำตัวอย่างเป็นข้อมูลที่จะเข้าไปทดสอบเพื่อหาแอททริบิวมาสร้างเป็นโหนดในต้นไม้การตัดสินใจ โดยการทำงานของอัลกอริธึม ID3 มีขั้นตอนต่าง ๆ ดังนี้

1. ทำการทดสอบว่าตัวอย่างทึกตัวที่รับเข้ามามีค่าของแอททริบิวเป้าหมายเป็นบวกทั้งหมดหรือไม่ ถ้าใช่ก็จะคืนโหนดที่มีค่าเป็นบวกให้ แล้วจบการทำงาน แต่ถ้าไม่ใช่ก็ทำต่อไป
2. ทำการทดสอบว่าตัวอย่างทุกตัวที่รับเข้ามามีค่าของแอททริบิวเป้าหมายเป็นลบทั้งหมดหรือไม่ ถ้าใช่ก็จะคืนโหนดที่มีค่าเป็นลบให้ แล้วจบการทำงาน แต่ถ้าไม่ใช่ก็ทำต่อไป
3. ทำการทดสอบว่าแอททริบิวอื่น ๆ ที่ไม่ใช่แอททริบิวเป้าหมายมีค่าหรือไม่ ถ้าไม่มีค่าก็จบการทำงาน แต่ถ้ายังมีค่าอยู่ก็ทำตามกระบวนการนี้

3.1 หาค่าอินฟอเมชันเกนของทุก ๆ แอททริบิวที่ไม่ใช่แอททริบิวเป้าหมายซึ่งแอททริบิวใดมีค่าเกน (Gain) สูงสุดจะถูกกำหนดให้เป็น โหนดรากตามสมการดังนี้

$$Gain(S, A) = Entropy(S) - S((|S_v| / |S|) * Entropy(S_v)) \quad (3.1)$$

Gain(S, A) คือค่า Information Gain ของ A

S คือค่าแต่ละค่าที่เป็นไปได้ของแอททริบิว A

S_v คือตัวอย่างทั้งหมดของค่าที่เป็นไปได้หนึ่ง ๆ ของแอททริบิว A ที่มีค่า v

$|S_v|$ คือจำนวน Elements ใน Sv

|S| คือจำนวน Elements ใน S

Entropy(S) คือเอนโทรปีของการรวบรวม S แบบเดิม

$Entropy(S_v)$ คือ ค่าที่ถูกคาดหวังของเอนโทรปีหลังจาก S ถูกแบ่งด้วยการใช้แอททริบิว A จากสมการ (1) สามารถหาค่าเอนโทรปีได้จากสมการ (2)

$$Entropy(S) = -p(+) \log2\ p(+) - p(-) \log2\ p(-) \quad (3.2)$$

p(+) คือ อัตราส่วนของตัวอย่างที่เป็นบวกใน S

p(-) คือ อัตราส่วนของตัวอย่างที่เป็นลบใน S

ถ้าค่าแอททริบิวเป้าหมายสามารถมีได้ c ค่า แล้ว Entropy ของ S จะเป็นดังนี้

$$\text{Entropy}(S) = \Sigma\ -p(I)\ \log_2\ p(I) \qquad (3.3)$$

โดยที่ p(I) คือ ความสัมพันธ์ของ S ที่คลาส I จะเห็นได้ว่าถ้าแอททริบิวเป้าหมายสามารถรับค่าได้ c ค่าแล้วเอนโทรปีจะเท่ากับ Log2c

4. กำหนดให้ A เท่ากับแอททริบิวที่มีค่าอินฟอเมชันเกนสูงสุดแล้วให้ A เป็นโหนดราก
5. ทำการวนลูปค่าต่าง ๆ ทั้งหมดที่เป็นไปได้ของ A โดยให้ v แทนค่าของ A ตัวที่ 1 ถึง I ตามขั้นตอนด้านล่างนี้ จนกว่าจะหมดแล้วทำการเลือกชุดตัวอย่างของ Attribute A ที่มีค่าเท่ากับ vi
6. ทำซ้ำตั้งแต่ข้อ 1 จนกว่าจะสร้างต้นไม้เสร็จ โดยทำการส่งค่าพารามิเตอร์ต่าง ๆ เข้าไป ดังนี้
   - ค่าตัวอย่างทั้งหมดที่เลือกมา
   - ค่าแอททริบิวเป้าหมาย
   - ค่าของแอททริบิวที่ลบเอาแอททริบิว A ออก

สรุปขั้นตอนการทำงาน โดยเริ่มจากให้อัลกอริธึมทำงานบนเซตของเรคคอร์ดที่ใช้ในการเทรนนิ่งซึ่งในที่นี้คือ S แล้วทำการตรวจสอบเงื่อนไขว่า ถ้าเซตของเรคคอร์ดทั้งหมด (All Instances) ใน S เป็นคลาส C ก็จะทำการสร้างโหนด C แล้วหยุด แต่ถ้ามีคลาสอื่นปนอยู่ด้วยก็ต้องทำการเลือกแอททริบิว A และสร้างโหนดการตัดสินใจ (Decision Node) ขึ้นมาแล้วทำการแบ่งส่วนเรคคอร์ดที่ใช้ในการเทรนนิ่งใน S ตามค่า V ของแอททริบิว A จากนั้นก็ทำซ้ำไปซ้ำมาในแต่ละเซตย่อยของ S_v ต่อไป

ในการเทรนนิ่งก็จะทำการแตกกิ่งไปเรื่อย ๆ จนถึงข้อมูลตัวสุดท้าย โดยที่ต้นไม้จะทำการแตกกิ่งไปเรื่อย ๆ แต่ในความเป็นจริงแล้วไม่ได้ง่ายอย่างที่คิดไว้ เพราะจะต้องพบกับอุปสรรคมากมายในระหว่างสร้างต้นไม้ โดยเฉพาะอย่างยิ่งเมื่อต้องการให้ความถูกต้องสูงขึ้นก็ต้องใช้ข้อมูลจำนวนมากขึ้นในการเทรนนิ่ง แต่ถ้าเป็นกรณีของการทดสอบข้อมูล ความถูกต้องจะเพิ่มขึ้นจนถึงจุด ๆ หนึ่งเท่านั้น แล้วถ้ามีการแตกกิ่งออกไปเรื่อย ๆ จะทำให้ความถูกต้องลดลง ซึ่งจุดที่ความถูกต้องถึงจุดสูงสุดของการทดสอบข้อมูล (Test Data) นี้เรียกว่า จุดโอเวอฟิต (Over fit) จึงเป็นผลให้เกิดการตัดกิ่งหรือแต่งกิ่ง (Pruning Tree) และสาเหตุที่ทำให้เกิดโอเวอฟิต อีกอย่างหนึ่งคือ เมื่อข้อมูลที่ใช้มีสิ่งรบกวน (Noise) หรือเมื่อข้อมูลที่ใช้ในการเทรน มีจำนวนน้อยเกินกว่าที่จะสร้างต้นไม้ได้ ซึ่งสิ่งเหล่านี้ทำให้อัลกอริธึมผลิตต้นไม้ที่เกิดการโอเวอฟิตได้ ซึ่ง ณ จุดที่เกิดการโอเวอฟิตของข้อมูลนั้นจะมีผลทำให้ความถูกต้องของแบบจำลองที่ได้ลดลง ดังนั้นทางที่ดีจึงควรใช้ข้อมูลจำนวนมากในการเทรนนิ่ง เพื่อที่จะทำให้ได้แบบจำลองที่ครอบคลุมมากที่สุด และใช้ข้อมูลจำนวนน้อย ๆ ในการทดสอบ

3.6 การหลีกเลี่ยงการเกิดโอเวอฟิต

การหลีกเลี่ยงการเกิดโอเวอฟิตสามารถทำได้ 2 แบบ คือ

3.6.1 พรี-พรุนทรี (Pre-Prune Tree)

หยุดต้นไม้ก่อนที่ต้นไม้จะแตกไปเรื่อย ๆ จนเกิดโอเวอฟิต คือในระหว่างการแตกต้นไม้ จะทำการทดสอบควบคู่ไปด้วยว่าเกิดโอเวอฟิตหรือไม่ ถ้าผลที่ได้จากการทดสอบทำให้ความถูกต้องแม่นยำลดลงก็หยุดทันที

3.6.2 โพส-พรุนทรี (Post-Prune Tree)

ทำการแตกต้นไม้จนหมดก่อนแล้วค่อยมาตัดแต่งกิ่งต้นไม้ในภายหลัง

จากทั้งหมดที่กล่าวมาจึงพอสรุปได้ว่า ID3 อยู่ในกลุ่มอัลกอริธึมการเรียนรู้ต้นไม้การตัดสินใจโดยใช้ทฤษฎีข่าวสาร (Information Theory) เพื่อตัดสินใจ ซึ่งแอททริบิวโดยการรวบรวมเรคคอร์ดเพื่อการแตกข้อมูลต่อไป แอททริบิวที่ถูกเลือกในวิธีนี้ถูกทำซ้ำไปซ้ำมาจนกระทั่งต้นไม้การตัดสินใจสมบูรณ์ ซึ่งคือการจำแนกทุก ๆ อินพุตที่เข้ามา ถ้าข้อมูลมีสิ่งรบกวนจะมีผลทำให้เรคคอร์ดอาจถูกจำแนกผิดพลาด อาจจะเป็นไปได้ที่จะทำการแต่งต้นไม้การตัดสินใจเพื่อที่จะลดความผิดพลาดในการจำแนกข้อมูลที่มีสิ่งรบกวนอยู่ได้ จากที่กล่าวมาทั้งหมดเป็นเพียงทฤษฎีของอัลกอริธึมนี้ อาจยังมองไม่เห็นถึงประโยชน์เท่าไร แต่ในบทต่อ ๆ ไป จะเป็นการนำทฤษฎีที่ได้ศึกษามาไปใช้ประโยชน์ได้กับงานจริง

บทที่ 4
การวิเคราะห์ระบบงาน

4.1 องค์ประกอบของระบบงาน

การทำงานของระบบนั้นมีหลายขั้นตอน สามารถที่จะอธิบายภาพรวมของระบบได้ดังนี้

4.1.1 Data Selection

การคัดเลือกข้อมูลเพื่อเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่งโดยจะต้องเลือกฐานข้อมูล ตาราง และฟิลด์ที่ผู้ใช้ต้องการ

4.1.2 Data Preparation

การเตรียมข้อมูลเพื่อเข้าสู่กระบวนการทำไมนิ่ง เมื่อคัดเลือกข้อมูลที่ต้องการแล้วจะนำข้อมูลมาทำการคลีนและการแปลงข้อมูล เพื่อเตรียมข้อมูลสำหรับการทำไมนิ่ง

4.1.3 Data Mining

การนำข้อมูลที่ได้จากขั้นตอนที่ 2 มาทำไมนิ่ง โดยจะแบ่งกลุ่มข้อมูลตามจำนวนกลุ่มที่ผู้ใช้กำหนด

4.1.4. Output

การแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล ในรูปแบบต่าง ๆ เช่น แสดงผลลัพธ์ทางหน้าจอ, บันทึกในรูปแบบของไฟล์ต่าง ๆ

4.2 Flow การทำงานของระบบที่พัฒนา

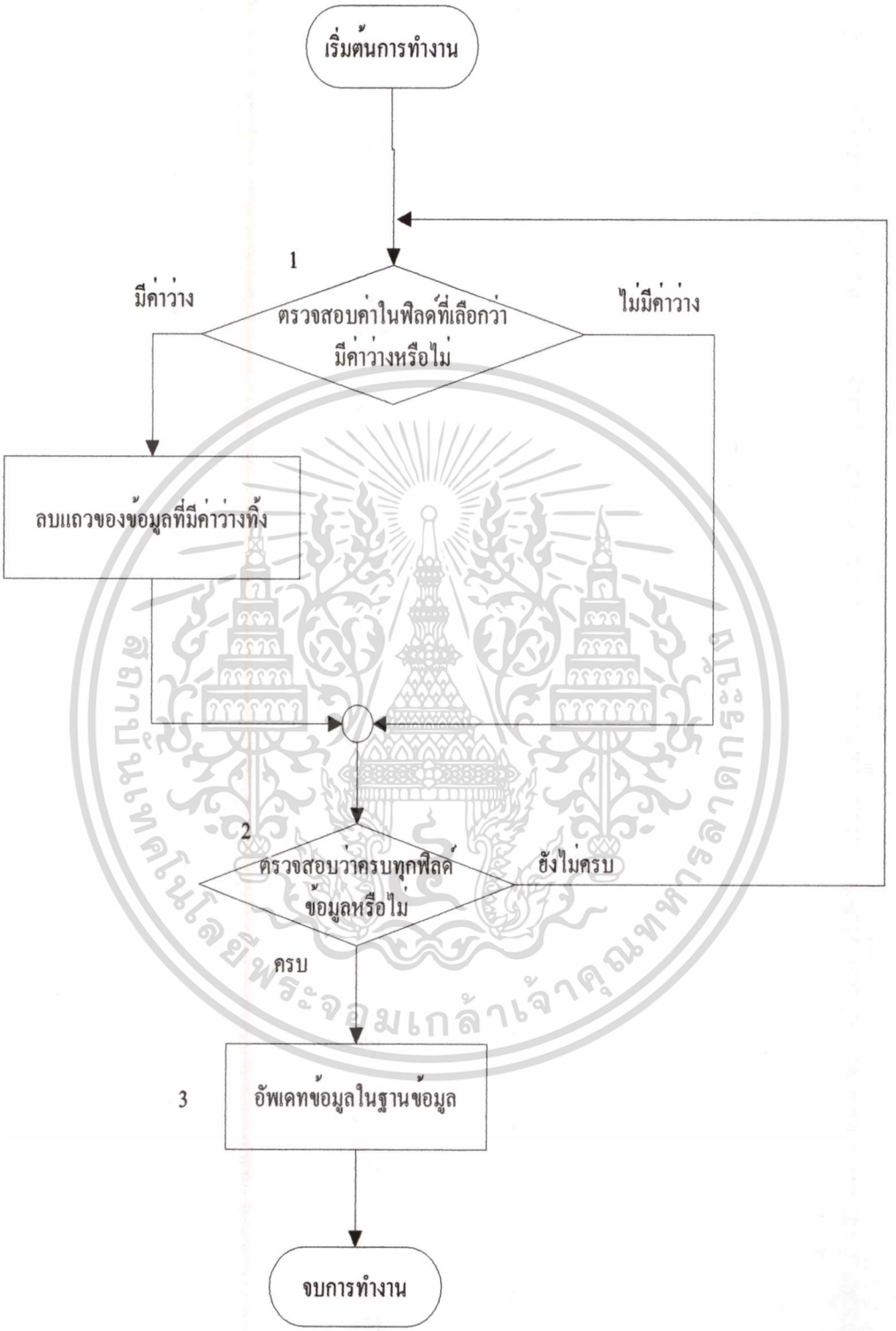
จากการทำงานของระบบที่กล่าวมาในบทข้างต้น สามารถนำมาวิเคราะห์และออกแบบโฟล์วการทำงาน (Flow Chart) ของระบบทั้งหมดได้ดังนี้

4.2.1 โฟล์วการทำงานภาพรวมของระบบ

รูปที่ 4.1 แสดงภาพรวมการทำงานของระบบทั้งหมด

1. ทำการติดต่อกับฐานข้อมูล Microsoft SQL Server โดยล็อกอินเพื่อเข้าใช้
2. เลือกไฟล์ฐานข้อมูลที่ต้องการ ซึ่งประกอบด้วยข้อมูลที่ต้องการนำมาจัดกลุ่ม โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูล Microsoft SQL Server เท่านั้น
3. เลือกตารางข้อมูลจากฐานข้อมูลที่ได้ทำการเลือกไว้แล้วโดยสามารถเลือกได้เพียง 1 ตาราง เท่านั้น
4. เลือกแอททริบิวที่ต้องการซึ่งสามารถเลือกได้ตามความต้องการของผู้ใช้งาน โดยข้อมูลจะอยู่ในรูปแบบที่เป็นตัวอักษรเท่านั้น
5. เลือกแอททริบิวเป้าหมาย เพื่อกำหนดเป้าหมายของการไมนิ่ง
6. การสร้างต้นไม้การตัดสินใจ โดยใช้ อัลกอริทึม ID3
7. แสดงผลการสร้างต้นไม้การตัดสินใจทางหน้าจอ

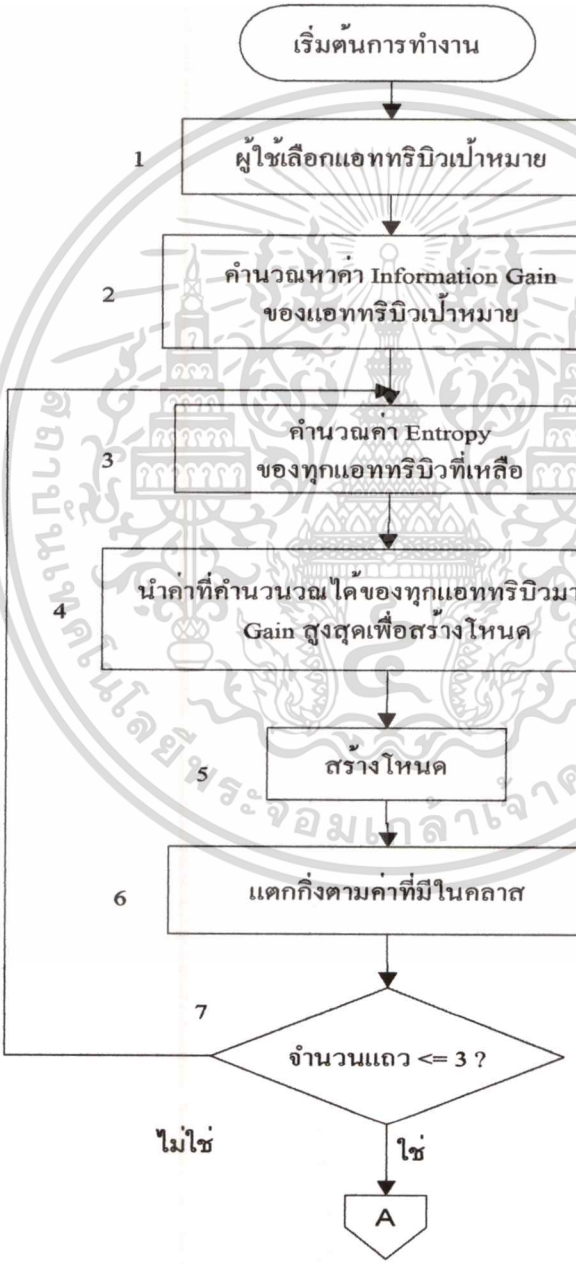
4.2.2 ขั้นตอนการลบดถวที่มีค่าว่างของข้อ มูล (Data Cleaning)



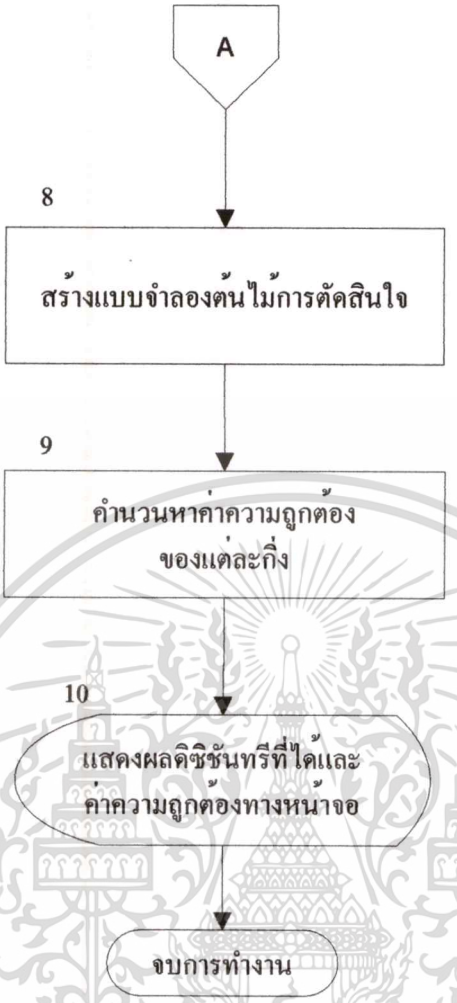
รูปที่ 4.2 แสดงขั้นตอนการคลีนข้อมูล

1. ตรวจสอบค่าในฟิลด์ข้อมูลที่เลือกว่ามีค่าว่าง (Missing Value)หรือไม่ ถ้าพบว่ามีค่าว่าง จะทำการลบแถวที่มีค่าว่างนั้นทิ้งไปทันที
2. ตรวจสอบข้อมูลว่าทำการคลีนข้อมูลในแถวที่พบว่ามีค่าว่างครบหรือไม่ ถ้ายังไม่ครบให้กลับไปทำข้อที่ 1ซ้ำ จนทุกแถวไม่มีค่าว่างจึงทำการอัพเดทค่าในตารางให้เป็นแบบที่คลีนแล้ว
3. ทำการอัพเดทข้อมูลในตารางฐานข้อมูล

4.2.3 ขั้นตอนการสร้างต้นไม้การตัดสินใจโดยใช้อัลกอริธึม ID3



รูปที่ 4.3 แสดงขั้นตอนการสร้างต้นไม้การตัดสินใจด้วยอัลกอริธึม ID3



รูปที่ 4.4 แสดงขั้นตอนการสร้างต้นไม้การตัดสินใจด้วยอัลกอริธึม ID3 (ต่อ)

1. ผู้ใช้เลือกแอททริบิวเป้าหมายเพื่อเข้าสู่กระบวนการไมนิ่ง
2. เข้าสูตรเพื่อคำนวณหาค่า Information Gain ของแอททริบิวเป้าหมาย
3. เข้าสูตรเพื่อหาค่า Entropy ของแอททริบวที่เหลือที่ไม่ได้เลือกมา
4. นำค่า Entropy กับ Information Gain ที่ได้มาค่า Gain สูงที่สุด
5. สร้างโหนดที่เกิดจากการวนหาค่า
6. แตกกิ่งของต้นไม้ตามค่าที่มีในคลาส
7. ตรวจสอบว่าค่าในตารางที่แตกออกแต่ละตารางมีจำนวนแถว น้อยกว่าหรือเท่ากับ 3 หรือไม่ ถ้ายังไม่ใช่ก็จะไปวนทำใหม่ตั้งแต่ (3)

8. สร้างแบบจำลองต้นไม้การตัดสินใจใน Tree View
9. คำนวนหาค่าความถูกต้องของแต่ละกิ่งของต้นไม้การตัดสินใจที่แตกออกมา
10. แสดงผลลัพธ์ของต้นไม้และค่าความถูกต้อทางหน้าจอ

4.3 การออกแบบส่วนการต่อประสานของระบบกับผู้ใช้

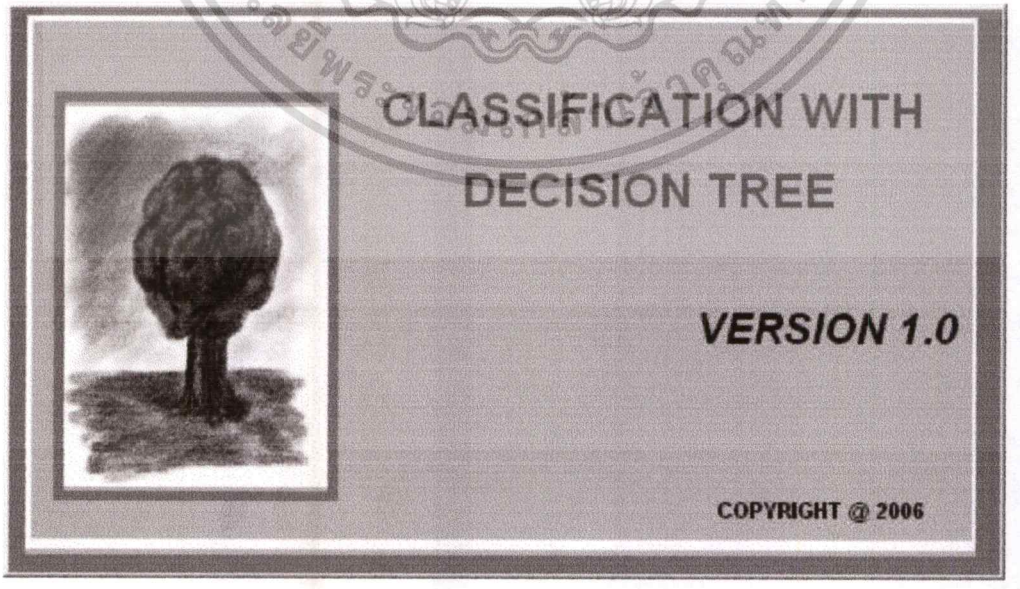
การออกแบบและพัฒนาระบบงาน มีเครื่องมือที่ใช้ในการพัฒนาและข้อจำกัดต่าง ๆ ดังนี้

- Application ที่ใช้ในการพัฒนาระบบคือ Microsoft Visual Studio.Net 2003
- ระบบฐานข้อมูล Microsoft SQL Server
- เครื่องคอมพิวเตอร์ที่นำมาใช้ในการพัฒนา คือ Microsoft Window XP Professional 2002, Pentium 3 Processor, CPU 666 Mhz, Ram 320 MB
- ตารางฐานข้อมูลที่สามารถเลือกมาใช้ในการไมนิ่งสามารถเลือกได้เพียง 1 ตารางเท่านั้น
- แอททริบิวทุกแอททริบิวที่เลือกไปใช้ในการไมนิ่งต้องเป็นชนิด categorical เท่านั้น
- หากพบแถวที่มีค่าว่างภายในฐานข้อมูลที่เลือกมาจะทำการตัดแถวนั้นทิ้งทันที
- การควบคุมการแตกกิ่งเพื่อเลี่ยงปัญหาโอเวอฟิตจะใช้วิธี Pre-prune tree เท่านั้น

การใช้งานเพื่อผลโดยมีข้อจำกัดของระบบที่ต่างไปจากนี้ อาจไม่สามารถหาผลลัพธ์ที่ถูกต้องได้ หรือระบบอาจไม่สามารถรองรับชุดข้อมูลดังกล่าวได้เลย

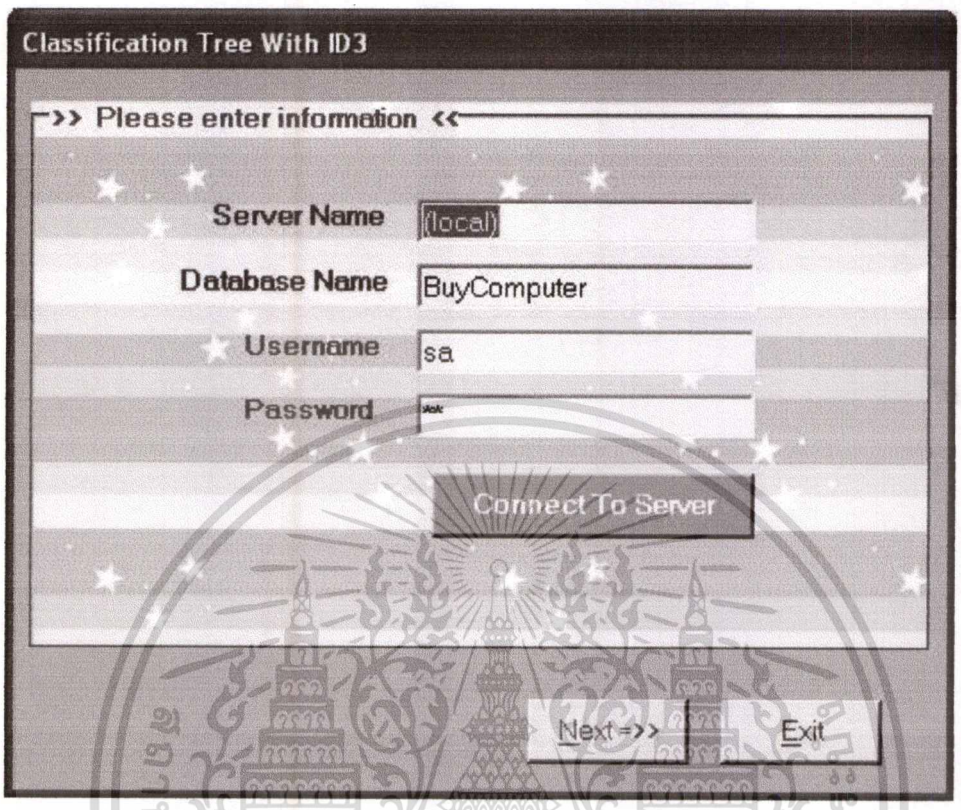
4.3.1 การออกแบบหน้าจอการทำงานของระบบ

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอต้อนรับเข้าสู่การใช้งาน



รูปที่ 4.5 แสดงหน้าจอต้อนรับเข้าสู่การใช้งานระบบ

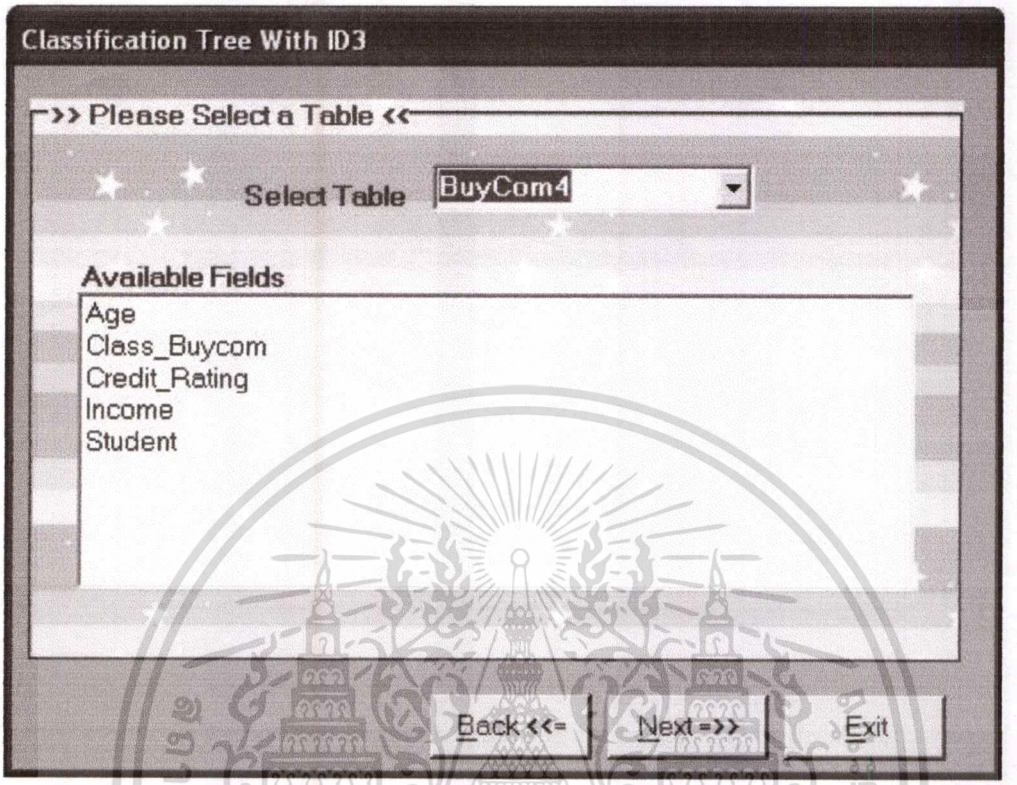
4.3.2 หน้าจอสำหรับล็อกอินเข้าไปใช้งานดาต้าเบสเซิฟเวอร์และระบุไฟล์ฐานข้อมูล



รูปที่ 4.6 แสดงหน้าจอล็อกอินเชื่อมต่อดาต้าเบสเซิฟเวอร์

- ในช่อง Server Name ผู้ใช้ต้องระบุชื่อของ Server ที่เก็บฐานข้อมูล Microsoft SQl Server
- ในช่อง Database Name ผู้ใช้ต้องระบุชื่อของไฟล์ฐานจข้อมูลที่ต้องการเรียกใช้
- ช่อง Username เป็นช่องให้ผู้ใช้ใส่ Username ของผู้ใช้เองเพื่อทำการล็อกอิน
- ช่อง Password เป็นช่องให้ผู้ใช้ใส่ Password ของผู้ใช้เองเพื่อทำการล็อกอินเข้าสู่ระบบ
- Connect To Server ปุ่มนี้ใช้กดเพื่อทำการล็อกอินไปยังฐานข้อมูลและ Server ที่เรียกขอใช้งาน หากการเชื่อมต่อนั้นสามารถเชื่อมได้สำเร็จจะขึ้นข้อความว่า “Connection Complete” แต่ถ้าหากไม่สามารถเชื่อต่อฐานข้อมูลนั้นได้ก็จะมีข้อความขึ้นมาบอกว่า “Connection Failed”
- Next =>> ปุ่มนี้มีเพื่อกดให้ไปทำงานยังหน้าต่างการทำงานต่อไป ปุ่มนี้จะสามารถกดได้ก็ต่อเมื่อสามารถเชื่อต่อกับฐานข้อมูลได้เป็นที่เรียบร้อยแล้วเท่านั้น
- Exit ปุ่มนี้เป็นปุ่มที่ใช้กดเมื่อต้องการออกจากโปรแกรม

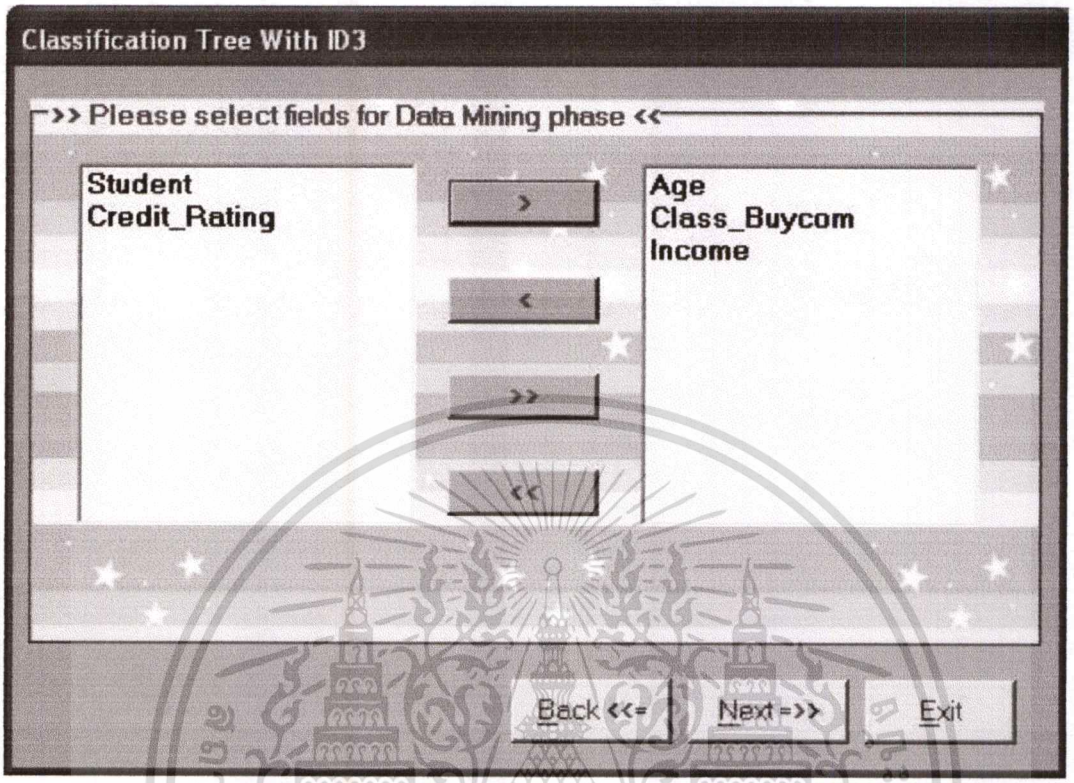
4.3.3 หน้าจอสำหรับเลือกตารางฐานข้อมูลที่ต้องการไปใช้งานในขั้นตอนต่อ ๆ ไป



รูปที่ 4.7 แสดงหน้าจอเลือกตารางฐานข้อมูล

- ในช่อง Select Table เป็นช่องเพื่อให้ผู้ใช้สามารถเลือกตารางจากแฟ้มฐานข้อมูลที่ได้ทำการเชื่อมต่อไว้แล้ว
- ในส่วนของ Available Fields จะแสดงถึงค่าแอททริบิวที่มีอยู่จริงทั้งหมดในตารางนั้น ๆ
- Back <<= ปุ่มเพื่อใช้ในการกดย้อนกลับไปยังหน้าต่างการทำงานของโปรแกรมหน้าก่อนหน้านี้ เพื่อทำการแก้ไขชื่อที่ใช้ในการล็อกอินทั้งหมดได้
- Next =>> ปุ่มนี้มีเพื่อกดให้ไปทำงานยังหน้าต่างการทำงานต่อไป
- Exit ปุ่มนี้เป็นปุ่มที่ใช้กดเมื่อต้องการออกจากโปรแกรม

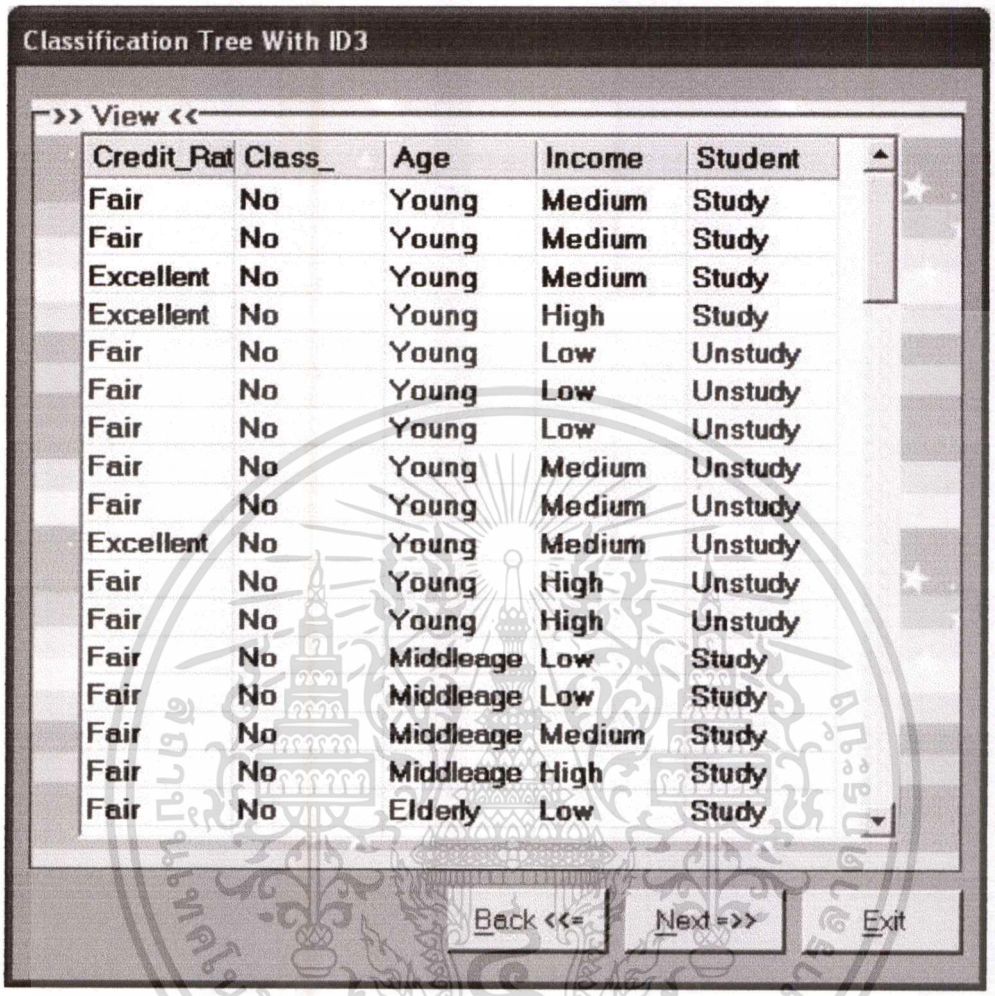
4.3.4 หน้าจอเลือกแอททริบิวที่จำเป็นต่อการทำไมนิ่ง



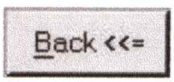
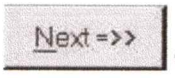
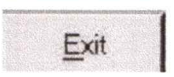
รูปที่ 4.8 แสดงหน้าจอการเลือกแอททริบิวที่สำหรับการทำไมนิ่ง

- ในช่องด้านซ้ายมือเป็นรายชื่อแอททริบิวที่มีทั้งหมดของตารางที่เลือกมา ส่วนช่องทางด้านขวาเป็นช่องที่แสดงถึงรายชื่อแอททริบิวที่ผู้ใช้เลือกเพื่อเข้าสู่กระบวนการไมนิ่ง
- > ปุ่มเพื่อเลือกแอททริบิวเข้าสู่ไมนิ่ง โดยเลือกได้ครั้งละ 1 แอททริบิว
- >> ปุ่มเพื่อเลือกแอททริบิวเข้าสู้ไมนิ่ง โดยเลือกทีเดียวได้ครบทุกแอททริบิว
- < ปุ่มเพื่อย้ายยกเลิกการเลือกแอททริบิว โดยเลือกได้ครั้งละ 1 แอททริบิว
- << ปุ่มเพื่อย้ายยกเลิกการเลือกแอททริบิว โดยเลือกครั้งเดียวย้ายกลับได้หมด
- Back <<= ปุ่มเพื่อใช้ในการกดย้อนกลับไปยังหน้าต่างการทำงานของโปรแกรมหน้าก่อนหน้านี้ เพื่อทำการแก้ไขชื่อตารางที่ต้องการเลือก
- Next =>> ปุ่มนี้มีเพื่อกดให้ไปทำงานยังหน้าต่างการทำงานต่อไป
- Exit ปุ่มนี้เป็นปุ่มที่ใช้กดเมื่อต้องการออกจากโปรแกรม

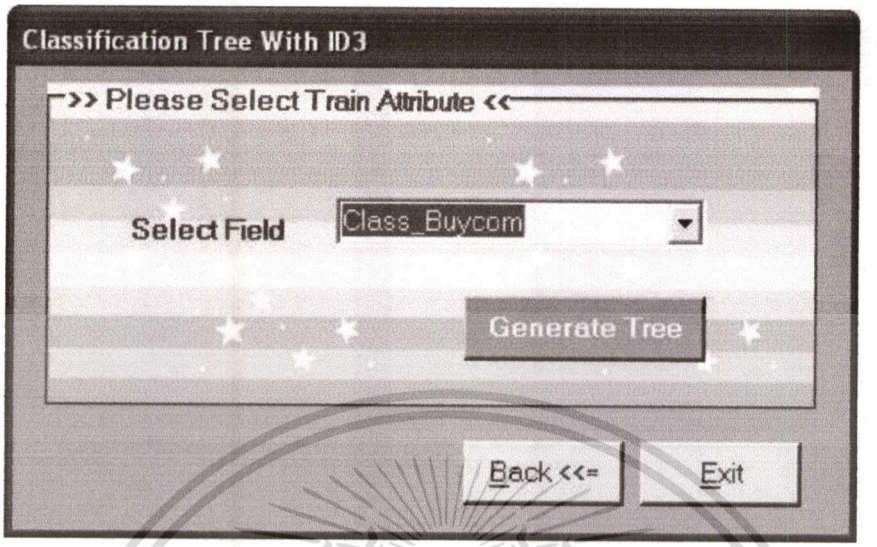
4.3.5 หน้าจอหน้าจอแสดงข้อมูลทั้งหมดที่เลือกมา



รูปที่ 4.9 หน้าจอแสดงข้อมูลทั้งหมดที่เลือกมาก่อนเข้าไมนิ่งเพื่อยืนยัน

- ในช่อง View จะแสดงให้เห็นถึงข้อมูลทั้งหมดที่เลือกมาก่อนที่จะเข้าสู่กระบวนการทำดาต้าไมนิ่ง เพื่อให้ผู้ใช้ตรวจสอบข้อมูลที่ตัวเองได้เลือกมาอีกทีว่าถูกต้องเหมาะสมหรือไม่
-  ปุ่มเพื่อใช้ในการกดย้อนกลับไปยังหน้าต่างการทำงานของโปรแกรมหน้าก่อนหน้านี้ เพื่อทำการแก้ไขแอททริบิวที่เลือกมาได้อีก
-  ปุ่มนี้มีเพื่อกดให้ไปทำงานยังหน้าต่างการทำงานต่อไป
-  ปุ่มนี้เป็นปุ่มที่ใช้กดเมื่อต้องการออกจากโปรแกรม

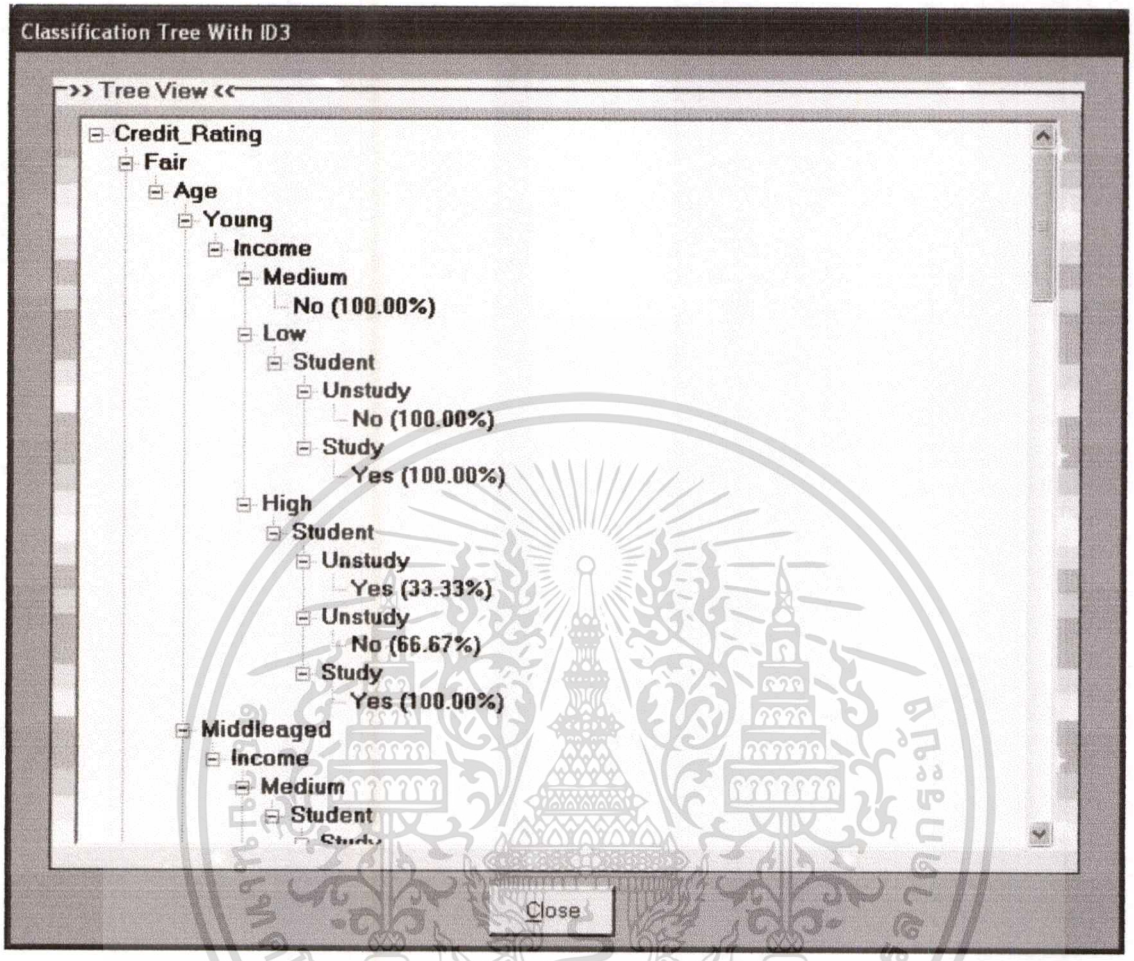
4.3.6 หน้าจอการกำหนดแอททริบิวเป้าหมาย



รูปที่ 4.10 แสดงหน้าจอการเลือกแอททริบิวเป้าหมายเพื่อสร้างต้นไม้

- แถบ Select Fields เป็นช่องที่ให้ผู้ใช้ทำการเลือกแอททริบิวเป้าหมายที่จะมีอิทธิพลต่อการทำนาย
- Generate Tree ปุ่มที่กดเพื่อทำการสร้างต้นไม้การตัดสินใจ เมื่อกดปุ่มนี้แล้วโปรแกรมจะทำการคำนวณแบบอัตโนมัติเพื่อสร้างต้นไม้และคำนวณค่าความถูกต้องของแต่ละกิ่งของต้นไม้
- Back <<= ปุ่มเพื่อใช้ในการกดย้อนกลับไปยังหน้าต่างการทำงานของโปรแกรมหน้าก่อนหน้านี้ เพื่อทำการแก้ไขการเลือกข้อมูลในเฟลก่อนหน้านี้
- Exit ปุ่มนี้เป็นปุ่มที่ใช้กดเมื่อต้องการออกจากโปรแกรม

4.3.7 หน้าจอสร้างต้นไม้การตัดสินใจ



รูปที่ 4.11 แสดงหน้าจอสร้างต้นไม้การตัดสินใจพร้อมแสดงค่าความถูกต้องของแต่ละกิ่งที่แตก

- ช่อง Tree View เป็นช่องแสดงให้เห็นถึงผลลัพธ์ของต้นไม้การตัดสินใจ ที่ปลายกิ่งแต่ละกิ่งจะมี ตัวเลขเป็นเปอร์เซ็นต์ บอกถึงค่าความถูกต้อง (Accuracy) ของแต่ละกิ่ง
- Close ปุ่มนี้มีไว้เพื่อปิดหน้าต่างการแสดงผลลัพธ์

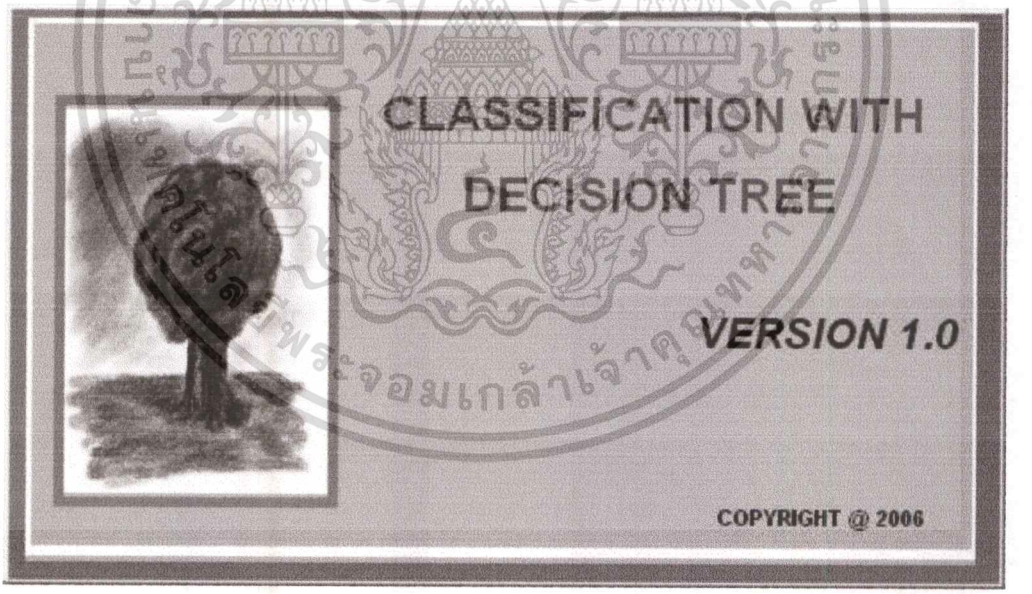
บทที่ 5
การประยุกต์ใช้งานของระบบ

โครงงานที่ได้พัฒนาขึ้นนี้สามารถนำไปใช้ได้กับทุกฐานข้อมูล แต่มีข้อจำกัดว่าฐานข้อมูลนั้นต้องเป็นฐานข้อมูลที่มีชนิดของข้อมูลที่เป็นตัวอักษร (Categorical) หรือกล่าวได้ว่าเป็นข้อมูลที่เตรียมมาและพร้อมที่จะเข้าสู่กระบวนการดาต้าไมนิ่งด้วยเทคนิค Classification ได้เลย

5.1 การประยุกต์ใช้งานของระบบ

ในบทนี้จะอธิบายถึงขั้นตอนการทำงานของระบบตั้งแต่ต้นจนจบว่ามีขั้นตอนและผลที่ออกมาเป็นอย่างไรบ้าง ในตอนท้ายของบทจะมีสรุปผลการทำงานของระบบว่ามีประโยชน์อย่างไรและช่วยในการพยากรณ์ได้อย่างไรบ้างต่อไป

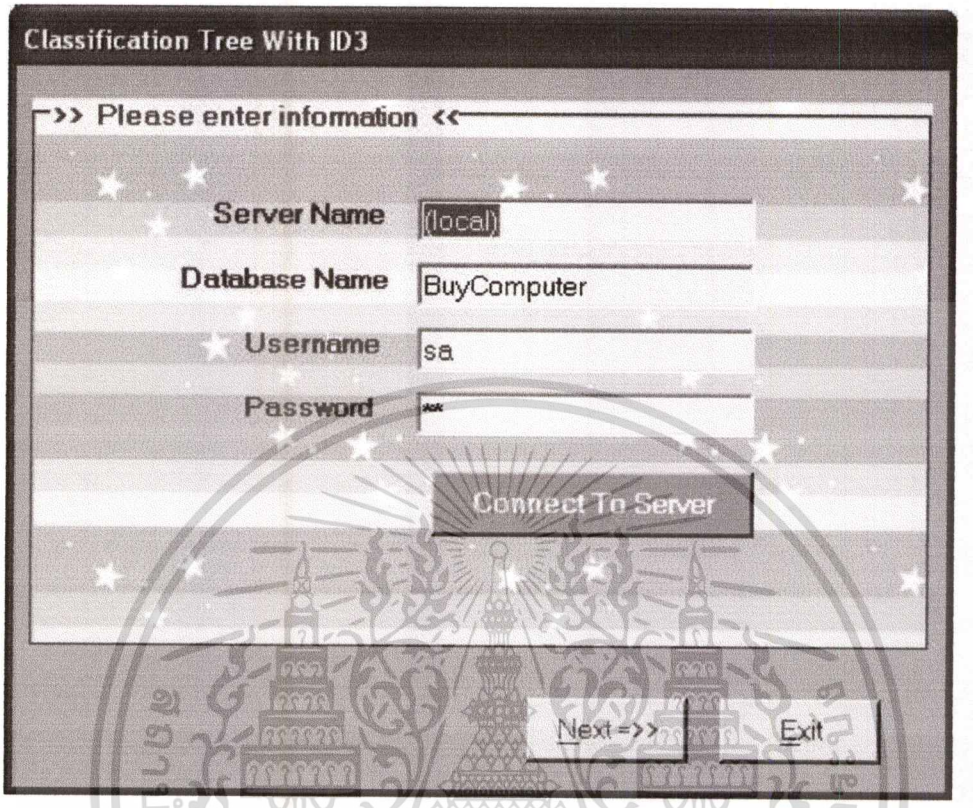
5.1.1 ต้อนรับเข้าสู่การใช้งานโปรแกรม



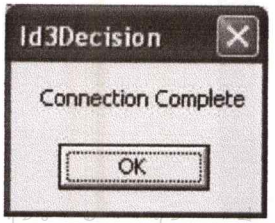
รูปที่ 5.1 หน้าจอต้อนรับเข้าสู่โปรแกรมการทำงาน

หน้าจอนี้แสดงหน้าแรกของโปรแกรม เป็นการต้อนรับผู้ใช้เข้าสู่การใช้งานระบบต่อไป

5.1.2 การล็อกอินเข้าสู่ระบบเพื่อเชื่อมต่อกับเซิฟเวอร์

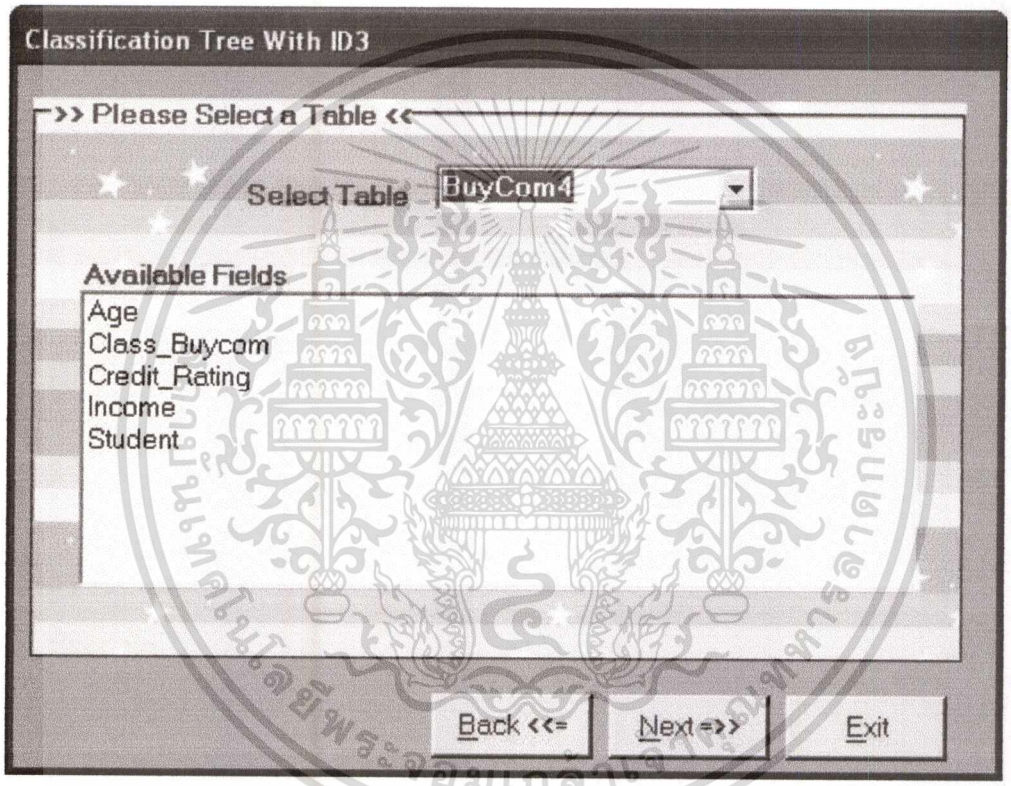


รูปที่ 5.2 หน้าจอล็อกอินเข้าใช้ฐานข้อมูล

- หน้าจอนี้ผู้ใช้ต้องกรอกข้อมูลทุกอย่างลงในช่องที่กำหนดไว้ให้ครบถ้วนเพื่อทำการติดต่อกับฐานข้อมูล Microsoft SQL Server
- ในช่อง Server Name ใส่ชื่อเซิฟเวอร์ที่เราเก็บข้อมูล ถ้าผู้ใช้ลงโปรแกรมตากปกติ ใช้ (local) ได้เหมือนกัน เพราะเป็นการเรียกเซิฟเวอร์ที่ใช้ภายในตัวเครื่องขึ้นมาใช้
- ในช่อง Database Name ใส่ชื่อแฟ้มข้อมูลที่ต้องการเรียกใช้งาน
- ช่อง UserName กับ Password เป็นการล็อกอินเข้าใช้ฐานข้อมูลที่เรียกอีกที ซึ่ง Username กับ Password นี้จำเป็นต้องระบุ ให้ดีอาจใช้เป็นชุดเดียวกับตอนติดตั้ง SQL Server ก็ได้ หรือจะสร้าง User account ขึ้นมาใหม่เพื่อเพิ่มสิทธิ์ให้ใช้ฐานข้อมูลที่เรียกก็ย่อมได้เช่นกัน
- จากนั้นกดปุ่ม Connect To Server เพื่อทำการเชื่อมต่อ ถ้าการเชื่อมต่อนั้นติดต่อได้สำเร็จ จะมีข้อความ  เป็นการบอกว่าการเชื่อมต่อนี้สำเร็จ

- ถ้าผู้ใช้เชื่อมต่อไม่สำเร็จจะปรากฏข้อความ “Connect Failed” ขึ้นซึ่งอาจมีสาเหตุมาจากหลายประการคือ ข้อมูลที่ใช้ในการเชื่อมต่อระบุผิด หรืออาจผิดพลาดที่ตัวฐานข้อมูลเอง
- ถ้าผู้ใช้กด Exit จะเป็นการออกจากโปรแกรม
- ผู้ใช้จะสามารถกดปุ่ม Next =>> ได้แล้วหากทำการเชื่อมต่อสำเร็จเพื่อทำงานในหน้าต่างต่อไป

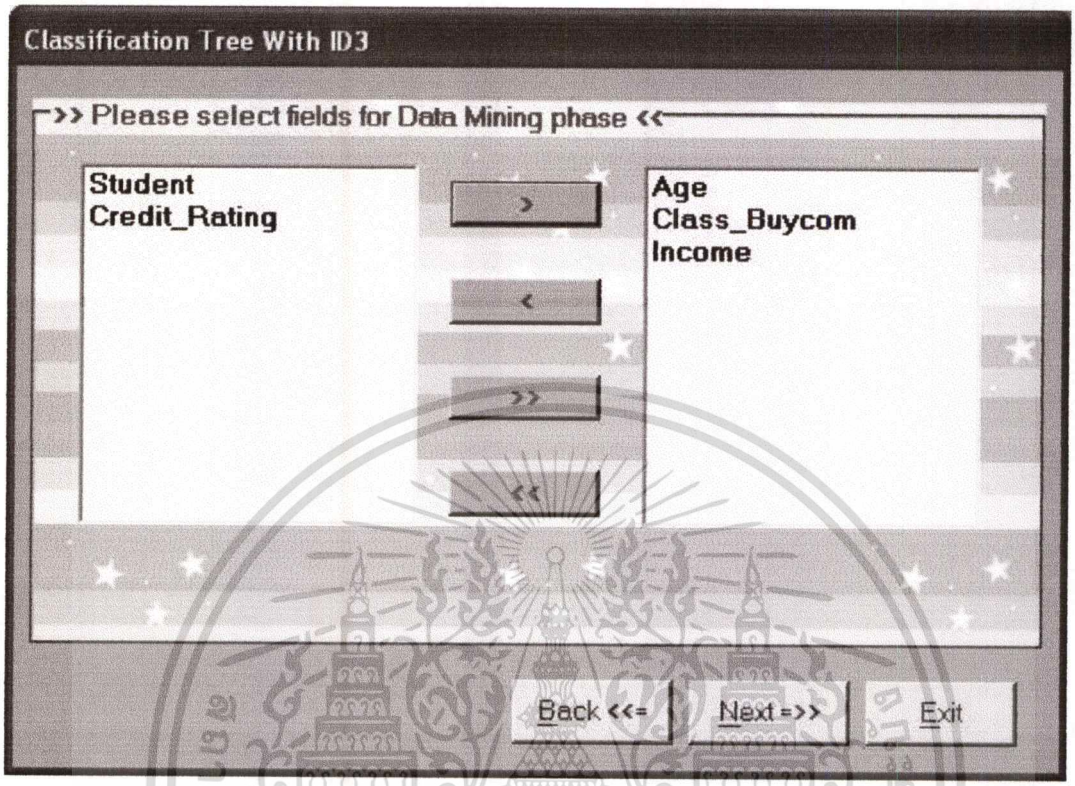
5.1.3 การเลือกตารางจากฐานข้อมูล



รูปที่ 5.3 หน้าจอให้ผู้ใช้เลือกตารางฐานข้อมูล

- หน้าต่างนี้จะเป็นหน้าต่างการทำงานเพื่อเลือกตารางฐานข้อมูลจากแฟ้มฐานข้อมูลที่ได้เรียกมาในเฟสก่อนหน้า
- ใน Drop down list box ประกอบไปด้วยชื่อตารางทั้งหมดที่มีอยู่ เพื่อให้ผู้ใช้เลือกใช้งานต่อไป
- แอททริบิวที่เป็นไปได้จากตารางฐานข้อมูลที่เลือกจะแสดงอยู่ในช่อง Available Fields
- จากตรงนี้เมื่อเลือกฐานข้อมูลแล้วหากกด Next =>> ก็จะไปทำงานในหน้าจอถัดไป
- หากกด Back <<= ก็จะสามารถกลับไปแก้ไขข้อมูลก่อนหน้าได้ เช่นชื่อฐานข้อมูลเป็นต้น
- หากกด Exit ก็จะสามารถออกจากโปรแกรมได้เลย

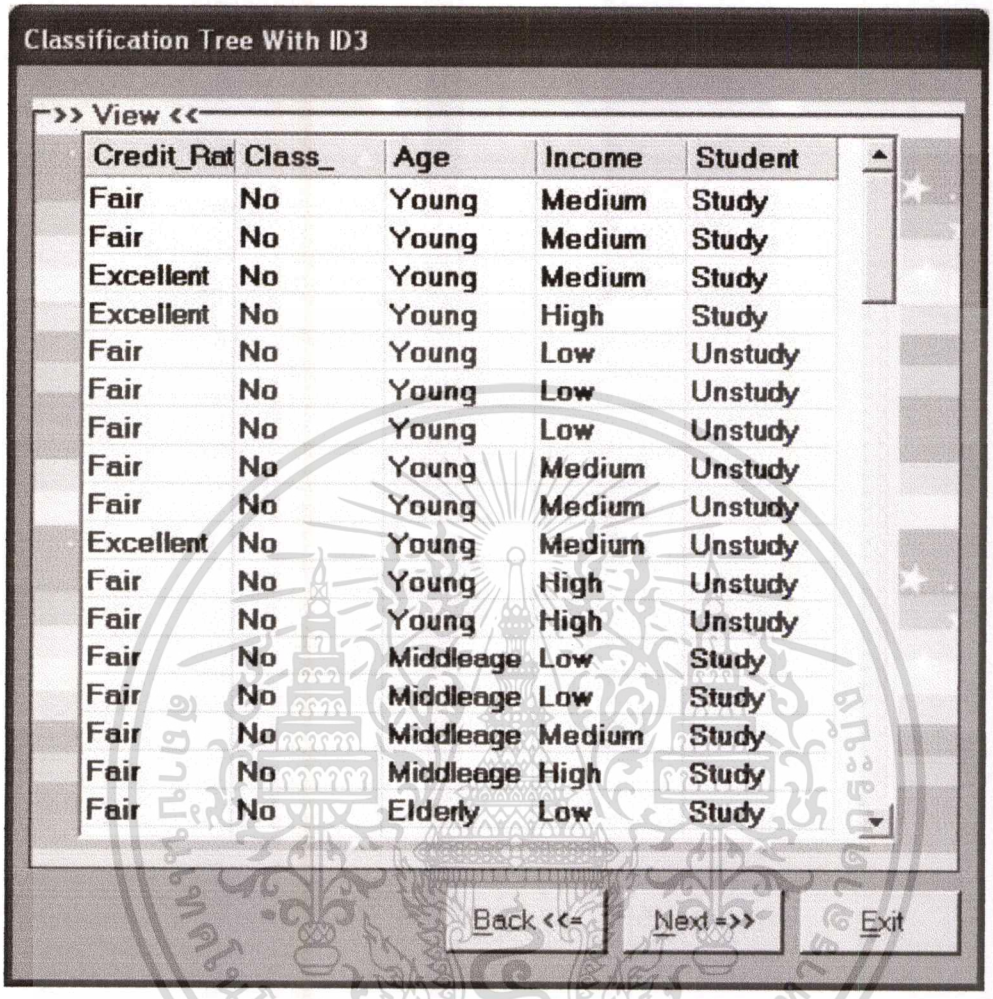
5.1.4 การเลือกใช้แอททริบิวที่ใช้ในไมนิ่ง



รูปที่ 5.4 หน้าจอเลือกแอททริบิวที่ใช้ในไมนิ่ง

- หน้าต่างนี้จะเป็นหน้าต่างการทำงานเพื่อเลือกแอททริบิวที่จำเป็นในการทำไมนิ่งโดยผู้ใช้สามารถเลือกมาได้เท่าที่ต้องการ อาจเลือกมาทีละ 1 แอททริบิว (>) หรือเลือกมาทั้งหมดเลยในครั้งเดียวก็ได้ (>>) หากต้องการย้ายออกไม่เลือกก็สามารถกด (<) ย้ายมาทีละ 1 หรือ (<<) ย้ายทีเดียวหมดเลยก็ได้
- เมื่อเลือกเสร็จแล้วจึงกดปุ่ม Next =>> เพื่อไปทำงานยังส่วนต่อไป
- หรือกด Back <<= กลับไปแก้สิ่งที่เลือกในเฟสก่อนหน้าได้
- หากกด Exit ก็จะสามารถออกจากโปรแกรมได้เลย

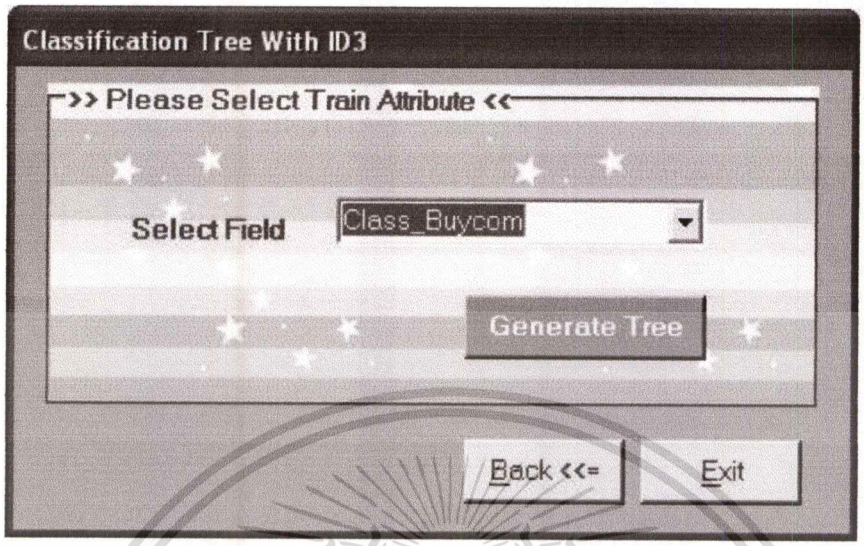
5.1.5 แสดงข้อมูลที่ทำการเลือกมา



รูปที่ 5.5 หน้าจอแสดงข้อมูลที่ทำการเลือกมา

- หน้าต่างนี้จะเป็นหน้าต่างแสดงข้อมูลทั้งหมดที่ได้ทำการเลือกมา ผู้ใช้สามารถตรวจสอบข้อมูลที่ตนเองเลือกขึ้นมาได้จากหน้านี้
- หากต้องการแก้ไขกด Back <<=
- หากต้องการออกจากโปรแกรมกด Exit
- หากข้อมูลถูกต้อง ต้องการทำงานต่อไปกด Next =>>
- หากกด Exit ก็จะสามารถออกจากโปรแกรมได้เลย

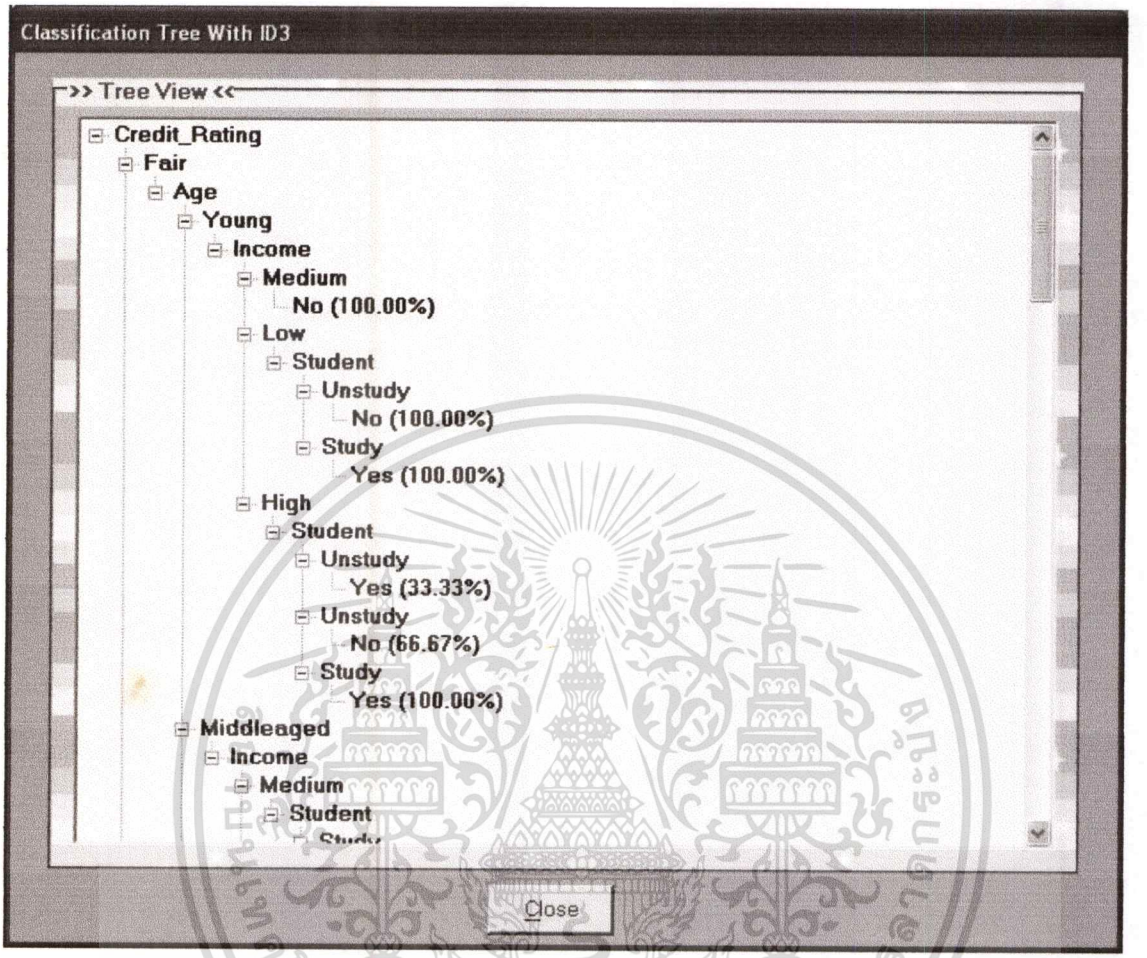
5.1.6 การเลือกแอททริบิวเป้าหมาย



รูปที่ 5.6 หน้าจอแสดงข้อมูลที่ทำการเลือกมา

- หน้าต่างนี้จะเป็นหน้าต่างเพื่อให้ผู้ใช้เลือกแอททริบิวเป้าหมายก่อนเข้ากระบวนการไมนิ่งด้วยอัลกอริธึม ID3
- ผู้ใช้จะทำการเลือกแอททริบิวเป้าหมายจาก drop down list box
- เมื่อทำการเลือกข้อมูลเสร็จแล้วให้กด Generate Tree เพื่อสร้างต้นไม้การตัดสินใจ หลังจากที่โปรแกรมทำงานเสร็จแล้วจะเปิดหน้าต่างอัตโนมัติมาอีก 1 บานเพื่อแสดงผลลัพธ์ต้นไม้การตัดสินใจและค่าความถูกต้อง
- หากกด Back จะกลับไปยังหน้าก่อนนี้เพื่อแก้ไขข้อมูลได้
- หากกด Exit ก็จะสามารถออกจากโปรแกรมได้เลย

5.1.7 การแสดงต้นไม้ประกอบการตัดสินใจ



รูปที่ 5.7 หน้าจอแสดงต้นไม้การตัดสินใจและค่าความถูกต้อง

- หน้าต่างนี้จะเป็นหน้าต่างแสดงต้นไม้การตัดสินใจและค่าความถูกต้อง
- หากกด Close จะปิดหน้าต่างแสดงต้นไม้การตัดสินจไป

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

โครงงานการพัฒนาระบบการจัดหมวดหมู่ของข้อมูล (Classification) ด้วยอัลกอรึธึม ID3 นั้น พัฒนาขึ้นมาเพื่อวิเคราะห์และพยากรณ์สิ่งที่คาดว่าน่าจะเป็นหรือน่าจะเกิดขึ้นจากฐานข้อมูล โครงงานนี้สามารถที่จะดึงเอาฐานข้อมูลที่ใช้อยู่ในปัจจุบันมาประยุกต์เพื่อเข้าใช้งานแล้วแปลงผลลัพธ์ให้อยู่ในรูปของต้นไม้การตัดสินใจที่เข้าใจได้ง่ายและเห็นถึงเปอร์เซ็นต์ความถูกต้องจากฐานข้อมูลนั้นๆ ได้ โดยผลที่ได้มานั้นผู้ใช้สามารถนำไปประยุกต์เพื่อประกอบการตัดสินใจในด้านต่าง ๆ ได้ ขึ้นอยู่กับว่าฐานข้อมูลที่เลือกมาใช้นั้นเป็นฐานข้อมูลเกี่ยวกับอะไร

จากตัวอย่างที่ได้ยกมาเป็นฐานข้อมูลเกี่ยวกับการซื้อคอมพิวเตอร์ ซึ่งประกอบไปด้วยแอททริบิวหลัก ๆ คือ อายุ, รายได้, สถานะการศึกษา, ระดับความน่าเชื่อถือของวงเงิน, ซื้อคอมพิวเตอร์ ผลของต้นไม้ที่ออกมาช่วยให้เห็นว่าปัจจัยที่สำคัญที่สุดของการซื้อคอมพิวเตอร์อยู่ที่แอททริบิวใด จากตัวอย่างนี้ก็ต้องตอบว่าอายุเป็นปัจจัยสำคัญในการซื้อคอมพิวเตอร์ ตาหากข้อมูลในฐานข้อมูลมีมากกว่านี้ หลากหลายกว่านี้ แน่นอนว่าข้อมูลของต้นไม้ก็จะเปลี่ยนไปด้วย

โครงงานนี้จึงสามารถนำไปปรับใช้กับฐานข้อมูลหลายด้านได้ และสามารถแบ่งกลุ่มของข้อมูลตามข้อมูลที่มีอยู่ในฐานข้อมูลได้เพื่อนำไปใช้ประโยชน์ต่อไป

6.2 ข้อเสนอแนะ

โครงงานนี้สามารถนำไปปรับให้สามารถทำงานให้ดียิ่งขึ้นได้อีก แนะนำได้ดังนี้

- โครงงานนี้สามารถทำงานกับตารางของฐานข้อมูลเพียง 1 ตาราง ซึ่งสามารถนำไปพัฒนาต่อให้รับค่าตารางให้เพิ่มขึ้นได้อีก
- โครงงานนี้ทำงานกับฐานข้อมูล Microsoft SQL Server เท่านั้น ซึ่งสามารถนำไปพัฒนาให้ปรับใช้กับฐานข้อมูลประเภทอื่น ๆ ได้อีก
- โครงงานนี้ใช้อัลกอริธึม ID3 ซึ่งทำงานได้แค่กับค่าที่เป็นตัวอักษรเท่านี้ หากนำไปพัฒนาเป็นอัลกอริธึมอื่น เช่น C4.5 ก็จะสามารถทำงานกับข้อมูลที่เป็นตัวเลขได้ด้วย

บรรณานุกรม

ณรงฤทธิ์ อนันต์ชัยพัทธนา.2545 "การค้นพบความรู้จากระบบสินค้าคงคลัง." **สัมนา 1 สารสนเทศลาดกระบัง.** 2(1) : 2-8

Daniel T. L. 2004. **Discovering Knowledge in Data**. United States of America. A John Wiley & Sons, INC.

Gary, B. S. et al. 2003. **System Analysis and Design**. United state of America. Thomson course technology.

Jiawei, H. and Kamber, M. 2001. **Data Mining Concepts And Techniques**. Morgan Kaufmann Publisher.

Peter C. et al. 1998. **Discovering Data Mining.** New Jersy. Prentice Hall PTR.

Richard, J. R. and Michael, W. G. 2003. **Data Mining a Tutorial – Based Primer.** United State of America. Addison Wesley.

Usama F. et al. 2002**. Information Visualization in Data Mining and Knowledge Discovery.** San Francisco. Morgan Kaufmann Publishers.

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวกรวิภา เกตุเรืองโรจน์
วันเดือนปีเกิด	10 มิถุนายน 2524
สถานที่เกิด	กรุงเทพมหานคร
ปริญญาตรี	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ คณะวิทยาศาสตร์และเทคโนโลยี สาขาวิทยาการคอมพิวเตอร์
ปีที่สำเร็จการศึกษา	2545