การพัฒนาระบบการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง

DEVELOPMENT OF DATA PREPARATION AND EXPLORATION FOR DATA MINING

โดย

อาทิตยา เชื้อจันอัด

ARTHITAYA CHUACHAN-AD

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 1 ปีการศึกษา 2549

DEVELOPMENT OF DATA PREPARATION AND EXPLORATION FOR DATA MINING

ARTHITAYA CHUACHAN-AD

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

1/ 2006

หัวข้อ	**การพัฒนาระบบการเตรียมข้อมูลและสำรวจ สำหรับการทำดาต้าไมนิ่ง**
นักศึกษา	**นางสาวอาทิตยา เชื้อจันอัด**
รหัสนักศึกษา	46066735
ปริญญา	**วิทยาศาสตรมหาบัณฑิต**
สาขาวิชา	**เทคโนโลยีสารสนเทศ**
แขนงวิชา	**วิทยาการสารสนเทศ**
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	**รศ.ดร.วรพจน์ กรีสุระเดช**

บทคัดย่อ

การดำเนินงานทางธุรกิจในยุคที่มีการแข่งขันกันสูง จำเป็นจะต้องมีการใช้เทคนิค กลยุทธ์ กลวิธีต่าง ๆ เพื่อช่วยให้การดำเนินธุรกิจอยู่เหนือคู่แข่งขัน จึงได้มีการนำเทคนิคดาต้าไมนิ่งมาใช้เพื่อวิเคราะห์ข้อมูลในฐานข้อมูลให้ได้สารสนเทศที่ซ่อนอยู่

กระบวนการที่มีบทบาทสำคัญต่อการทำดาต้าไมนิ่ง คือ การเตรียมข้อมูล เพื่อให้นำข้อมูลเข้าสู่อัลกอริทึมของดาต้าไมนิ่งได้ หากมีการเตรียมข้อมูลที่ไม่ดีแล้ว อาจทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งไม่มีคุณภาพด้วยเช่นกัน โครงการพัฒนาระบบงานการพัฒนาระบบการเตรียมข้อมูลและสำรวจสำหรับการทำดาต้าไมนิ่งนี้ ได้นำเทคนิคในการเตรียมข้อมูลมาใช้เพื่อปรับปรุงคุณภาพของข้อมูลให้เหมาะสมที่จะนำเข้าสู่กระบวนการดาต้าไมนิ่งต่อไปได้

Title	Development of Data Preparation and Exploration for Data Mining
Student	Miss.Arthitaya Chuachan-ad
Student ID.	46066735
Degree	Master of Science
Programme	Information Science
Academic Year	2006
Advisor	Assoc. Prof. Dr.Worapoj Kreesuradej

ABSTRACT

Doing business in an era of high competitiveness, it is necessary to using techniques, strategies and artifices helps to improve business gaining advantage over competitors. This is the initiative to manipulate data mining techniques for database analysis as provide concealed information.

Data preparation is held as the key to successful data mining. In order to access data through data mining algorithms, if data preparation is defective; data mining results maybe non-qualify. Development of data preparation and exploration for data mining system is applied data preparation techniques to provide data streams of suitable quality for data mining process.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ รศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงาน ที่ได้กรุณาให้ความรู้ ให้คำปรึกษาและคำแนะนำทางเทคนิคต่างๆ ที่เป็นประโยชน์ต่อการพัฒนาระบบ และสละเวลาในการ ตรวจสอบแก้ไขข้อบกพร่องของโครงการนี้

ขอกราบขอบพระคุณมารดาที่ให้โอกาสทางการศึกษากับข้าพเจ้า และเป็นกำลังใจหลักในการทำงานครั้งนี้ และขอบคุณทุก ๆ กำลังใจจากคนในครอบครัวที่ทำให้การพัฒนาระบบงานชิ้นนี้บรรลุผลสำเร็จได้เป็นอย่างดี

ขอบคุณเพื่อนๆ IS16.1 และ IT06 ที่เป็นกำลังใจและเป็นที่ปรึกษาในการพัฒนาระบบงานนี้

ขอบคุณคณาจารย์คณะเทคโนโลยีสารสนเทศที่ได้ประสิทธิประสาทวิชาความรู้ให้

ท้ายที่สุดนี้ คุณความดีและกุศลที่พึงบังเกิดมีจากโครงการพัฒนาระบบนี้ ข้าพเจ้าขออุทิศให้แก่นายชาญชัย เชื้อจันอัด ที่หยิบยื่นโอกาสทางการศึกษาให้กับข้าพเจ้า และสอนข้าพเจ้าว่าโอกาสทางการศึกษาไม่ได้มีกันทุกคน และอย่าละทิ้งโอกาสนั้น

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

รูปที่ หน้า

สารบัญรูป(ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมา

การทำดาต้าไมนิ่ง(Data mining) ให้มีประสิทธิภาพนั้นเวลาส่วนหนึ่งต้องถูกนำมาจัดการกับกระบวนการในการเตรียมข้อมูล (Data Preparation) เพื่อทำการจัดการกับข้อมูลเหล่านั้นให้นำเข้าสู่กระบวนการของดาต้าไมนิ่งได้ หากข้อมูลที่จะนำมาวิเคราะห์ด้วยเทคนิคดาต้าไมนิ่งเป็นข้อมูลที่ไม่สมบูรณ์ จะส่งผลให้ผลลัพธ์ที่ได้จากการวิเคราะห์ในกระบวนการไมนิ่งไม่มีประสิทธิภาพดีพอต่อองค์กร

ข้อมูลที่นำมาใช้ในกระบวนการวิเคราะห์อาจเป็นข้อมูลที่รวบรวมมาจากหลายแหล่งข้อมูล ส่งผลให้ข้อมูลเหล่านั้นมีความซ้ำซ้อน ขาดความเป็นมาตรฐานเดียวกัน ค่าของข้อมูลในแต่ละแอตทริบิวมีความหลากหลาย มีขอบเขตที่กว้างเกินไป เป็นต้น ซึ่งเทคนิคในการเตรียมข้อมูล จะช่วยเพิ่มคุณภาพของข้อมูล ความถูกต้องแม่นยำ และประสิทธิภาพในกระบวนการไมนิ่ง

1.2 วัตถุประสงค์

โครงการพัฒนาระบบงานเรื่องการพัฒนาระบบการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง มีวัตถุประสงค์คือ ลดระยะเวลาในการเตรียมข้อมูลในการทำดาต้าไมนิ่ง และพัฒนาโปรแกรมเพื่อใช้เป็นเครื่องมือช่วยในการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง ซึ่งทำให้ผลลัพธ์ที่ได้จากการเตรียมข้อมูลมีคุณภาพกว่าการเตรียมข้อมูลจากการแทนค่าข้อมูลโดยไม่มีขอบเขตหรือใช้วิธีการแทนค่าข้อมูลแบบสุ่มเดา ช่วยเพิ่มมาตรฐานให้กับข้อมูลในฐานข้อมูลที่ไม่สมบูรณ์ ซึ่งจะส่งผลให้ผลลัพธ์ของการทำดาต้าไมนิ่งมีประสิทธิภาพดีเพิ่มขึ้น

1.3 ขอบเขตการดำเนินงาน

1. ข้อมูลที่จะนำมาใช้ต้องเป็นข้อมูลที่จัดเก็บใน Microsoft SQL Server 2000 เท่านั้น หากเป็นฐานข้อมูลที่เก็บอยู่ใน DBMS(Database Management System) ระบบอื่นจะต้องทำการนำข้อมูลเข้า(Import) ผ่านทาง DTS Import/Export Wizard ของ Microsoft SQL Server 2000
2. โปรแกรมพัฒนาระบบนี้สามารถเลือกข้อมูลจากหลายตารางภายในฐานข้อมูลเดียวกันได้ โดยผู้ใช้ระบบจะต้องทราบความสัมพันธ์ของข้อมูลในแต่ละตาราง และสามารถใช้คำสั่ง SQL พื้นฐานในการเชื่อมความสัมพันธ์เหล่านั้น เพื่อเลือกข้อมูล(Data Selection) มาสร้างตารางใหม่ตามรูปแบบของโปรแกรม

3. ขั้นตอนการปรับแต่งข้อมูลที่ใช้ทำดาต้าไมนิ่ง(Data Cleaning) จะทำการกับเรคคอร์ดที่มีค่าว่าง (Null)
4. ขั้นตอนการแปลงข้อมูล (Data Transformation) สามารถแปลงค่าที่เป็น Numericให้เป็น Categorical และแปลงค่าที่เป็น Categorical ให้เป็น Numeric ได้
5. แสดงการสำรวจข้อมูลในรูปแบบของกราฟแท่ง(Bar Chart) และกราฟวงกลม(Pie Chart) ได้

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการดำเนินงานไว้ ดังนี้

1. ศึกษาทฤษฎีที่เกี่ยวข้องกับการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง (Data Preparation for Data Mining)
2. กำหนดวัตถุประสงค์ในการพัฒนาระบบ
3. ออกแบบระบบเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง
4. พัฒนาระบบเตรียมข้อมูลเพื่อการทำดาต้าไมนิ่ง
5. ทดสอบการใช้งานระบบ
6. สรุปผลการศึกษาและข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถลดระยะเวลาในการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง
2. ได้เครื่องมือในการเตรียมข้อมูลที่มีคุณภาพ และข้อมูลมีความถูกต้องเหมาะสมต่อการนำไปใช้งาน
3. ได้เครื่องมือในการเตรียมข้อมูลที่สามารถเตรียมข้อมูลได้หลากหลาย ให้เหมาะกับแต่ละอัลกอริทึมของดาต้าไมนิ่งที่ต้องการข้อมูลในการวิเคราะห์แตกต่างกัน

1.6 เครื่องมือที่ใช้ในการพัฒนาระบบ

- Visual Basic.NET หรือ VB.NET เป็นเครื่องมือที่ใช้พัฒนาโปรแกรม Visual Programming บนระบบปฏิบัติการ Windows ซึ่งได้รับการพัฒนามาจากภาษา BASIC (Beginners All Purpose Symbolic Instruction Code) ซึ่งเป็นภาษาโปรแกรมที่ได้รับความนิยมอย่างแพร่หลายสำหรับผู้เริ่มต้นหัดเขียนโปรแกรมคอมพิวเตอร์ เนื่องจากภาษา BASIC เป็นโปรแกรมที่สามารถทำความเข้าใจได้ง่าย

- Microsoft SQL Server 2000 เป็นโปรแกรมการจัดการฐานข้อมูลในตระกูล Micrọsoft ถูกพัฒนาขึ้นภายใต้การใช้ภาษา SQL ที่เป็นสากล

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่งเป็นวิธีการที่ใช้ในการวิเคราะห์ข้อมูลจำนวนมากเพื่อหาแนวโน้มหรือความสัมพันธ์ของข้อมูลที่มีอยู่ข้อมูลที่ได้มาจากทำดาต้าไมนิ่งไม่ได้เกิดจากการคาดคะเนหรือจากการสมมติฐานแต่เป็นข้อมูลที่มีความสัมพันธ์ที่ซ่อนอยู่ภายใต้ข้อมูลที่เรามีอยู่ดังนั้นในการทำดาต้าไมนิ่งจึงไม่ได้เป็นการตั้งสมมติฐานแต่เป็นการดูผลลัพธ์ที่ได้จากการทำงานมากกว่าจะเห็นได้ว่าการทำดาต้าไมนิ่งนั้นเป็นวิธีการที่แตกต่างไปจากวิธีการวิเคราะห์ข้อมูลทางสถิติในแบบอื่นๆในการทำดาต้าไมนิ่งนั้นผลลัพธ์ที่เกิดขึ้นถือได้ว่าเป็นข้อมูลที่มีประโยชน์เป็นอย่างมากโดยสามารถที่จะนำข้อมูลเหล่านี้ไปใช้เป็นแนวทางในการตัดสินใจที่ก่อให้เกิดผลดีในการทำธุรกิจ

โดยทั่วไปแล้วในการทำ ดาต้าไมนิ่งนั้นมีด้วยกันอยู่สองบรรทัดฐานด้วยกัน คือ การค้นหาความรู้ (Knowledge Discovery: KD) และการสร้างแบบจำลองการคาดการณ์ (Predictive Modeling: PM) ในทางปฏิบัติแล้วจะทำการประยุกต์ใช้ AI หรือ เทคโนโลยีในการเรียนรู้ ของเครื่องจักร ในการวิเคราะห์ข้อมูลในฐานข้อมูลขนาดใหญ่ จุดประสงค์ของทั้งสองบรรทัดฐานนี้ก็คือ พยายามที่จะสร้างกระบวนการที่เป็นแบบอัตโนมัติให้มากที่ สุดเท่าที่จะเป็นไปได้ ซึ่งในทางปฏิบัติแล้ว การทำดาต้าไมนิ่งไม่ใช่ระบบอัตโนมัติอย่างสมบูรณ์ทั้งหมด แต่เป็นกระบวนการแบบกึ่งอัตโนมัติเท่านั้น

ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ใช้งานกันอย่างกว้างขวางเป็นผลเนื่องมาจาก

1. ขนาดของข้อมูลมีขนาดใหญ่และขยายตัวอย่างรวดเร็ว การสืบค้นข้อมูลจะมีประโยชน์ก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก

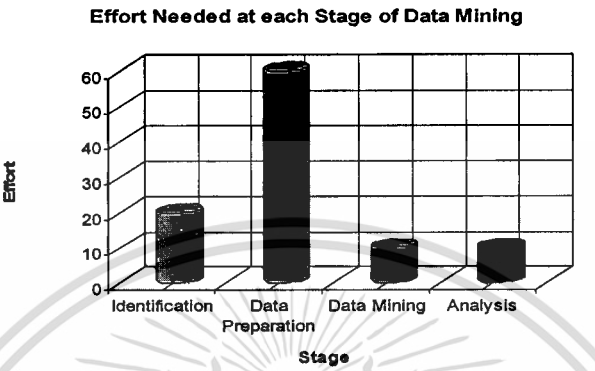
2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบสนับสนุนการตัดสินใจ เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์ ส่วนมากข้อมูลจะถูกจัดเก็บอยู่ในรูปของ Data Warehouse ซึ่งเป็นการง่ายต่อการนำไปใช้ในการสืบค้นความรู้

3. เทคนิคดาต้าไมนิ่งประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนจึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงด้วย ซึ่งในปัจจุบันระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงในท้องตลาดก็มีราคาถูกลงมาก จึงเป็นสาเหตุให้มีการนิยมใช้ดาต้าไมนิ่งกันมากขึ้น

4. ในวงการธุรกิจมีการแข่งขันกันสูง จึงมีข้อมูลเกิดขึ้นเป็นจำนวนมากแต่ไม่มีการนำมาใช้ให้เกิดประโยชน์ จึงมีความจำเป็นที่จะต้องนำเทคนิคดาต้าไมนิ่งมาใช้เพื่อให้ได้ความรู้นั้น

2.2 ขั้นตอนการทำดาต้าไมนิ่ง

กระบวนการของการทำดาต้าไมนิ่งประกอบไปด้วยขั้นตอนหลักที่สามารถแบ่งได้ 5 ขั้นตอน ดังนี้



รูปที่ 2.1 แสดงขั้นตอนในการทำดาต้าไมนิ่ง

ขั้นตอนที่1: กำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจจะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะจะเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่งและสามารถกำหนดได้ว่าเมื่อไหร่จะใช้ดาต้าไมนิ่งในการแก้ปัญหา เนื่องจากในทุกปัญหาไม่สามารถแก้ไขได้ด้วยหลักการดาต้าไมนิ่งทั้งหมด ซึ่งในส่วนนี้จะประกอบไปด้วยการวิเคราะห์ทางธุรกิจและวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลอะไรและต้องการอะไรจากข้อมูลที่มีอยู่

ขั้นตอนที่ 2: การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ต้องใช้เวลานานที่สุดประมาณ 60% ของการทำดาต้าไมนิ่ง เพราะเป็นส่วนที่มีความสำคัญที่สุดในการทำดาต้าไมนิ่ง เนื่องจากข้อมูลที่นำมาใช้ในการทำดาต้าไมนิ่งเป็นข้อมูลที่มาจากฐานข้อมูลขนาดใหญ่ที่อาจมาจากฐานข้อมูลหลายๆแหล่งมารวมกัน ข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความถูกต้องเพื่อส่งผลให้ผลลัพธ์ในการทำไมนิ่งมีประสิทธิภาพ รายละเอียดของการเตรียมข้อมูลอยู่ในบทที่ 3 การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

ขั้นตอนที่ 3: การทำดาต้าไมนิ่ง (Data Mining)

เป็นขั้นตอนการประมวลผลข้อมูลตามอัลกอริทึมที่กำหนดไว้ ซึ่งเกี่ยวข้องกับการเลือกอัลกอริทึมในการทำดาต้าไมนิ่งซึ่งจะต้องพิจารณาลักษณะของปัญหาเป็นหลัก เพราะในแต่ละปัญหาต้องเลือกใช้อัลกอริทึมที่เหมาะสมจึงจะได้ผลการวิเคราะห์ที่ถูกต้อง ซึ่งอาจใช้หลายอัลกอริทึมเพื่อเปรียบเทียบผลลัพธ์ได้

ขั้นตอนที่ 4: การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง (Analysis of Result)

เป็นการวิเคราะห์และตีความหมายจากผลที่ได้ เช่น ศึกษาพฤติกรรมของลูกค้าไม่ให้ผลออกมาตรงๆ แต่จะได้ความสัมพันธ์จำนวนมาก ผู้ใช้ต้องนำมาวิเคราะห์และประเมินกฎเหล่านี้เอง ตัวอย่างเช่น การแบ่งส่วนข้อมูล ผลที่ได้จะรู้ข้อมูลกลุ่มไหนๆแต่ต้องวิเคราะห์เองว่าแต่ละกลุ่มหมายถึงอะไร ซึ่งวิธีการนี้ผลที่ได้จะแปลความหมายยากและใช้เวลานาน วิธีเลือกสิ่งที่น่าสนใจจากผลลัพธ์ของดาต้าไมนิ่งเป็นวิธีที่วัดว่าผลที่ได้นั้นน่าสนใจแค่ไหนจะเลือกโดยดูจากผลที่ได้ง่ายต่อความเข้าใจ และ เป็นสารสนเทศที่ใหม่ สมเหตุสมผล

ขั้นตอนที่ 5: การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of Knowledge)

การนำความรู้ที่ได้ไปใช้เป็นขั้นตอนสุดท้ายของกระบวนการทั้งหมด ซึ่งเป็นการรวบรวมความเข้าใจในแบบจำลองที่เป็นผลมาจากขั้นตอนการวิเคราะห์ผลลัพธ์ที่ได้ มารวมเข้ากับส่วนความรู้ทางธุรกิจเพื่อที่จะนำเสนอถึงวิธีการที่จะนำผลที่ได้นี้ไปใช้ให้เกิดประโยชน์ ในขั้นตอนนี้จะมีหลักอยู่ 2 ประการคือ

1. แสดงแนวคิดทางธุรกิจที่ค้นพบใหม่
2. กฎเกณฑ์ที่จะใช้ความรู้ใหม่ที่พบให้ได้ประโยชน์สูงสุด

2.3 เทคนิคการทำดาต้าไมนิ่ง

ดาต้าไมนิ่งมีเทคนิคและอัลกอริทึมที่สามารถนำมาใช้งานหลายประเภท ขึ้นอยู่กับเอพลิเคชั่น (Application) ที่ต้องการนำมาใช้งาน แบ่งออกเป็นรูปแบบต่าง ๆ ได้ดังตารางที่ 2.1

ตารางที่ 2.1 เทคนิคของดาต้าไมนิ่ง

Predictive Modeling	Classification
	Value Prediction
Database Segmentation	Demographic Clustering
	Neural Clustering
Link Analysis	Associations Discovery
	Sequential Pattern Discovery
Deviation Detection	Visualization
	Statistics

2.3.1 การสร้างแบบจำลองพยากรณ์ (Predictive Modeling)

เป็นการทำนายความเป็นไปได้ โดยใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ คือจะใช้วิธีนี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อตัดสินใจเลือกลักษณะข้อมูลที่ต้องการ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้ แล้วจึงนำไปวิเคราะห์กลุ่มข้อมูลที่ต้องการ ซึ่งวิธีนี้เรียกว่า Supervised Learning ดังนั้นข้อมูลที่มีอยู่ต้องสมบูรณ์ จึงจะทำให้ผลลัพธ์ออกมาถูกต้อง เพราะเราต้องนำข้อมูลในอดีตมาสร้างแบบจำลอง การทำงานจะแบ่งออกเป็น 2 ขั้นตอน คือ

Training Phase คือขั้นตอนการสร้างแบบจำลองขึ้นมาโดยใช้ข้อมูลในอดีต ซึ่งจะใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมด

Testing Phase คือขั้นตอนที่ใช้ทำการทดสอบแบบจำลองที่สร้างว่ามีความเหมาะสมหรือไม่โดยจะนำข้อมูลส่วนที่เหลือ 20% มาใช้ทดสอบแบบจำลองที่สร้างขึ้น

การสร้างแบบจำลองพยากรณ์ สามารถแบ่งย่อยได้อีก เป็น 2 ประเภท คือ

- Classification เป็นการทำนายว่าสิ่งนั้นควรอยู่ในกลุ่มไหน ซึ่งเป็นการแบ่งกลุ่มของข้อมูลตามชนิดของกลุ่มข้อมูลที่ควรจะเป็น และสามารถแบ่งกลุ่มข้อมูลได้อย่างชัดเจน เช่น การจัดกลุ่มของลูกค้าเพื่อพิจารณาว่าควรจะให้วงเงินสินเชื่อเพิ่มขึ้นหรือไม่ เป็นต้น ซึ่งวิธีที่นิยมใช้คือ Tree Induction และ Neural Induction

- Value Prediction เป็นการทำนายถึง ค่าความต่อเนื่องของข้อมูล เป็นการทำนายค่าที่เป็น Numeric เช่น การทำนายราคาหุ้น เป็นต้น โดยมีวิธีที่ใช้คือ Linear Regression และ Nonlinear Regression

2.3.2 การแบ่งส่วนฐานข้อมูล (Database Segmentation)

จะเป็นการแบ่งหรือจัดกลุ่มของข้อมูลที่มีลักษณะคล้ายกัน หรือมีคุณสมบัติใกล้เคียงกันในหลายๆ ด้าน ให้เป็นข้อมูลกลุ่มเดียวกัน ซึ่งแต่ละกลุ่มจะถูกเรียกว่าเซกเมนต์ (Segments) หรือคลัสเตอร์ (Clusters) การแบ่งกลุ่มข้อมูลนี้เราจะไม่สามารถกำหนดได้ว่าข้อมูลใดควรจะอยู่กลุ่มใด แต่จะเป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเอง ไม่ได้ใช้ความรู้สึกหรือประสบการณ์ในการตัดสินใจแบ่งกลุ่มข้อมูล และข้อมูลจะถูกจัดการโดยอัลกอลิทึมที่เหมาะสม จึงเรียกว่าเป็นรูปแบบของ Unsupervised Learning ซึ่งสามารถแบ่งย่อยตามวิธีที่ใช้ เช่น Demographic Clustering และ Neural Clustering ยกตัวอย่าง เช่น บริษัทจำหน่ายรถยนต์แห่งหนึ่งได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1. กลุ่มผู้มีรายได้สูง (>$80,000)
2. กลุ่มผู้มีรายได้ปานกลาง ($25,000 to $ 80,000)
3. กลุ่มผู้มีรายได้ต่ำ (less than $25,000)

บทที่ 3

การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

เป้าหมายในการเตรียมข้อมูล เพื่อทำให้ข้อมูลมีความเหมาะสมกับอัลกอริทึมและเพื่อเพิ่มประสิทธิภาพในการทำดาต้าไมนิ่ง ดังนั้นจะเห็นได้ว่ากระบวนการเตรียมข้อมูลมีความสำคัญเป็นอย่างมากและเวลาส่วนใหญ่จึงถูกใช้ไปในกระบวนการนี้

3.1 การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

เป็นขั้นตอนในการทำข้อมูลดิบที่เราได้รับมา ซึ่งจะอยู่ในรูปแบบที่หลากหลายแตกต่างกันไปให้อยู่ในรูปที่พร้อมจะใช้งาน เพื่อให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมีความถูกต้องแม่นยำมากยิ่งขึ้น เป็นขั้นตอนที่ใช้เวลานาน เนื่องจากปริมาณข้อมูลมีเป็นจำนวนมากและข้อมูลที่ได้รับมาจากหลายแหล่ง รูปแบบของข้อมูลจึงแตกต่างกัน จึงต้องมีการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกันเพื่อให้พร้อมใช้งาน โดยขั้นตอนในการเตรียมข้อมูลนี้แบ่งออกเป็น 3 ขั้นตอน ได้ดังนี้

1. การเลือกข้อมูล: Data Selection
2. การเตรียมข้อมูล: Data Preprocessing
3. การแปลงข้อมูล: Data Transformation

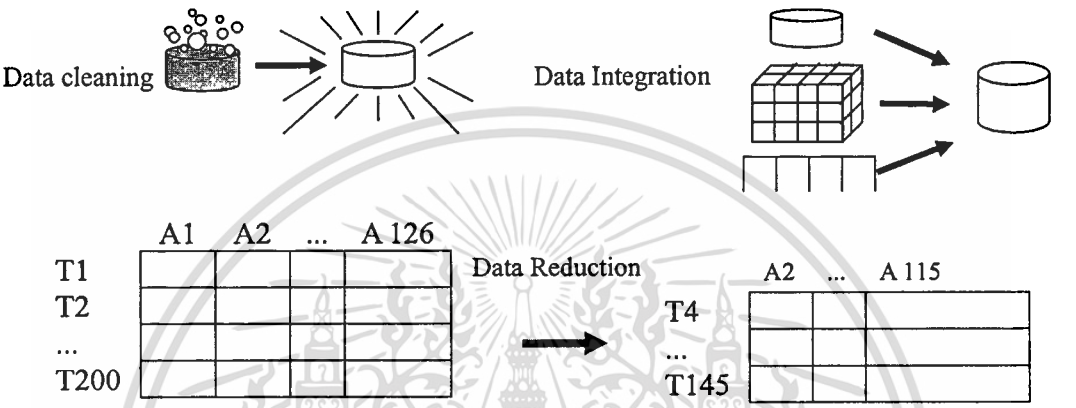
3.2 การเลือกข้อมูล: Data Selection

การเลือกข้อมูล เป็นการเลือกข้อมูลที่สำคัญออกมาจากฐานข้อมูลเพื่อทำการวิเคราะห์ในขั้นตอนต่อไป ข้อมูลที่นำมาวิเคราะห์นั้นต้องขึ้นอยู่กับวัตถุประสงค์ทางธุรกิจขององค์กรที่ได้กำหนดไว้ การเลือกข้อมูลจำเป็นจะต้องเข้าใจความหมายประเภทข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบและลักษณะอื่นๆ ของข้อมูล โดยแบ่งลักษณะของข้อมูลได้เป็น 2 ลักษณะ คือ

- ข้อมูลแบบแบ่งประเภท (Categorical)
    - Nominal: ตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น เพศ (ชาย, หญิง)
    - Ordinal: ตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น ลำดับของสินค้า (ดี, ไม่ดี)
- ข้อมูลแบบปริมาณ (Quantitative)
    - Continuous: ค่าที่เก็บเป็นเลขจำนวนจริง หรือเป็นค่าต่อเนื่อง เช่น จำนวนเงิน
    - Discrete: ค่าที่เก็บเป็นเลขจำนวนเต็ม เช่น จำนวนบุตร

3.3 การเตรียมข้อมูล: Data Preprocessing

วัตถุประสงค์ของการเตรียมข้อมูล คือ การแก้ไขปัญหาที่พบในข้อมูลเพื่อทำให้ข้อมูลมีคุณภาพก่อนที่จะนำข้อมูลไปประมวลผล ซึ่งในขั้นตอนนี้จะแบ่งออกเป็น 3 ขั้นตอนย่อย ดังนี้ Data cleaning ที่ใช้ขจัดข้อมูลที่มีค่าผิดจากค่าที่ควรจะเป็น (noise) และขจัดข้อมูลที่ขัดแย้งกัน (inconsistencies), Data integration ใช้เพื่อรวมข้อมูลจากหลายๆ แหล่งให้เป็นข้อมูลก้อนเดียวกัน, Data Reduction ทำการลดขนาดของข้อมูลที่จะใช้ทำไมนิ่ง



รูปที่ 3.1 รูปแบบของการเตรียมข้อมูล

3.3.1 Data Cleaning

เป็นขั้นตอนในการเลือกข้อมูลที่ต้องการและเอาข้อมูลที่ไม่ต้องการออกจากแหล่งข้อมูล ซึ่งข้อมูลส่วนใหญ่ในฐานข้อมูลนั้นมักจะไม่สมบูรณ์ (Incomplete) โดยการเติมข้อมูลใหม่แทนข้อมูลเดิมที่ขาดหาย (missing values), ขจัดข้อมูลที่ผิดจากค่าที่ควรจะเป็น (noisy data) หรือ ลบข้อมูลที่อยู่นอกเหนือขอบเขตของข้อมูล (remove outliers), ขจัดปัญหาข้อมูลที่ขัดแย้งกัน (resolve inconsistencies) ซึ่งวิธีการของ Data Cleaning มีดังนี้

- เติมค่าที่ขาดหายไป (Missing Value)
- กำหนดค่าที่เกินขอบเขต เพื่อลดความคลาดเคลื่อน (Noisy Data)
- ทำให้ข้อมูลสอดคล้องตรงกัน

3.3.1.1 Missing Value

ข้อมูลที่ขาดหาย ไม่มีข้อมูลในบางแอทริบิว ซึ่งอาจเกิดจากการบันทึกข้อมูลผิดพลาด หรือกรอกข้อมูลไม่ครบถ้วน วิธีการในการเติมข้อมูลให้กับค่าที่ขาดหายไปนั้น มีดังนี้

1. Ignore the tuple: โดยไม่สนใจทิ้งข้อมูลแถว (record) นั้นไปเลย ซึ่งเป็นวิธีที่ไม่มีคุณภาพนัก นอกเสียจากว่าข้อมูลในแถวนั้นมีแอทริบิวว่างเป็นจำนวนมาก

2. Fill in the missing value manually: วิธีการนี้คือเติมค่าของแอทริบิวลงไปเอง ซึ่งเป็นวิธีที่ไม่เหมาะกับฐานข้อมูลขนาดใหญ่ๆ ที่มีข้อมูลขาดหายเป็นจำนวนมาก

3. Use a global constant to fill in the missing value: วิธีการนี้จะเติมค่าคงที่หรือตัวแทนของค่าที่กำหนดขึ้นลงในแอทริบิวเช่น “Unknown” หรือ “∞” แต่การเติมข้อมูลเหล่านี้อาจส่งผลให้ผลจากการทำดาต้าไมนิ่งเกิดความผิดพลาดและไม่มีประสิทธิภาพได้

4. Use the attribute mean to fill in the missing value: คือการใช้ค่าเฉลี่ยของแอทริบิวนั้นมาเติมลงในค่าที่ขาดหายไป เช่น การนำค่าเฉลี่ยรวมของรายได้มาเติม

5. Use the attribute mean for all samples belonging to the same class as the given tuple: ตัวอย่างเช่น หากมีการจัดกลุ่มของลูกค้าตาม credit_risk แล้วก็ให้นำค่าเฉลี่ยรวมของรายได้ของลูกค้าในแต่ละกลุ่มเติมลงในค่าที่ขาดหายไปในลูกค้ากลุ่มเดียวกัน

6. Use the most probable value to fill in the missing value: โดยวิธีการเติมค่าที่คาดว่าจะเป็นไปได้ลงในแอทริบิวโดยใช้ Decision tree ในการทำนายค่าที่ขาดหายไป

3.3.1.2 Noisy Data

การเกิด Noise หมายถึงค่าของข้อมูลที่ผิดไปจากค่าที่ควรจะเป็น มีได้หลายทาง รวมไปถึงค่าที่ไม่ถูกต้องในบางแอทริบิวเช่นเกิด ข้อผิดพลาดระหว่างการรับส่งข้อมูลในเครือข่าย ข้อจำกัดทางเทคโนโลยี ขนาดของที่พักข้อมูล (buffer) ที่เล็กกว่าขนาดของข้อมูล การตั้งชื่อแอทริบิวที่ขัดแย้งกัน เป็นต้น ค่าเหล่านี้หากนำไปพิจารณาแล้วอาจส่งผลให้ผลลัพธ์ที่ได้จากการไมนิ่งเบี่ยงเบนไปจากที่ควรจะเป็นได้ ซึ่งเทคนิคในการขจัด noise มีดังนี้

1.) Binning: เป็นวิธีจัดหมวดหมู่ของข้อมูลโดยดูจากค่าใกล้เคียงของข้อมูล และในแต่ละกลุ่มก็จะแทนด้วยค่าเดียวกัน ดังตัวอย่าง

ข้อมูลเรียงจากน้อยไปมาก 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equidepth) bins: จัดเป็นกลุ่มกลุ่มละเท่าๆ กัน

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means: นำค่าเฉลี่ยของกลุ่มมาแทนค่า

Bin 1: 9, 9, 9

Bin 2: 21, 21, 21

Bin 3: 29, 29, 29

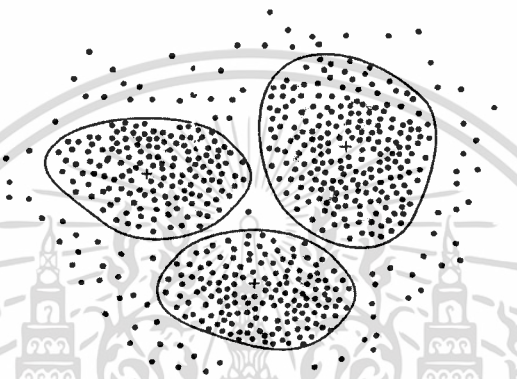
Smoothing by bin boundaries: นำค่าขอบของข้อมูลในกลุ่มมาแทนค่า

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

2.) Clustering: เป็นการแบ่งกลุ่มข้อมูลซึ่งข้อมูลที่มีลักษณะคล้ายกันจะอยู่ในกลุ่มเดียวกัน และข้อมูลที่มีลักษณะต่างกันจะอยู่ต่างกลุ่มออกไป ซึ่งค่าที่อยู่นอกกลุ่มนั้นเราจะเรียกว่า Outlier ซึ่งแสดงได้ดังรูปที่ 3.2



รูปที่ 3.2 ภาพแสดงผลที่ได้จากการทำ clustering

3.) Combined computer and human inspection: ใช้ทั้งคอมพิวเตอร์และคนในการหาความคลาดเคลื่อน

4.) Regression: ซึ่งอาจนำ Linear Regression หรือ Multiple Regression มาใช้ Regression นั้นเป็นการสร้างโมเดลและนำค่าที่ได้จากโมเดลมาใช้แทนค่าจริง ซึ่งการคำนวณค่า Regression มี 2 รูปแบบ ดังนี้

1. Linear Regression เป็นสมการเส้นตรงที่หาความสัมพันธ์ระหว่างตัวแปร2 ตัว โดยมีสูตรดังนี้

$$Y = \alpha + \beta X \tag{3.1}$$

โดย α, β : เป็นพารามิเตอร์

จากสูตรเราจะต้องหาว่า Y นั้นสัมพันธ์กับตัวแปรตัวใด จึงจะนำ Linear Regression มาใช้ได้

2. Multiple Regression เป็นการหาความสัมพันธ์ของสมการหลายตัวแปร โดยมีสูตรดังนี้

$$Y = b0 + b1X1 + b2X2 \tag{3.2}$$

3.3.2 Data Integration

ในการทำดาต้าไมนิ่ง จะต้องมีการทำ Data Integration เพื่อทำการรวมข้อมูลจากหลายแหล่งมาไว้รวมกัน และต้องมีการปรับข้อมูลให้เหมาะสมเพื่อใช้ในการทำดาต้าไมนิ่ง ซึ่งข้อมูลนั้นจะต้องถูกปรับให้อยู่ในรูปแบบที่เหมาะสมในการทำไมนิ่ง สิ่งที่ต้องคำนึงถึงก็คือ

- Schema integration คือการวางแผนในการรวมข้อมูล ปัญหาที่พบได้คือในการวิเคราะห์ข้อมูลทำอย่างไรโปรแกรมในการวิเคราะห์ข้อมูลจึงจะทราบว่า Customer_id กับ Cust_number ในแต่ละฐานข้อมูลหมายถึงแอทริบิวเดียวกัน

- การตรวจสอบและแก้ปัญหาความขัดแย้งกันของข้อมูล เช่น น้ำหนักหน่วยวัดของไทยเป็นกิโลกรัม แต่ของอเมริกาเป็นปอนด์ เป็นต้น

- ความซ้ำซ้อนของข้อมูล (Redundancy) เช่น ข้อมูลอาจเป็นข้อมูลตัวเดียวกัน แต่ชื่อแอทริบิวต่างกัน ซึ่งเราสามารถลดความซ้ำซ้อนของข้อมูลได้โดยใช้การวิเคราะห์ Correlation

การคำนวณ Correlation มีสูตรดังนี้

$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A \sigma_B} \tag{3.3}$$

n : จำนวนของ Tuple

$\bar{A}, \bar{B}$: ค่าเฉลี่ยของ A และ B โดยที่ $\bar{A} = \frac{\sum A}{n}$

σ_A, σ_B : ค่าความแปรปรวนมาตรฐานของ A และ B

โดยที่ $$\sigma_A = \sqrt{\frac{\sum (A-\bar{A})^2}{n-1}} \tag{3.4}$$

เราสามารถวิเคราะห์ผลได้ดังนี้

- ถ้าค่า r มากกว่า 0 แสดงว่าตัวแปร A และ B มีความสัมพันธ์กันในทางบวก ซึ่งหมายถึงถ้าค่าของตัวแปร A เพิ่มขึ้น ตัวแปร B จะมีค่าเพิ่มขึ้นด้วย ซึ่งแสดงว่าแอทริบิวนั้นมีนัยสำคัญต่อกัน ยิ่งค่า r มีค่าสูงแสดงว่าเราสามารถตัด A หรือ B ออกไปได้ เพราะข้อมูลอาจจะมีความซ้ำซ้อนกัน

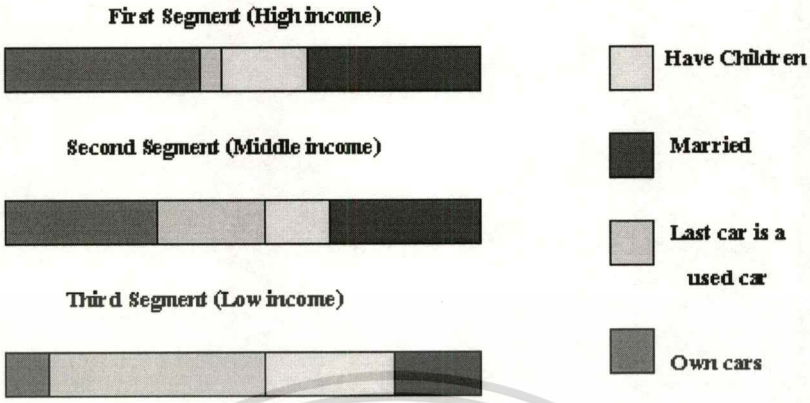
- ถ้าค่า r เท่ากับ 0 แสดงว่าตัวแปร A และ B นั้นเป็นอิสระกัน

- ถ้าค่า r น้อยกว่า 0 แสดงว่าตัวแปร A และ B มีความสัมพันธ์กันในทางลบ ซึ่งหมายถึงถ้าค่าของตัวแปร A เพิ่มขึ้น จะทำให้ค่าของตัวแปร B มีค่าลดลง

3.3.3 Data Reduction

หากมีการเลือกข้อมูลจาก Data Warehouse เพื่อทำการวิเคราะห์นั้นจะพบได้ว่าข้อมูลมีขนาดใหญ่มาก ซึ่งจะทำให้การวิเคราะห์ข้อมูลและการทำไมนิ่งนั้นใช้เวลานาน ส่งผลให้ผลลัพธ์

และภายในแต่ละกลุ่มยังแยกออกเป็น Have Children, Married, Last car, is a used car, Own cars



รูปที่ 2.2 การแยกกลุ่มลูกค้าของบริษัทรถยนต์แห่งหนึ่ง

จากข้อมูลข้างต้นทำให้ทางบริษัททราบว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

2.3.3 การวิเคราะห์ความสัมพันธ์ (Link Analysis)

เป็นการหาความสัมพันธ์ของข้อมูล เช่น ลูกค้าเข้าร้านซื้อสินค้าอะไรบ้าง, วันที่เข้าร้าน (มักจะนำดาต้าไมนิ่งไปใช้กับพวกธุรกิจค้าปลีก) ซึ่งจะมีเทคนิค ได้แก่

- Associations Discovery เป็นหลักการค้นหาสิ่งที่มีความสัมพันธ์กัน
- Sequential Pattern Discovery เป็นการศึกษาว่าเหตุการณ์ใดเกิดแล้วเหตุการณ์ใดจะเกิดตามมา เช่น การกู้จะกู้เพื่อการศึกษาก่อน จากนั้นแล้วจะกู้เพื่อแต่งงานเป็นต้น หาลำดับว่ามีรูปแบบ (Pattern) เหล่านี้ เช่น กู้ซื้อบ้านแล้วต้องกู้ซื้อรถด้วย เป็นต้น
- Similar Time Sequence Discovery เป็นการศึกษาพฤติกรรมของข้อมูลที่เกิดขึ้นทั้งหมดหรือเกิดขึ้นในช่วงเวลาเดียวกัน เพื่อหาความสัมพันธ์ระหว่างกลุ่มของข้อมูลเหล่านี้

2.3.4 การตรวจสอบค่าเบี่ยงเบน (Deviation Detection)

เป็นเทคนิคที่ใช้ทำการหาค่าที่มีความแตกต่างไปจากค่ามาตรฐานว่ามีค่ามากน้อยเพียงใด เป็นแบบจำลองที่ใช้เทคนิคทางสถิติ (Statistics) เพื่อใช้วัดความน่าเชื่อถือของข้อมูล และการแสดงให้เห็นภาพ (Visualization) ซึ่งเป็นการสรุปข้อมูลให้แสดงผลออกมาในรูปแบบกราฟิก เช่น แผนภูมิแท่ง หรือ แผนภูมิวงกลม เป็นต้น เพื่อให้สามารถเข้าใจได้ง่าย นอกจากนี้ยังสามารถนำไปใช้ร่วมกับเทคนิคอื่นๆ โดยใช้ในการแสดงผลที่ได้ในรูปแบบของกราฟิก

จากการวิเคราะห์ไม่มีประสิทธิภาพดีพอ เทคนิค Data Reduction ช่วยลดขนาดของ data set ลงทำให้การทำไมนิ่งกับข้อมูลที่ผ่านกระบวนการ Data Reduction มีประสิทธิภาพมากยิ่งขึ้น

Data cube aggregation:

เป็นวิธีในการลดขนาดของข้อมูลโดยการนำข้อมูลมารวมกัน (ทำ Data Aggregation) เช่น เรามีข้อมูลยอดขายสินค้าอิเลคทรอนิกส์ทุกประเภทต่อไตรมาส ในปี 1977-1999 แต่ในความเป็นจริงแล้ว เราจะสนใจยอดขายประจำปีมากกว่ารายละเอียดของยอดขายแต่ละไตรมาส ดังนั้นสามารถลดขนาดข้อมูลโดยการสรุปยอดการขายประจำปีจากผลรวมของแต่ละไตรมาส (ทำ data aggregation) ได้ดังรูปที่ 3.3

Year 1999

Year 1998

Year 1997

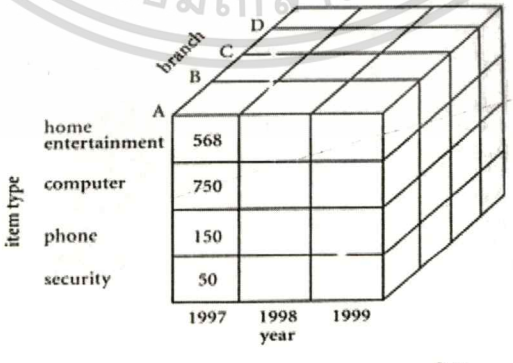
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

→

Year	Sales
1997	$1,568,000
1998	$2,356,000
1999	$3,594,000

รูปที่ 3.3 แสดงการทำ Data aggregated

จากรูปที่ 3.3 แสดงให้เห็นว่าขนาดของ Data Set ของข้อมูลจะมีขนาดเล็กลงแต่สารสนเทศที่จำเป็นในการใช้วิเคราะห์ยังอยู่ครบถ้วน นอกจากนี้เรายังสามารถสร้าง Data Cube ได้จากการนำ Multidimensional Aggregated Information หลายๆอันมารวมกัน ดังรูปที่ 3.4 ที่แสดงยอดขายสินค้าอิเลคทรอนิกส์ ทุกประเภท ในแต่ละปี ของแต่ละสาขา



รูปที่ 3.4 แสดง Data cube ของยอดขายของบริษัทแห่งหนึ่ง

Dimension reduction:

Heuristic methods สามารถลดขนาดของข้อมูลที่สนใจได้

1. Step-wise forward selection มีหลักการ คือ
   - กำหนดค่าเริ่มต้นของ Reduce Set ที่ต้องการให้เท่ากับ Set ว่าง { }
   - พิจารณาค่าแอทริบิวแต่ละตัวเพื่อหาค่าที่ดีที่สุดใส่ใน Reduce set แสดงได้นี้

     Initial attribute set:

     [A1, A2, A3, A4, A5, A6]

     Initial Reduce set:

     { }

     → {A1}

     → {A1, A4}

     → Reduce attribute set: {A1, A4, A6}

2. Step-wise backward elimination มีหลักการ คือ
   - กำหนดค่าเริ่มต้นของ Reduce Set ที่ต้องการให้มีจำนวนเท่ากับแอทริบิวทั้งหมดที่มี
   - พิจารณาค่าแอทริบิวแต่ละตัวเพื่อหาค่าที่แย่ที่สุด แล้วทำการดึงออกมาจาก Reduce Set แสดงได้ดังนี้

     Initial attribute set:

     [A1, A2, A3, A4, A5, A6]

     → {A1, A3, A4, A5, A6}

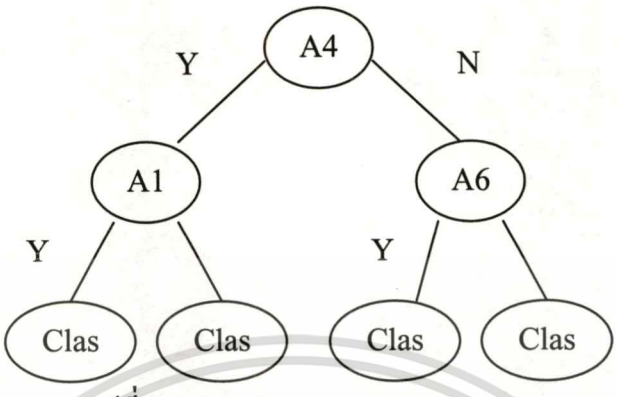
     → {A1, A4, A5, A6}

     → Reduce attribute set:

     → {A1, A4, A6}

3. Combining forward selection and backward elimination มีหลักการคือ ในแต่ละขั้นตอนจะทำการเลือกแอทริบิวที่ดีที่สุดไว้ แล้วดึงแอทริบิวที่แย่ที่สุดออกไป

Decision Tree Induction: เป็นวิธีในการลดขนาดของข้อมูลโดยนำแอทริบิวแต่ละตัวใส่ใน Decision Tree Algorithm จากนั้นตัวอัลกอริทึมทำการเลือกแอทริบิวที่ดีที่สุดเพื่อใส่ในกลุ่ม (Class) แต่ละกลุ่มเอง ดังนี้



รูปที่ 3.5 แสดง Decision Tree Induction

Initial attribute set: [A1, A2, A3, A4, A5, A6]

→ Reduce attribute set:

→ {A1, A4, A6}

Data compression: เป็นวิธีในการลดขนาดของข้อมูลโดยใช้การ encoding หรือ Transformation Data โดยข้อมูลที่ถูกนำมา Encoding หรือ Transformation จะเรียกว่า Compress Data เทคนิคในการทำ Data Compression แบ่งออกเป็น 2 แบบ ได้แก่

- Lossless เป็นวิธีการทำ Data Compression โดยเราสามารถ Reconstruction Original Data จาก Compress Data ได้โดยไม่มีสารสนเทศใดๆสูญหายไป

- Lossy เป็นวิธีการทำ Data Compression โดยเราสามารถใช้ Compress Data ทำ Reconstruction ค่าโดยประมาณของ Original Data เท่านั้น

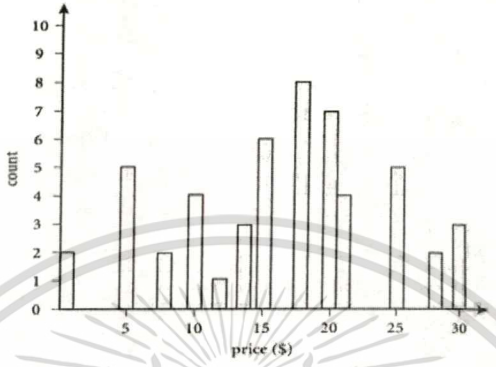
Numerosity reduction:

เป็นวิธีในการลดขนาดของข้อมูลโดยการเลือกข้อมูลขนาดเล็ก (Smaller Data) มาใช้แทนข้อมูลทั้งหมด แบ่งออกเป็น 2 วิธี ได้แก่

- Parameter model มีวิธีการคือ จะใช้ Model Parameter แทนข้อมูลจริง เช่น Regression and Log-Linear model
- Non Parameter method เช่น

Histograms เป็นเทคนิคการลดข้อมูลโดยแบ่งข้อมูลออกเป็นชุดๆเรียกว่า buckets ตัวอย่างเช่น มีจำนวนข้อมูลการขายสินค้าของทุกรายการ และราคาที่ใช้ในการขายสินค้า ดังนี้

1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30 สามารถนำมาลดขนาดโดยใช้ Histogram ได้ดังรูปที่ 3.6



รูปที่ 3.6 แสดง Histograms ของราคาสินค้าโดยใช้เทคนิค buckets

Clustering เป็นเทคนิคการลดข้อมูลโดยจะแบ่งข้อมูลออกเป็นกลุ่มๆเรียกว่า Cluster โดยข้อมูลแต่ละตัวใน Cluster เดียวกันจะมีคุณสมบัติคล้ายกัน แต่แตกต่างกันในคนละ Cluster

Sampling เป็นเทคนิคการลดข้อมูล โดยสุ่มเลือกข้อมูลกลุ่มเล็กๆขึ้นมาเพื่อใช้แทนข้อมูลกลุ่มใหญ่ แบ่งออกเป็น 4 วิธี ได้แก่

- Simple Random Sample Without Replacement (SRSWOR) ทำการเลือก Sample Data ออกมาโดยไม่มีการคืนเข้าไปยัง Data Set เดิม
- Simple Random Sample With Replacement (SRSWR) ทำการเลือก Sample Data ออกมาโดยสามารถใส่กลับคืนไปยัง Data Set เพื่อเลือกใหม่ได้
- Cluster sample จะทำการแบ่งข้อมูลออกเป็น Cluster ก่อน จากนั้นจึงทำการเลือก Sample Data แบบ SRSWOR จาก Cluster
- Stratified sample จะทำการแบ่งข้อมูลออกเป็นกลุ่มๆก่อน จากนั้นจะทำการเลือกข้อมูล Sample Data จากกลุ่มของข้อมูลที่เหมือนกัน
- ในการเลือก Sampling Data จะสามารถคำนวณจำนวนขนาดของ Sample Data ที่เลือกออกมาได้จากสูตรต่อไปนี้

$$n = \frac{N}{1 + N.e^2} \tag{3.5}$$

โดย n = จำนวน sample data

N = จำนวน data ใน data set ทั้งหมด

e = ขนาดของ error ที่จะยอมให้เกิดขึ้นได้

3.4 การแปลงข้อมูล: Data Transformation

การแปลงข้อมูลมีวัตถุประสงค์ 2 อย่างคือ ทำให้ไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำมาใช้ เนื่องจากข้อมูลที่จะนำมาใช้ทำดาต้าไมนิ่งในบางครั้งอยู่ในรูปแบบที่ไม่เหมาะสมกับอัลกอริทึมที่เลือกใช้ ดังนั้นจึงจำเป็นที่จะต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมนั้นๆก่อน โดยวิธีการแปลงข้อมูลมีอยู่หลายวิธีซึ่งขึ้นอยู่กับปัญหาของข้อมูล

3.4.1 Smoothing: ทำการลบเอาค่าความคลาดเคลื่อน (noise) ออกจากข้อมูล โดยใช้เทคนิค binning, clustering และ regression

3.4.2 Aggregation: คือวิธีในการรวมหรือสรุปข้อมูล เช่นข้อมูลการขายประจำวันเราอาจจะทำการสรุปข้อมูลให้เป็นรายปีหรือรายเดือน

3.4.3 Generalization: การแปลงข้อมูลดิบหรือข้อมูลที่อยู่ในระดับต่ำให้อยู่ในระดับที่สูงกว่าตามลำดับขั้น (Hierarchies) เช่น การนำข้อมูลในแอทริบิว Street รวมกับข้อมูลในแอทริบิว City, Country หรือการทำให้แอทริบิว age เป็น high-level คือ young, middle-aged, senior

3.4.4 Normalization: การทำให้ข้อมูลในแอทริบิวมีค่าไม่เกินขอบเขตที่กำหนด เช่น -1.0 ถึง 1.0, 0.0 ถึง 1.0

3.4.5 Attribute construction: สร้างแอทริบิวใหม่เพิ่มในแอทริบิวเซต เพื่อช่วยในกระบวนการดาต้าไมนิ่ง

Attribute ที่ทำการ ด้วยการกำหนดค่าให้อยู่ในขอบเขตที่กำหนด เช่น 0.0 ถึง 1.0 ซึ่งการทำ normalize นั้นเป็นประโยชน์ต่อการจัดกลุ่มเพื่อใช้ใน Algorithm Neural Network, Nearest Neighbor Classification และ Clustering การ normalize ค่าของอินพุท จะทำให้กระบวนการในการหาความรู้ (Learning Phase) ทำได้เร็วขึ้น

3.5 เทคนิคในการ Normalization

การ Normalize เป็นวิธีการแปลงข้อมูลให้อยู่ในช่วงหนึ่ง ๆ ช่วยทำให้ค่าในแอทริบิวมีขอบเขตที่ไม่กว้าง และหลากหลายเกินไป กระบวนการในการ Normalize คือ

3.5.1 Min-max normalization:

เป็นการเปลี่ยนช่วงของข้อมูลให้แสดงเป็นแบบ linear สูตรในการคำนวณหาค่าของข้อมูล ที่มีค่า $\min_A$ และ $\max_A$ เป็นค่าที่น้อยที่สุดและค่าที่มากที่สุดในแอทริบิวนั้น โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \qquad (3.6)$$

ตัวอย่างเช่น มีค่า Minimum และ Maximum ของแอทริบิว income เป็น 12000$ และ 98000$ ตามลำดับ และมีค่า income เป็น 73600$ และเราต้องการแปลงค่า income ให้อยู่ในช่วง [0.0, 1.0] โดยใช้ Min-Max Normalization สามารถคำนวณค่า income ใหม่ (v') ได้จากสูตร

$$v' = \frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716 \tag{3.7}$$

3.5.2 Z-score normalization:

หรือ Zero-mean Normalize การทำ Z-score Normalization จะเหมาะกับเหตุการณ์ที่เราไม่สามารถรู้ค่า min – max ที่แท้จริงได้โดยค่าที่ได้สำหรับแอทริบิว A เป็นค่ากึ่งกลางและค่าเบี่ยงเบนมาตรฐานของ A กำหนดให้ A' เป็นค่าเฉลี่ย และ σ_A เป็นค่าเบี่ยงเบนมาตรฐาน โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v - \bar{A}}{\sigma_A} \tag{3.8}$$

ตัวอย่างเช่น ให้ค่าเฉลี่ย (mean) ของแอทริบิว income = 54000$, Standard Deviation = 16000$, income = 73600$ จะสามารถคำนวณค่า income ใหม่ v' ได้จากสูตร

$$v' = \frac{73{,}600 - 54{,}000}{16{,}000} = 1.225 \tag{3.9}$$

3.5.3 Normalization by decimal scaling:

เป็นการเติมจุดทศนิยมให้กับแอทริบิว A ซึ่งขึ้นอยู่กับค่าสูงสุดของ $|A|$ โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v}{10^j} \tag{3.10}$$

*Where j is the smallest integer such that $Max(|A'|) < 1$.

บทที่ 4

การออกแบบระบบงาน

4.1 ระบบงานของการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง

1. การเลือกข้อมูล (Data Selection) เป็นขั้นตอนแรกในเตรียมข้อมูลสำหรับการทำดาต้า ไมนิ่ง ซึ่งผู้ใช้สามารถเลือกข้อมูลที่ต้องการโดยต้องผ่านขั้นตอนการติดต่อกับ Microsoft SQL Server 2000

2. การแก้ไขข้อมูล (Data Cleaning) เป็นขั้นตอนที่สองสำหรับการเตรียมข้อมูลสำหรับ ทำดาต้าไมนิ่ง ซึ่งจะให้ผู้ใช้ทำการแก้ไขค่าว่างโดยวิธีต่างๆ เพื่อให้ข้อมูลที่จะนำเข้าการทำดาต้าไมนิ่งนั้นมีประสิทธิภาพ

3. การปรับเปลี่ยนข้อมูล (Data Transformation) เป็นขั้นตอนสุดท้ายสำหรับการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่ง จะเป็นการปรับเปลี่ยนข้อมูลของ Numerical ให้อยู่ในช่วงๆหนึ่ง โดยผู้ใช้จะทำการกำหนดเองว่าต้องการปรับให้ข้อมูล Numerical อยู่ในช่วงใด เพื่อให้ผลลัพธ์ข้อมูลระหว่างแอตทริบิวต์ต่างๆ หลังจากประมวลผลดาต้าไมนิ่งแล้วค่าจะไม่ต่างกันเกินไป

4. การสำรวจข้อมูล (Data Exploration) เป็นการสำรวจข้อมูลขั้นสุดท้าย ก่อนที่จะนำเข้าการทำดาต้าไมนิ่งแสดงผลในรูปแบบของกราฟแท่ง กราฟวงกลม

4.2 ขั้นตอนการทำงานของระบบ

การทำงานของระบบ การเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง มีขั้นตอนการทำงาน ดังนี้

4.2.1 การเลือกข้อมูล (Data Selection)

1. ผู้ใช้ระบบระบุชื่อเซิฟเวอร์ และ ชื่อดาต้าเบส ที่ต้องการติดต่อ
2. ผู้ใช้ระบุวิธีการเลือกข้อมูล
   - เลือกข้อมูลจากหนึ่งตาราง
   - เลือกข้อมูลจากหลายตาราง
3. เลือกแอตทริบิวที่ต้องการทำดาต้าไมนิ่ง
   - ข้อมูลจากหนึ่งตาราง เลือกแอตทริบิวของตารางที่เลือกได้เลย
   - ข้อมูลจากหลายตาราง ผู้ใช้ต้องใช้คำสั่ง SQL ในการ Join ตารางเข้าด้วยกัน
4. คลิกที่ปุ่ม Execute เพื่อเข้าสู่การแก้ไขข้อมูล

4.2.2 การแก้ไขข้อมูลที่มีค่าว่าง (Data Cleaning)

1. เลือกแอตทริบิวที่ต้องการ ระบบจะแสดงรายละเอียดของแอตทริบิวบนหน้าจอ
2. เลือกวิธีในการแก้ไขข้อมูล

ข้อมูลที่เป็น Categorical

- ทางเลือกในการแก้ไขข้อมูล คือ
  - เติมค่าฐานนิยม (Mode)
  - เติม Unknown
  - ลบเรคคอร์ดที่มีค่า Null

ข้อมูลที่เป็น Numerical

- ทางเลือกในการแก้ไขข้อมูล คือ
  - เติมค่าเฉลี่ย
  - เติมค่าที่ต้องการเอง
  - ลบเรคคอร์ดที่มีค่า Null

3. ผู้ใช้สามารถคลิกที่ปุ่ม Auto Clean เพื่อดำเนินการลบค่าว่างในทุกเรคคอร์ด ที่เลือกมาทำดาต้าไมนิ่ง
4. ทำการแก้ไขข้อมูลจนครบทุกแอตทริบิว

4.2.3 การแปลงข้อมูล (Data Transformation)

ข้อมูลที่เป็น Categorical

1. หากไม่ต้องการแปลงค่าให้เลือกแอตทริบิวแล้วเลือกที่ No Transform
2. เลือกวิธีในการแปลงข้อมูลที่เป็น Categorical มีสองวิธี คือ
   - One of N Coding
   - แปลงตัวอักษรเป็นตัวเลข
3. ระบบจะสร้างแอตทริบิวใหม่ที่ได้จากการแปลงข้อมูล

ข้อมูลที่เป็น Numerical

1. หากไม่ต้องการแปลงค่าให้เลือกแอตทริบิวแล้วเลือกที่ No Transform
2. วิธีในการแปลงข้อมูลที่เป็น Numerical มีสามวิธีหลัก คือ

2.1 Normalization คือการแปลงค่าให้อยู่ในช่วงเล็กๆ ที่กำหนด

- Min-Max Normalization กำหนดช่วงสูงสุด-ต่ำสุดให้กับข้อมูล

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \qquad (4.1)$$

- Z score Normalization

$$v' = \frac{v - \overline{A}}{\sigma_A} \tag{4.2}$$

- Decimal Scaling การเติมทศนิยมให้กับข้อมูล ตามสูตร

$$v' = \frac{v}{10^j} \tag{4.3}$$

2.2 Construct Attribute คือการสร้างแอตทริบิวใหม่ที่ได้จากการคำนวณ เช่นสร้างแอตทริบิวพื้นที่ ที่ได้จากการนำ แอตทริบิวความกว้าง คูณกับแอตทริบิวความสูง

2.3 Numerical to Categorical คือการกำหนดข้อมูลตัวเลขตามช่วงที่กำหนด แล้วแทนค่าด้วยข้อมูลที่ไม่ใช่ตัวเลข เช่น ข้อมูลรายได้

- 5,000 ถึง 10,000 แทนค่าด้วย Low
- 10,001 ถึง 15,000 แทนค่าด้วย Medium
- 15,001 ถึง 20,000 แทนค่าด้วย High

4.2.4 การสำรวจข้อมูล (Data Exploration)

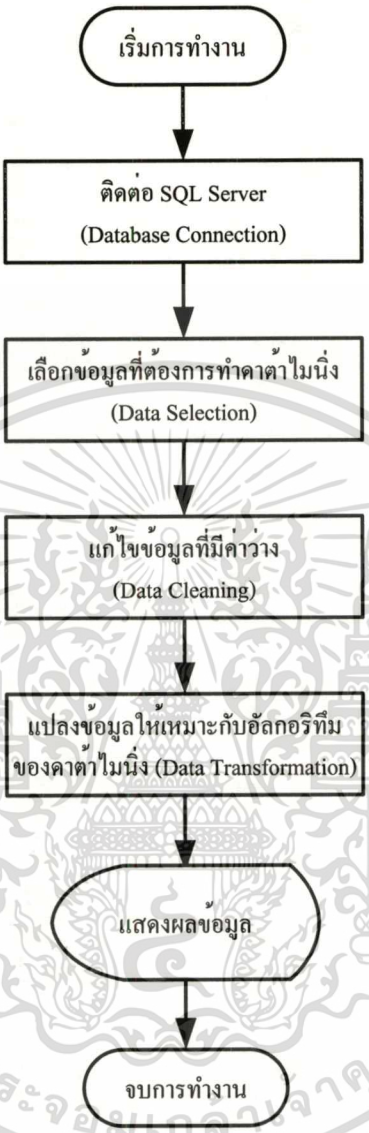
ข้อมูลที่เป็น Categorical

1. ระบบจะเลือกค่าที่แตกต่างกัน (distinct) ออกมา
2. นับจำนวนของแต่ละค่า
3. นำมาแสดงผลในรูปแบบกราฟแท่ง กราฟวงกลม

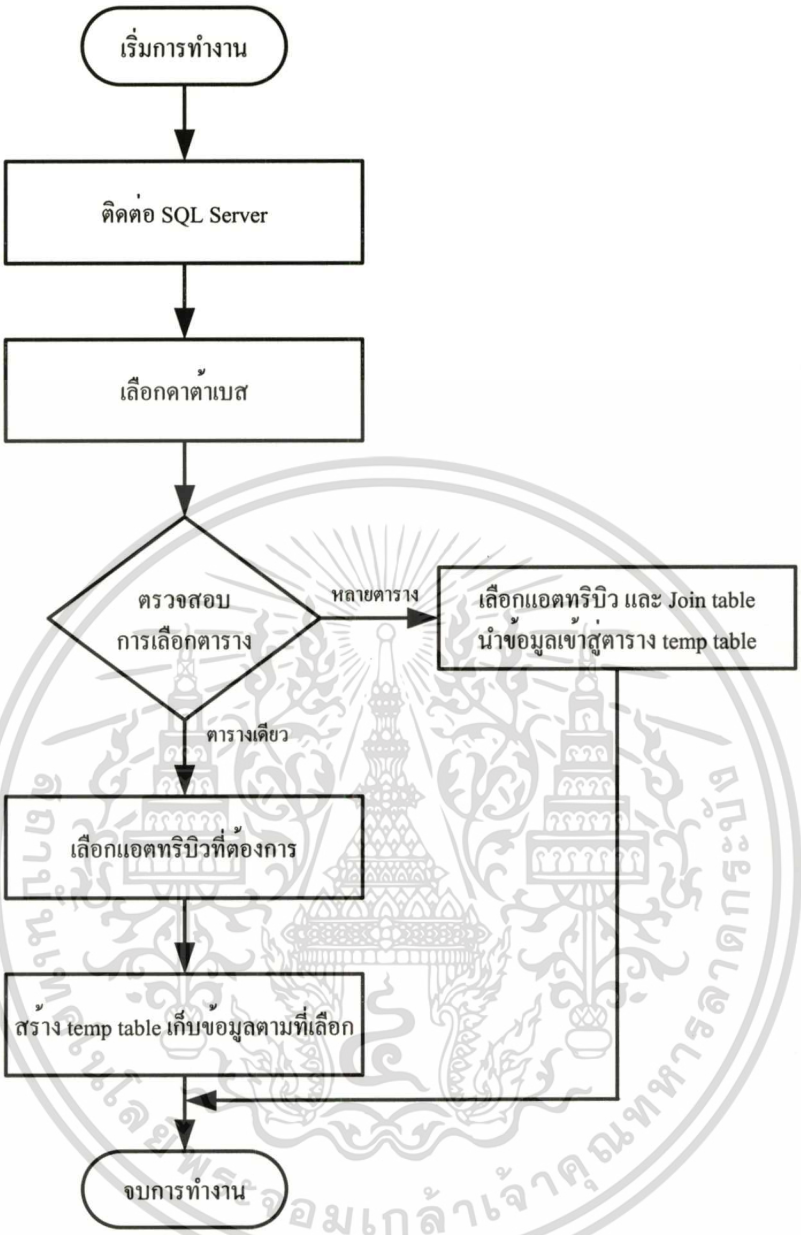
ข้อมูลที่เป็น Numerical

1. ระบบจะแบ่งข้อมูลออกเป็น 10 ช่วง
2. นับจำนวนข้อมูลในแต่ละช่วง
3. นำมาแสดงผลในรูปแบบกราฟแท่ง กราฟวงกลม

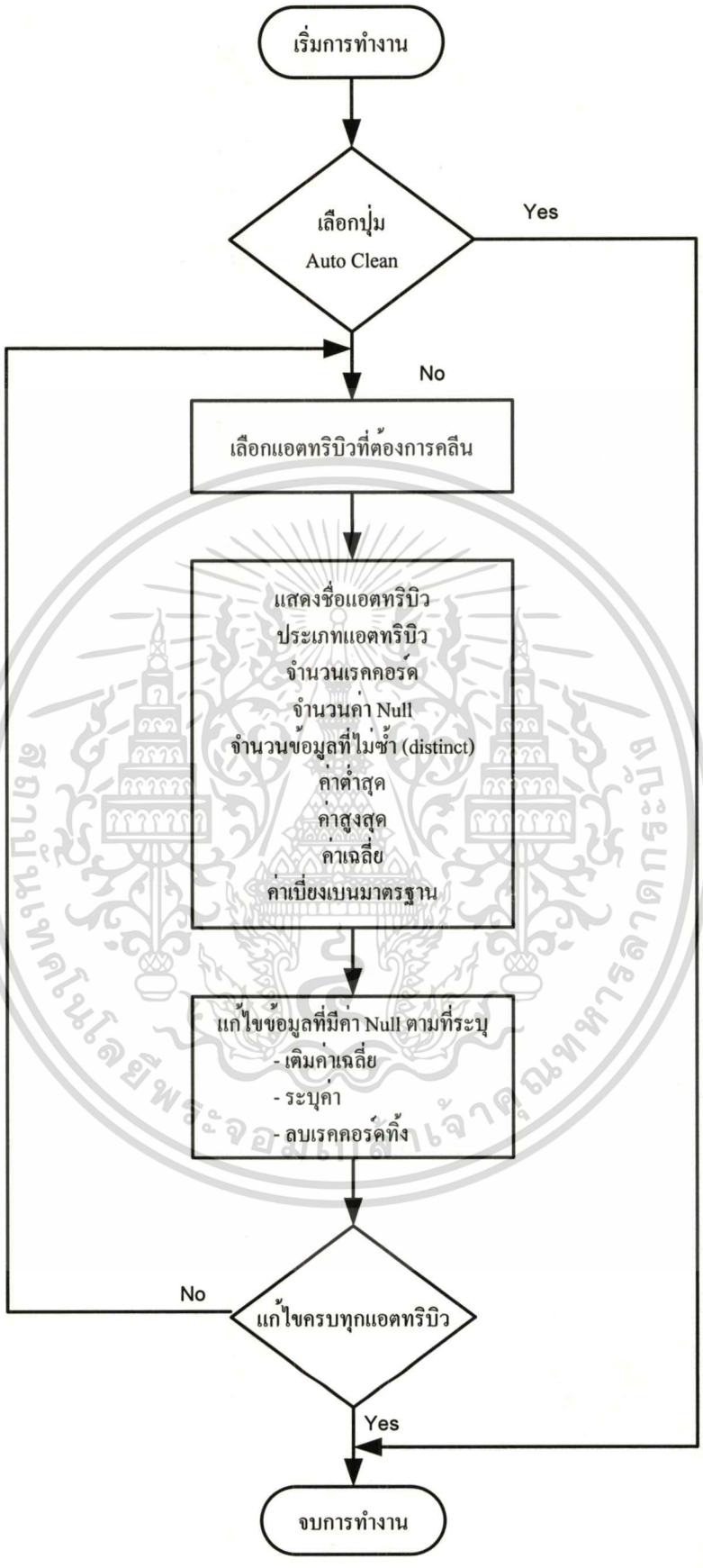
4.3 ผังการทำงานของระบบ



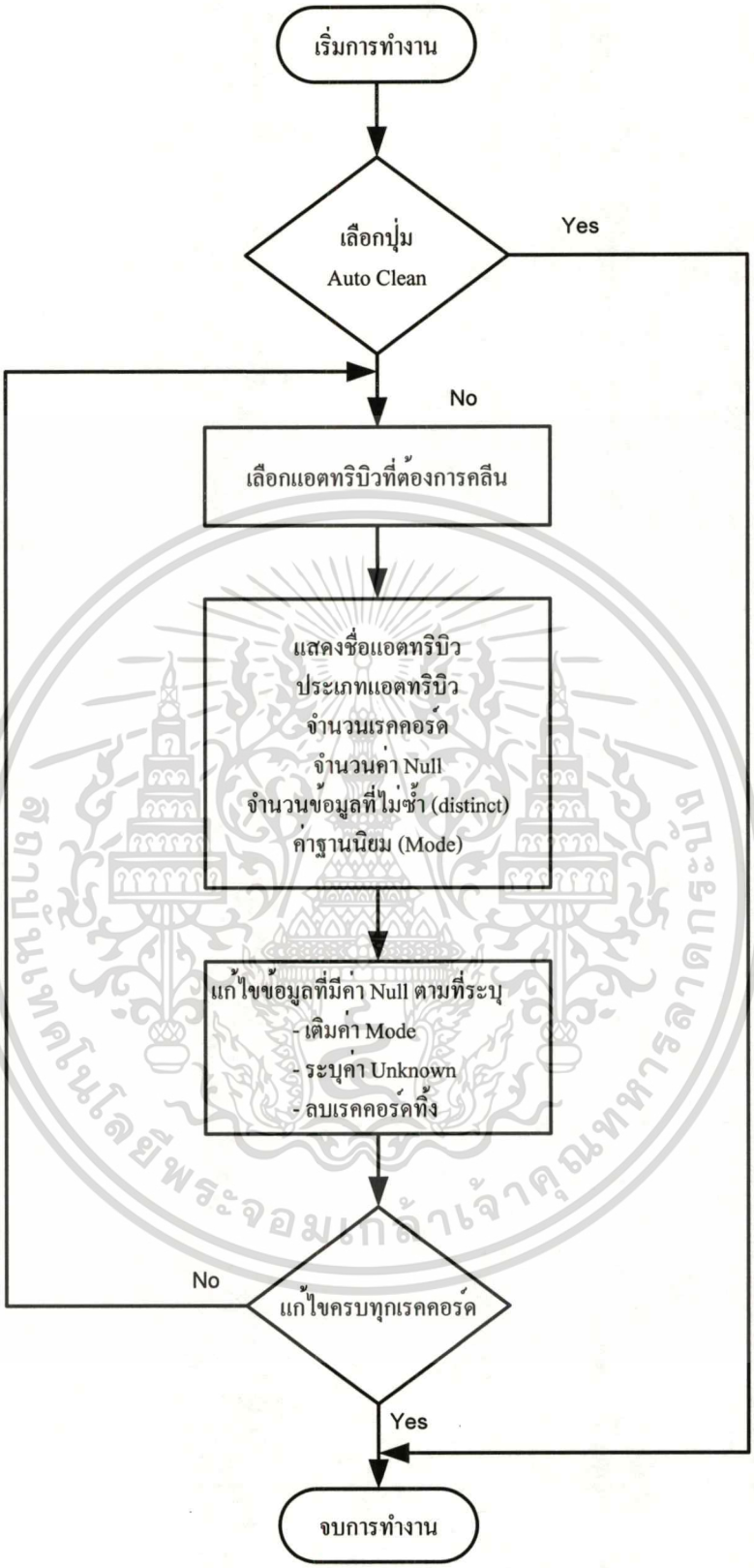
รูปที่ 4.1 ผังงานการทำงานหลักของระบบ



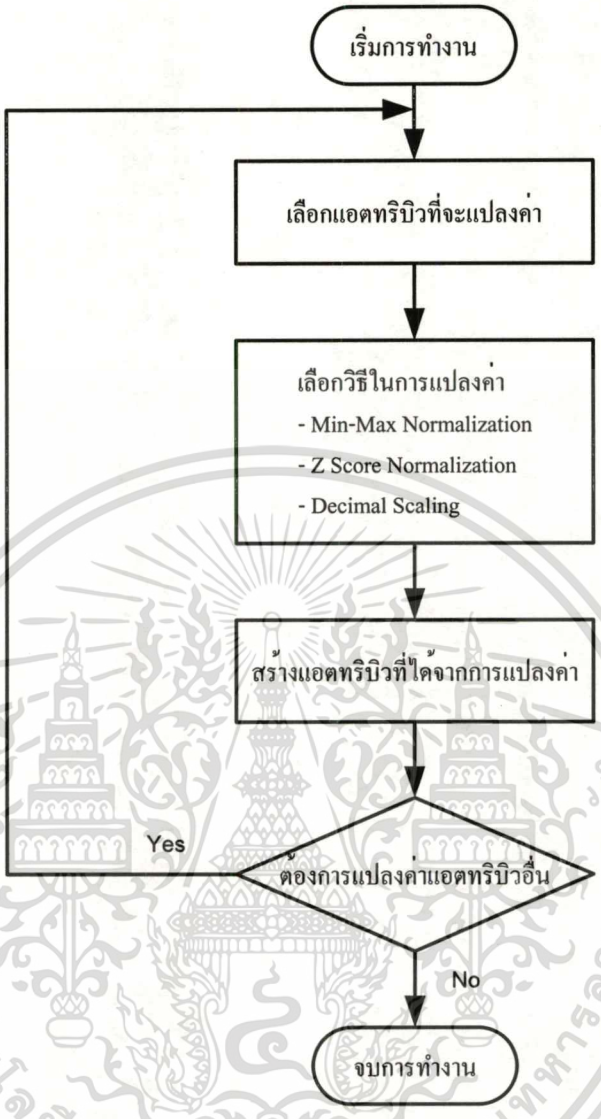
รูปที่ 4.2 ผังงานการทำงานการเลือกข้อมูล



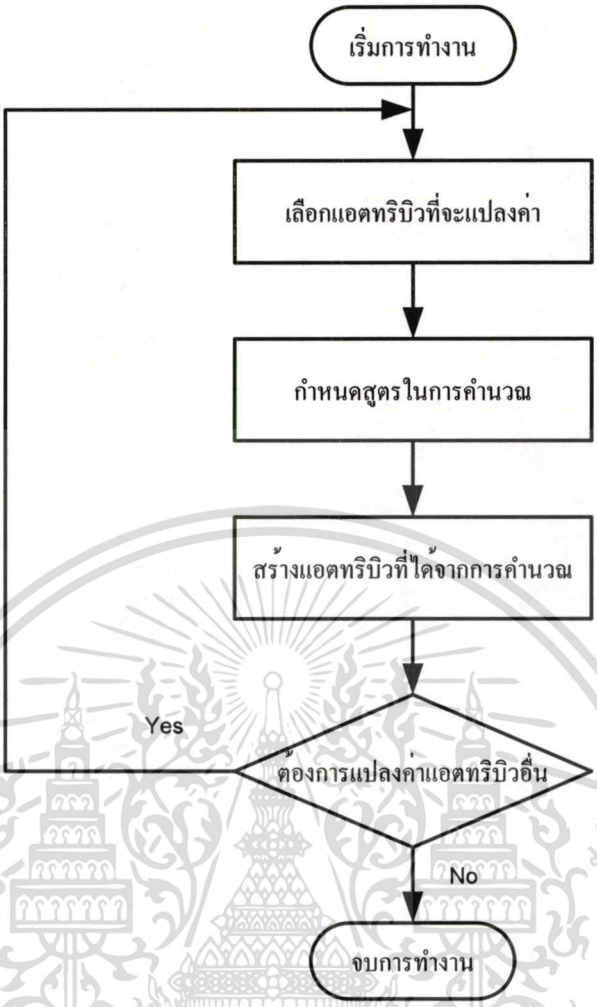
รูปที่ 4.3 ผังงานการทำงานการแก้ไขข้อมูลตัวเลข ที่มีค่าว่าง



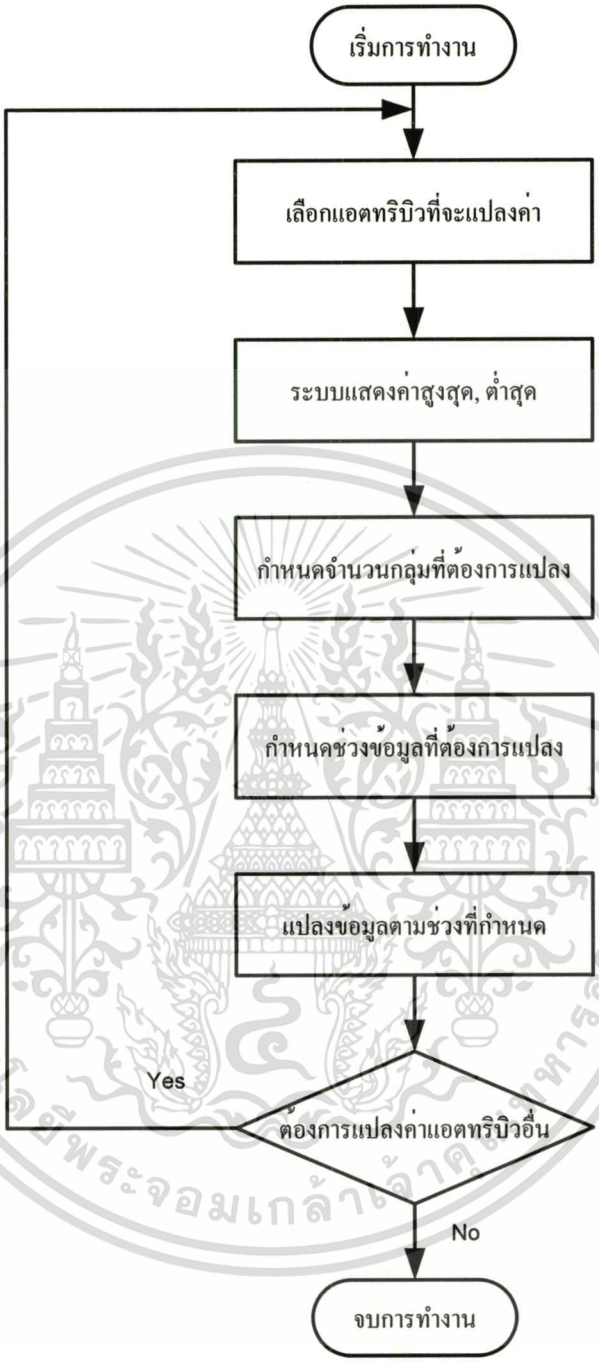
รูปที่ 4.4 ผังงานการทำงานการแก้ไขข้อมูลที่ไม่ใช่ตัวเลข ที่มีค่าว่าง



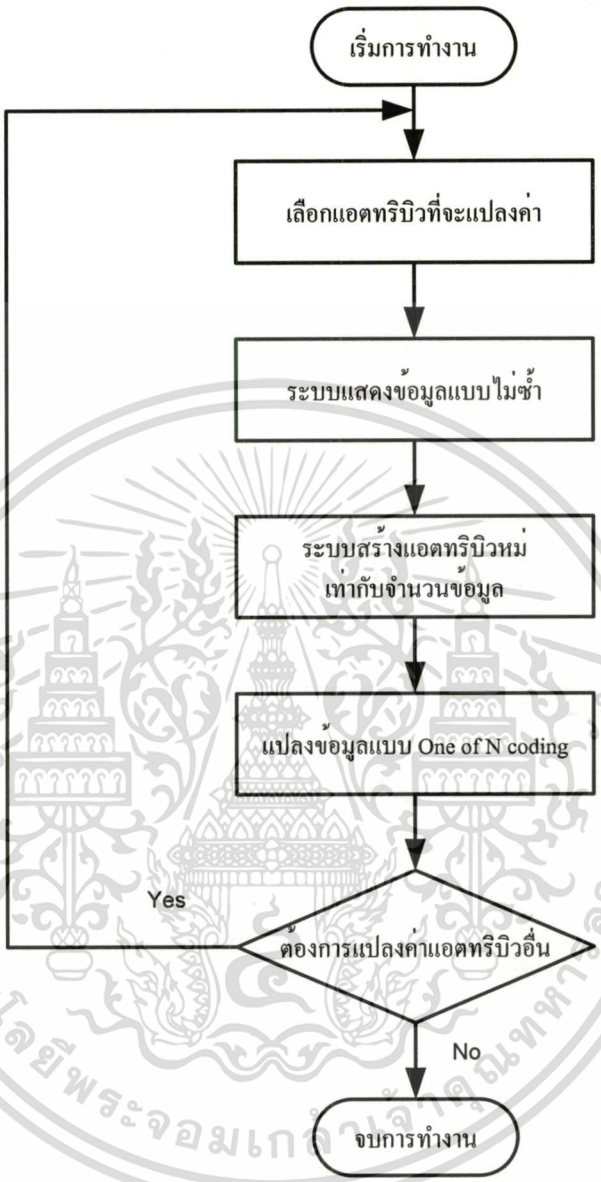
รูปที่ 4.5 ผังงานการแปลงข้อมูลที่เป็น Numeric โดยวิธี Normalization



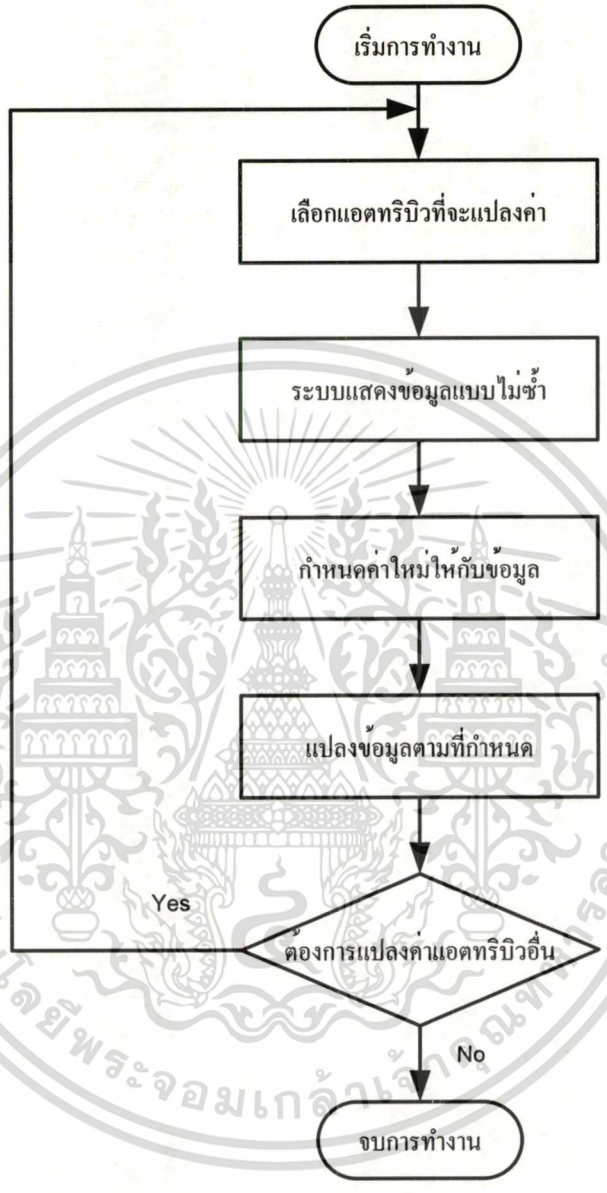
รูปที่ 4.6 ผังงานการแปลงข้อมูลโดยการสร้างแอตทริบิวใหม่ที่ได้จากการคำนวณ



รูปที่ 4.7 ผังงานการแปลงข้อมูลตัวเลข ให้เป็นตัวอักษร โดยการกำหนดช่วงข้อมูล



รูปที่ 4.8 ผังงานการแปลงข้อมูลที่ไม่ใช่ตัวเลข ให้เป็นตัวเลขโดยวิธี One of N Coding



รูปที่ 4.9 ผังงานการแปลงข้อมูลที่เป็น Category ให้เป็นตัวเลข

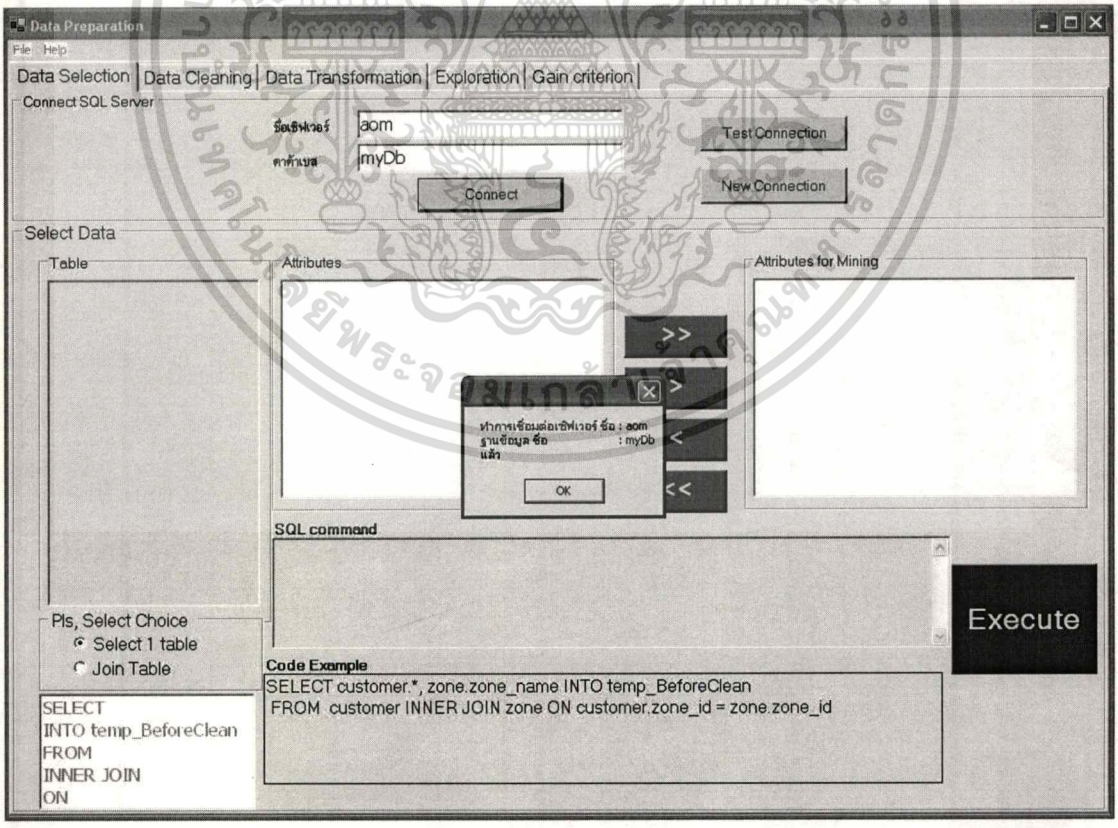
บทที่ 5

การประยุกต์ใช้โปรแกรม

5.1 การติดต่อกับฐานข้อมูล

ขั้นตอนแรกก่อนที่จะเข้าสู่กระบวนการทำดาต้าไมนิ่งต้องทำการติดต่อกับฐานข้อมูล โดยฐานข้อมูลที่จะทำการติดต่อด้วยคือ Microsoft SQL Server 2000

1. ผู้ใช้ระบบต้องทำการกรอกข้อมูลคือ ชื่อเซิฟเวอร์ และ ชื่อดาต้าเบส
2. กดปุ่ม Test Connection เพื่อทดสอบการเชื่อมต่อกับฐานข้อมูล
3. กดปุ่ม New Connection เพื่อเปลี่ยนการเชื่อมต่อดาต้าเบส
4. กดปุ่ม Connect เพื่อเชื่อมต่อกับฐานข้อมูลตามที่เราระบุ เพื่อเข้าสู่ขั้นตอนการเลือกข้อมูลต่อไป
5. ระบบแสดงข้อความให้ทราบว่าได้ทำการเชื่อมต่อกับฐานข้อมูลเรียบร้อยแล้ว

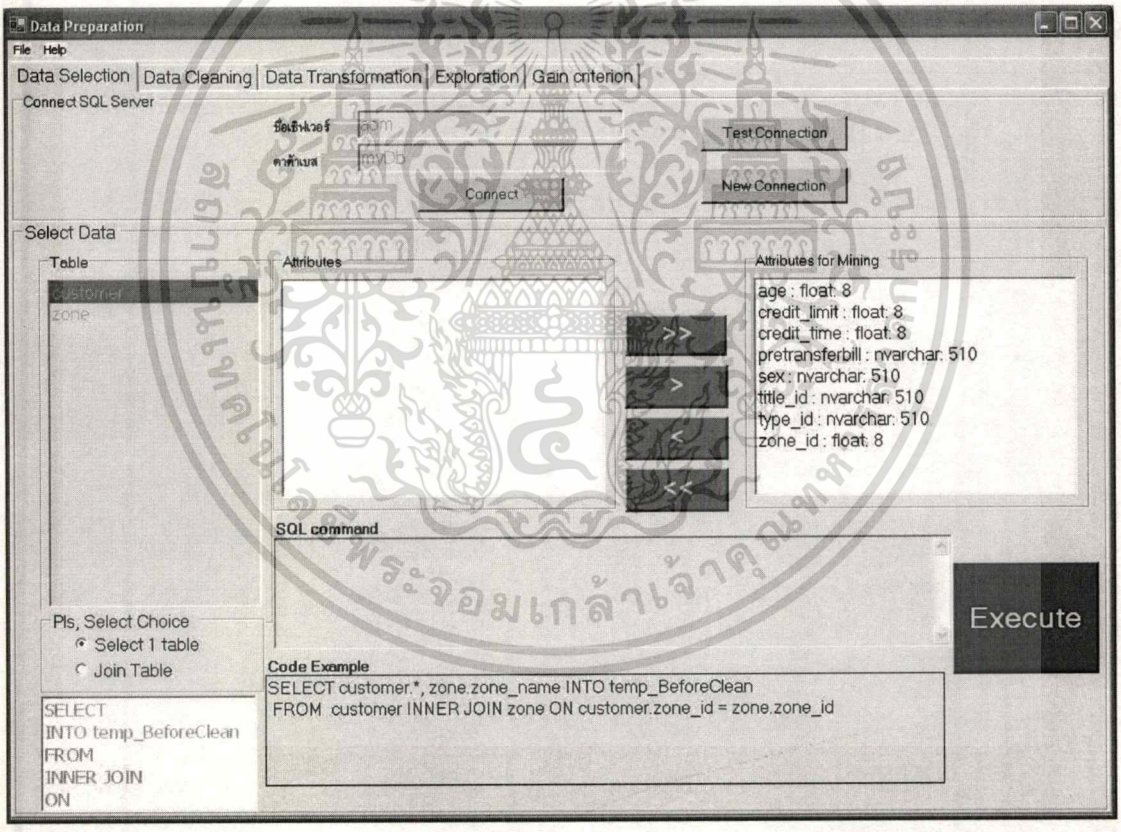


รูปที่ 5.1 ขั้นตอนการติดต่อกับฐานข้อมูล

5.2 การเลือกข้อมูล (Data Selection)

5.2.1 การเลือกข้อมูลจากหนึ่งตาราง

1. หลังจากติดต่อกับฐานข้อมูลที่เราระบุเรียบร้อยแล้ว หน้าจอระบบจะแสดงรายชื่อตารางที่อยู่ในฐานข้อมูลในช่อง Table คลิกที่ชื่อของตาราง ภายในช่อง Attributes จะแสดงรายชื่อแอตทริบิวของตารางนั้น
2. ช่อง Attributes ด้านซ้ายแสดงรายชื่อของแอตทริบิวจากตารางที่เลือก
3. ช่อง Attributes for Mining แสดงรายชื่อแอตทริบิวที่ผู้ใช้ระบบต้องการนำไปใช้ในการทำดาต้าไมนิ่งต่อไป
4. คลิกที่ปุ่ม Execute สิ้นสุดขั้นตอนการเลือกข้อมูล เพื่อเข้าสู่การเตรียมข้อมูลในขั้นตอนต่อไป

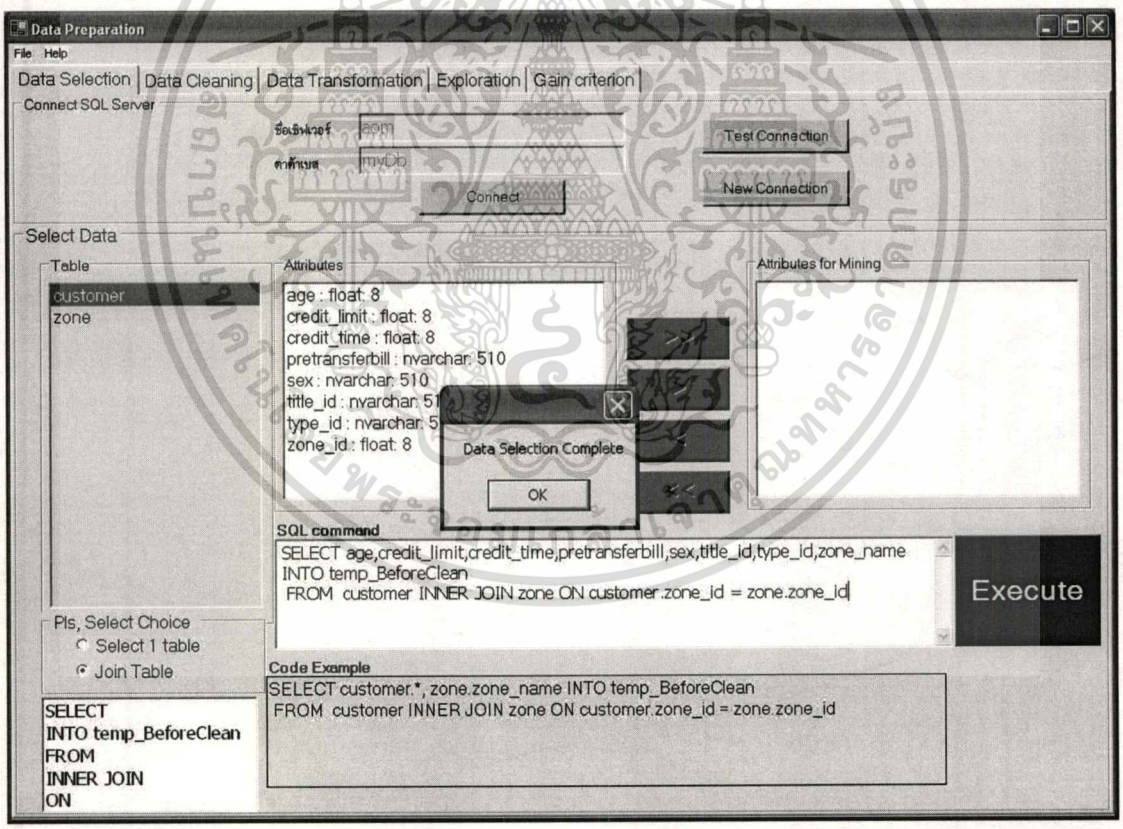


รูปที่ 5.2 ขั้นตอนการเลือกข้อมูลจากหนึ่งตาราง

5.2.2 การเลือกข้อมูลจากหลายตาราง

ในการเลือกข้อมูลสามารถเลือกข้อมูลสามารถเลือกข้อมูลที่มาจากหลายตารางภายในฐานข้อมูลเดียวกันได้ดังรูปที่ 5.3 โดยมีขั้นตอนการทำงานดังนี้คือ

1. ในช่อง Pls, Select Choice ให้เลือกตัวเลือกที่สอง Join Table
2. พิมพ์คำสั่งของ SQL ลงใน SQL Command เพื่อเลือกข้อมูลที่ต้องการ โดยผู้ใช้ระบบจะต้องทราบความสัมพันธ์ของข้อมูลในแต่ละตาราง และสามารถใช้คำสั่ง SQL พื้นฐานในการเชื่อมความสัมพันธ์เหล่านั้น เพื่อเลือกข้อมูลที่จะนำมาทำดาต้าไมนิ่งได้
3. คลิกที่ปุ่ม Execute สิ้นสุดขั้นตอนการเลือกข้อมูล เพื่อเข้าสู่การเตรียมข้อมูลในขั้นตอนต่อไป
4. ระบบแสดงกล่องข้อความบอกผู้ใช้ว่าเสร็จสิ้นขั้นตอนการเลือกข้อมูลแล้ว
5. การเลือกข้อมูลในการทำดาต้าไมนิ่งผู้ใช้ระบบ จะเลือกข้อมูลจากหนึ่งตาราง หรือหลายตารางจากการ Join ได้วิธีใดวิธีหนึ่งเท่านั้น



รูปที่ 5.3 ขั้นตอนการเลือกข้อมูลจากหลายตาราง

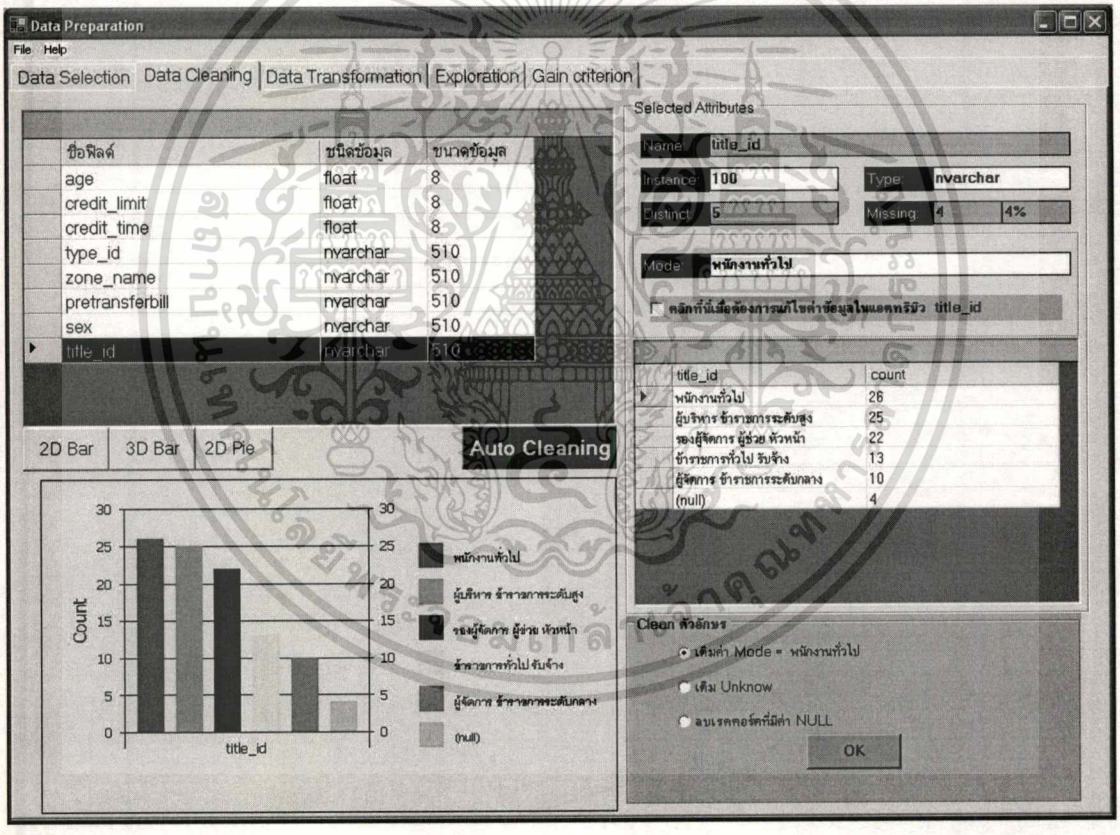
5.3 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลคือ การแก้ไขปัญหาที่พบในข้อมูลเพื่อทำให้ข้อมูลมีคุณภาพก่อนที่จะนำข้อมูลไปประมวลผล ซึ่งในขั้นตอนนี้เป็นการทำ Data Cleaning

5.3.1 Data Cleaning ข้อมูลที่เป็น Categorical

เป็นขั้นตอนในการจัดการข้อมูลที่ต้องการ ซึ่งข้อมูลส่วนใหญ่ในฐานข้อมูลนั้นมักจะไม่สมบูรณ์ ซึ่งมีการทำงานของระบบ ดังนี้

1. หลังจากขั้นตอนการเลือกข้อมูลเสร็จสิ้น ระบบจะแสดงแท็บ Data Cleaning ขึ้นมาโดยอัตโนมัติ
2. เลือกชื่อแอตทริบิวที่ใช้ทำดาต้าไมนิ่ง ระบบจะแสดงรายละเอียดของแอตทริบิวและแสดงกราฟของจำนวนข้อมูลในแอตทริบิวที่ผู้ใช้เลือก



รูปที่ 5.4 ขั้นตอนการขจัดข้อมูลที่เป็น Categorical ในเรคคอร์ดที่เป็น Null

จากรูปที่ 5.5 แสดงรายละเอียดของแอตทริบิวที่ผู้ใช้ระบบคลิกเลือก โดยแสดงรายละเอียดดังนี้

Name: ชื่อแอตทริบิว

Instance: จำนวนเรคคอร์ด

Type: ชนิดข้อมูล

Distinct: จำนวนข้อมูลที่ไม่ซ้ำกัน

Missing: จำนวนเรคคอร์ดที่ค่าหายไป และคิดเป็นเปอร์เซ็นต์

Mode: ค่าฐานนิยม

รูปที่ 5.5 รายละเอียดของแอตทริบิวที่เลือก

จากรูปที่ 5.6 แสดงรายละเอียดของข้อมูลในแอตทริบิว โดยแสดงค่าที่ไม่ซ้ำและเรียงลำดับการนับจำนวนเรคคอร์ดจากจำนวนมากไปหาน้อย และแสดงเรคคอร์ดที่มีค่า Null

	title_id	count
▶	พนักงานทั่วไป	26
	ผู้บริหาร ข้าราชการระดับสูง	25
	รองผู้จัดการ ผู้ช่วย หัวหน้า	22
	ข้าราชการทั่วไป รับจ้าง	13
	ผู้จัดการ ข้าราชการระดับกลาง	10
	(null)	4

รูปที่ 5.6 รายละเอียดของข้อมูลจากการนับจำนวนเรคคอร์ด

จากรูปที่ 5.7 ระบบแสดงทางเลือกในการขจัดข้อมูลที่ขาดหายไป 3 ทางเลือกคือ

- เติมค่าฐานนิยม (mode) ลงในข้อมูลที่หายไป
- เติม Unknown ลงในข้อมูลที่หายไป
- ลบเรคคอร์ดที่มีค่า Null ทิ้งไป

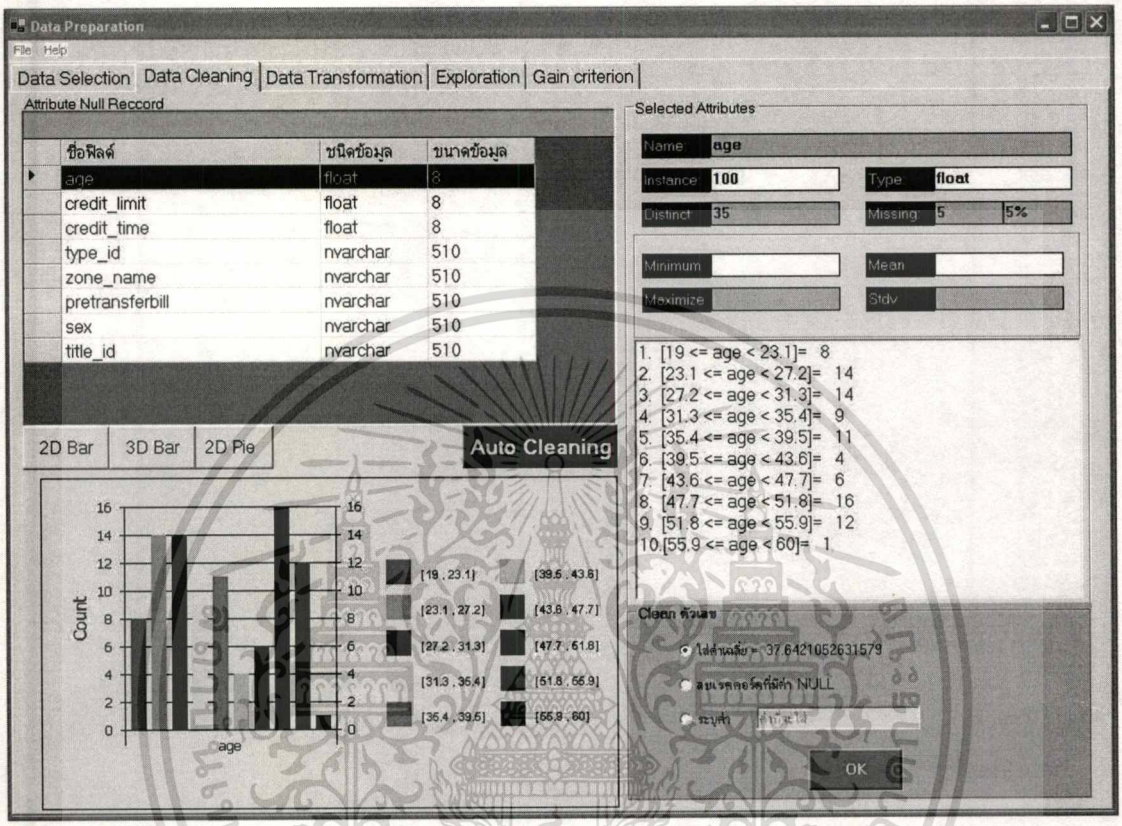
ผู้ใช้สามารถคลิกเลือกวิธีตามที่ต้องการได้ จากนั้นคลิก OK เพื่อทำการคลีนข้อมูลตามวิธีที่ได้เลือกไว้ ระบบแสดงข้อความบอกว่าได้ทำการแก้ไขข้อมูลเรียบร้อยแล้ว ดังรูปที่ 5.8 แล้วทำการแก้ไขข้อมูลต่อไปจนครบทุกแอตทริบิว หากผู้ใช้ไม่ต้องการที่จะทำการแก้ไขข้อมูลครั้งละแอตทริบิว ให้คลิกที่ปุ่ม Auto Cleaning เพื่อให้ระบบลบเรคคอร์ดที่มี Null ทิ้งไปเป็นการเสร็จสิ้นขั้นตอนการทำ Data Cleaning

รูปที่ 5.7 ทางเลือกในการ clean ข้อมูลที่เป็น Category

รูปที่ 5.8 ข้อความแสดงการแก้ไขข้อมูลแล้ว

5.3.2 Data Cleaning ข้อมูลที่เป็น Numerical

เลือกชื่อแอตทริบิวที่ใช้ทำคาต้าไมนิ่ง จากตัวอย่างนี้ให้คลิกเลือกแอตทริบิวที่เป็น Numeric ระบบจะแสดงรายละเอียดของแอตทริบิว และแสดงกราฟข้อมูลในแอตทริบิว



รูปที่ 5.9 ขั้นตอนการขจัดข้อมูลที่เป็น Numeric ในเรคคอร์ดที่เป็น Null

จากรูปที่ 5.10 แสดงรายละเอียดของแอตทริบิวที่ผู้ใช้ระบบคลิกเลือก โดยแสดงรายละเอียดดังนี้

Name: ชื่อแอตทริบิว

Instance: จำนวนเรคคอร์ด

Type: ชนิดข้อมูล

Distinct: จำนวนข้อมูลที่ไม่ซ้ำกัน

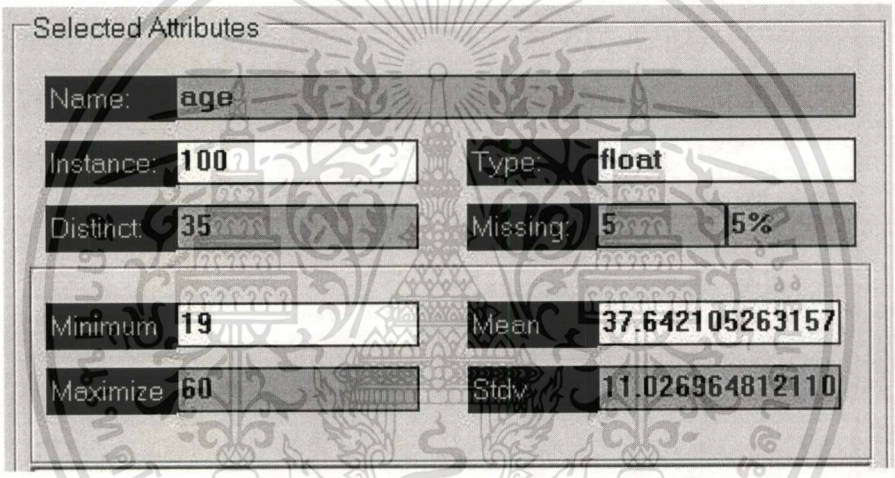
Missing: จำนวนเรคคอร์ดที่ค่าหายไป และคิดเป็นเปอร์เซ็นต์

Minimum: ค่าที่น้อยที่สุด

Maximize: ค่าที่มากที่สุด

Mean: ค่าเฉลี่ย หรือค่ากลาง

Stdv: ค่าเบี่ยงเบนมาตรฐาน

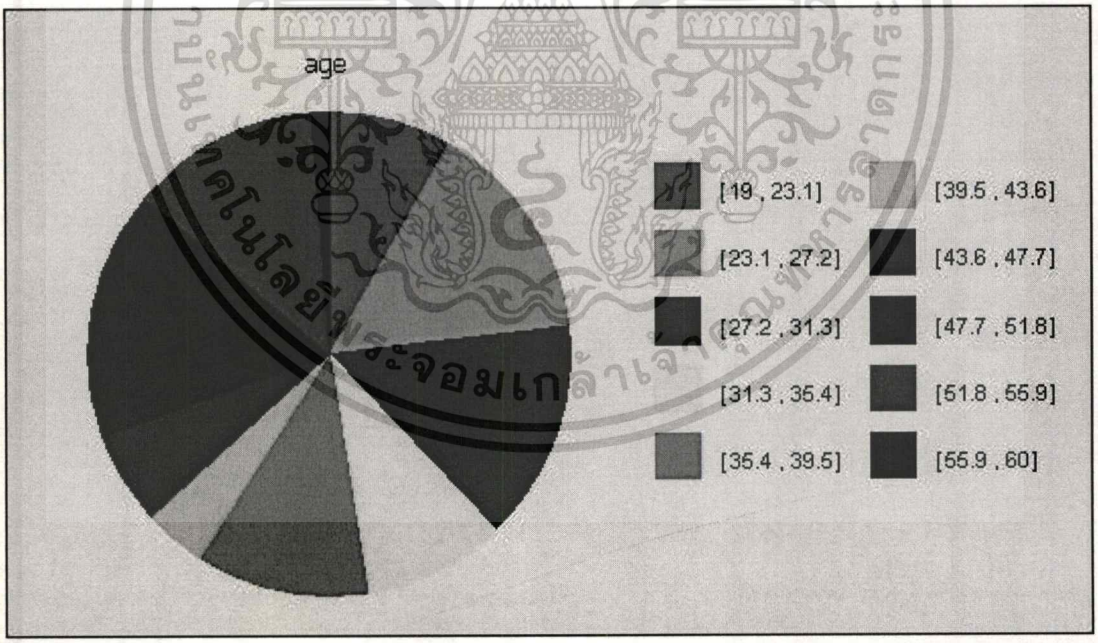


รูปที่ 5.10 รายละเอียดของแอตทริบิวที่เป็น Category

รูปที่ 5.11 แสดงช่วงของข้อมูลในแอตทริบิว age มีชนิดข้อมูลแบบ float ที่เป็น Numeric ซึ่งมีค่า Min = 19 และมีค่า Max =60 โดยกำหนดไว้ 10 ช่วง ช่วงละเท่า ๆ กัน เรียงจากน้อยไปมาก ตามลำดับ และแสดงจำนวนเรคคอร์ดที่นับได้ในแต่ละช่วง ซึ่งข้อมูลที่แสดงนี้จะสัมพันธ์กับกราฟ ดังรูปที่ 5.12

1. [19 <= age < 23.1]= 8
2. [23.1 <= age < 27.2]= 14
3. [27.2 <= age < 31.3]= 14
4. [31.3 <= age < 35.4]= 9
5. [35.4 <= age < 39.5]= 11
6. [39.5 <= age < 43.6]= 4
7. [43.6 <= age < 47.7]= 6
8. [47.7 <= age < 51.8]= 16
9. [51.8 <= age < 55.9]= 12
10. [55.9 <= age < 60]= 1

รูปที่ 5.11 ช่วงของข้อมูลในแอตทริบิว age



รูปที่ 5.12 กราฟข้อมูลในแอตทริบิว age

จากรูปที่ 5.13 ระบบแสดงทางเลือกในการขจัดข้อมูลที่ขาดหายไป 3 ทางเลือกคือ

- เติมค่าเฉลี่ย (mean) ลงในข้อมูลที่หายไป
- ลบเรคคอร์ดที่มีค่า Null ทิ้งไป
- ผู้ใช้ระบุตัวเลขที่ต้องการลงไป

ผู้ใช้สามารถคลิกเลือกวิธีตามที่ต้องการได้ จากนั้นคลิก OK เพื่อทำการคลีนข้อมูลตามวิธีที่ได้เลือกไว้ ระบบแสดงข้อความบอกว่าได้ทำการแก้ไขข้อมูลเรียบร้อยแล้ว ดังรูปที่ 5.14 แล้วทำการแก้ไขข้อมูลต่อไปจนครบทุกแอตทริบิว หรือคลิกที่ปุ่ม Auto Cleaning ก็ได้เช่นเดียวกัน

รูปที่ 5.13 ทางเลือกในการ clean ข้อมูลที่เป็น Numeric

รูปที่ 5.14 ข้อความแสดงการแก้ไขข้อมูลแล้ว

5.4 การแปลงข้อมูล (Data Transformation)

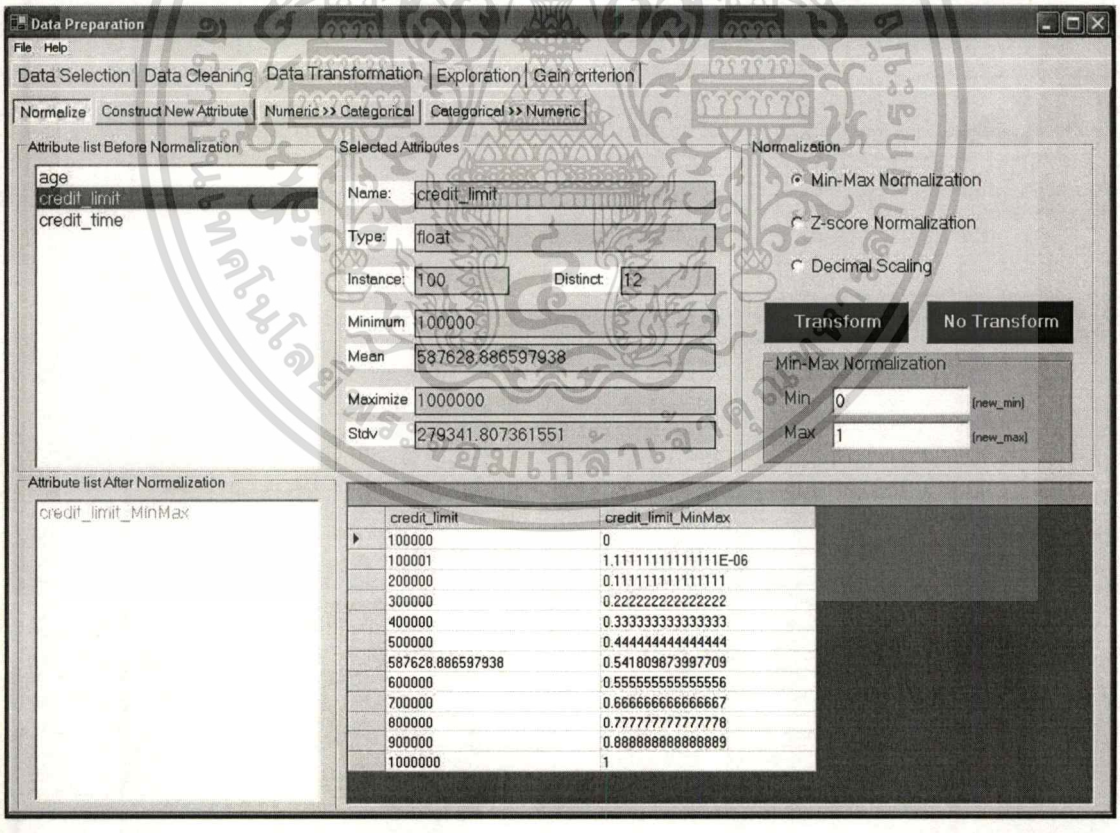
ขั้นตอนการแปลงข้อมูลคือ การทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำมาใช้

5.4.1 Normalization

การแปลงค่าทำให้ข้อมูลในแอทริบิวมีค่าไม่เกินขอบเขตที่กำหนด

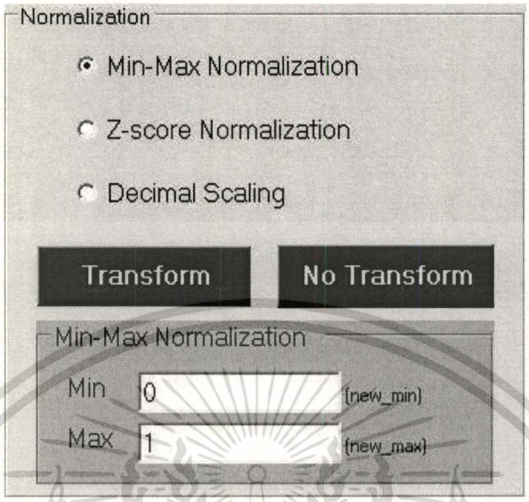
5.4.1.1 Min-Max Normalization

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Min- Max Normalization
3. กำหนดขอบเขตของค่าที่ต้องการแปลง Min คือช่วงต่ำสุดของค่าที่ต้องการ, Max คือช่วงสูงสุดของค่าที่ต้องการ
4. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
5. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Min-Max Normalization



รูปที่ 5.15 ขั้นตอนการแปลงข้อมูลแบบ Min- Max Normalization

จากรูปที่ 5.16 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Min- Max Normalization ผู้ใช้ต้องระบุขอบเขตของข้อมูลก่อน จากตัวอย่างกำหนด Min = 0 และ Max = 1 ค่าที่ได้จากการแปลงจะอยู่ในขอบเขตนี้



รูปที่ 5.16 การเลือกวิธี Min- Max Normalization

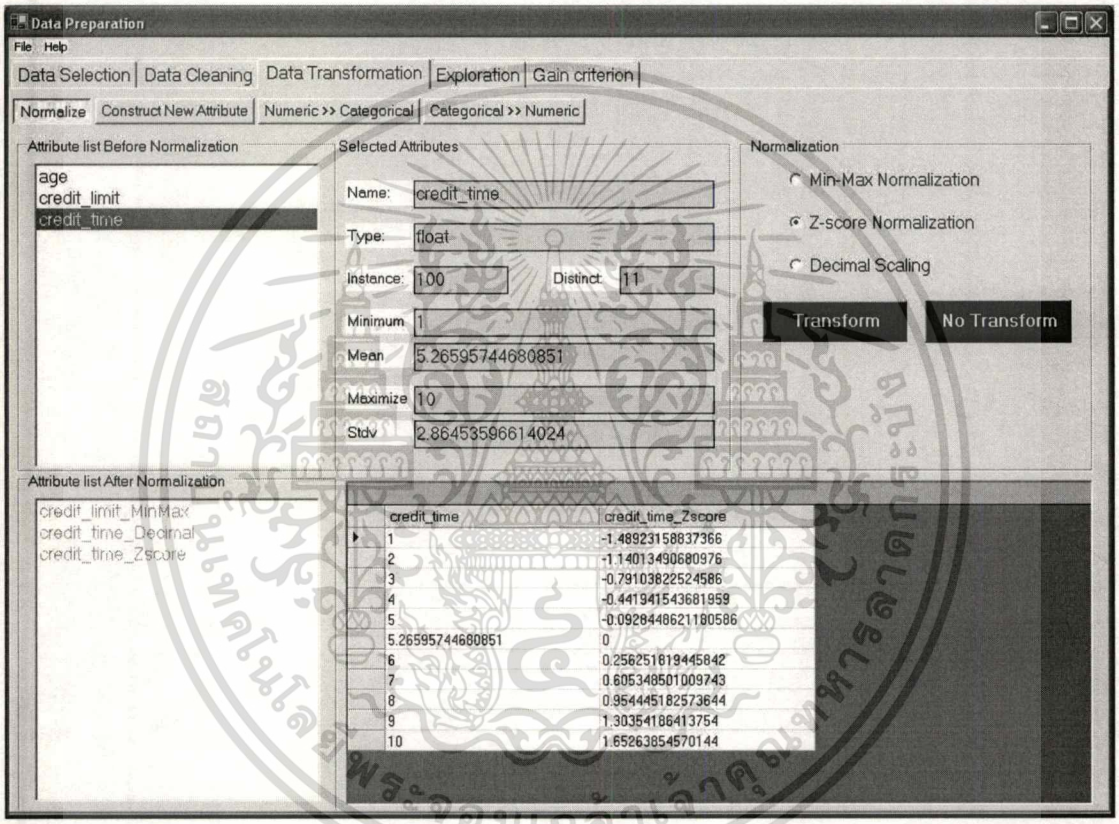
รูปที่ 5.17 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว credit_limit ก่อนทำการแปลงค่าแบบ Min- Max Normalization โดยค่าที่แสดงเรียงลำดับจากน้อยไปมาก คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว credit_limit_MinMax โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

	credit_limit	credit_limit_MinMax
▶	100000	0
	100001	1.11111111111111E-06
	200000	0.111111111111111
	300000	0.222222222222222
	400000	0.333333333333333
	500000	0.444444444444444
	587628.886597938	0.541809873997709
	600000	0.555555555555556
	700000	0.666666666666667
	800000	0.777777777777778
	900000	0.888888888888889
	1000000	1

รูปที่ 5.17 ข้อมูลที่ได้จากการแปลงโดยวิธี Min- Max Normalization

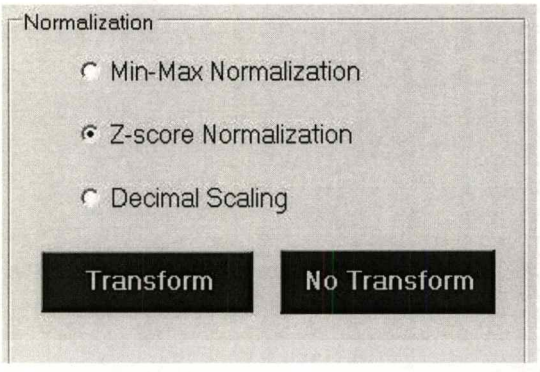
5.4.1.2 Z-score Normalization

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Z-score Normalization
3. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
4. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Z-score Normalization



รูปที่ 5.18 ขั้นตอนการแปลงข้อมูลแบบ Z-score Normalization

จากรูปที่ 5.19 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Z-score Normalization



รูปที่ 5.19 การเลือกวิธี Z-score Normalization

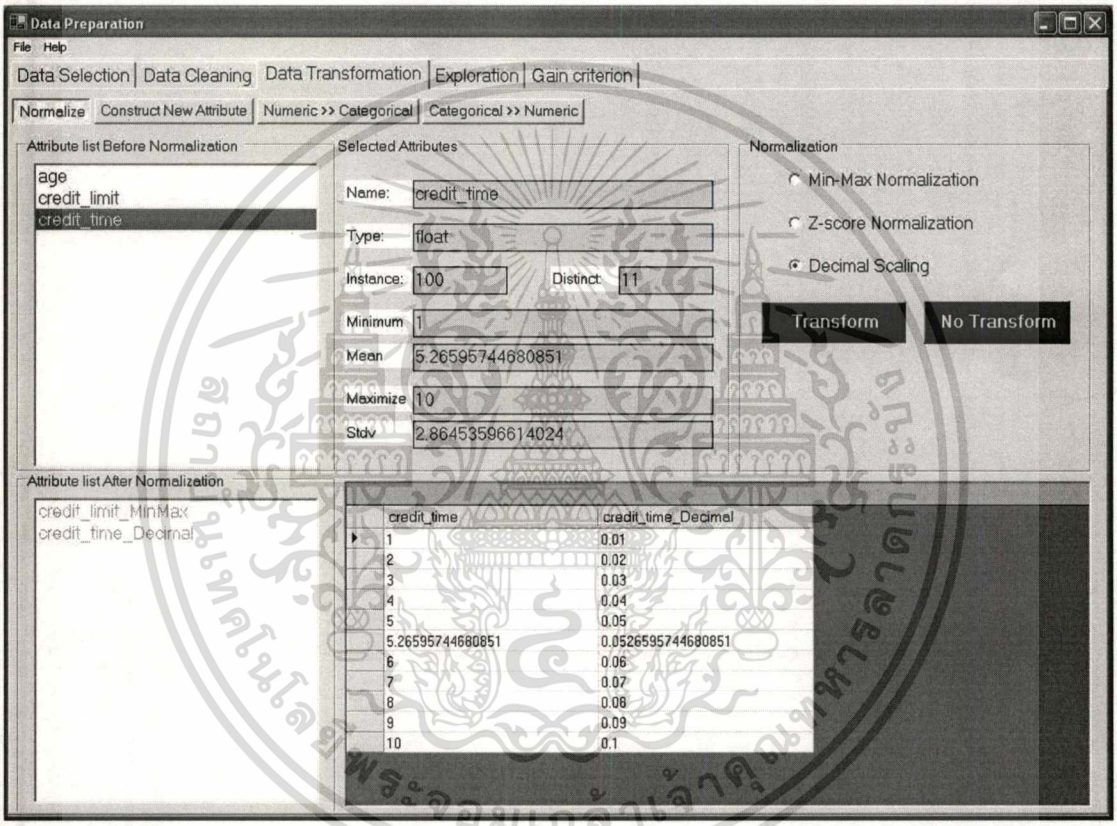
รูปที่ 5.20 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว credit_time ก่อนทำการแปลงค่าแบบ Z-score Normalization คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว credit_time_Zscore โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

credit_time	credit_time_Zscore
1	-1.48923158837366
2	-1.14013490680976
3	-0.79103822524586
4	-0.441941543681959
5	-0.0928448621180586
5.26595744680851	0
6	0.256251819445842
7	0.605348501009743
8	0.954445182573644
9	1.30354186413754
10	1.65263854570144

รูปที่ 5.20 ข้อมูลที่ได้จากการแปลงโดยวิธี Z-score Normalization

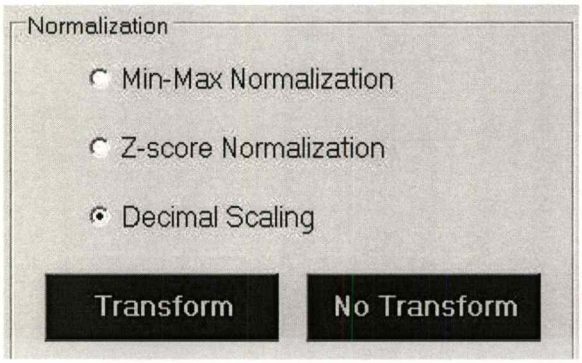
5.4.1.3 Decimal Scaling

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Decimal Scaling
3. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
4. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Decimal Scaling



รูปที่ 5.21 ขั้นตอนการแปลงข้อมูลแบบ Decimal Scaling

จากรูปที่ 5.22 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Decimal Scaling



รูปที่ 5.22 ขั้นตอนการแปลงข้อมูลแบบ Decimal Scaling

รูปที่ 5.23 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว credit_time ก่อนทำการแปลงค่าแบบ Decimal Scaling คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว credit_time_Decimal โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล ซึ่งวิธีนี้เป็นการเติมทศนิยมให้กับข้อมูล

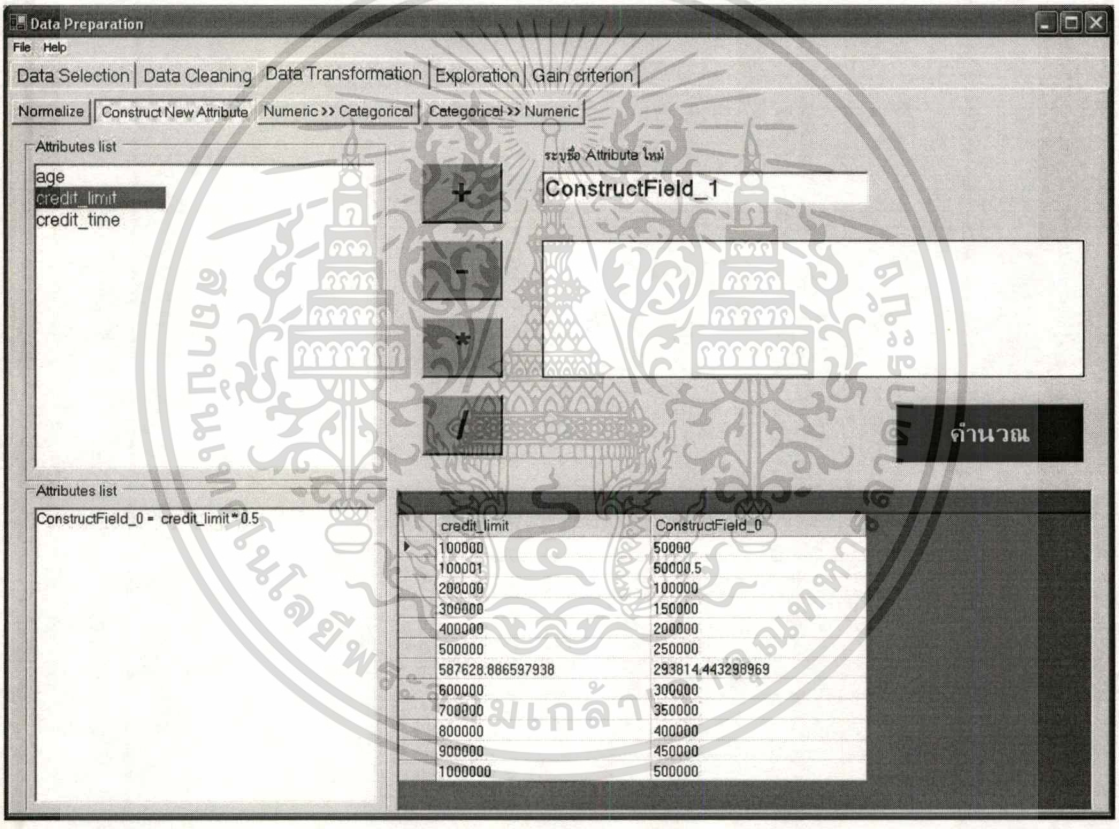
credit_time	credit_time_Decimal
1	0.01
2	0.02
3	0.03
4	0.04
5	0.05
5.26595744680851	0.0526595744680851
6	0.06
7	0.07
8	0.08
9	0.09
10	0.1

รูปที่ 5.23 ข้อมูลหลังจากการแปลงแบบ Decimal Scaling

5.4.2 Construct New Attribute

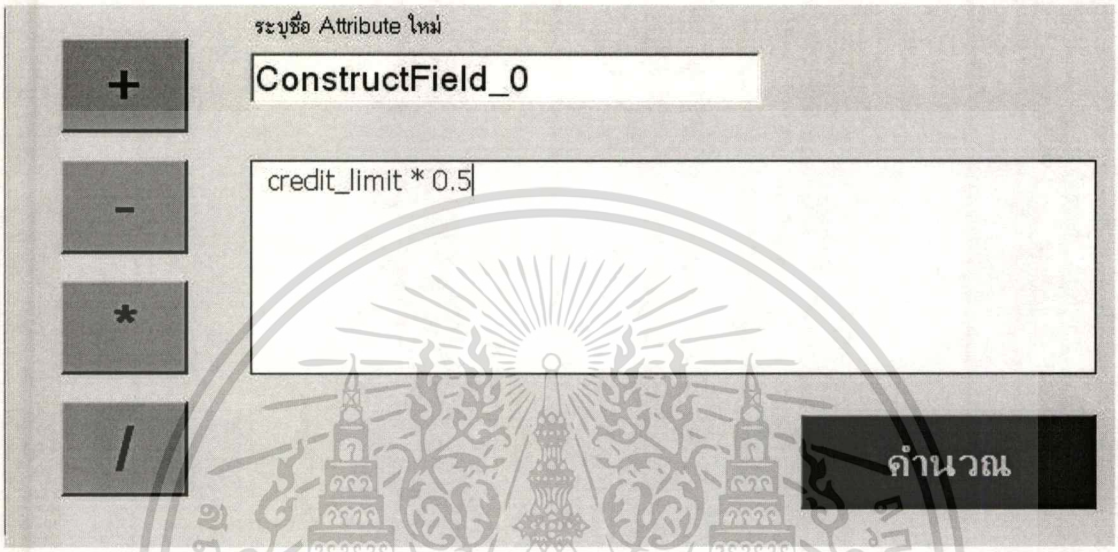
การสร้างแอตทริบิวใหม่ที่ได้จากการคำนวณ มีขั้นตอนการทำงานดังนี้

1. คลิกที่แท็บ Construct New Attribute
2. หากผู้ใช้ระบบต้องการระบุชื่อ แอตทริบิวใหม่ ให้พิมพ์ลงในช่อง ระบุชื่อ Attribute ใหม่ หรือใช้ชื่อตามที่ระบบตั้งให้
3. พิมพ์สูตรที่ต้องการคำนวณ
4. กดปุ่ม คำนวณ
5. ระบบแสดงแอตทริบิวใหม่พร้อมข้อมูลที่ได้จากการคำนวณ
6. ในช่อง Attribute list ด้านล่างแสดง ชื่อแอตทริบิวใหม่และสูตรที่คำนวณ



รูปที่ 5.24 ขั้นตอนการแปลงข้อมูลแบบ Construct New Attribute

จากรูปที่ 5.25 คือการสร้างแอตทริบิวที่ได้จากการคำนวณ ผู้ใช้สามารถตั้งชื่อของแอตทริบิวได้เองในช่องระบุชื่อ Attribute ใหม่ หรือระบบตั้งให้ซึ่งขึ้นต้นด้วยคำว่า ConstructField แล้วตามด้วยหมายเลข ช่องด้านล่างให้ผู้ใช้ระบุสูตรที่ใช้ในการคำนวณแอตทริบิวที่สร้างขึ้นมาใหม่โดยผู้ใช้จะต้องระบุสูตรที่ถูกต้องลงไปเอง ขั้นสุดท้ายเมื่อระบุสูตรแล้ว คลิกที่ปุ่มคำนวณเป็นการเสร็จสิ้นการสร้างแอตทริบิวที่ได้จากการคำนวณ



รูปที่ 5.25 ขั้นตอนการแปลงข้อมูลแบบ Construct New Attribute

รูปที่ 5.26 ในคอลัมน์แรกแสดงข้อมูลของแอตทริบิว credit_limit คอลัมน์ที่สองแสดงข้อมูลของแอตทริบิว ConstructField_0 ที่ได้จากการคำนวณตามสูตร ConstructField_0 = credit_limit * 0.5

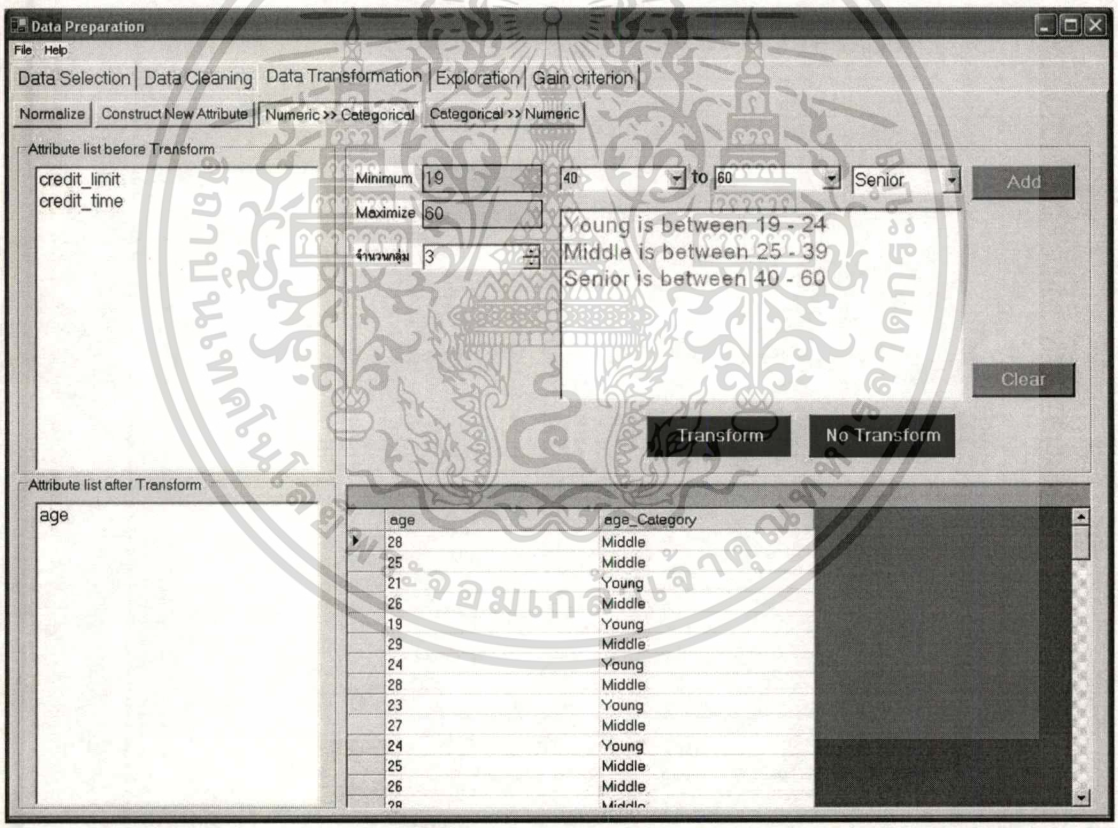
	credit_limit	ConstructField_0
▶	100000	50000
	100001	50000.5
	200000	100000
	300000	150000
	400000	200000
	500000	250000
	587628.886597938	293814.443298969
	600000	300000
	700000	350000
	800000	400000
	900000	450000
	1000000	500000

รูปที่ 5.26 ข้อมูลหลังการแปลงแบบ Construct New Attribute

54.3 Numeric to Categorical

การแปลงข้อมูลที่เป็น Numeric ให้เป็นข้อมูลที่เป็น Category มีขั้นตอนการแปลงข้อมูลดังนี้

1. คลิกที่แท็บ Numeric >> Categorical
2. คลิกเลือกจำนวนกลุ่มที่ต้องการ
3. ระบุช่วงที่ต้องการแปลงข้อมูล
4. ระบุคำที่ต้องการแปลง สำหรับช่วงที่กำหนด คลิกปุ่ม Add เพิ่มลงในระบบ
5. ทำซ้ำข้อ 3-5 จนครบจำนวนกลุ่มที่ต้องการแบ่ง
6. คลิกปุ่ม Transform เพื่อแปลงข้อมูล
7. ระบบแสดงข้อมูลของแอตทริบิวที่ต้องการแปลง และข้อมูลที่ได้หลังการแปลงแบบ Numeric to Categorical



รูปที่ 5.27 ขั้นตอนการแปลงข้อมูลจากตัวเลขเป็นตัวอักษร

จากรูปที่ 5.28 ระบบแสดงค่า Min และค่า Max ของแอตทริบิวที่ต้องการแปลง ให้ผู้ใช้ระบุจำนวนกลุ่มที่ต้องการ จากนั้นให้กำหนดช่วงของตัวเลขที่ต้องการ แล้วกำหนดข้อความที่จะแทนลงไป คลิกที่ปุ่ม Add แล้วกำหนดแบบเดิมจนครบตามจำนวนกลุ่มที่ระบุ ขั้นตอนสุดท้ายให้คลิกที่ปุ่ม Transform เพื่อทำการแปลงค่าตามที่กำหนด

ตัวอย่าง การแปลงข้อมูลอายุให้เป็นข้อความ 3 ข้อความตามช่วงอายุ เช่น

- อายุช่วง 19-24 แปลงค่าเป็น Young
- อายุช่วง 25-39 แปลงค่าเป็น Middle
- อายุช่วง 40-60 แปลงค่าเป็น Senior

รูปที่ 5.28 ตัวอย่างการกำหนดข้อความให้กับข้อมูลที่ต้องการแปลง

รูปที่ 5.29 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว age ก่อนที่จะแปลงข้อมูล คอลัมน์ที่สองแสดงข้อมูลที่ได้หลังจากการแปลงจากข้อมูลแบบ Numeric เป็นข้อมูลแบบ Category โดยชื่อแอตทริบิวจะขึ้นต้นด้วยชื่อแอตทริบิวเดิมต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

	age	age_Category
	28	Middle
	20	Young
	27	Middle
	37.6421052631579	Middle
	40	Senior
	31	Middle
	37	Middle
	40	Senior
	32	Middle
	32	Middle
	30	Middle
	39	Middle
	38	Middle

รูปที่ 5.29 ข้อมูลหลังการแปลงแบบ Numeric to Category

5.4.4 Categorical to Numeric

5.4.4.1 One of N Coding

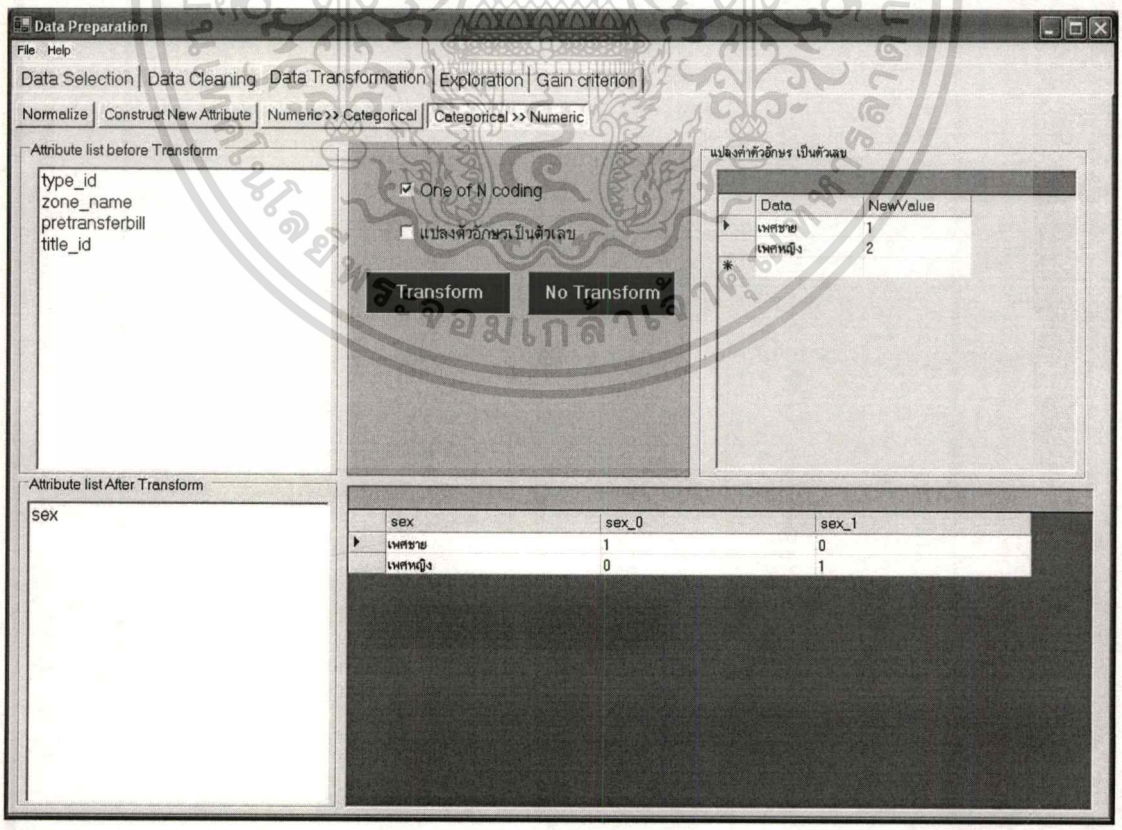
ตัวอย่าง การแปลงค่าในแอตทริบิว sex ที่มีค่า ชาย, หญิง ให้เป็นตัวเลข ได้ข้อมูลดังตารางที่ 5.1

ตารางที่ 5.1 ตัวอย่างของการแปลงค่าแบบ One of N Coding

sex	sex_0	sex_1
ชาย	1	0
หญิง	0	1

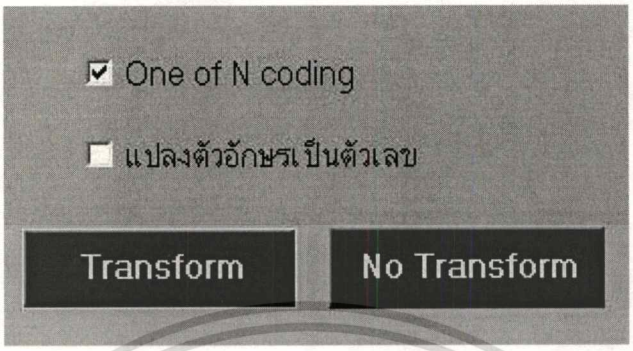
ขั้นตอนการใช้งานโปรแกรม มีดังนี้

1. คลิกที่แท็บ Categorical to Numeric
2. คลิกเลือกที่แอตทริบิวที่ต้องการแปลงค่า
3. คลิกเลือก One of N Coding
4. คลิกที่ปุ่ม Transform
5. ระบบแสดงค่าของแอตทริบิวที่ได้หลังการแปลงแบบ One of N Coding



รูปที่ 5.30 การแปลงข้อมูลจากตัวอักษรเป็นตัวเลขวิธี One of N Coding

รูปที่ 5.11 แสดงการเลือกวิธีการแปลงข้อมูลแบบ One of N coding โดยคลิกเลือกที่วิธีแรก ผู้ใช้สามารถเลือกวิธีการแปลงข้อมูลได้ทั้งสองวิธีพร้อมกัน ซึ่งการแปลงข้อมูลให้เป็นตัวเลขดูรายละเอียดได้ในหัวข้อ 5.4.4.2



รูปที่ 5.31 การแปลงข้อมูลจากตัวอักษรเป็นตัวเลขวิธี One of N Coding

รูปที่ 5.11 คอลัมน์แรกแสดงข้อมูลของแอตทริบิว sex ก่อนทำการแปลงข้อมูลแบบ One of N Coding คอลัมน์ที่สองและคอลัมน์ที่สามแสดงข้อมูลที่ได้หลังจากการแปลงข้อมูลแล้ว

	sex	sex_0	sex_1
▶	เพศชาย	1	0
	เพศหญิง	0	1

รูปที่ 5.32 การแปลงข้อมูลจากตัวอักษรเป็นตัวเลขวิธี One of N Coding

5.4.4.2 การแปลงข้อมูลให้เป็นตัวเลข

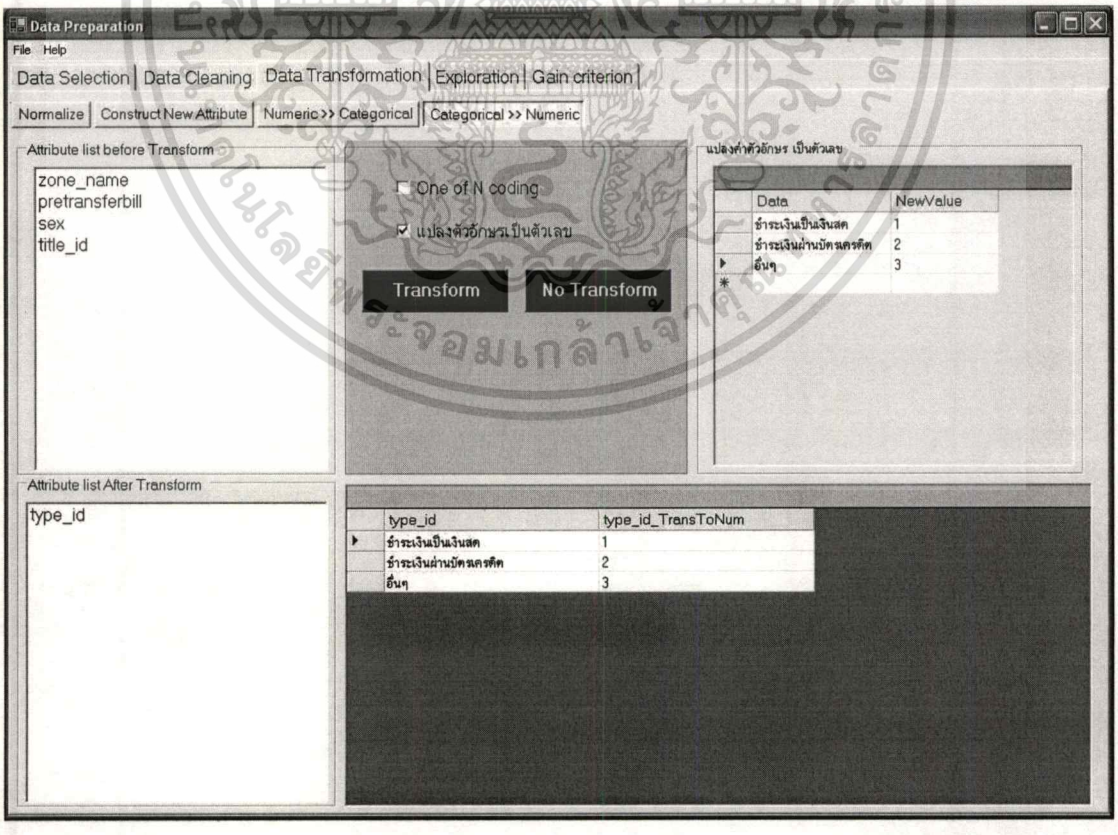
การแปลงข้อมูลที่ไม่ใช่ตัวเลขให้เป็นตัวเลข ตัวอย่างการแปลงค่าในแอตทริบิว sex ที่มีค่า ชาย, หญิง ให้เป็นตัวเลข ได้ข้อมูลดังตารางที่ 5.2

ตารางที่ 5.2 ตัวอย่างของการแปลงค่าให้เป็นตัวเลข

sex	sex_TransToNom
ชาย	1
หญิง	2

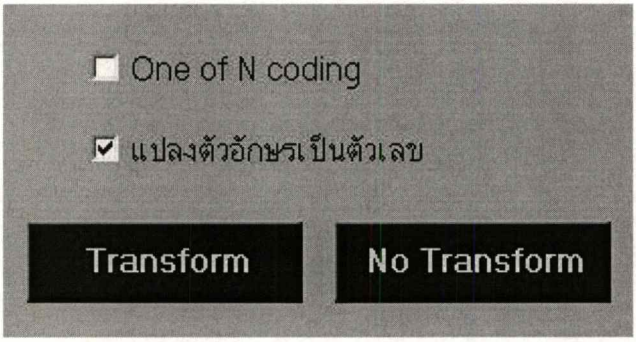
ขั้นตอนการใช้งานโปรแกรม มีดังนี้

1. คลิกที่แท็บ Categorical to Numeric
2. คลิกเลือกที่แอตทริบิวที่ต้องการแปลงค่า
3. คลิกเลือก แปลงตัวอักรเป็นตัวเลข
4. ระบบจะแสดงค่า DATA และ NEW VALUE โดยผู้ใช้ระบบสามารถแก้ไขค่า NEW VALUE ให้เป็นค่าตัวเลขตามที่ต้องการ ได้ คลิกที่ปุ่ม Transform
5. ระบบแสดงค่าของแอตทริบิวก่อนการแปลงข้อมูล และหลังการแปลงให้ทราบ



รูปที่ 5.33 ขั้นตอนการแปลงข้อมูลจากตัวอักษรเป็นตัวเลข

รูปที่ 5.34 แสดงการเลือกวิธีแปลงข้อมูลแบบ แปลงตัวอักษรเป็นตัวเลข



รูปที่ 5.34 ขั้นตอนการแปลงข้อมูลจากตัวอักษรเป็นตัวเลข

รูปที่ 5.35 ในคอลัมน์ Data แสดงข้อมูลแบบไม่ซ้ำของแอตทริบิว Type_id ก่อนทำการแปลงข้อมูล ในคอลัมน์ NewValue ระบบทำการระบุตัวเลขให้กับข้อมูลโดยอัตโนมัติ ซึ่งผู้ใช้สามารถกำหนดตัวเลขอื่นให้กับข้อมูลได้

	Data	NewValue
	ชำระเงินเป็นเงินสด	1
	ชำระเงินผ่านบัตรเครดิต	2
▶	อื่นๆ	3
*		

รูปที่ 5.35 ข้อมูลจากแอตทริบิวที่ต้องการแปลงและค่าใหม่ที่เป็นตัวเลข

รูปที่ 5.36 ในคอลัมน์แรก type_id แสดงข้อมูลเดิมก่อนทำการแปลงข้อมูล คอลัมน์ที่สอง type_id_TransToNum แสดงข้อมูลที่ได้หลังทำการแปลงแล้ว

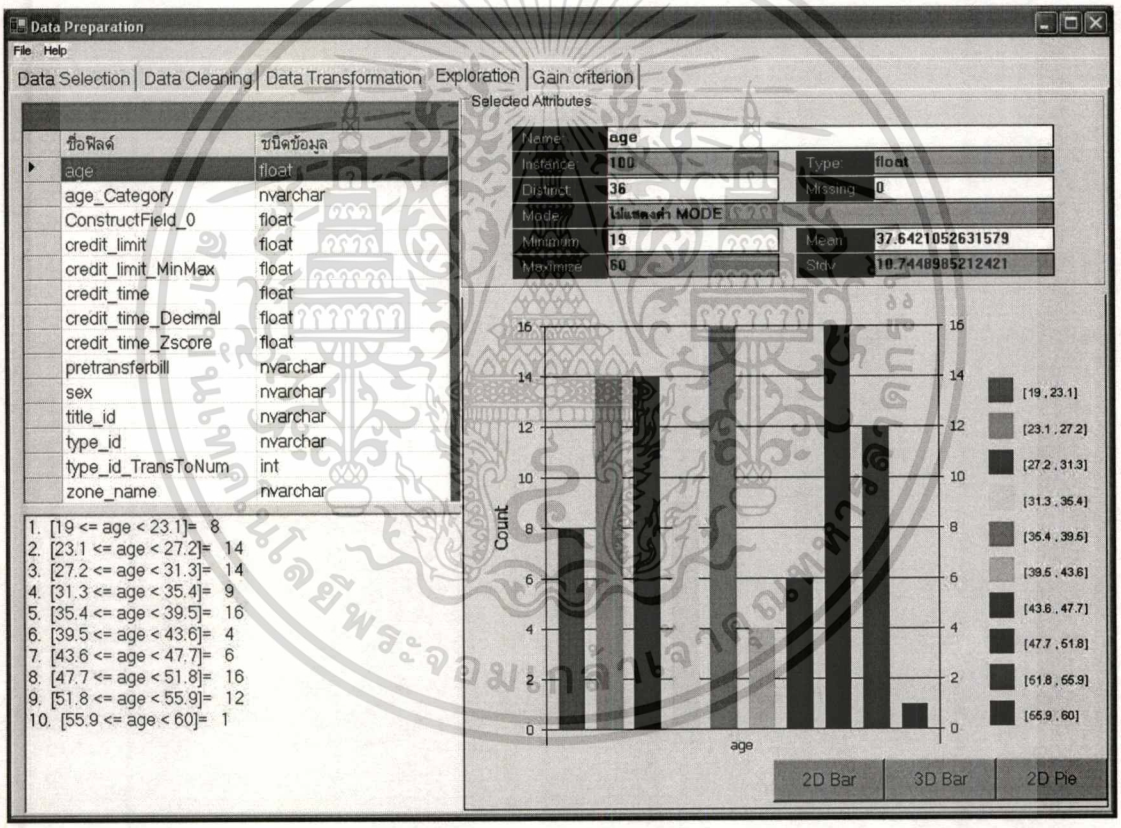
	type_id	type_id_TransToNum
▶	ชำระเงินเป็นเงินสด	1
	ชำระเงินผ่านบัตรเครดิต	2
	อื่นๆ	3

รูปที่ 5.36 ข้อมูลจากการแปลงข้อมูล Category เป็น Numeric

5.5 การสำรวจข้อมูล (Data Exploration)

5.5.1 การสำรวจข้อมูลที่เป็น Numeric

1. หลังจากทำขั้นตอนการแปลงข้อมูลเรียบร้อยแล้ว คลิกที่แท็บ Exploration
2. ระบบแสดงรายชื่อแอตทริบิวทั้งหมดที่ได้จากการแปลงข้อมูล
3. คลิกเลือกชื่อแอตทริบิว
4. ระบบจะแสดง จำนวนเรคคอร์ด ประเภทของแอตทริบิว ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน
5. เลือกรูปแบบการแสดงกราฟได้ 3 แบบคือ แบบกราฟแท่ง 2 มิติ, กราฟแท่ง 3 มิติ และแบบวงกลม



รูปที่ 5.37 การสำรวจข้อมูลที่เป็น Numeric

จากรูปที่ 5.38 ระบบแสดงรายชื่อฟิลด์ทั้งหมดที่ได้จากขั้นตอน Data Selection และแอตทริบิวที่ได้หลังจากขั้นตอน Data Transformation

	ชื่อฟิลด์	ชนิดข้อมูล
▶	age	float
	age_Category	nvarchar
	ConstructField_0	float
	credit_limit	float
	credit_limit_MinMax	float
	credit_time	float
	credit_time_Decimal	float
	credit_time_Zscore	float
	pretransferbill	nvarchar
	sex	nvarchar
	title_id	nvarchar
	type_id	nvarchar
	type_id_TransToNum	int
	zone_name	nvarchar

รูปที่ 5.38 รายชื่อแอตทริบิวทั้งหมดหลังขั้นตอน Data Transformation

รูปที่ 5.39 เมื่อคลิกที่ชื่อแอตทริบิวที่ต้องการดูข้อมูลแล้ว ในส่วนนี้จะเป็นส่วนแสดงรายละเอียดของแอตทริบิวให้ผู้ใช้ทราบ

Selected Attributes

Name:	age		
Instance:	100	Type:	float
Distinct:	36	Missing:	0
Mode:	ไม่แสดงค่า MODE		
Minimum	19	Mean	37.6421052631579
Maximize	60	Stdv	10.7448985212421

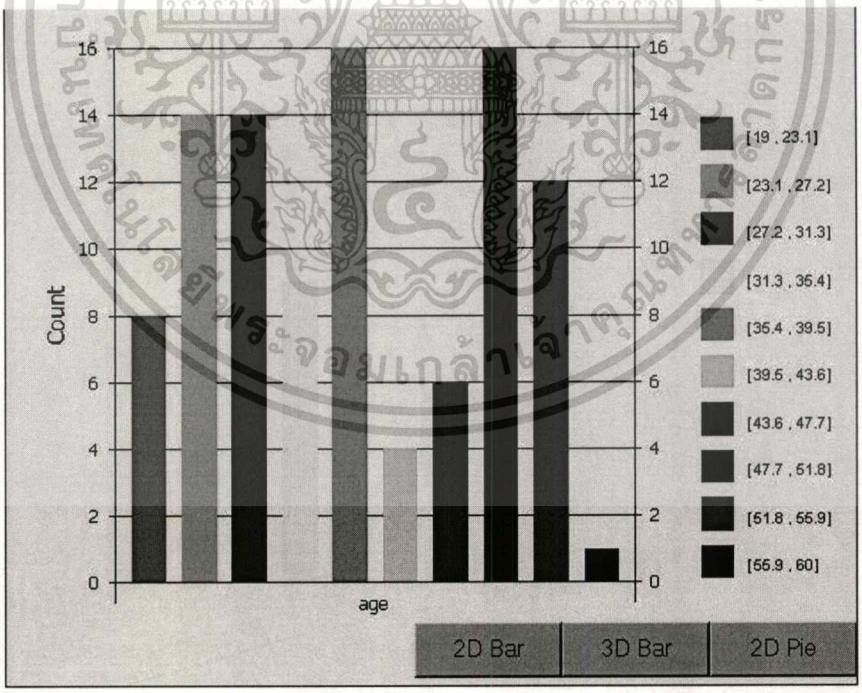
รูปที่ 5.39 รายชื่อแอตทริบิวทั้งหมดหลังขั้นตอน Data Transformation

รูปที่ 5.40 แสดงจำนวนเรคคอร์ดที่นับได้ในแต่ละช่วงข้อมูล โดยระบบกำหนดให้ข้อมูลที่เป็น Numeric มี 10 ช่วงดังรูป ซึ่งข้อมูลที่แสดงในส่วนนี้จะสัมพันธ์กับกราฟในรูปที่ 5.41

```
1.  [19 <= age < 23.1]=  8
2.  [23.1 <= age < 27.2]=  14
3.  [27.2 <= age < 31.3]=  14
4.  [31.3 <= age < 35.4]=  9
5.  [35.4 <= age < 39.5]=  16
6.  [39.5 <= age < 43.6]=  4
7.  [43.6 <= age < 47.7]=  6
8.  [47.7 <= age < 51.8]=  16
9.  [51.8 <= age < 55.9]=  12
10.  [55.9 <= age < 60]=  1
```

รูปที่ 5.40 รายละเอียดของข้อมูลที่เป็น Numeric

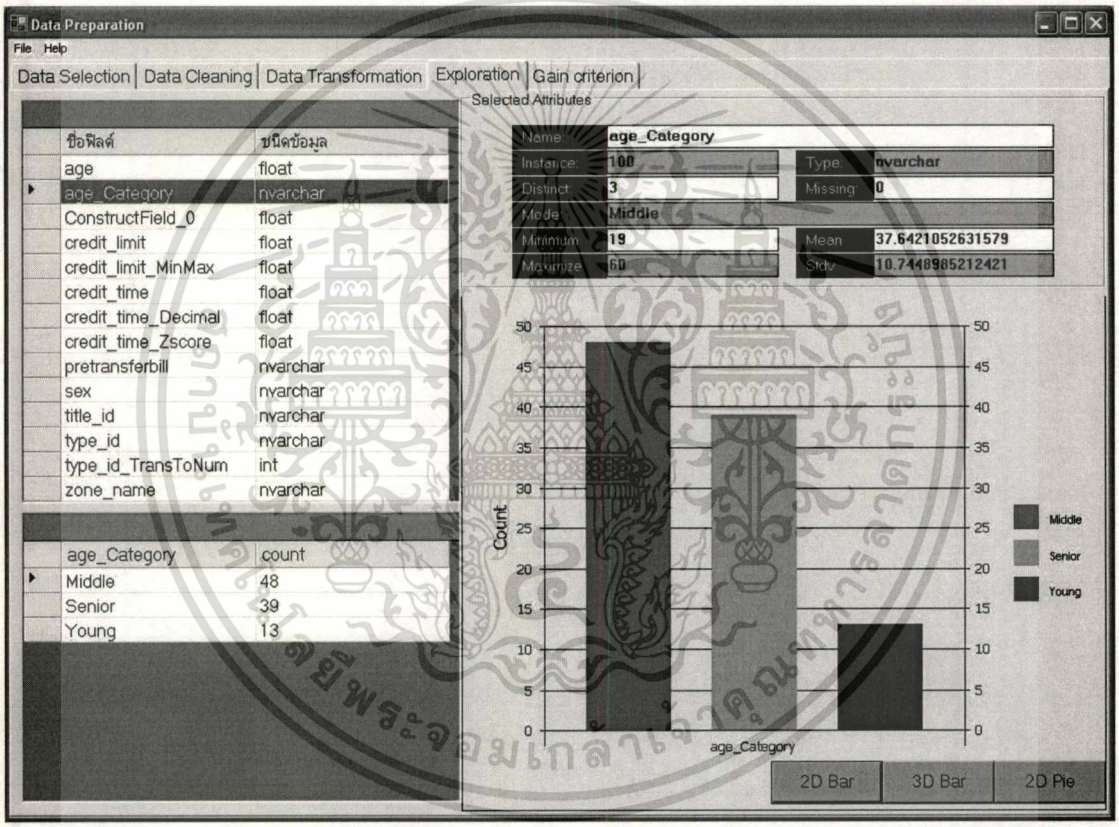
รูปที่ 5.41 แสดงกราฟแท่งของแอตทริบิว age ซึ่งสามารถเปลี่ยนการแสดงผลได้ 3 รูปแบบ คือ กราฟแท่ง 2 มิติ กราฟแท่ง 3 มิติ และกราฟวงกลม โดยคลิกที่ปุ่มด้านล่าง



รูปที่ 5.41 กราฟแท่งแสดงข้อมูลของแอตทริบิว age

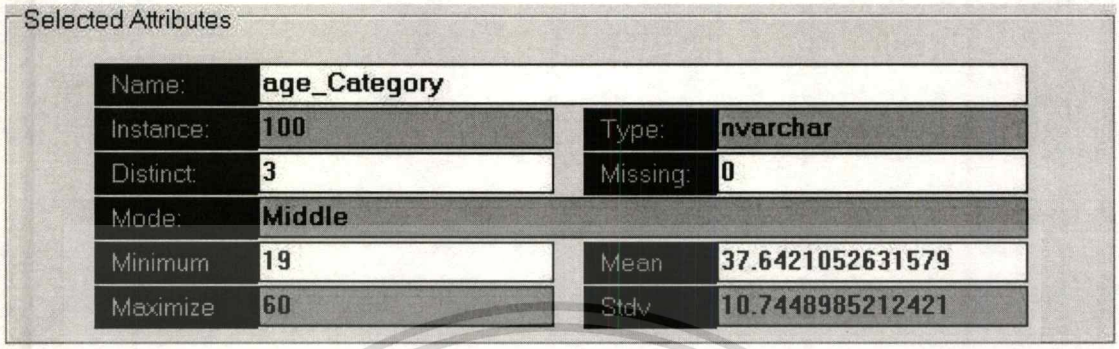
5.5.2 การสำรวจข้อมูลที่เป็น Category

1. คลิกเลือกชื่อแอตทริบิว
2. กราฟที่แสดงได้จากการหาค่าที่ต่างกัน แล้วนับจำนวนข้อมูลแต่ละตัว จากรูปที่ 4.14 เป็นการคลิกเลือกที่แอตทริบิว age_Category ที่ได้จากการแปลงค่าในขั้นตอน Data Transformation โดยค่าที่แปลงจากอายุ เป็นตัวอักษรแบ่งได้ 3 ค่า คือ
   - Middle จำนวน 48 ค่า
   - Senior จำนวน 39 ค่า
   - Young จำนวน 13 ค่า



รูปที่ 5.42 การสำรวจข้อมูลที่เป็น Categoryแสดงกราฟแท่ง

รูปที่ 5.43 เมื่อคลิกที่ชื่อแอตทริบิวที่ต้องการดูข้อมูลแล้ว ในส่วนนี้จะเป็นส่วนแสดงรายละเอียดของแอตทริบิวให้ผู้ใช้ทราบ รูปที่ 5.43 แสดงรายละเอียดของแอตทริบิว age_Category ที่ได้จากขั้นตอน Data Transformation



รูปที่ 5.43 รายละเอียดของแอตทริ age_Category

รูปที่ 5.44 แสดงข้อมูลในแอตทริบิว age_Category ที่เลือกมาแบบไม่ซ้ำและจำนวนเรคคอร์ดที่นับได้

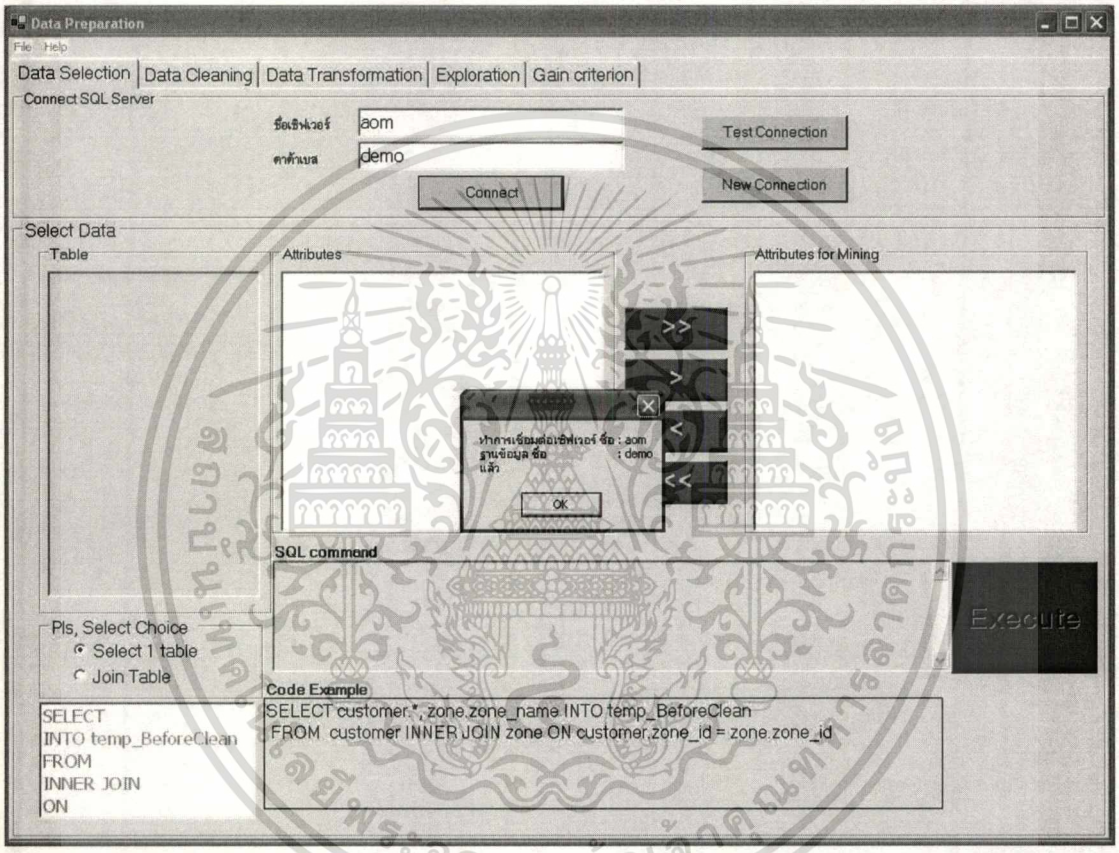
	age_Category	count
▶	Middle	48
	Senior	39
	Young	13

รูปที่ 5.44 ข้อมูลในแอตทริบิว age_Category

5.6 การหาค่า Information Gain สำหรับข้อมูลที่เป็น Categorical

5.6.1 การติดต่อกับฐานข้อมูล

1. ผู้ใช้ระบบต้องทำการกรอกข้อมูลคือ ชื่อเซิฟเวอร์ และ ชื่อดาต้าเบส
2. กดปุ่ม Connect เพื่อเชื่อมต่อกับฐานข้อมูลตามที่เราระบุ เพื่อเข้าสู่ขั้นตอนการเลือกข้อมูลต่อไป
3. ระบบแสดงข้อความให้ทราบว่าได้ทำการเชื่อมต่อกับฐานข้อมูลเรียบร้อยแล้ว



รูปที่ 5.45 ขั้นตอนการติดต่อกับฐานข้อมูล

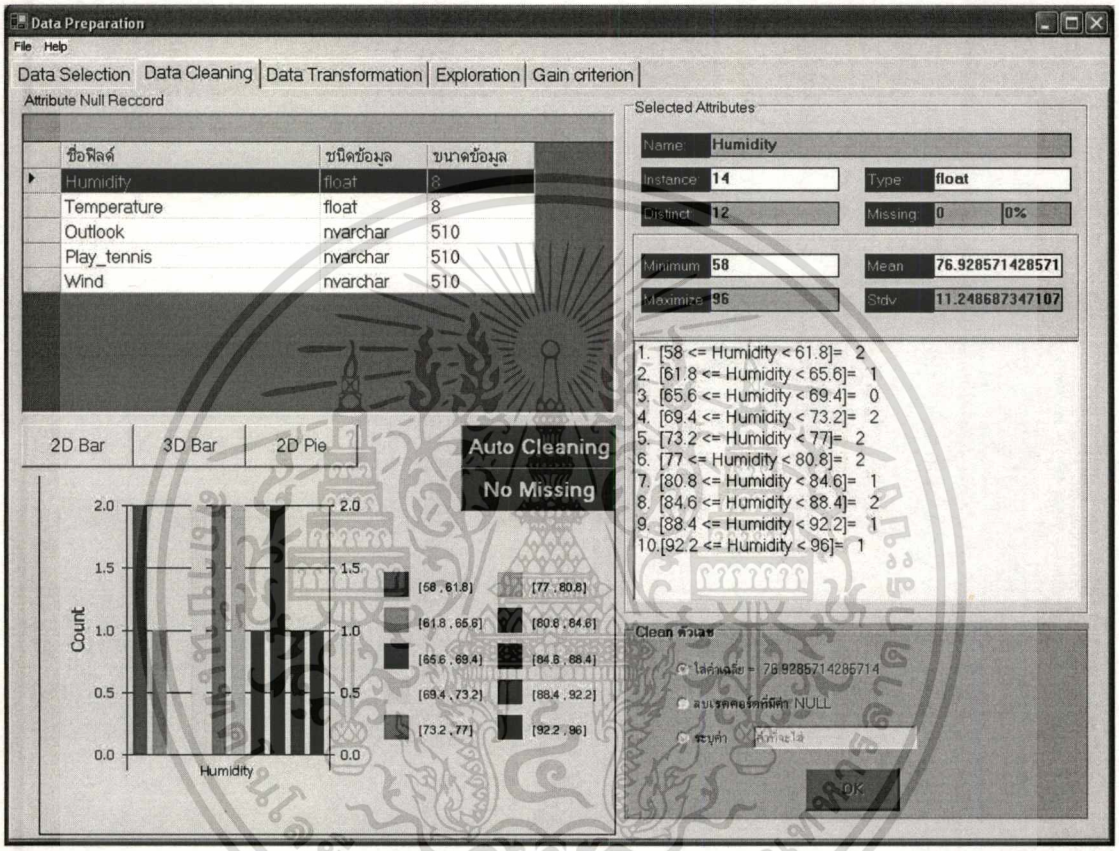
5.6.2 การเลือกข้อมูล

1. เมื่อติดต่อกับฐานข้อมูลแล้ว ตัวอย่างนี้ให้เลือกตาราง Play_tennis
2. เลือกแอตทริบิวในการทำดาต้าไมนิ่ง ดังนี้ Humidity, Outlook, Play_tennis, Wind
3. คลิกที่ปุ่ม Excute สิ้นสุดขั้นตอนการเลือกข้อมูล เพื่อเข้าสู่การเตรียมข้อมูลในขั้นตอนต่อไป

รูปที่ 5.46 ขั้นตอนการเลือกข้อมูล

5.6.3 การคลีนข้อมูล

1. ในตัวอย่าง การหาค่า Information Gain นี้ให้คลิกที่ปุ่ม Auto Cleaning เพื่อเป็นการทำ Data Cleaning โดยการลบเรคคอร์ดที่มีค่า Null ทิ้ง
2. ขั้นตอนการคลีนข้อมูลที่เป็น Numeric และ Categorical ทีละแอตทริบิว ใช้วิธี ดังรูปที่ 5.4 และรูปที่ 5.9



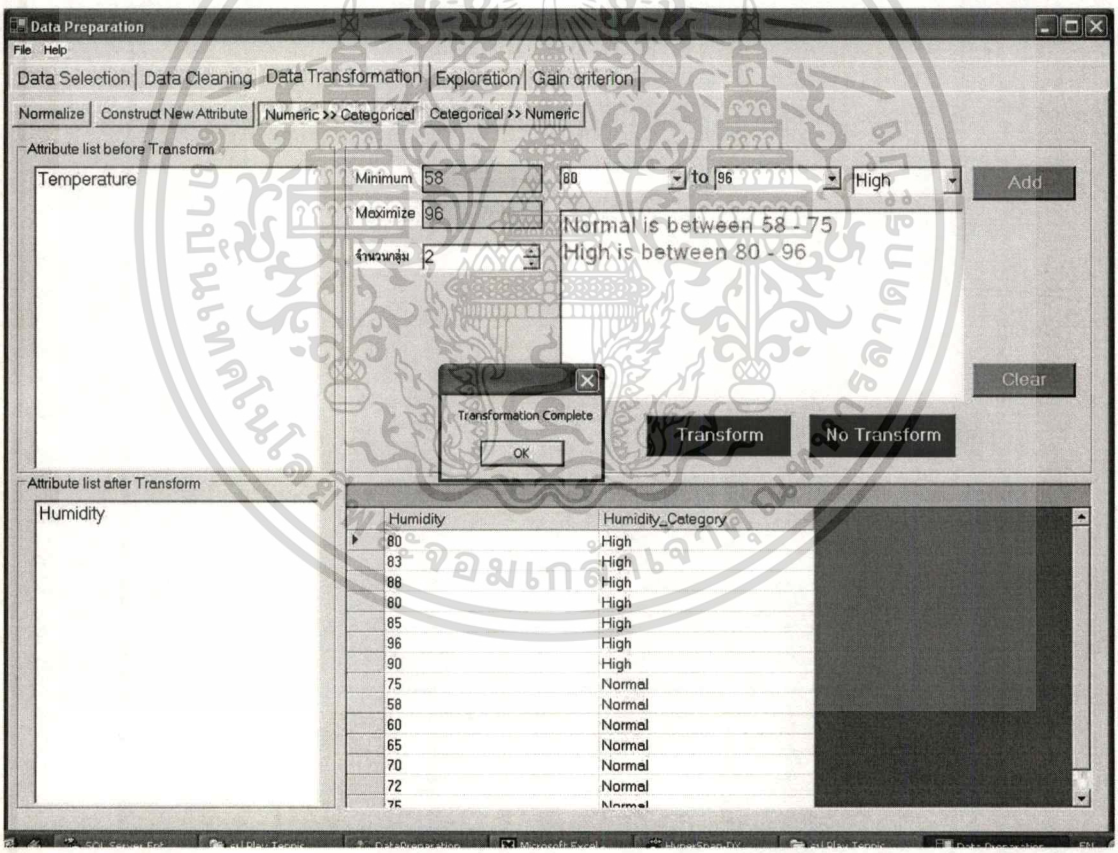
รูปที่ 5.47 ขั้นตอนการคลีนข้อมูล

รูปที่ 5.48 เมื่อคลิกที่ปุ่ม Auto Cleaning แล้วระบบจะแสดงข้อความบอกให้ผู้ใช้ทราบว่าได้ทำการคลีนข้อมูลเรียบร้อยแล้ว พร้อมเข้าสู่ขั้นตอนต่อไป

รูปที่ 5.48 ข้อความแสดงหลังจากคลิกปุ่ม Auto Clean

5.6.4 การแปลงข้อมูล

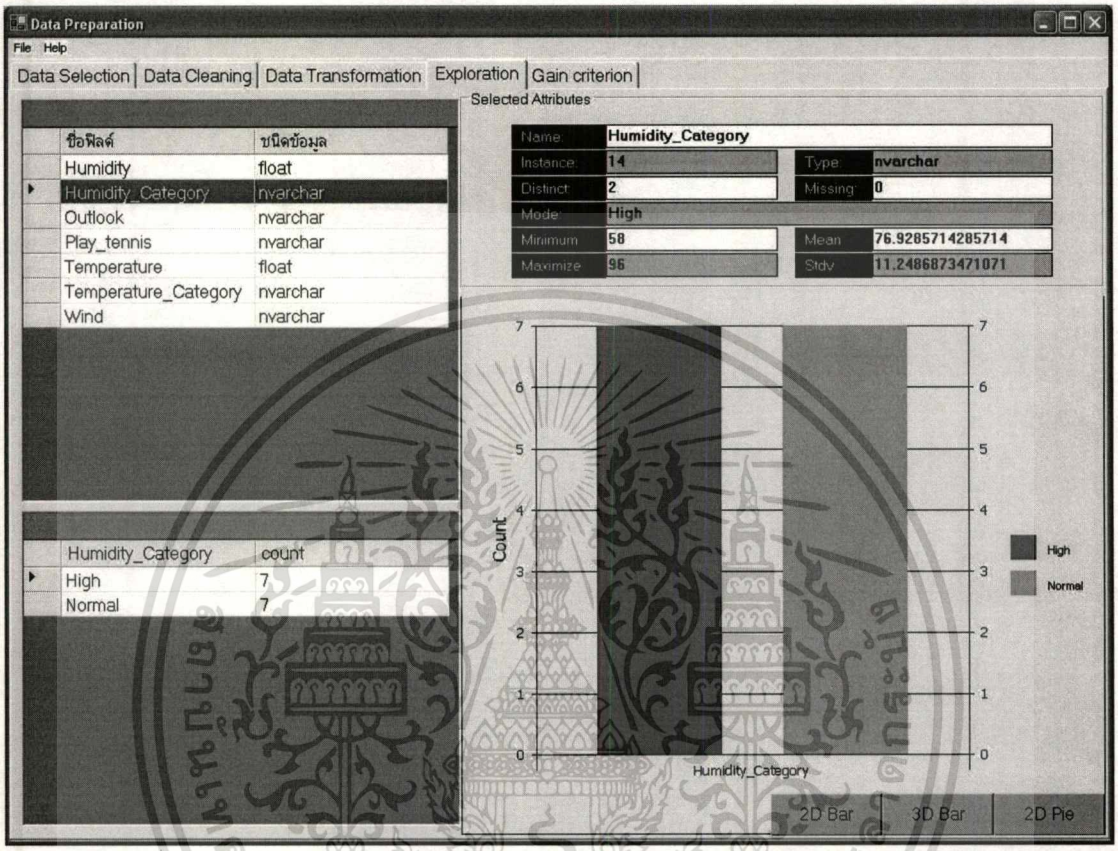
1. การหาค่า Information Gain ให้กับแต่ละแอตทริบิว สามารถทำได้กับข้อมูลที่เป็น Categorical เท่านั้น ดังตัวอย่างนี้จึงต้องทำการแปลงข้อมูลของ แอตทริบิว Humidity และ Temperature ให้เป็น Categorical ก่อน
2. คลิกที่ชื่อแอตทริบิวที่ต้องการแปลงค่า
3. กำหนดจำนวนกลุ่มที่ต้องการแปลงค่า
4. ระบุค่า Min และค่า Max ของช่วงที่ต้องการแปลงค่า
5. ระบุข้อความที่ต้องการแทนข้อมูลที่เป็น Numeric
6. คลิกที่ปุ่ม Add เพื่อแสดงช่วงข้อมูลที่ต้องการแปลงค่า
7. ระบุช่วงข้อมูลจนครบตามกลุ่มที่กำหนด
8. คลิกที่ปุ่ม Transform เพื่อทำการแปลงข้อมูล



รูปที่ 5.49 ขั้นตอนการแปลงข้อมูล

5.6.5 การสำรวจข้อมูล

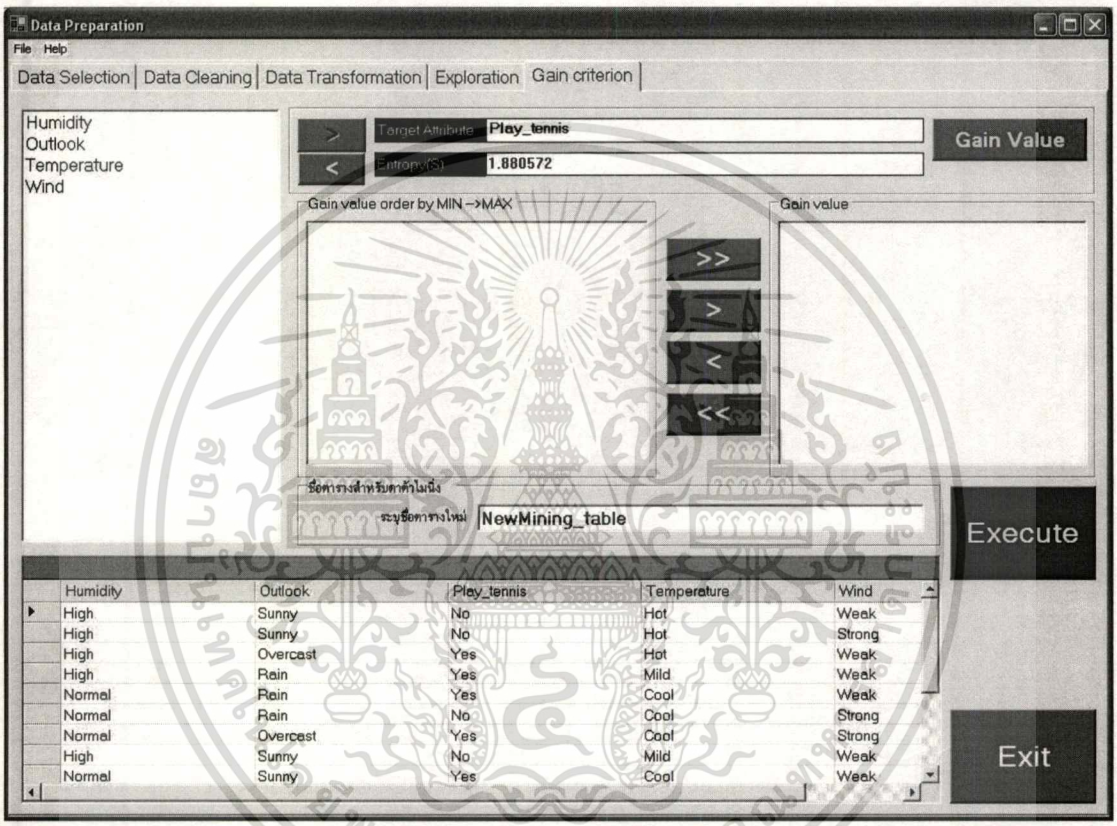
รูปที่ 5.50 คลิกที่ชื่อแอตทริบิวที่ต้องการสำรวจข้อมูล จากตัวอย่างคลิกที่ Humidity_Category ซึ่งเป็นแอตทริบิวที่ได้จากการแปลงมาจากแอตทริบิว Humidity



รูปที่ 5.50 การสำรวจข้อมูลที่ได้หลังจากการแปลงค่า

5.6.6 การหาค่า Entropy(s)

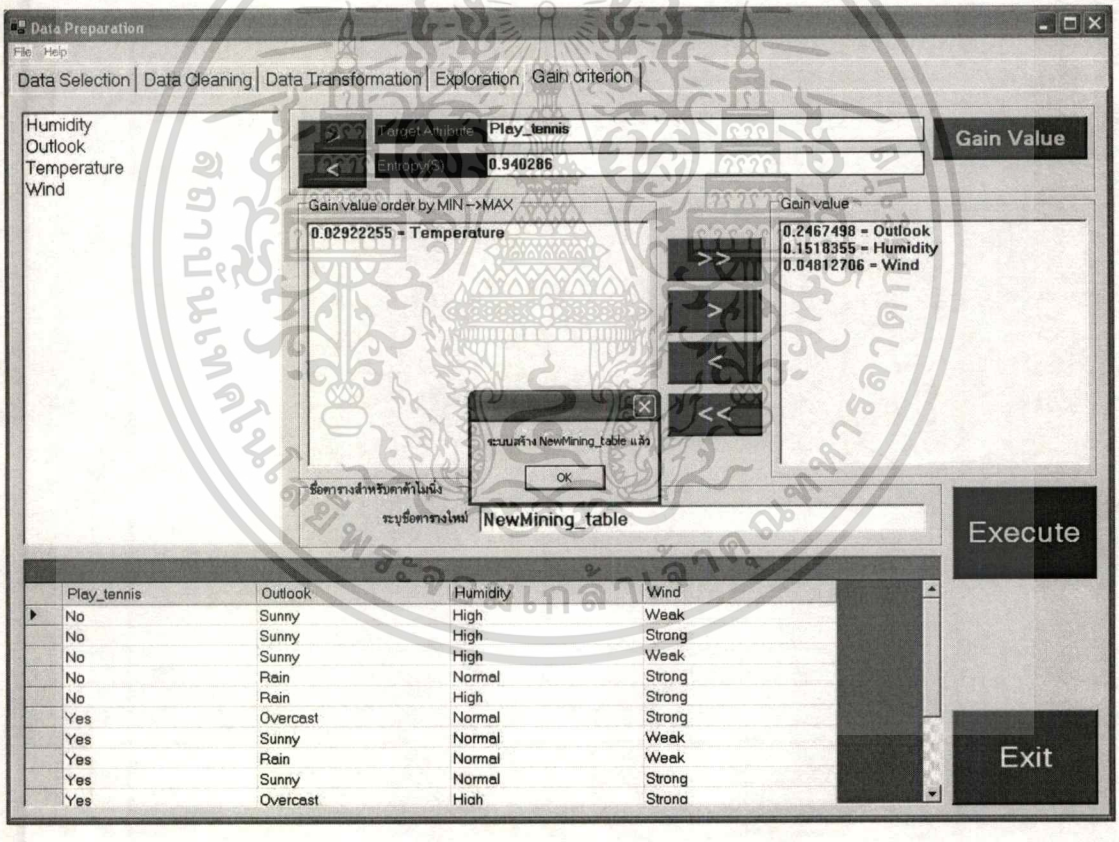
1. การหาค่า Information Gain สามารถหาได้จากแอตทริบิวที่เป็น Categorical
2. ผู้ใช้เลือก Target Attribute จากรายชื่อแอตทริบิวด้านซ้าย ระบบจะคำนวณค่า Entropy แล้วแสดงให้ทราบ
3. จากตัวอย่าง เลือก Play_tennis เป็น Target Attribute คำนวณค่า Entropy(s) ได้ 0.90286



รูปที่ 5.51 การหาค่า Entropy(S)

5.6.6 การหาค่า Information Gain

1. คลิกที่ปุ่ม Gain Value ระบบจะคำนวณค่า Gain ของแต่ละแอตทริบิวแล้วแสดงโดยเรียงลำดับค่า Gain จากน้อยไปมาก
2. ผู้ใช้สามารถนำค่า Gain ที่ได้มาพิจารณาประกอบในการเลือกข้อมูลเพื่อนำไปใช้ในการทำดาต้าไมนิ่งต่อไป โดยพิจารณาจากค่า Gain ที่มีค่ามาก มาน้อย
3. ผู้ใช้เลือก แอตทริบิวด้านซ้าย โดยพิจารณาจากค่า Gain ที่มีค่ามากก่อน
4. จากตัวอย่าง เลือก Outlook, Humidity_category และ Wind ที่มีค่า Gain จากมากมาน้อย ตามลำดับ
5. คลิกที่ปุ่ม Execute เพื่อสร้างตารางใหม่จากแอตทริบิวที่เลือกโดยผู้ใช้สามารถระบุชื่อตารางได้ตามต้องการ
6. ระบบแสดงข้อความว่า ได้สร้าง table_Mining แล้ว



รูปที่ 5.52 การเลือกแอตทริบิวที่ได้จากการหาค่า Gain

5.6.7 การออกจากโปรแกรม

1. คลิกที่ปุ่ม Exit เพื่อออกจากโปรแกรม หรือคลิกที่เมนูบาร์ File แล้วเลือก Exit
2. ระบบแสดงข้อความเตือนการออกจากระบบ
3. กด Yes เพื่อยืนยันการออกจากระบบ

รูปที่ 5.53 ข้อความยืนยันการออกจากระบบ

บทที่ 6

สรุปผลการศึกษาและข้อเสนอแนะ

6.1 สรุปผลการศึกษา

จากการศึกษาทฤษฎีของดาต้าไมนิ่ง ทำให้ได้เรียนรู้ว่าดาต้าไมนิ่งเป็นกระบวนการที่ค้นหาข้อมูลที่เป็นประโยชน์จากภายในฐานข้อมูลที่มีอยู่ ทำให้ได้รับสารสนเทศที่เป็นประโยชน์ และสามารถนำสารสนเทศนั้นไปช่วยสนับสนุนการตัดสินใจและการประยุกต์นำไปใช้งานกับธุรกิจต่างๆได้ ซึ่งกระบวนการดังกล่าวจะเริ่มตั้งแต่การกำหนดวัตถุประสงค์ของการทำดาต้าไมนิ่ง จากนั้นก็มีขั้นตอนการเตรียมข้อมูลมาวิเคราะห์ ซึ่งจะประกอบไปด้วยการคัดเลือกข้อมูล การทำความสะอาดข้อมูล และการแปลงข้อมูลให้เหมาะสม หลังจากนั้นก็จะทำดาต้าไมนิ่งเมื่อข้อมูลผ่านการทำไมนิ่ง ก็จะได้ผลลัพธ์ที่เกิดประโยชน์ในทางธุรกิจ

ก่อนที่จะเข้าสู่ขั้นตอนของการทำดาต้าไมนิ่งได้นั้น การเตรียมข้อมูลถือได้ว่าเป็นขั้นตอนที่ใช้ระยะเวลาในการดำเนินการมากกว่าขั้นตอนอื่นๆ ของการทำดาต้าไมนิ่ง เนื่องจากปริมาณข้อมูลมีเป็นจำนวนมากและข้อมูลที่ได้รับมาจากหลายแหล่ง รูปแบบของข้อมูลแตกต่างกัน จึงต้องมีการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกัน เพื่อให้พร้อมใช้งาน แต่ละอัลกอริทึมของดาต้าไมนิ่งก็ต้องการการนำเข้าข้อมูลแตกต่างกัน อัลกอริทึมบางประเภทใช้เพื่อวิเคราะห์ข้อมูลที่เป็น Numeric เท่านั้น จึงต้องมีการแปลงข้อมูลเหล่านั้นให้เหมาะสมกับอัลกอริทึมแต่ละแบบ

6.2 ข้อเสนอแนะ

1. ระบบนี้เลือกข้อมูลได้ครั้งละหนึ่งฐานข้อมูลเท่านั้น จึงควรเพิ่มเติมในส่วนของการเลือกข้อมูลให้สามารถเชื่อมต่อได้มากกว่าหนึ่งฐานข้อมูล
2. ในการเลือกข้อมูลจากหลายตาราง ผู้ใช้ระบบจะต้องเข้าใจความสัมพันธ์ของแต่ละตาราง แต่ละแอตทริบิว เพื่อใช้คำสั่ง SQL ในการเลือกข้อมูลเข้าสู่ระบบนี้ได้อย่างถูกต้อง
3. ระบบที่พัฒนาขึ้นสามารถเชื่อมต่อกับฐานข้อมูลในระบบ Microsoft SQL Server 2000 เท่านั้น หากผู้ใช้ต้องการทำงานกับโปรแกรมฐานข้อมูลระบบอื่น จะต้องทำการนำเข้าข้อมูลโดยผ่าน DTS Import/Export Wizard ที่ Microsoft SQL Server 2000 รองรับ

บรรณานุกรม

ธาริน สิทธิธรรมชารี. **Microsoft SQL Server 2000 ฉบับสมบูรณ์.** ซัคเซส มีเดีย

บัญชา ปะสีละเตสัง. 2545. **เริ่มต้นการเขียนโปรแกรม VISUAL BASIC .NET.** กรุงเทพฯ : ซีเอ็ด ยูคชั่น.

สมพร จิวรสกุล. 2545. **คู่มือการติดตั้งและใช้งาน Microsoft SQL Server 2000 ฉบับสมบูรณ์.** กรุงเทพฯ : อินโฟเพรส.

สุรสิทธิ์ คิวประสพศักดิ์ และนันทนี แขวงโสภา. 2546. **อินไซท์ Visual Basic .NET ฉบับสมบูรณ์.** กรุงเทพฯ : โปรวิชัน.

Dorain Pyle. 1999. **Data Preparation for Data Mining.** Morgan Kaufmann.

Ian H. Witten and Eibe Frank. 2005. **Data Mining Practical Machine Learning Tools and Techniques 2nd Edition.** Morgan Kaufmann.

Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada. **Data Mining: Concepts and Techniques Chapter 3.** [Online]. Available: www.cs.sfu.ca/~han/bk/3prep.ppt

Jiawei Han and Micheline Kamber. 2001. **Data Mining: Concepts and Techniques.** USA : Academic Press.

Kira Tarapanoff. 2001. **Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil.** [Online]. Available: http://informationr.net/ir/7-1/paper117.html

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวอาทิตยา เชื้อจันอัด
วันเกิด	27 พฤษภาคม 2522
สถานที่เกิด	นครราชสีมา
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต
สถานที่สำเร็จการศึกษา	คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
ปีที่สำเร็จการศึกษา	2545