การพัฒนาเครื่องมือสำหรับการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm

Developing tool for clustering data using K-Prototypes Algorithm

โดย

สาลินี รุ่งเรือง

รหัสประจำตัว 47066103

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคฤดูร้อน ปีการศึกษา 2548

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาเครื่องมือสำหรับการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm
นักศึกษา	นางสาวสาลินี รุ่งเรือง
อาจารย์ที่ปรึกษา	ผศ.ดร.พรฤดี เนติโสภากุล
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

ในสภาวการณ์ปัจจุบันมีการแข่งขันกันสูงมากในงานทางด้านธุรกิจการค้า บริษัทต่าง ๆ พยายามที่จะหาช่องทางในการเพิ่มโอกาสทางธุรกิจให้กับตนเอง ซึ่งแนวทางหนึ่งคือ การนำข้อมูลที่มีอยู่ในระบบมาวิเคราะห์ และจัดกลุ่มประเภทของข้อมูลตามลักษณะและคุณสมบัติที่คล้ายคลึงกัน ก่อให้เกิดข้อมูลใหม่ ๆ ที่สามารถช่วยในการคาดคะเนและสนับสนุนการตัดสินใจของผู้บริหาร โดยการจัดกลุ่มข้อมูลที่ได้กล่าวมานั้น จะใช้อัลกอริทึม K-Prototypes เพื่อนำไปพัฒนาเครื่องมือในการจัดกลุ่มข้อมูล ซึ่งสามารถใช้ได้กับข้อมูลทั้งประเภทที่เป็นตัวเลขและไม่ใช่ตัวเลข ซึ่งเหมาะสมกับการข้อมูลที่ใช้ในการทำงานจริงในปัจจุบัน

Title	Developing tool for cluster data using K-Prototypes Algorithm
Student	Miss Salinee Rungruang
Advisor	Asst.Prof.Dr. Ponrudee Natisopakul
Level of study	Master of science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

There are a lot of business competitions in this circumstance nowadays. Many companies are trying to find the best way to earn more opportunities in their business. One way is to analyze all data in the system and cluster similar data by their characteristics. These actions will create new information that will help executives to make right decisions. In this project, we study K-Prototypes algorithm. K-Prototypes algorithm is a tool for clustering data. Using K-Prototypes algorithm will help clustering data in both text and number which appropriate to the data that we use in the real business these days.

กิตติกรรมประกาศ

ในการทำโครงงานเรื่องการพัฒนาเครื่องมือสำหรับการจัดกลุ่มข้อมูลโดยใช้K-Prototypes Algorithm (Developing tool for cluster data using K-Prototypes Algorithm) เป็นระบบที่มีการทำงานหลายขั้นตอนโดยมีความช่วยเหลือจากบุคคลหลาย ๆ ท่าน ทำให้การพัฒนาระบบนั้นสำเร็จลุล่วงไปได้ด้วยดี ข้าพเจ้าจึงขอขอบพระคุณ มา ณ ที่นี้ด้วย

ข้าพเจ้าต้องขอขอบพระคุณ ผศ.ดร.พรฤดี เนติโสภากุล อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานที่กรุณาให้คำแนะนำและเป็นที่ปรึกษา อันเป็นประโยชน์ต่อการพัฒนาระบบ รวมทั้งเป็นผู้ตรวจสอบความถูกต้องของโครงการฉบับนี้

ขอขอบพระคุณ บิดา มารดา และบุคคลในครอบครัว ที่ได้ให้ความสนับสนุนทาง ด้านกำลังใจในการเรียนจนการทำโครงการพัฒนาระบบนี้สำเร็จด้วยดี รวมทั้งขอขอบคุณพี่อั๋น และเพื่อน ๆ ห้อง KMAKE ที่แนะนำด้านเทคนิคต่าง ๆ, เพื่อน ๆ สามแม่ครัวกับโต้ง ที่มีกำลังใจให้เสมอมาไม่เคยขาด และ IS 17.1 ทุกคน ที่ให้ความช่วยเหลือในเรื่ องต่าง ๆ ที่เกี่ยวข้องกับโครงการทั้งหลายไว้ ณ ที่นี้ด้วย

สาลินี รุ่งเรือง

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....I
บทคัดย่อภาษาอังกฤษ..... II
กิตติกรรมประกาศ..... III
สารบัญ..... IV
สารบัญตาราง..... VII
สารบัญรูป..... IX
บทที่
1. บทนำ.....1
1.1 ความเป็นมาของโครงการ.....1
1.2 วัตถุประสงค์ของโครงการ.....1
1.3 ขอบเขตของการดำเนินงาน.....2
1.4 ขั้นตอนการศึกษา.....2
1.5 ประโยชน์ที่คาดว่าจะได้รับ.....2
2. ทฤษฎีที่เกี่ยวข้อง 3
2.1 ความหมายของดาต้าไมนิ่ง.....3
2.2 ขั้นตอนการทำดาต้าไมนิ่ง.....3
2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination).....3
2.2.2 การเตรียมข้อมูล (Data Preparation).....3
2.2.3 การทำดาต้าทำไมนิ่ง (Data Mining Operation).....6
2.2.4 การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้.....7
2.3 โมเดลการแบ่งกลุ่มข้อมูล.....8
2.4 อัลกอริทึม K-Prototypes.....9
2.4.1 หลักคณิตศาสตร์เบื้องต้นที่ใช้ใน K-Prototypes 9
2.4.2 การวัดความเหมือนกัน.....12

สารบัญ (ต่อ)

หน้า

2.4.3 การทำงานของอัลกอริทึม K-Prototypes Algorithm.....12

2.4.4 ตัวอย่างการนำข้อมูลมาใช้กับ K-Prototypes Algorithm.....16

3. การออกแบบและอิมพลีเมนต์ระบบ.....22

3.1 องค์ประกอบของระบบงาน 22

3.2 อัลกอริทึมของ K-Prototypes.....26

3.2.1 ขั้นตอนการทำงานหลักของระบบ (Main program).....26

3.2.2 ขั้นตอนการแก้ไขข้อมูล (Data Cleaning).....28

3.2.3 ขั้นตอนการปรับเปลี่ยนข้อมูล (Data Transformation).....29

3.2.4 ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm.....30

3.2.5 ขั้นตอนการคำนวณหาค่า Mindistance.....33

3.2.6 ขั้นตอนการคำนวณหาค่า Distance.....34

3.2.7 ขั้นตอนการคำนวณจุดศูนย์กลางกลุ่มใหม่.....36

3.3 เครื่องมือที่ใช้ในการพัฒนาระบบ.....37

3.4 การออกแบบหน้าจอการทำงานของระบบ.....38

3.4.1 การเลือกข้อมูล (Data Selection).....38

3.4.2 การทำความสะอาดข้อมูล (Data Cleaning).....39

3.4.3 การปรับเปลี่ยนข้อมูล (Data Transformation).....43

3.4.4 การกำหนดจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม.....44

3.4.5 การแสดงผลการจัดกลุ่มข้อมูล.....44

3.4.6 การส่งออกข้อมูล.....45

4. การทดลองใช้งานโปรแกรม.....46

4.1 คู่มือและตัวอย่างการใช้งาน.....46

4.1.1 ส่วนนำเข้าข้อมูล.....46

4.1.2 ส่วนประมวลผล.....46

สารบัญ (ต่อ)

หน้า

4.1.3 ส่วนแสดงผล……47
4.2 ขั้นตอนการประยุกต์ใช้งานระบบ……...47
4.3 การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้ ……...55
5. สรุปผลการศึกษาและข้อเสนอแนะ ……...57
5.1 สรุปผลการศึกษา……57
5.2 ข้อเสนอแนะ……...57
5.3 การประยุกต์ทดลองใช้งานโปรแกรมในรูปแบบต่าง ๆ ……....58
บรรณานุกรม……...68
ภาคผนวก ก คู่มือการใช้งานโปรแกรม……69
ประวัติผู้เขียน……...88

สารบัญตาราง

หน้า

ตารางที่

2.1 แสดงประเภทของข้อมูลที่จะนำมาจัดกลุ่ม.....17

2.2 แสดงข้อมูลที่จะนำมาจัดกลุ่ม.....17

2.3 แสดงสมาชิกข้อมูลในกลุ่มที่ 1.....18

2.4 แสดงสมาชิกข้อมูลในกลุ่มที่ 2.....18

2.5 แสดงสมาชิกข้อมูลในกลุ่มที่ 3.....18

2.6 แสดงสมาชิกของกลุ่มข้อมูลที่ 1 หลังจากการคำนวณ.....19

2.7 แสดงสมาชิกของกลุ่มข้อมูลที่ 2 หลังจากการคำนวณ.....20

2.8 แสดงสมาชิกของกลุ่มข้อมูลที่ 3 หลังจากการคำนวณ.....20

4.1 แสดงข้อมูลตาราง Customer ที่นำมาทำการคัดเลือกข้อมูลเพื่อจัดกลุ่ม.....48

5.1 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Actors.....58

5.2 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Bank.....59

5.3 แสดงฟิลด์และประเภทของข้อมูลจากตาราง SuperMKT.....59

5.4 แสดงฟิลด์และประเภทของข้อมูลจากตาราง GolfBall.....60

5.5 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Video.....60

5.6 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical ตาราง Actors.....60

5.7 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง Bank.....61

5.8 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง SuperMKT.....61

5.9 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง GolfBall.....61

5.10 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง Video.....61

5.11 แสดงผลการจัดกลุ่มข้อมูลประเภท Categorical.....62

5.12 แสดงค่าที่ match กันของข้อมูลประเภท Categorical.....61

5.13 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Actors.....63

5.14 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Bank.....63

5.15 แสดงฟิลด์และประเภทของข้อมูลจากตาราง SuperMKT.....63

สารบัญตาราง(ต่อ)

หน้า

ตารางที่

5.16 แสดงฟิลด์และประเภทของข้อมูลจากตาราง GolfBall……..…..64
5.17 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Video……...64
5.18 แสดงผลการจัดกลุ่มข้อมูลประเภท Numerical…….....64
5.19 แสดงค่าที่ match กันของข้อมูลประเภท Numerical……...65
5.20 แสดงผลการจัดกลุ่มข้อมูลแบบผสม……..65
5.21 แสดงค่าที่ match กันของข้อมูลแบบผสม……..66

สารบัญรูป

หน้า

รูปที่

2.1 แสดงภาพรวมเทคโนโลยีของกระบวนการคาต้าไมนิ่ง.....7
2.2 ลำดับการทำงานของอัลกอริทึมการจัดกลุ่มข้อมูล.....8
2.3 แสดงผลกระทบจาก weight γ_2 ในการแบ่งกลุ่มข้อมูล..... 12
2.4 แสดงกระบวนการเริ่มต้นการจัดกลุ่มโดย K-PrototypesAlgorithm.....14
2.5 แสดงกระบวนการจัดกลุ่มใหม่ของ K-Prototypes Algorithm.....15
2.6 แสดงกราฟการทำงานของ K-Prototypes Algorithm.....16
3.1 รูปแสดงภาพรวมของระบบ.....22
3.2 แสดงการทำงานในส่วนของ Data Selection 23
3.3 แสดงการทำงานในส่วนของ Data Preparation.....23
3.4 แสดงการทำงานในส่วนของ Data Cleaning.....24
3.5 แสดงการทำงานในส่วนของ Data Transformation..... 25
3.6 แสดงการทำงานในส่วนของ Data Mining.....25
3.7 แสดงการทำงานในส่วนของการแสดงผล.....26
3.8 ผังงานแสดงขั้นตอนการทำงานหลักของระบบ.....26
3.9 ผังงานแสดงขั้นตอนการทำความสะอาดข้อมูล.....28
3.10 ผังงานแสดงขั้นตอนการปรับเปลี่ยนข้อมูล.....29
3.11 ผังงานแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm 31
3.12 (ต่อ) ผังงานแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm.....32
3.13 ผังงานแสดงขั้นตอนการคำนวณหาค่า Mindistance.....33
3.14 ผังงานแสดงขั้นตอนการคำนวณหาค่า Distance..... 35
3.15 ผังงานแสดงขั้นตอนการคำนวณหาจุดศูนย์กลางกลุ่มใหม่..... 36
3.16 แสดงหน้าจอสำหรับเลือกฐานข้อมูลที่ต้องการนำมาจัดกลุ่ม.....35
3.17 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า Null..... 39
3.18 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่มีค่า Null.....40

สารบัญรูป (ต่อ)

หน้า

รูปที่

3.19 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null.....41
3.20 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null.....42
3.21 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่มีค่า Null.....43
3.22 แสดงหน้าจอการกำหนดจำนวนสำหรับการจัดกลุ่มข้อมูล.....44
3.23 แสดงหน้าจอผลที่ได้จากการจัดกลุ่มข้อมูล.....45
4.1 แสดงขั้นตอนการประยุกต์ใช้งานระบบ.....47
4.2 แสดงหน้าจอเลือกตาราง, ฟิลด์ข้อมูลที่ต้องการจัดกลุ่ม..... 49
4.3 แสดงหน้าจอการทำความสะอาดข้อมูลสำหรับข้อมูลประเภท Numerical.....50
4.4 แสดงหน้าจอการทำความสะอาดข้อมูลสำหรับข้อมูลประเภท Categorical.....51
4.5 แสดงหน้าจอการปรับเปลี่ยนข้อมูล.....52
4.6 แสดงหน้าจอการกำหนดจำนวนกลุ่มข้อมูล.....53
4.7 แสดงหน้าจอแสดงผลการจัดกลุ่มข้อมูล..... 54
5.1 แสดงหน้าจอการคัดเลือกข้อมูลประเภท Categorical.....58
5.2 แสดงหน้าจอการทำความสะอาดข้อมูลที่เป็นแบบ Categorical ที่ไม่มีค่าว่าง.....59
5.3 แสดงหน้าจอการทำความสะอาดข้อมูลที่เป็นแบบ Categorical ที่มีค่าว่าง.....60
5.4 แสดงหน้าจอการปรับเปลี่ยนข้อมูลประเภท Categorical 60
5.5 แสดงหน้าจอการกำหนดค่าในการจัดกลุ่มข้อมูลประเภท Categorical..... 61
5.6 แสดงหน้าจอแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลประเภท Categorical.....62
5.7 แสดงหน้าจอการคัดเลือกข้อมูลประเภท Numerical.....62
5.8 แสดงหน้าจอการทำความสะอาดข้อมูลที่เป็นแบบ Numerical ที่ไม่มีค่าว่าง.....63
5.9 แสดงหน้าจอการทำความสะอาดข้อมูลที่เป็นแบบ Numerical ที่มีค่าว่าง..... 64
5.10 แสดงหน้าจอการปรับเปลี่ยนข้อมูลประเภท Numerical 64
5.11 แสดงหน้าจอการกำหนดค่าในการจัดกลุ่มข้อมูลประเภท Numerical.....65
5.12 แสดงหน้าจอแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลประเภท Numerical.....66

สารบัญรูป (ต่อ)

หน้า

รูปที่

ก.1 หน้าจอแรกเมื่อเข้าสู่การติดตั้งโปรแกรม.....63
ก.2 หน้าจอที่สองของการติดตั้งโปรแกรม.....64
ก.3 หน้าจอแสดงกระบวนการติดตั้งระบบ.....64
ก.4 หน้าจอแสดงความก้าวหน้าในการติดตั้งโปรแกรม.....65
ก.5 แสดงหน้าจอการติดตั้งโปรแกรมเสร็จสมบูรณ์.....65
ก.6 แสดงไดเรกทอรี่ที่โปรแกรมได้ถูกติดตั้ง.....65
ก.7 การเข้าสู่โปรแกรมโดยเลือกจาก Start Menu.....66
ก.8 หน้าจอเริ่มต้นของโปรแกรม 66
ก.9 หน้าจอแสดงผล เมนู File.....67
ก.10 หน้าจอแสดงผล เมนู View.....67
ก.11 หน้าจอเลือกฐานข้อมูล.....68
ก.12 หน้าจอเลือกตารางและฟิลด์.....68
ก.13 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า Null.....69
ก.14 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null.....70
ก.15 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่มีค่า Null.....71
ก.16 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่มีค่า Null.....72
ก.17 หน้าจอแสดงผลการปรับเปลี่ยนข้อมูล.....73
ก.18 หน้าจอแสดงผลการกำหนดจำนวนกลุ่มข้อมูล.....74
ก.19 หน้าจอแสดงผลการจัดกลุ่มข้อมูล.....75
ก.20 หน้าจอแสดงผลโดยเลือกเมนู File > Export to > Ms Excel.....75
ก.21 หน้าจอแสดงผลการบันทึกข้อมูลแบบ Excel.....76
ก.22 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Centroid.....76
ก.23 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Member.....77
ก.24 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Graph.....77

สารบัญรูป (ต่อ)

หน้า

รูปที่

ก.25 หน้าจอแสดงผลโดยเลือกเมนู File > Export To > Text.....78
ก.26 หน้าจอแสดงผลการบันทึกข้อมูลแบบ Text.....78
ก.27 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Text.....79
ก.28 หน้าจอการปิดโปรแกรมโดยเลือกเมนู File > Exit.....79
ก.29 ข้อความเตือนเมื่อไม่ได้ทำการเลือกฐานข้อมูล.....79
ก.30 ข้อความเตือนเมื่อไม่ได้ค่าลงในช่องรับข้อความ.....80
ก.31 ข้อความเตือนเมื่อใส่ค่า weight ไม่ถูกต้อง.....80
ก.32 ข้อความเตือนเมื่อใส่ค่าจำนวนกลุ่มที่ต้องการแบ่งมากกว่าค่าจำนวนข้อมูล.....80

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

หลักการตลาดถือว่าเป็นหัวใจของการประกอบธุรกิจในปัจจุบัน เนื่องจากสินค้า บริการ ถูกผลิตออกมาสู่ท้องตลาดอย่างมากมาย เพื่อตอบสนองต่อความต้องการของลูกค้า ที่มีปริมาณสูงขึ้น ประกอบกับตลาดมีลักษณะการแข่งขันที่เป็นไปอย่างเสรี ดังนั้นการลงทุนเพื่อสร้างสินค้าและบริการออกมาสู่ท้องตลาดจะประสบความสำเร็จได้ ต้องสามารถเข้าถึงความต้องการของลูกค้าได้มากที่สุด

เทคโนโลยีสารสนเทศทางด้านฐานข้อมูลได้ถูกพัฒนาขึ้น เพื่อใช้ประโยชน์อย่างกว้างขวาง โดยเฉพาะการจัดเก็บข้อมูล อาจประกอบด้วยข้อมูลที่มีสำคัญและไม่มีสำคัญรวมอยู่ด้วยกัน แต่ข้อมูลที่สามารถนำไปประยุกต์ใช้ได้นั้นมีจำนวนน้อยมาก ทำให้มีการศึกษาทางด้านดาต้าไมนิ่ง (Data Mining) เนื่องจากความก้าวหน้าทางด้านเทคโนโลยีในการประมวลผลข้อมูล พร้อมทั้งมีการวิจัยและพัฒนาอัลกอริทึมของการทำดาต้าไมนิ่งอย่างต่อเนื่อง ดาต้าไมนิ่งสามารถที่จะดึงความรู้ออกจากฐานข้อมูลขนาดใหญ่ โดยทำการสำรวจและวิเคราะห์ข้อมูลในปริมาณมากนั้นได้อย่างอัตโนมัติ เพื่อทำให้เกิดความคิดใหม่ ๆ ผลที่ได้จากการทำดาต้าไมนิ่งจึงเป็นที่ยอมรับและสามารถนำไปประยุกต์ใช้งานได้ดี โดยเฉพาะในภาคธุรกิจ เพื่อจุดประสงค์ในการได้ส่วนแบ่งทางการตลาดเพิ่มขึ้น ดังนั้น จึงมีแนวความคิดว่า ถ้ามีการพัฒนาระบบที่ช่วยในการจัดกลุ่มข้อมูล จะช่วยให้สามารถนำข้อมูลที่ได้จากการจัดกลุ่มไปใช้ในการวิเคราะห์และสนับสนุนการตัดสินใจได้ดีมากยิ่งขึ้น

1.2 วัตถุประสงค์ของโครงการ

1. เพื่อศึกษาเทคนิคการทำดาต้าไมนิ่ง
2. เพื่อศึกษา K-Prototypes Algorithm ซึ่งเป็นอัลกอริทึมหนึ่งในการจัดกลุ่มข้อมูล
3. จัดกลุ่มข้อมูลต่าง ๆ เพื่อที่จะนำไปวิเคราะห์และใช้ประโยชน์ต่อไป เช่น การแบ่งกลุ่มประเภทลูกค้า

1.3 ขอบเขตของการดำเนินงาน

1. ทำการจัดกลุ่มฐานข้อมูลโดยใช้ K-Prototypes Algorithm
2. ข้อมูลที่สามารถทำการวิเคราะห์ในการจัดกลุ่มข้อมูลต้องเป็นฐานข้อมูล Microsoft Access 2000 , Access 2003 เท่านั้น
3. สามารถส่งออกไฟล์ในรูปแบบของเท็กซ์ไฟล์ (.txt) และเอกสาร Microsoft Excel (.xls)

1.4 ขั้นตอนการศึกษา

1. ศึกษาขั้นตอนและวิธีการทำดาต้าไมนิ่งในการจัดกลุ่มข้อมูล
2. ศึกษากระบวนการทำงานของ K-Prototypes Algorithm
3. ออกแบบและพัฒนาระบบงานโดยใช้ K-Prototypes Algorithm
4. ตรวจสอบและแก้ไขระบบให้สมบูรณ์และถูกต้อง
5. สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำแนวทางด้านดาต้าไมนิ่งมาประยุกต์ใช้กับการดำเนินธุรกิจ
2. สามารถแปลงข้อมูลดิบจากฐานข้อมูล ให้เป็นข้อมูลที่มีประโยชน์ได้
3. สร้างเครื่องมือที่ใช้ในการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm
4. สามารถใช้ประโยชน์ในการนำข้อมูลที่ได้ไปสนับสนุนการวิเคราะห์เพื่อการตัดสินใจ

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ความหมายของดาต้าไมนิ่ง

ดาต้าไมนิ่ง คือ การค้นหาความสัมพันธ์และรูปแบบทั่วไปของข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่ แต่ถูกซ่อนไว้ในกลุ่มของข้อมูลจำนวนมาก กล่าวได้ว่าเป็นวิธีที่ช่วยในการวิเคราะห์ข้อมูลและหาความสัมพันธ์ระหว่างข้อมูลต่าง ๆ ซึ่งข้อมูลที่นำมาใช้ จะต้องมีจำนวนมากและถูกจัดเก็บอย่างเป็นระเบียบ เป็นการหาความหมายที่แอบแฝงอยู่ในฐานข้อมูลและความสัมพันธ์ของข้อมูล เช่น ความสัมพันธ์ระหว่างข้อมูลของคนไข้กับการบำบัดรักษาทางยา ความสัมพันธ์เหล่านี้จะแสดงถึงความรู้และสิ่งที่อยู่ในฐานข้อมูล เพื่อให้ได้สารสนเทศที่มีประโยชน์ ช่วยในการสนับสนุนการตัดสินใจ จุดมุ่งหมายหลักของการทำดาต้าไมนิ่งคือ การพยายามที่จะค้นพบสารสนเทศที่เป็นความรู้ใหม่ โดยได้จากฐานข้อมูลหรือข้อมูล

2.2 ขั้นตอนการทำดาต้าไมนิ่ง

กระบวนการทำดาต้าไมนิ่งประกอบด้วย 4 ขั้นตอนหลัก ๆ คือ การกำหนดจุดประสงค์ทางธุรกิจ การเตรียมข้อมูล การทำไมนิ่ง และการวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่ง

2.2.1 การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

เป็นขั้นตอนแรกของการทำดาต้าไมนิ่ง ขั้นตอนนี้จะเป็นการกำหนดขอบเขตและเป้าหมายของการทำดาต้าไมนิ่ง โดยจะต้องทำความเข้าใจกับปัญหาและความต้องการทางธุรกิจ เพื่อทำให้การแก้ปัญหาเป็นไปอย่างถูกต้อง

2.2.2 การเตรียมข้อมูล (Data Preparation)

ถือได้ว่าเป็นส่วนที่สำคัญมากและใช้เวลาในการทำงานของกระบวนการต่าง ๆ มากที่สุด เพื่อให้ได้ข้อมูลมาทำการวิเคราะห์ เนื่องจากขั้นตอนนี้จะเป็นการจัดเตรียมข้อมูลเพื่อส่งต่อไปยังกระบวนการไมนิ่ง หากเกิดข้อผิดพลาดจากการเตรียมข้อมูล จะส่งผลให้การทำไมนิ่งนั้นผิดไปจากวัตถุประสงค์ที่ตั้งไว้ แบ่งออกเป็นขั้นตอนย่อยได้ถึง 3 ขั้นตอนดังนี้ (Richard J. Roiger and Micheal W. Geatz, 2003)

ขั้นตอนที่1 การคัดเลือกข้อมูล (Data Selection)

การระบุหรือคัดเลือกแหล่งข้อมูลที่มีอยู่ เป็นการแบ่งข้อมูลตามเงื่อนไขบางอย่าง เช่น บุคคลที่เป็นเจ้าของรถหนึ่งคัน ซึ่งสิ่งที่อยู่ในข้อมูลเหล่านั้นจะถูกกรองออกมา จุดประสงค์หลักในการเลือกข้อมูลคือ การกำหนดแหล่งข้อมูลต่าง ๆ โดยจะทำการระบุลักษณะและเลือกข้อมูลที่ต้องการ ซึ่งการเลือกจะเปลี่ยนแปลงตามวัตถุประสงค์ และจะต้องไม่เป็นข้อมูลที่อาจส่งผลให้เกิดปัญหาทางด้านกฎหมาย ข้อมูลที่เราสามารถนำมาใช้ทำดาต้าไมนิ่งนั้นมีหลายประเภท เช่น ฐานข้อมูลที่มีการเก็บรวบรวมจากหลาย ๆ แหล่งมาไว้ในรูปแบบเดียวกัน เรียกกันว่า " คลังข้อมูล " (Data warehouse) เป็นต้น

รูปแบบชนิดของข้อมูลนั้นสามารถจำแนกออกเป็น 2 กลุ่มใหญ่ได้ดังนี้

- Categorical คือ กลุ่มหรือข้อมูลที่สามารถบอกถึงระดับความสำคัญของสิ่งนั้นได้ชัดเจน แบ่งได้ออกเป็น 2 ประเภทย่อยดังนี้
    - Nominal ค่าที่เป็นลำดับไม่มีความสำคัญเช่น เพศ
    - Ordinal ค่าที่เป็นลำดับมีความสำคัญ เช่น ระดับความพึงพอใจของลูกค้า (ปานกลาง, ดี, ดีมาก) ถ้าแปลงเป็นตัวเลขความหมายยังเหมือนเดิม
- Quantitative บอกถึงขนาดหรือปริมาณ ค่าความแตกต่างที่สามารถเป็นไปได้ แบ่งได้ 2 ชนิด คือ
    - Discrete เป็นค่าที่เป็นจำนวนเต็ม เช่น จำนวนพนักงาน
    - Continuous เป็นค่าที่เป็นจำนวนจริง เช่น รายได้

<u>ขั้นตอนที่ 2</u> การจัดเตรียมข้อมูลให้เหมาะสมก่อนการประมวลผล (Data Preprocessing)

ขั้นตอนการจัดเตรียมข้อมูลให้เหมาะสมก่อนการประมวลผล เป็นการทำ Data Cleaning คือ การตรวจสอบข้อมูลที่มีการคัดเลือกมาว่า เป็นข้อมูลที่มีความเหมาะสมหรือไม่ เพื่อแก้ไขปัญหาต่าง ๆ ที่เกี่ยวกับข้อมูลเช่น ข้อมูลแบบ Categorical เป็นต้น หรืออาจทำการวัดการกระจายของข้อมูล เพื่อให้เข้าใจข้อมูลที่มีอยู่มากยิ่งขึ้น ทำให้สามารถหาแนวโน้มของข้อมูลที่อาจเกิดขึ้นได้ ส่วนข้อมูลที่เป็นแบบ Quantitative อาจเป็นข้อมูลที่ได้จากการวิเคราะห์ข้อมูลโดยการหาค่าสูงสุด ต่ำสุด ค่าเฉลี่ยได้ เป็นต้น เป็นการปรับรูปแบบของข้อมูลจากแหล่งต่าง ๆ ให้สอดคล้องกัน สามารถนำมาแก้ปัญหาในส่วนต่าง ๆ เหล่านี้ได้

- Noisy data คือ ข้อมูลมีความคลาดเคลื่อนไป ความคลาดเคลื่อนที่เกิดขึ้นนี้อาจเกิดได้จากหลายสาเหตุ เช่น จากการเก็บข้อมูลผิดพลาด อาจจะเกิดจากการป้อนหรือบันทึกข้อมูลผิดพลาด ส่งผลให้ข้อมูลคลาดเคลื่อนไป (Outlier) ควรพิจารณาและแก้ไขให้ถูกต้องโดยการทำ regression หรือวิธีการ binning เพื่อทำการตัดค่าที่ผิดพลาดนั้นทิ้งไป

- Missing data คือ การที่ข้อมูลบาง attribute ในบาง record ขาดหายไป กรณีที่ค่าน้อยมาก อาจจะตัดทิ้ง หรืออาจใช้ค่าเฉลี่ยแทนค่าที่หายไป เช่น ข้อมูลขาดหายไปประมาณ 20-30 เปอร์เซ็นต์ และ attribute นั้นไม่ค่อยมีความจำเป็นอาจจะใช้วิธีการเติมค่าที่ขาดหายไป เช่น ค่า unknown หรือจะเติมด้วยค่าเฉลี่ยของ attribute ก็ได้ (กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และธนาวินท์ รักธรรมานนท์.,2544)

ข้อมูลอาจมีการซ้ำซ้อนจากการถูกส่งมาจากหลายแหล่ง เกิดปัญหาขึ้นหลายรูปแบบ เช่น ข้อมูลที่เป็น attribute เดียวกันแต่ใช้ชื่อ attribute ต่างกัน ข้อมูลที่มีชื่อ attribute เดียวกันแต่ตัวข้อมูลต่างกัน หรือข้อมูลเมื่อมารวมกันแล้วทำให้เกิดความซ้ำซ้อน ข้อมูลเหล่านั้นควรนำมาทำ data Integration or Enrichment โดยก่อนที่จะทำการรวบรวมข้อมูลให้ดูที่ meta data ประกอบข้อมูลที่จะนำมาทำไมนิ่งนั้น ส่วนใหญ่จะได้มาจากคลังข้อมูล ซึ่งข้อมูลที่อยู่ในคลังข้อมูลนั้นมีขนาดใหญ่และมีความซับซ้อนมาก จึงจำเป็นที่จะต้องทำการลดจำนวนข้อมูล (Data reduction) ซึ่งสามารถทำได้ 2 แบบ คือ

- การลดจำนวน attribute (ลดตามแนว column) วิธีการลดจำนวน attribute นั้น ทำได้ 3 วิธี
  วิธีที่ 1 (Step-wise forward selection) เริ่มจาก 1 attribute แล้วค่อย ๆ เพิ่มข้อมูลเข้าไปทีละ attribute ไปเรื่อยจนกว่าค่า error จะเกินกว่าที่จะยอมรับได้
  วิธีที่ 2 (Step-wise backward selection) เริ่มจากทุก attribute แล้วค่อย ๆ ตัดออกทีละ attribute จนกระทั่งค่า error จะเกินกว่าที่จะยอมรับได้
  วิธีที่ 3 (Decision-tree induction) ใช้ decision tree ในการทำนายว่า attribute ใดไม่จำเป็นต้องใช้แล้วจึงตัด attribute นั้นออก
- ลดปริมาณข้อมูล (ลดตามแนว row) วิธีการลดปริมาณข้อมูลจะทำโดยใช้วิธี Sampling ข้อมูลขึ้นมา

ขั้นตอนที่ 3 การปรับรูปแบบข้อมูลให้เหมาะสมตามแบบ Algorithm (Data Transformation)

การปรับรูปแบบข้อมูลให้เหมาะสมตามแบบ Algorithm ของดาต้าไมนิ่งที่เลือกใช้ มีวัตถุประสงค์คือ ทำให้ไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลและอัลกอริทึมที่จะนำมาใช้ เนื่องจากข้อมูลที่จะนำมาทำไมนิ่ง โดยวิธีการแปลงข้อมูลมีอยู่หลายวิธีซึ่งขึ้นอยู่กับปัญหาของข้อมูล โดยมีเทคนิคการแปลงข้อมูลได้หลายรูปแบบ (Daniel T. Larose. ,2005) ได้แก่

- วิธี Normalization เป็นวิธี ที่แปลงข้อมูลข้อมูลให้อยู่ในช่วง ๆ หนึ่ง เช่น
  Min - Max normalization มีสูตรการคำนวณดังนี้

$$v' = \frac{v\text{-}min_A}{max_A\text{-}min_A}(new\_max_A - new\_min_A) + new\_min_A$$

V' = ค่าข้อมูลที่ได้หลังจากการแปลง

V = ข้อมูลที่จะนำมาทำการแปลง

min_A = ค่าต่ำสุดของข้อมูล attribute A

max_A = ค่าสูงสุดของข้อมูลใน attribute A

new_min_A = ค่าต่ำสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

new_max_A = ค่าสูงสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

- วิธี Discretization เป็นวิธีการแปลงข้อมูลที่ต่อเนื่องให้เป็นข้อมูลที่ไม่ต่อเนื่อง เช่น อุณหภูมิเป็นข้อมูลที่ต่อเนื่อง เราอาจจะจัดเป็นช่วง ๆ คือ ช่วง 0-20 องศาเซลเซียส เป็นช่วงอากาศเย็น ช่วง 21-30 องศาเซลเซียส เป็นช่วงอากาศอุ่น ถ้าอุณหภูมิ 20.1 องศาเซลเซียส จะถูกปัดไปเป็น 21 องศาเซลเซียส และถูกจัดให้อยู่ในกลุ่มอากาศอุ่นซึ่งความจริงแล้ว 20.1 องศาเซลเซียส ไม่ต่างกับ 20 องศาเซลเซียส ควรจัดอยู่ในกลุ่มอากาศเย็นมากกว่า ดังนั้นการแก้ไขปัญหานี้สามารถทำได้โดยการแบ่งช่วงให้ละเอียดมากขึ้น แต่ก็ไม่ควรที่จะละเอียดเกินไป
- วิธี Generalization เป็นวิธีที่แปลงข้อมูลโดยมองเป็นภาพรวม เช่น จัดกลุ่มถนนเป็นเขต จัดกลุ่มเขตเป็นจังหวัด จัดกลุ่มจังหวัดเป็นประเทศ เป็นต้น
- วิธี Attribute/Feature construction เป็นวิธีแปลงข้อมูลโดยการสร้างข้อมูลใหม่จากข้อมูลเดิม เช่น พื้นที่หาจากกว้าง x ยาว

2.2.3 การทำดาต้าไมนิ่ง (Data Mining Operation)

กระบวนการนี้ขึ้นอยู่กับว่าเทคนิคดาต้าไมนิ่งที่เลือกใช้ เพื่อให้ตรงกับวัตถุประสงค์ที่ต้องการศึกษา ขั้นตอนการทำไมนิ่งนี้เป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูล โดยการทำไมนิ่งนั้น มีโมเดลอยู่หลายแบบซึ่งการเลือกใช้โมเดลนั้น ขึ้นอยู่กับวัตถุประสงค์ในการทำไมนิ่ง ซึ่งโมเดลในการทำไมนิ่งมีดังนี้

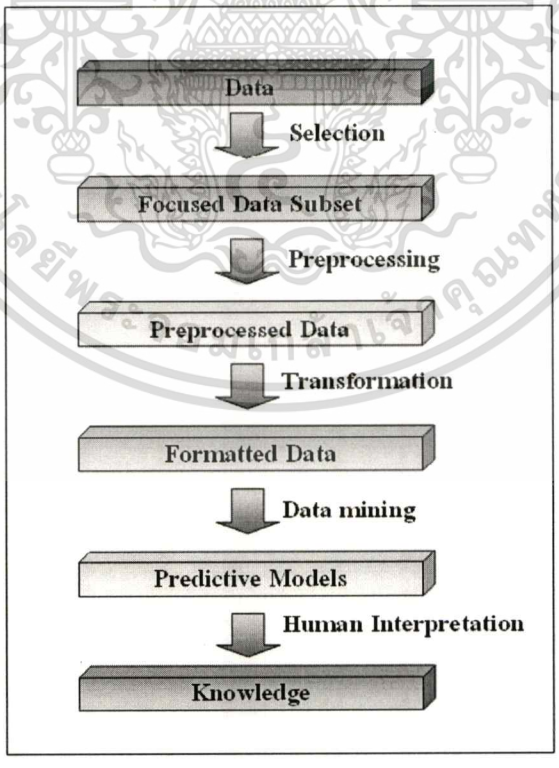
- Predictive Modeling เป็นโมเดลที่ใช้ในการสร้างแบบจำลองพยากรณ์ โดยจะมีลักษณะที่คล้ายกับการเรียนรู้ของมนุษย์ คือการใช้การสังเกตเพื่อที่จะสร้างแบบจำลองของคุณลักษณะที่สำคัญของปรากฏการณ์บางอย่าง โดยข้อมูลที่มีความถูกต้องและสมบูรณ์ จะทำให้แบบจำลองสามารถทำนายผลได้อย่างถูกต้องเช่น ทำนายยอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่
- Database Segmentation หรือ Database Clustering เป็นโมเดลที่ใช้ในการแบ่งกลุ่มข้อมูล โดยการแบ่งกลุ่มข้อมูลจะแบ่งตามลักษณะที่เหมือนกันของข้อมูล ส่วนใหญ่การแบ่งกลุ่มข้อมูลใช้

กับข้อมูลลูกค้า หรือกลุ่มตลาดเป้าหมาย เช่น เป็นกลุ่มลูกค้าตามรายได้ของลูกค้า หรือแบ่งกลุ่มลูกค้าตามลักษณะการชำระเงินของลูกค้า เป็นต้น

- Link Analysis เป็นโมเดลที่ใช้วิเคราะห์หาความสัมพันธ์ (Association) ระหว่างข้อมูลว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันอย่างไร เช่น ต้องการหาความสัมพันธ์ของสินค้าที่ลูกค้ามักจะซื้อร่วมกัน หรือเมื่อลูกค้าซื้อสินค้าชนิดนี้แล้วจะต้องซื้อสินค้าอีกประเภทต่อเนื่องกัน
- Deviation Detection เป็นวิธีหาค่าที่แตกต่างกันไปจากค่ามาตรฐาน โดยทั่วไปมักใช้วิธีการทางสถิติหรืออาศัยการวาดกราฟ แล้วดูการกระจายของข้อมูลว่ามีการกระจายออกไปจากกลุ่มหรือไม่ มักใช้ในการตรวจจับสิ่งผิดปกติต่าง ๆ เช่น การจับเท็จ

2.2.4 การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้ (Human Interpretation)

เป็นขั้นตอนที่นักวิเคราะห์ข้อมูลและนักวิเคราะห์นำผลที่ได้จากดาต้าไมนิ่งมาแปลความหมาย และประเมินค่าผลที่ได้ เพื่อนำสารสนเทศที่ถูกต้อง มาผสมผสานกับประสบการณ์ของนักวิเคราะห์ ที่สามารถนำมาใช้สนับสนุนการตัดสินใจของมนุษย์ เพื่อหาวิธีในตอบสนองต่อความต้องการหรือวัตถุประสงค์ที่ต้องการค้นหา (Jiawei Han and Micheline Kamber., 2001)



รูปที่ 2.1 แสดงภาพรวมเทคโนโลยีของกระบวนการดาต้าไมนิ่ง

2.3 โมเดลการแบ่งกลุ่มข้อมูล (Data Mining Operation: Database Segmentation)

อัลกอริทึมการจัดกลุ่มข้อมูล เป็นอัลกอริทึมที่พยายามค้นหากลุ่มของข้อมูลที่มีความคล้ายคลึงกัน โดยจะหาศูนย์กลางของกลุ่มข้อมูลเพื่อที่จะระบุสมาชิกของกลุ่มข้อมูลซึ่งจะประเมินจากระยะห่างระหว่างข้อมูล และศูนย์กลางของกลุ่มข้อมูล ข้อมูลที่มีระยะทางใกล้ศูนย์กลางของแต่ละกลุ่มข้อมูลมากที่สุดจะถูกนำมารวมกลุ่ม (Jain A.K., Murty M.N., and Flynn P.J. ,1999). ซึ่งผลที่ได้อธิบายได้ดังรูป

รูปที่ 2.2 ลำดับการทำงานของอัลกอริทึมการจัดกลุ่มข้อมูล

วัตถุประสงค์หลักในการทำ Database Segmentation คือ การแบ่งส่วนข้อมูลที่มีลักษณะคล้ายคลึงกันในฐานข้อมูลไว้เป็นกลุ่มเดียวกัน โดยในตอนเริ่มต้นเราไม่รู้ว่าจะแบ่งข้อมูลออกเป็นกี่กลุ่ม นั่นคือข้อมูลเหล่านี้จะมีคุณสมบัติหนึ่งเหมือนกัน และจะถูกพิจารณาเป็นกลุ่มข้อมูลเดียวกัน โดยคุณสมบัติที่เหมือนกันหมายถึงข้อมูลต่าง ๆ ในกลุ่มเดียวกันจะมีลักษณะใกล้เคียงกัน ซึ่งลักษณะใกล้เคียงกันนี้ สามารถวัดได้จากความแตกต่างของข้อมูลกับจุดศูนย์กลางกลุ่ม (Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees และ Alessando Zanasi ,1998)

ผลจากการเจริญเติบโตของฐานข้อมูลและข้อมูลที่มีอยู่หลายชนิดในฐานข้อมูล ทำให้จำเป็นต้องแบ่งข้อมูลออกเป็นส่วน ๆ ที่เหมาะสม โดยการใช้เทคนิคของการแบ่งส่วนฐานข้อมูลเพื่อจับกลุ่มของข้อมูลที่มีความคล้ายคลึงกันและสัมพันธ์กัน วิธีในการแบ่งกลุ่มข้อมูลนั้นมีหลายวิธีดังนี้

- Partitioning Algorithms วิธีการนี้จะทำการแบ่งส่วนข้อมูลให้เป็นกลุ่ม ๆ เริ่มต้นจากการกำหนดกลุ่มที่ต้องการจะแบ่งโดยไม่จำเป็นว่าข้อมูลจะคล้ายกัน สมมติว่าในฐานข้อมูลมีข้อมูล อยู่ n ตัว ทำการแบ่งกลุ่มออกเป็น k ส่วน โดยที่ $k \leq n$ หลังจากนั้นทำการแบ่งข้อมูล ทีละตัวเข้ากับแต่ละส่วนที่เรากำหนดไว้โดยจะพยายามปรับปรุงการแบ่งส่วนข้อมูล จากนั้นจึงนำข้อมูลแต่ละตัวมาลองทดสอบย้ายกลุ่มไปแต่ละกลุ่ม และพิจารณาค่าแตกต่างจากจุดศูนย์กลางแล้วให้ข้อมูลตัวนั้นอยู่ที่กลุ่มที่ให้ค่าแตกต่างจากจุดศูนย์กลางน้อยที่สุด ในวิธี Partitioning นี้ยังมีอัลกอริทึมให้

เลือกใช้อีกหลายอัลกอริทึม เช่น K-Means, K-Modes และ K-Prototypes อัลกอริทึม K-Means นั้นเหมาะกับข้อมูลประเภท Numerical อย่างเดียว ส่วน K-Modes เหมาะกับข้อมูลประเภท Categorical ตัวอย่างเช่น เพศ สถานะ การแต่งงาน โดยอัลกอริทึม K-Modes จะแทนค่า Means ของกลุ่มข้อมูล (Cluster) ด้วยค่าฐานนิยม (Modes) และอัลกอริทึม K-Prototypes ใช้ได้กับข้อมูลทั้งแบบ Numerical และ Categorical

- Hierarchy Algorithms เป็นการจัดกลุ่มข้อมูลโดยสร้างเป็นชั้น ๆ แบบลำดับขั้น ซึ่งวิธีนี้ไม่เป็นที่นิยม เพราะมีค่าใช้จ่ายในการคำนวณสูงและเสียเวลา ส่วนใหญ่จะใช้กับข้อมูลที่ไม่ใช่ตัวเลข เช่น สัญลักษณ์
- Neural Network เป็นเทคโนโลยีที่มาจากงานวิจัยทางด้านปัญญาประดิษฐ์ Artificial Intelligence: AI เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูลและมีพื้นฐานมาจากสมองของมนุษย์ เป็นเทคนิคที่เหมาะสมกับข้อมูลเข้าที่เป็นตัวเลข โดยมีอัลกอริทึมในการทำงานอยู่หลายชนิดด้วยกัน เช่น Kohan Feature Map เป็นต้น หลักการทำงานของ Neural Network จะมี 2 ขั้นตอนหลัก คือ ขั้นตอนการเรียนรู้ (Training) และขั้นตอนการนำไปใช้งาน (Deploying)

โดยวิธีการจะทำให้เครื่องเรียนรู้จากตัวอย่างต้นแบบแล้วฝึก (train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของนิวรอลเน็ตจะแประกอบไปด้วยโหนด (node) สำหรับ Input-Output และการประมวลผลกระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ input layer, output layer และ hidden layer การประมวลผลของนิวรอลเน็ตจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ใน layer นี้

2.4 อัลกอริทึม K- Prototypes

ข้อมูลจะถูกแบ่งกลุ่มตามจุดศูนย์กลางของกลุ่มข้อมูล ซึ่งจะคล้ายกับอัลกอริทึม K-Means แต่แทนที่จะใช้ค่าเฉลี่ยของกลุ่มข้อมูล K- Prototypes ได้มีการพัฒนาวิธีการปรับเปลี่ยนค่าจุดศูนย์กลางของกลุ่มข้อมูล เพื่อที่จะทำให้ข้อมูลภายในกลุ่มข้อมูลเดียวกันมีความคล้ายกันมากที่สุด ซึ่งการวัดความคล้ายกันของข้อมูลจะได้มาจากแบบ numeric และแบบ categorical ซึ่งเป็นการรวมอัลกอริทึมระหว่าง K-Means และ K-Modes เข้าด้วยกัน (Zhexue Huang, 1998)

2.4.1 หลักคณิตศาสตร์เบื้องต้นที่ใช้ใน K- Prototypes

ให้ $X = \{X_1, X_2, \ldots, X_n\}$ แทน เซตของจำนวนข้อมูล n ตัว และ $X_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ แทนข้อมูลที่นำมาจัดกลุ่มโดยที่ m คือ ค่าจำนวน attribute ของข้อมูล

วัตถุประสงค์ของการจัดกลุ่มข้อมูล X คือการแบ่งแยกข้อมูลใน X ออกเป็น k กลุ่ม โดยที่ k เป็นจำนวนเต็มบวก สำหรับ n คือจำนวนในการแบ่งกลุ่มที่เป็นไปได้ซึ่งมีจำนวนมาก วิธีที่ใช้เป็นแนวทางในการแบ่งกลุ่มข้อมูล ใช้หลักการแบ่งกลุ่มที่เรียกว่า "Cost Function"

<u>การคำนวณหาค่า Cost Function</u>

โดยทั่วไปมักใช้ค่า Cost Function ในการวัดค่าความเหมือนกันของข้อมูล ซึ่งทางหนึ่งที่ใช้ในการกำหนด Cost Function คือ

$$E = \sum_{l=1}^{k}\sum_{i=1}^{n} y_{il}\, d(X_i, Q_l) \qquad (2.1)$$

โดยที่ $Q_l = [q_{l1}, q_{l2}, \ldots, q_{lm}]$ คือจุดศูนย์กลางของกลุ่มข้อมูล l และ y_{il} คือสมาชิกของ partition matrix Y_{nxl} ส่วน d คือหน่วยที่ใช้วัดความคล้ายกันของข้อมูลซึ่งจะหาได้จากการคำนวณแบบ Square Euclidean Distance

Y จะมีคุณสมบัติ 2 ข้อ คือ 1) $0 \le y_{il} \le 1$ และ 2) $\sum_{l=1}^{k} y_{il} = 1$. ซึ่งถ้า $y_{il} \in \{0,1\}$ จะเรียก Y ว่าเป็น Hard partition นอกเหนือจากนี้จะเรียกว่า Fuzzy partition ที่ Hard partition ค่า $y_{il} = 1$ ซึ่งเป็นการระบุว่าข้อมูล X_i ถูกกำหนดให้อยู่ในกลุ่มข้อมูล l แต่เราจะพิจารณาเพียง Hard partition เท่านั้น

จากสมการที่ (2.1) ค่า $E_l = \sum_{i=1}^{n} y_{il}\, d(X_i, Q_l)$ คือ ค่า cost ทั้งหมดของข้อมูล X ที่อยู่ในกลุ่มข้อมูล l ตัวอย่างเช่น การกระจายข้อมูลในกลุ่มข้อมูล l จากจุดศูนย์กลางของกลุ่มข้อมูล (Q_l) ซึ่งค่า E_l จะมีค่าต่ำสุดก็ต่อเมื่อ

$$q_{lj} = \frac{1}{n_l}\sum_{i=1}^{n} y_{il}\, x_{ij} \qquad \text{for } j=1,\ldots,m \qquad (2.2)$$

โดยที่ $j = 1, \ldots, m$ และ $n_l = \sum_{i=1}^{n} y_{il}$ เป็นจำนวนข้อมูลในกลุ่มข้อมูล l

ถ้าข้อมูล X มี attribute แบบ Categorical แล้วจะสามารถวัดความคล้ายกันได้โดย

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c) \qquad (2.3)$$

โดย $\delta(p,q) = 0$ แล้ว p = q

$\delta(p,q) = 1$ แล้ว $p \neq q$

x_{ij}^r เป็นค่า Numeric attributes ของข้อมูล i และ q_{il}^r เป็นจุดศูนย์กลางของกลุ่มข้อมูล l

q_{ij}^r เป็นค่า Categorical attributes ของข้อมูล i และ q_{il}^c เป็นจุดศูนย์กลางของกลุ่มข้อมูล l

m_r เป็นจำนวน attributes ของ Numeric

m_c เป็นจำนวน attributes ของ Categorical

γ_l เป็น weight ของ Categorical attributes ของกลุ่มข้อมูล l

ดังนั้นสามารถเขียน E_l ใหม่ได้เป็น

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c)$$
$$= E_l^r + E_l^c \qquad (2.4)$$

โดย E_l^r เป็นค่า cost ทั้งหมดของ Numerical attribute ของข้อมูลที่อยู่ในกลุ่มข้อมูล l ซึ่งจะเป็นค่าต่ำสุด ถ้า q_{lj}^r หามาจากสมการที่ (2.2)

ให้ C_j เป็นเซตที่ประกอบไปด้วยที่เป็นหนึ่งเดียวใน Categorical attribute j และ $p(c_j \in C_j|l)$ เป็นความน่าจะเป็นของค่า C_j ที่ปรากฏในกลุ่มข้อมูล l ซึ่ง E_l^c ในสมการที่ (2.4) สามารถเขียนใหม่ได้เป็น

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l (1 - p(q_{lj}^c \in C_j|l)) \qquad (2.5)$$

โดย n_l เป็นจำนวนของข้อมูลในกลุ่ม l ซึ่งสามารถหาค่าต่ำสุดของ E_l^c ได้จากบทแทรกที่ 1

บทแทรกที่ 1 : สำหรับกลุ่มข้อมูล l แล้วค่า E_l^c มีค่าต่ำสุดก็ต่อเมื่อ $p(q_{lj}^c \in C_j|l) \geq p(c_j \in C_j|l)$ โดย $q_{lj}^c \neq c_j$ สำหรับทุก Categorical attributes

ดังนั้นเราสามารถเขียน E ใหม่ได้เป็น

$$E = \sum_{l=1}^{k} (E_l^r + E_l^c) = \sum_{l=1}^{k} E_l^r + \sum_{l=1}^{k} E_l^c = E^r + E^c \qquad (2.6)$$

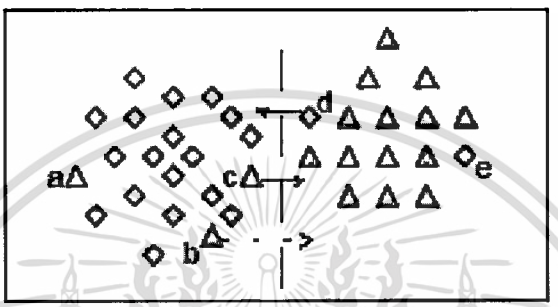
สมการที่(2.6) เป็น Cost function ของกลุ่มข้อมูลทั้ง Numeric และ Categorical ทั้งค่า E^r และ E^c ไม่เป็นค่าติดลบ ค่าต่ำสุดของ E สามารถหาได้จากค่าต่ำสุดของ E^r และ E^c ซึ่ง

- ค่าต่ำสุดของ E^r หาได้จากสมการที่ (2.2)
- ค่าต่ำสุดของ E^c หาได้จากบทแทรกที่ 1

ดังนั้นสมการที่(2.2) และบทแทรกที่ 1 จะกำหนดแนวทางในการเลือกจุดศูนย์กลางของกลุ่มข้อมูลเพื่อทำให้ค่า Cost function ที่สมการที่ (2.6) มีค่าต่ำสุด

2.4.2 การวัดความเหมือนกัน

การวัดความเหมือนกันของข้อมูลแบบ Numeric attribute คือการใช้ Square Euclidean Distance ส่วนข้อมูลแบบ Categorical attributes จะทำการพิจารณาจำนวนที่ไม่เข้าคู่กันระหว่างข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูล ซึ่งผลกระทบจาก weight γ_l ในการแบ่งกลุ่มข้อมูลสามารถอธิบายได้ดังรูป



รูปที่ 2.3 แสดงผลกระทบจาก weight γ_l ในการแบ่งกลุ่มข้อมูล (Zhexue Huang, 1998)

จากรูปที่ 2.2 รูปสามเหลี่ยมแทนค่าของ Categorical attributes และรูปเพชรแทนค่าของ Numeric attributes โดยข้อมูลเหล่านี้จะถูกแบ่งเป็นกลุ่มข้อมูล 2 กลุ่ม

ถ้า $\gamma_l = 0$ หมายถึงว่าการแบ่งข้อมูลนั้นขึ้นอยู่กับ Numeric attributes ซึ่งจะแสดงให้เห็นจากตำแหน่งของข้อมูล ซึ่งผลที่ได้จากการแบ่งกลุ่มข้อมูลคือ จะแบ่งกลุ่มออกเป็น 2 กลุ่ม

ถ้า $\gamma_l > 0$ หมายถึงการแบ่งกลุ่มข้อมูลนั้นขึ้นอยู่กับทั้งข้อมูลของ Numerical attributes และข้อมูลของ Categorical attributes ซึ่งความโน้มเอียงของข้อมูลจะขึ้นอยู่กับค่า γ_l และตำแน่งที่อยู่ของข้อมูลนั้น ๆ ซึ่งจะแสดงให้เห็นจากตำแหน่งของข้อมูล

- ข้อมูล c อาจจะเปลี่ยนไปอยู่ที่กลุ่มข้อมูลด้านขวา เพราะอยู่ใกล้กลุ่มข้อมูล ด้านขวา
- ข้อมูล d อาจจะเปลี่ยนไปอยู่ที่กลุ่มข้อมูลด้านซ้าย
- ข้อมูล a และข้อมูล e อาจจะยังคงอยู่ที่กลุ่มข้อมูลเดิมเพราะอยู่ไกลจากกลุ่มข้อมูลที่จะย้ายไป
- ข้อมูล b ขึ้นอยู่กับค่า γ_l ถ้า γ_l โน้มอียงไปทางข้อมูลแบบ Categorical ข้อมูล b อาจจะเปลี่ยนไปอยู่กลุ่มข้อมูลด้านขวา นอกเหนือจากนี้จะยังคงอยู่กลุ่มทางด้านซ้าย

2.4.3 การทำงานของอัลกอริทึม K-Prototypes

ขั้นตอนการทำงานของอัลกอริทึม K-Prototypes (Zhexue Huang, 1998)

1. เลือกค่า k เริ่มต้น เป็นจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่มของข้อมูล X

2. จัดข้อมูลแต่ละตัวที่อยู่ใน X ให้อยู่ในกลุ่มข้อมูลใดกลุ่มข้อมูลหนึ่ง ซึ่งข้อมูลนั้นอยู่ใกล้กับจุดศูนย์กลางของกลุ่มข้อมูลมากที่สุดสามารถหาได้จากสมการที่ (2.3) และทำการปรับเปลี่ยนค่าจุดศูนย์กลางของกลุ่มข้อมูลหลังจากจัดข้อมูลเสร็จแล้ว
3. หลังจากที่ทำการจัดข้อมูล ให้อยู่ในกลุ่มข้อมูลหนึ่ง ๆ แล้ว ให้ทำการทดสอบความคล้ายกันของข้อมูลอีกครั้ง กับจุดศูนย์กลาง ถ้าพบว่าข้อมูลนั้น อยู่ใกล้กับจุดศูนย์กลางของกลุ่มอื่นมากกว่า ให้ทำการย้ายข้อมูลไปยังกลุ่มข้อมูลนั้น และทำการปรับเปลี่ยนจุดศูนย์กลางของกลุ่มข้อมูลทั้ง 2 กลุ่มใหม่อีกครั้ง
4. ทำซ้ำข้อ 3 จนกระทั่งข้อมูลใน x ทุกตัวไม่เปลี่ยนกลุ่ม

จะเห็นได้ว่าอัลกอริทึม K-prototypes มีการทำงานหลักอยู่ 3 กระบวนการคือ การเลือก การกำหนดค่าจำนวนกลุ่มข้อมูลที่จะจัดกลุ่ม จัดกลุ่มข้อมูลเริ่มต้นและจัดกลุ่มข้อมูลใหม่ ในรูปที่ 2.4 และ 2.5 เป็นกระบวนการบางส่วนที่อยู่ในอัลกอริทึม K-Prototypes ซึ่งประกอบไปด้วยตัวแปรและฟังก์ชันดังต่อไปนี้

X[i] = ข้อมูล

X[i,j] = ค่าของ attribute ของจุดศูนย์กลางกลุ่มข้อมูล

O_prototypes[ ] = เก็บ Numeric attribute ของจุดศูนย์กลางกลุ่มข้อมูล

C_prototypes[ ] = เก็บ Categorical attribute ของจุดศูนย์กลางกลุ่มข้อมูล

O_prototypes[i,j] = เก็บ Numeric attribute ของจุดศูนย์กลางกลุ่มข้อมูล i

C_prototypes[i,j] = เก็บ Categorical attribute ของจุดศูนย์กลางกลุ่มข้อมูล i

Distance( ) = เป็นฟังก์ชันในการหาค่า Square Euclidean Distance

Sigma( ) = เป็นฟังก์ชันในการหาค่า $\delta(\,)$ ในสมการที่ (2.3)

Clustership[ ] = เก็บกลุ่มข้อมูลที่อยู่ในข้อมูลนั้น

ClusterCount[ ] = เก็บจำนวนข้อมูลที่อยู่ในกลุ่มข้อมูล

SumInCluster[ ] = ผลรวมค่า Numeric ของข้อมูลในกลุ่มข้อมูล และใช้ในการปรับเปลี่ยนค่า Numeric attribuite

FrequencyInCluster[ ] = เก็บความถี่ของค่าที่เปลี่ยนแปลงของ Categorical attribute ในกลุ่มข้อมูล

HighestFreq( ) = เป็นฟังก์ชันในการหาค่าของบทแทรกที่ 1 เพื่อปรับเปลี่ยน Categorical attribute ของกลุ่มข้อมูล

```
FOR i = 1 TO NumberOfObjects
    Mindistance= Distance(X[i],O_prototypes[1])+ gamma* Sigma(X[i],C_prototypes[1])
    FOR j = 1 TO NumberOfClusters
        distance= Distance(X[i],O_prototypes[j])+ gamma * Sigma(X[i],C_prototypes[j])
        IF (distance < Mindistance)
            Mindistance=distance
            cluster=j
        ENDIF
    ENDFOR
    Clustership[i]=cluster
    ClusterCount[cluster] + 1
    FOR j=1 TO NumberOfNumericAttributes
        SumInCluster[cluster,j] + X[i,j]
        O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
    ENDFOR
    FOR j=1 TO NumberOfCategoricAttributes
        FrequencyInCluster[cluster,j,X[i,j]] + 1
        C_prototypes[cluster,j]=HighestFreq(FrequencyInCluster,cluster,j)
    ENDFOR
ENDFOR
```

รูปที่ 2.4 แสดงกระบวนการเริ่มต้นการจัดกลุ่ม โดย K-Prototypes Algorithm (Zhexue Huang, 1998)

จากรูป 2.4 แสดงอัลกอริทึมในส่วนของกระบวนการเริ่มต้นจัดกลุ่มข้อมูล โดยมีขั้นตอนการทำงานดังนี้

1. กำหนดค่าต่ำสุดของ Cost function (Mindistance)
2. คำนวณค่า Cost function ของข้อมูล i ที่อยู่ในกลุ่มข้อมูล j (distance)
3. เปรียบเทียบค่า Cost function ในข้อที่1 และข้อที่ 2 ถ้าค่า distance < Mindistance ให้ค่า Mindistance มีค่าเท่ากับ distance และเก็บค่ากลุ่มข้อมูลที่ค่าต่ำสุดนั้นอยู่
4. กลับไปทำข้อ 2 ซ้ำจนกระทั่งทำครบทั้งกลุ่มข้อมูล
5. เก็บกลุ่มข้อมูลที่ข้อมูลนั้นอยู่ซึ่งมีค่า Cost function ต่ำสุด
6. เก็บค่าจำนวนข้อมูลในกลุ่มข้อมูลเพิ่มขึ้นหนึ่ง
7. วนลูปเพื่อหาผลรวมของค่า Numeric ของข้อมูล I ในกลุ่มข้อมูล cluster และหาค่าต่ำสุดของค่า E^r
8. วนลูปเพื่อหาค่า FrequencyInCluster และหาค่าต่ำสุด E^c

   กลับไปทำข้อ 1 ซ้ำจนกระทั่งทำครบทุกข้อมูล

```
moves=0
FOR i = 1 TO NumberOfObjects
    ...
    (To find the cluster whose prototype is the nearest to object i.  Same as Figure 2)
    ...
    IF (Clustership[i]<>cluster)
        moves+1
        oldcluster=Clustership[i]
        ClusterCount[cluster] + 1
        ClusterCount[oldcluster] - 1
        FOR j=1 TO NumberOfNumericAttributes
            SumInCluster[cluster,j] + X[i,j]
            SumInCluster[oldcluster,j] - X[i,j]
            O_prototypes[cluster,j]=SumInCluster[cluster,j]/ClusterCount[cluster]
            O_prototypes[oldcluster,j]= SumInCluster[oldcluster,j]/ClusterCount[oldcluster]
        ENDFOR
        FOR j=1 TO NumberOfCategoricAttributes
            FrequencyInCluster[cluster,j,X[i,j]] + 1
            FrequencyInCluster[oldcluster,j,X[i,j]] - 1
            C_prototypes[cluster,j]=HighestFreq(cluster,j)
            C_prototypes[oldcluster,j]=HighestFreq(oldcluster,j)
        ENDFOR
    ENDIF
ENDFOR
```

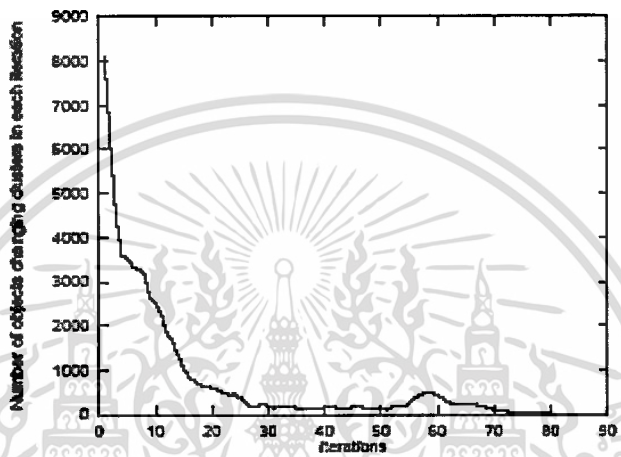
รูปที่ 2.5 แสดงกระบวนการจัดกลุ่มใหม่ของ K-Prototypes Algorithm (Zhexue Huang, 1998)

รูปที่ 2.5 แสดงอัลกอริทึมในส่วนของกระบวนการจัดกลุ่มข้อมูลใหม่ โดยมีขั้นตอนการทำงานดังนี้

1. กำหนดค่า moves เริ่มต้นให้เท่ากับ 0
2. จัดข้อมูลให้อยู่ในกลุ่มข้อมูลซึ่งข้อมูลอยู่ใกล้กับจุดศูนย์กลางกลุ่มข้อมูลมากที่สุด (ตามกระบวนการในรูปที่ 2.5)
3. เปรียบเทียบ Clustership[i] ว่าเท่ากับ cluster หรือไม่ถ้าไม่เท่ากัน
   - ให้เพิ่มค่า moves ขึ้น 1
   - เพิ่มจำนวนข้อมูลในกลุ่มข้อมูลใหม่ขึ้น 1
   - ลดจำนวนข้อมูลในกลุ่มข้อมูลเดิมลง1
   - วนลูปเพื่อหาผลรวมของค่า Numeric ของข้อมูล i ในกลุ่มข้อมูล cluster และหาค่าต่ำสุดของ E^r ของทั้ง 2 กลุ่มข้อมูล (กลุ่มข้อมูลใหม่และกลุ่มข้อมูลเดิม)
   - วนลูปเพื่อหาFrequencyInCluster และหาค่าต่ำสุดของ E^c ของทั้ง 2 กลุ่มข้อมูล (กลุ่มข้อมูลใหม่และกลุ่มข้อมูลเดิม)

4. กลับไปทำข้อที่ 2 ซ้ำ จนกระทั่งครบข้อมูลทุกตัว

อัลกอริทึมจะทำซ้ำจนกว่าข้อมูลจะไม่เปลี่ยนกลุ่ม ในรูปที่ 2.5 แสดงให้เห็นกราฟของอัลกอริทึมนี้ ซึ่งทำการทดสอบกับข้อมูล 75,808 records จำนวน 20 attributes และแบ่งข้อมูลออกเป็น 64 กลุ่มข้อมูล



รูปที่ 2.6 แสดงกราฟการทำงานของ K-Prototypes Algorithm (Zhexue Huang, 1998)

จากรูปที่ 2.6 จะเห็นได้ว่า จำนวนการเปลี่ยนแปลงของข้อมูลลดลงอย่างรวดเร็วในช่วงเริ่มต้นและในช่วงท้ายมีการเปลี่ยนแปลงเพียงเล็กน้อยเท่านั้น ก่อนที่จะไม่มีการเปลี่ยนแปลง นั่นคือข้อมูลจะมีการเปลี่ยนกลุ่มไปเรื่อย ๆ จนกระทั่งข้อมูลถูกจัดอยู่ในกลุ่มข้อมูลที่เหมาะสมแล้ว ค่าการเปลี่ยนแปลงของข้อมูลจะเป็นศูนย์

ค่า cost ของอัลกอริทึมนี้มีค่าเท่ากับ O((t+1)kn) โดยที่ค่า n เป็นจำนวนข้อมูลทั้งหมด ส่วนค่า k เป็นจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม และสุดท้ายค่า t เป็นจำนวนครั้งในกระบวนการจัดกลุ่มใหม่ซ้ำ โดยทั่วไปค่า k จะต้องน้อยกว่าค่า n มาก และค่า t จะไม่เกิน 100 ดังนั้นอัลกอริทึม นี้จะมีประสิทธิภาพอย่างดีกับการจัดกลุ่มข้อมูลขนาดใหญ่

2.4.4 ตัวอย่างการนำข้อมูลมาใช้กับ K-Prototypes Algorithm

สมมติข้อมูลขึ้นมาเพื่อทำการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm ในการจัดกลุ่มข้อมูล จะสมมติข้อมูลที่มี attribute ทั้งหมด 3 attribute คือ รหัสลูกค้า (cus_id), ชื่อบริษัทที่ทำการค้าด้วย (company_name) และวงเงินเครดิตที่กำหนดให้ (creditlimit) โดยข้อมูล cus_id และ company_name จะเป็นข้อมูลประเภท Categorical ส่วนข้อมูล creditlimit เป็นข้อมูลประเภท Numerical ซึ่งจะแสดงดังตารางที่ 2.1

ตารางที่ 2.1 แสดงประเภทของข้อมูลที่จะนำมาจัดกลุ่ม

ชื่อข้อมูล	ประเภทของข้อมูล
cus_id	char(8)
company_name	char(50)
creditlimit	int

ตารางที่ 2.2 แสดงข้อมูลที่จะนำมาจัดกลุ่ม

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	AAA	10
00000101	CCC	20
00000101	BBB	15
00000150	AAA	15
00000150	BBB	20
00000216	CCC	10
00000216	BBB	15
00000354	AAA	10
00000460	CCC	20

ขั้นตอนในการจัดกลุ่มข้อมูลตารางที่ 2.2 มีดังนี้

1. กำหนดกลุ่มข้อมูล ในที่นี้จะจัดกลุ่มข้อมูลออกเป็น 3 กลุ่ม
2. จัดกลุ่มข้อมูลให้อยู่ในแต่ละกลุ่ม จะได้ดังตาราง

ตารางที่ 2.3 แสดงสมาชิกข้อมูลกลุ่มที่ 1

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	**AAA**	**10**
00000150	BBB	20
00000354	AAA	10

ตารางที่ 2.4 แสดงสมาชิกข้อมูลกลุ่มที่ 2

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	**CCC**	**20**
00000216	CCC	10
00000460	CCC	20

ตารางที่ 2.5 แสดงสมาชิกข้อมูลกลุ่มที่ 3

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	BBB	15
00000150	AAA	15
00000216	**BBB**	**15**

1. กำหนดจุดศูนย์กลางกลุ่มข้อมูลในแต่ละกลุ่ม ซึ่งกำหนดได้ดังนี้ (ที่ตาราง 2.3, 2.4, 2.5 **บรรทัดตัวหนาหมายถึงจุดศูนย์กลางของกลุ่มข้อมูล**)
   กลุ่มข้อมูลที่ 1 คือ cus_id =00000101, company_name= AAA
   กลุ่มข้อมูลที่ 2 คือ cus_id =00000101, company_name= CCC
   กลุ่มข้อมูลที่ 3 คือ cus_id =00000216, company_name= BBB
2. นำข้อมูลแต่ละตัวมาหาค่า distance ในแต่ละกลุ่มข้อมูล ในที่นี้จะยกตัวอย่างให้เห็นโดยสมมติว่าทำการจัดกลุ่มข้อมูลจากตารางที่ 2.4 โดยเลือกข้อมูลที่ cus_id = 00000216 และ company_name = CCC ดังนั้นต้องทำการหาค่า distance จากตัวข้อมูลกับจุดศูนย์กลางของกลุ่มข้อมูลในแต่ละกลุ่ม เพื่อเลือกว่าควรจะจัดข้อมูลนี้ไว้ที่กลุ่มใดมากที่สุดโดยจะกำหนดให้
   - ค่า weight ของ cus_id = 0.5
   - ค่า weight ของ company_name = 0.3

เริ่มพิจารณาทีละกลุ่ม

<u>กลุ่มที่ 1</u> หาค่า distance ของข้อมูล

$$\begin{pmatrix} 00000101 \\ AAA \\ 10 \end{pmatrix} \text{ กับ } \begin{pmatrix} 00000216 \\ CCC \\ 10 \end{pmatrix}$$

ซึ่งสามารถหาค่าได้จากสมการที่ (2.3) จะได้เป็น

Distance = $(10-10)^2 + 0.5(1) + 0.3(1)$

$= 0 + 0.5 + 0.3$

<u>กลุ่มที่ 2</u> หาค่า distance ของข้อมูล

00000101	กับ	00000216
CCC		CCC
20		10

ซึ่งสามารถหาค่าได้จากสมการที่ (2.3) จะได้เป็น

Distance = $(15-10)^2 + 0.5(1) + 0.3(0)$

$= 100 + 0.5 + 0$

$= 100.5$

<u>กลุ่มที่ 3</u> หาค่า distance ของข้อมูล

00000216	กับ	00000216
BBB		CCC
15		10

ซึ่งสามารถหาค่าได้จากสมการที่ (2.3) จะได้เป็น

Distance = $(15-10)^2 + 0.5(0) + 0.3(1)$

$= 25 + 0 + 0.3$

$= 25.3$

จากการคำนวณค่า distance ในแต่ละกลุ่มจะเห็นว่า ถ้าจัดข้อมูลไว้ที่กลุ่มที่ 1 จะหาค่า distance ได้ค่าต่ำที่สุด ดังนั้นจึงจัดข้อมูลไว้ที่กลุ่มที่ 1 ซึ่งข้อมูลที่จัดได้จะแสดงในตารางที่ 2.6

ตารางที่ 2.6 แสดงสมาชิกกลุ่มที่ 1 หลังจากการคำนวณ

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	AAA	10
00000150	BBB	20
00000216	CCC	10
00000354	AAA	10

ตารางที่ 2.7 แสดงสมาชิกกลุ่มที่ 2 หลังจากการคำนวณ

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	CCC	20
00000460	CCC	20

ตารางที่ 2.8 แสดงสมาชิกกลุ่มที่ 3 หลังจากการคำนวณ

cus_id	company_name	creditlimit (หมื่นบาท)
00000101	BBB	15
00000150	CCC	15
00000126	BBB	15

3. ทำการคำนวณหาค่าจุดศูนย์กลางของกลุ่มที่ 1 และกลุ่มที่ 2 โดยถ้าเป็นข้อมูลประเภท Numerical ให้ใช้วิธีการหาค่าเฉลี่ย และถ้าเป็นข้อมูลประเภท Categorical จะใช้วิธีฐานนิยม ดังนั้นในกลุ่มที่ 1 และกลุ่มที่ 2 จะหาค่าจุดศูนย์กลางกลุ่มข้อมูลใหม่ได้เป็น

<u>กลุ่มที่ 1</u>

- cus_id เลือกตัวไหนก็ได้เพราะมีค่าฐานนิยมเป็น 1 ทั้งหมด
- company_name เลือก AAA เพราะมีค่าฐานนิยมมากที่สุดคือเท่ากับ 2
- creditlimit ใช้วิธีหาค่าเฉลี่ยจะได้เป็น (10+20+10+10)/4 = 10.25

ดังนั้นจุดศูนย์กลางกลุ่มที่ 1 คือ = $\begin{pmatrix} 00000101 \\ AAA \\ 10.25 \end{pmatrix}$

<u>กลุ่มที่ 2</u>

- cus_id เลือกตัวไหนก็ได้เพราะมีค่าฐานนิยมเป็น 1 ทั้งหมด
- company_name เลือก CCC
- creditlimit ใช้วิธีหาค่าเฉลี่ยจะได้เป็น (20+20)/2 = 20

ดังนั้นจุดศูนย์กลางกลุ่มที่ 2 คือ = $\begin{pmatrix} 00000101 \\ CCC \\ 20 \end{pmatrix}$

<u>กลุ่มที่3</u>

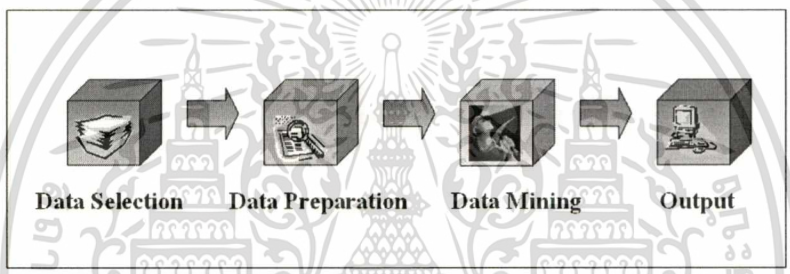
- cus_id เลือกตัวไหนก็ได้เพราะมีค่าฐานนิยมเป็น 1 ทั้งหมด
- company_name เลือก BBB เพราะมีค่าฐานนิยมมากที่สุดคือเท่ากับ 2
- creditlimit ใช้วิธีหาค่าเฉลี่ยจะได้เป็น (15+15+15)/3 =15

ดังนั้นจุดศูนย์กลางกลุ่มที่ 3 คือ $= \begin{pmatrix} 00000101 \\ BBB \\ 15 \end{pmatrix}$

บทที่ 3
การออกแบบและอิมพลีเมนท์ระบบ

3.1 องค์ประกอบของระบบงาน

การทำงานของระบบนั้นมีหลายขั้นตอน สามารถที่จะอธิบายภาพรวมของระบบได้ดังรูปที่ 3.1



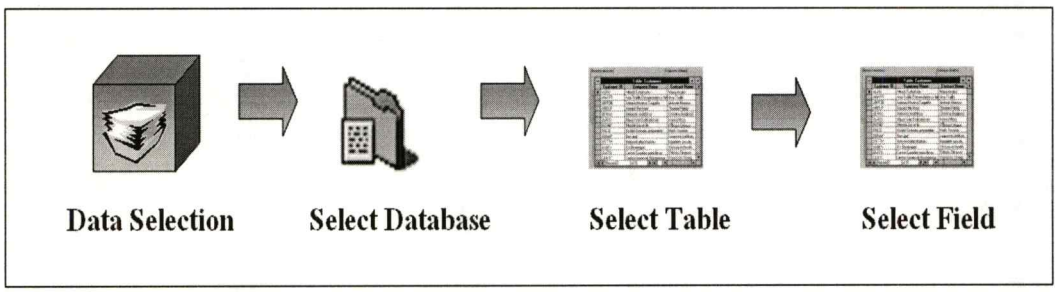
รูปที่ 3.1 แสดงภาพรวมของระบบ

ภาพรวมของระบบจะประกอบด้วยส่วนต่าง ๆ ดังนี้

1. Data Selection หมายถึง การคัดเลือกข้อมูลเพื่อเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง โดยจะต้องเลือกฐานข้อมูล ตาราง และฟิลด์ที่ผู้ใช้ต้องการ
2. Data Preparation หมายถึง การเตรียมข้อมูลเพื่อเข้าสู่กระบวนการทำไมนิ่ง เมื่อคัดเลือกข้อมูลที่ต้องการแล้วจะนำข้อมูลมาทำการ Clean และการ Transformation เพื่อเตรียมข้อมูลสำหรับการทำไมนิ่ง
3. Data Mining หมายถึง การนำข้อมูลที่ได้จากขั้นตอนที่ 2 มาทำไมนิ่ง โดยจะแบ่งกลุ่มข้อมูลตามจำนวนกลุ่มที่ผู้ใช้กำหนด
4. Output หมายถึง การแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล ในรูปแบบต่าง ๆ เช่น แสดงผลลัพธ์ทางหน้าจอ, บันทึกในรูปแบบของไฟล์ต่าง ๆ

ซึ่งมีรายละเอียดต่าง ๆ ดังต่อไปนี้

1. Data Selection เป็นส่วนของการเลือกข้อมูล ซึ่งเป็นขั้นตอนแรกในเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง



รูปที่ 3.2 แสดงการทำงานในส่วนของ Data Selection

จากรูปที่ 3.2 สามารถอธิบายได้ดังนี้

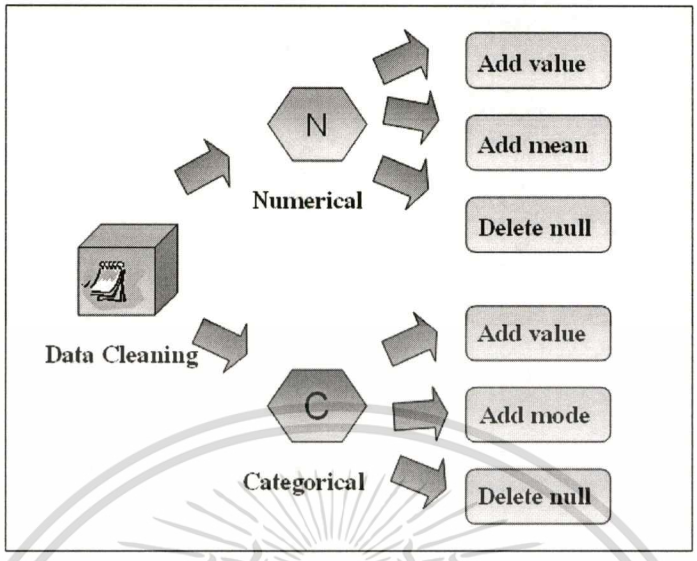
- Select Database หมายถึง เลือกฐานข้อมูลของ Microsoft Access 2000
- Select Table หมายถึง เลือกตารางจากฐานข้อมูลที่ได้เลือกไว้
- Select Field หมายถึง เลือกฟิลด์ที่ผู้ใช้ต้องการจากตารางที่ได้เลือกไว้แล้ว ซึ่งเลือกทำได้กับฐานข้อมูลและตารางเดียวเท่านั้น ส่วนฟิลด์สามารถเลือกตามความต้องการของผู้ใช้

2. Data Preparation เป็นขั้นตอนการเตรียมข้อมูลก่อนจะนำข้อมูลเหล่านั้นมาทำการไมนิ่ง

รูปที่ 3.3 แสดงการทำงานในส่วนของ Data Preparation

จากรูปที่ 3.3 สามารถอธิบายได้ดังนี้

- Data Cleaning หมายถึง การทำความสะอาดข้อมูล เป็นการแก้ไขค่าว่างที่มีอยู่ในฐานข้อมูล เพื่อที่จะทำให้ผลลัพธ์ของการทำดาต้าไมนิ่งนั้นมีประสิทธิภาพ

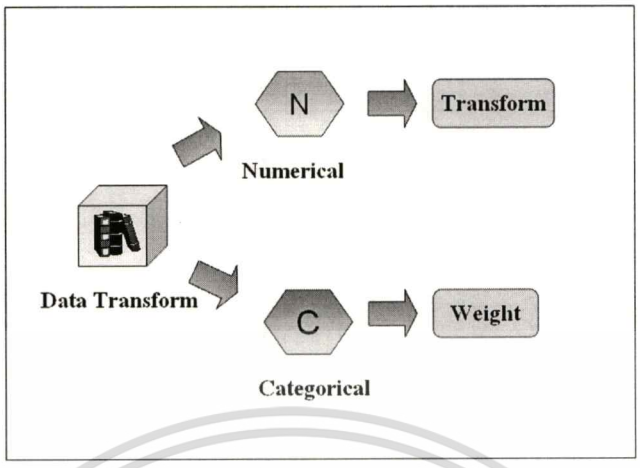


รูปที่ 3.4 แสดงการทำงานในส่วนของ Data Cleaning

จากรูปที่ 3.4 สามารถอธิบายได้ดังนี้

- ข้อมูลแบบ Numerical
  - Add value หมายถึง ใส่ค่าโดยที่ผู้ใช้กำหนดเอง
  - Add mean หมายถึง ใส่ค่าค่าเฉลี่ยแทนข้อมูลที่เป็นค่า null
  - Delete null หมายถึง ลบเรคอร์ดที่มีค่า null
- ข้อมูลแบบ Categorical
  - Add value หมายถึง ใส่ค่าโดยที่ผู้ใช้กำหนดเอง
  - Add mode หมายถึง ใส่ค่าฐานนิยมแทนข้อมูลที่เป็นค่า null
  - Delete null หมายถึง ลบเรคอร์ดที่มีค่า null

- Data Transformation การปรับเปลี่ยนข้อมูลให้อยู่ในช่วงใด ๆ เพื่อให้ผลลัพธ์ข้อมูลมีค่าไม่ต่างกันมากจนเกินไป เพื่อที่จะทำให้ผลลัพธ์ของการทำดาต้าไมนิ่งนั้นมีประสิทธิภาพ

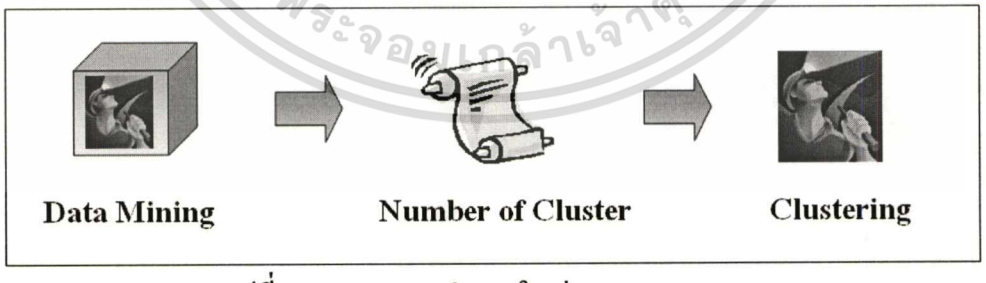


รูปที่ 3.5 แสดงการทำงานในส่วนของ Data Transformation

จากรูปที่ 3.5 สามารถอธิบายได้ดังนี้

- ข้อมูล Numerical
  - Transform หมายถึง การปรับเปลี่ยนค่าให้อยู่ในช่วงใด ๆ เพื่อให้ผลลัพธ์ข้อมูลระหว่าง แอตทริบิวต์ต่าง ๆ หลังจากประมวลผลดาต้าไมนิ่งแล้วค่าจะไม่ต่างกันเกินไป
- ข้อมูล Categorical
  - Weight หมายถึง ทำการใส่ค่า weight ให้กับฟิลด์เหล่านั้น

3. Data Mining ทำการไมนิ่งข้อมูล โดยจะแบ่งกลุ่มข้อมูลตามจำนวนกลุ่มที่ผู้ใช้กำหนด



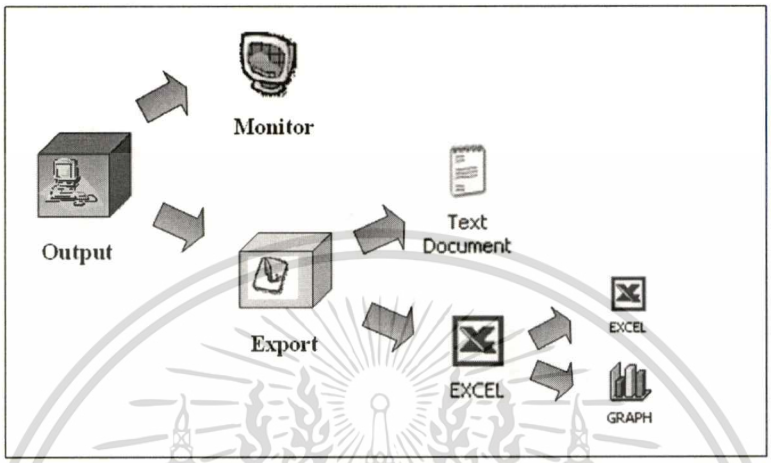
รูปที่ 3.6 แสดงการทำงานในส่วนของ Data Mining

จากรูปที่ 3.6 สามารถอธิบายได้ดังนี้

- Number of Cluster หมายถึง การกำหนดจำนวนกลุ่มในการทำดาต้าไมนิ่งของการจัดกลุ่มโดยใช้ K-Prototypes Algorithm โดยจำนวนกลุ่มต้องไม่มากกว่าจำนวนข้อมูลที่ผู้ใช้ต้องการนำมาจัดกลุ่ม

- Clustering หมายถึง การทำดาต้าไมนิ่งของการจัดกลุ่มโดยใช้ K-Prototypes Algorithm

4. Output ส่วนของการแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล



รูปที่ 3.7 แสดงการทำงานในส่วนของ Output

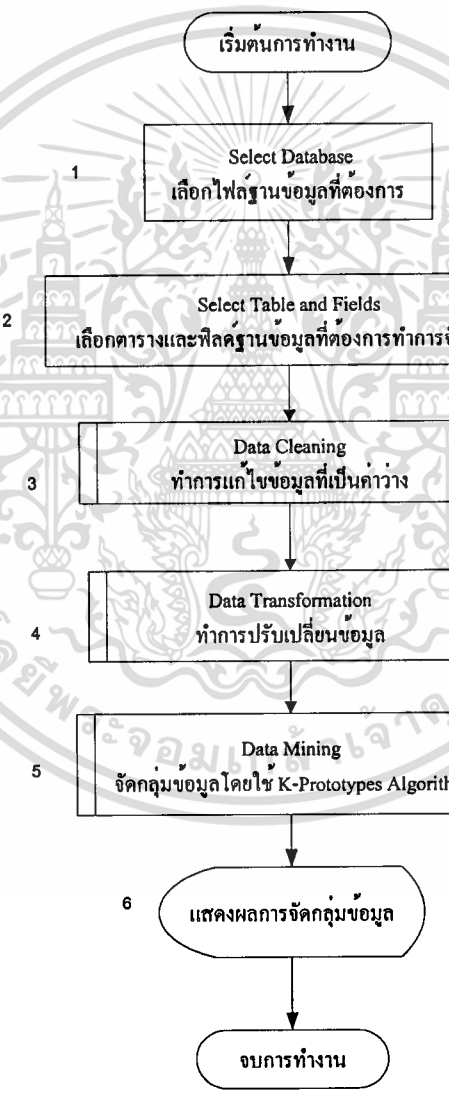
จากรูปที่ 3.7 สามารถอธิบายได้ดังนี้

- Monitor หมายถึง ระบบจะแสดงผลการจัดกลุ่มทางหน้าจอ โดยแสดงรายละเอียดดังนี้
  - แสดงค่าจุดศูนย์กลางของกลุ่มข้อมูล และจำนวนสมาชิกในกลุ่ม
  - แสดงรายละเอียดของสมาชิกแต่ละตัวในกลุ่มข้อมูลและระยะห่างจากจุดศูนย์กลาง
- Export หมายถึง การส่งออกข้อมูล ทำให้สะดวกต่อการนำผลลัพธ์ที่ได้ไปวิเคราะห์เพื่อใช้ในการสนับสนุนการตัดสินใจต่อไป สามารถส่งออกข้อมูลในรูปแบบต่าง ๆ ดังนี้
  - Text Document หมายถึง เท็กซ์ไฟล์ (.txt)
  - Excel หมายถึง เอกสาร Microsoft Excel (.xls) และกราฟแสดงผล

3.2 อัลกอริทึมของ K-Prototypes

การทำงานของอัลกอริทึม K-Prototypes นั้น ข้อมูลจะถูกแบ่งกลุ่มตามจุดศูนย์กลางของกลุ่มข้อมูล ซึ่งจะคล้ายกับอัลกอริทึม K-Means แต่แทนที่จะใช้ค่าเฉลี่ยของกลุ่มข้อมูลอัลกอริทึม K- Prototypes ได้มีการพัฒนาวิธีการปรับเปลี่ยนค่าจุดศูนย์กลางของกลุ่มข้อมูล เพื่อที่จะทำให้ข้อมูลภายในกลุ่มข้อมูลเดียวกันมีความคล้ายกันมากที่สุด

3.2.1 ขั้นตอนการทำงานหลักของระบบ (Main program)

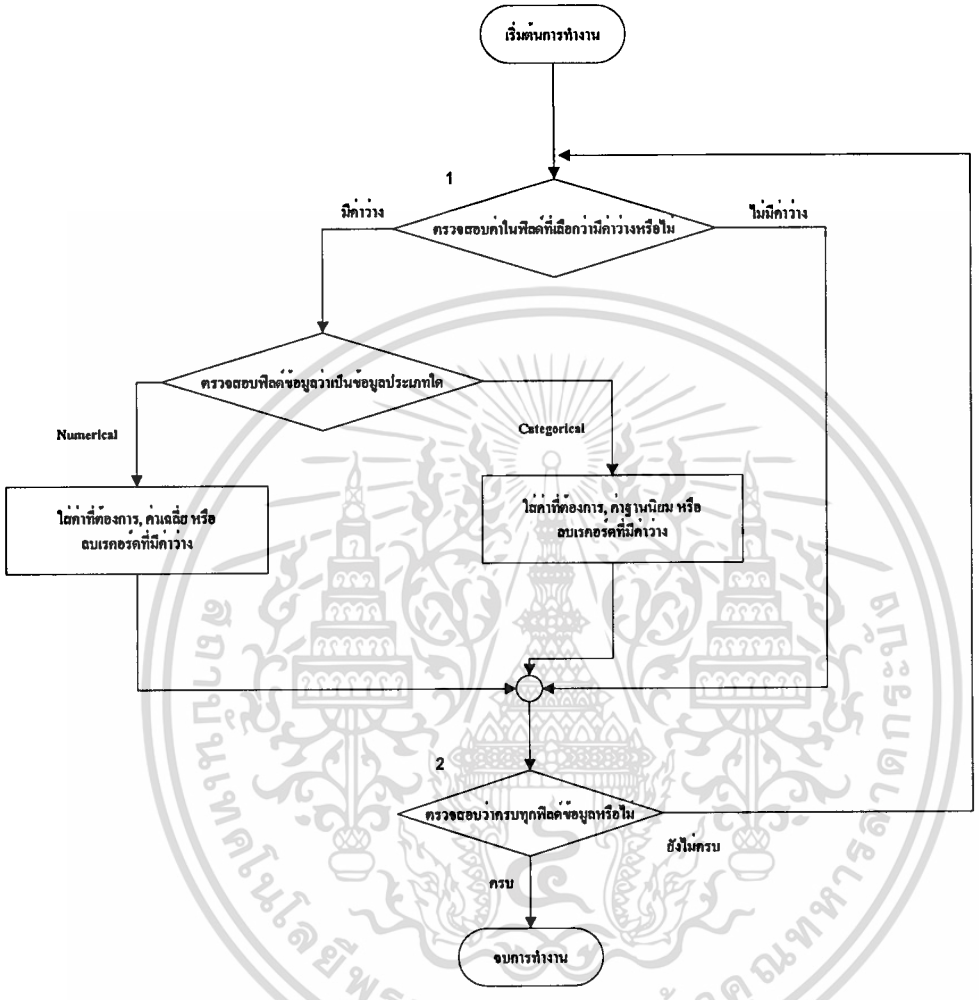


รูปที่ 3.8 ผังงานแสดงขั้นตอนการทำงานหลักของระบบ (Main program)

จากรูปที่ 3.8 สามารถอธิบายขั้นตอนการทำงานหลักของระบบได้ดังนี้

1. Select Database : เลือกไฟล์ฐานข้อมูลที่ต้องการ ซึ่งประกอบด้วยข้อมูลที่ต้องการนำมาจัดกลุ่ม โดยฐานข้อมูลที่เลือกเข้ามานั้นจะต้องเป็นฐานข้อมูล Microsoft Access 2000หรือ Microsoft Access XP เท่านั้น
2. Select Table and Fields : เลือกตารางข้อมูลจากฐานข้อมูลที่ได้ทำการเลือกไว้แล้วโดยสามารถเลือกได้เพียง 1 ตารางเท่านั้น และเลือกฟิลด์ข้อมูลที่ต้องการซึ่งสามารถเลือกได้ตามความต้องการของผู้ใช้งาน
3. Data Cleaning : ทำการ Clean ข้อมูล โดยจะมีการตรวจสอบหาค่าที่ขาดหายไป(Missing Value) และให้ผู้ใช้งานทำการแก้ไขให้เรียบร้อย
4. Data Transformation : ทำการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบที่เหมาะสม
5. Data Mining using K-Prototypes Algorithm : ทำการจัดกลุ่มข้อมูลตาม อัลกอริทึม K-Prototypes
6. Output : แสดงผลการจัดกลุ่มข้อมูล

3.2.2 ขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning)



รูปที่ 3.9 ผังงานแสดงขั้นตอนการทำความสะอาดข้อมูล (Data Cleaning)

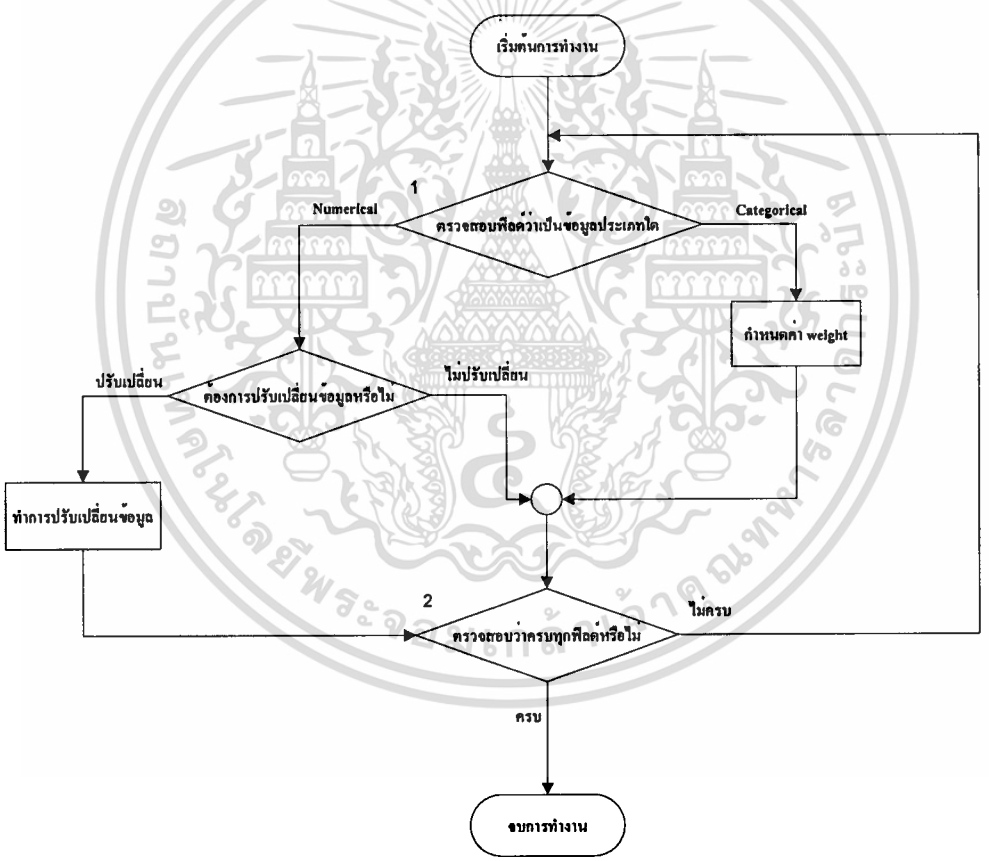
จากรูปที่ 3.9 สามารถอธิบายขั้นตอนการทำความสะอาดข้อมูลได้ดังนี้

1. ตรวจสอบค่าในฟิลด์ข้อมูลที่เลือกว่ามีค่าว่าง (Missing Value)หรือไม่
    - ถ้าพบว่ามีค่าว่าง จะทำการตรวจสอบว่าฟิลด์ข้อมูลนั้นเป็นข้อมูลประเภทใด
        - ถ้าเป็นข้อมูลประเภท Categorical สามารถทำการ Clean ข้อมูลได้โดย
            - ใส่ค่าที่ต้องการแทนที่ข้อมูลที่เป็นค่าว่าง
            - ใส่ค่าฐานนิยมของข้อมูลในฟิลด์นั้น ๆ
            - ลบเรคอร์ดที่มีค่าว่างนั้นออก

- ถ้าเป็นข้อมูลประเภท Numerical
  - ใส่ค่าที่ต้องการแทนที่ข้อมูลที่เป็นค่าว่าง
  - ใส่ค่าเฉลี่ยของข้อมูลในฟิลด์นั้น ๆ
  - ลบเรคอร์ดที่มีค่าว่างนั้นออก
- ไม่พบค่าว่าง

2. ตรวจสอบข้อมูลว่าทำการ Clean ข้อมูลในฟิลด์ที่พบว่ามีค่าว่างครบหรือไม่ ถ้ายังไม่ครบให้กลับไปทำข้อที่ 1

3.2.3 ขั้นตอนการปรับเปลี่ยนข้อมูล (Data Transformation)



รูปที่ 3.10 ผังงานแสดงขั้นตอนการทำการปรับเปลี่ยนข้อมูล (Data Transformation)

จากรูปที่ 3.10 สามารถอธิบายขั้นตอนการทำการปรับเปลี่ยนข้อมูล(Data Transformation) ได้ดังนี้

1. ทำการตรวจสอบฟิลด์ข้อมูลว่าเป็นข้อมูลประเภทใด
   - ถ้าเป็นข้อมูลประเภท Categorical สามารถใส่ค่า weight ให้กับข้อมูลได้

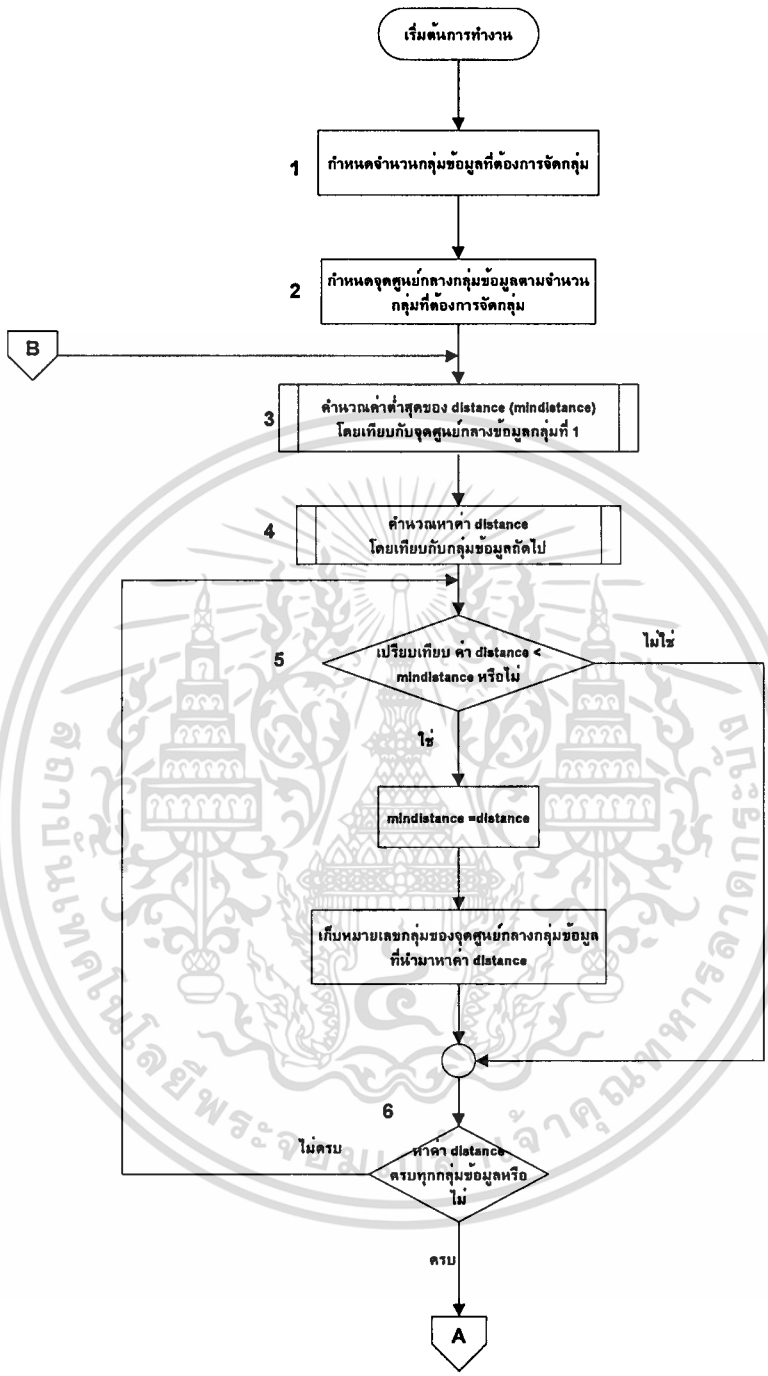
- ถ้าเป็นข้อมูลประเภท Numerical สามารถปรับเปลี่ยนค่าให้อยู่ในช่วงใดช่วงหนึ่งได้ โดยการใส่ค่า Min, Max ให้กับข้อมูลที่ต้องการปรับเปลี่ยนหรือ เลือกที่จะไม่ปรับเปลี่ยนข้อมูลก็ได้เช่นกัน

2. ตรวจสอบว่าได้ทำการปรับเปลี่ยนครบทุกฟิลด์ข้อมูลหรือยัง ถ้ายังไม่ครบให้กลับไปทำที่ข้อ 1 จนกว่าจะครบทุกฟิลด์ข้อมูล

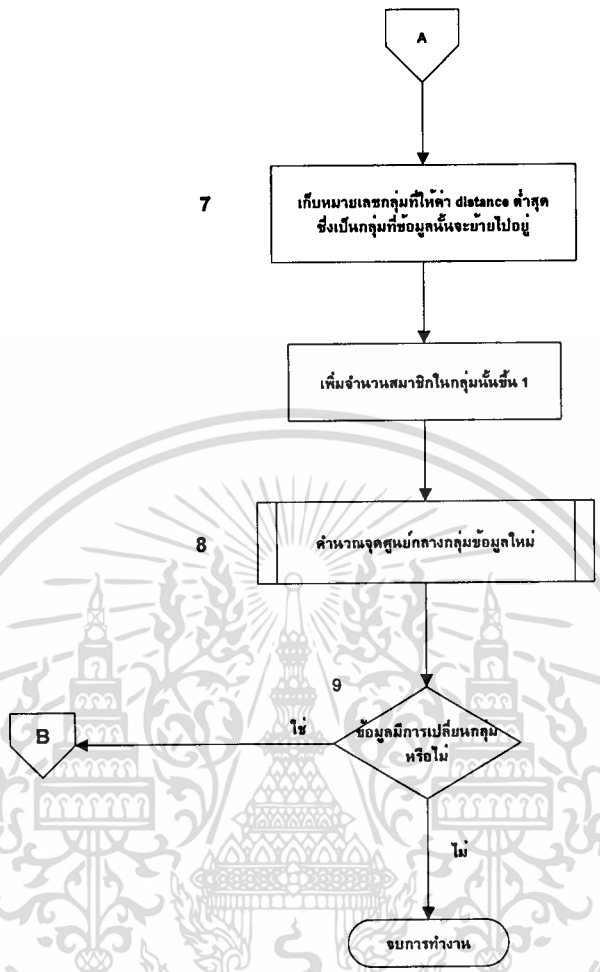
3.2.4. ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K- Prototypes Algorithm

จากรูปที่ 3.11-3.12 อธิบายขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm

1. กำหนดจำนวนกลุ่มข้อมูลที่ต้องการทำการจัดกลุ่ม
2. กำหนดค่าเริ่มต้นของจุดศูนย์กลางกลุ่มข้อมูลตามจำนวนกลุ่มข้อมูลที่ได้กำหนดไว้
3. ทำการคำนวณหาค่าต่ำสุดของ distance (Mindistance) โดยนำมาทำการคำนวณกับจุดศูนย์กลางกลุ่มที่ 1
4. คำนวณหาค่า distance โดยเทียบกับจุดศูนย์กลางกลุ่มอื่น ๆ
5. เปรียบเทียบค่า cost function ที่คำนวณได้ใหม่กับค่า cost function ที่ต่ำสุด โดยที่ถ้าค่า cost function ที่คำนวณได้ใหม่มีค่าน้อยกว่าค่าต่ำสุดของ cost function ให้ค่าต่ำสุดมีค่าเท่ากับค่า cost function ที่คำนวณได้ใหม่ และเก็บค่า cluster ที่ค่าต่ำสุดนั้นอยู่
6. ตรวจสอบว่าทำการเปรียบเทียบกับจุดศูนย์กลางกลุ่มข้อมูลครบทุกกลุ่มหรือยัง ถ้ายังไม่ครบให้กลับไปทำในข้อที่ 4
7. เก็บ cluster ที่ข้อมูลนั้นอยู่ที่มีค่า cost function ต่ำสุดและเก็บค่าจำนวนข้อมูลใน cluster เพิ่มขึ้นหนึ่ง
8. คำนวณจุดศูนย์กลางของกลุ่มข้อมูลของกลุ่มใหม่ และกลุ่มเดิม
9. ตรวจสอบว่าทำครบทุกเรคอร์ดหรือยัง ถ้ายังให้กลับไปทำที่ 3 จนกว่าจะครบทุกข้อมูลและข้อมูลจะไม่ทำการย้ายกลุ่ม

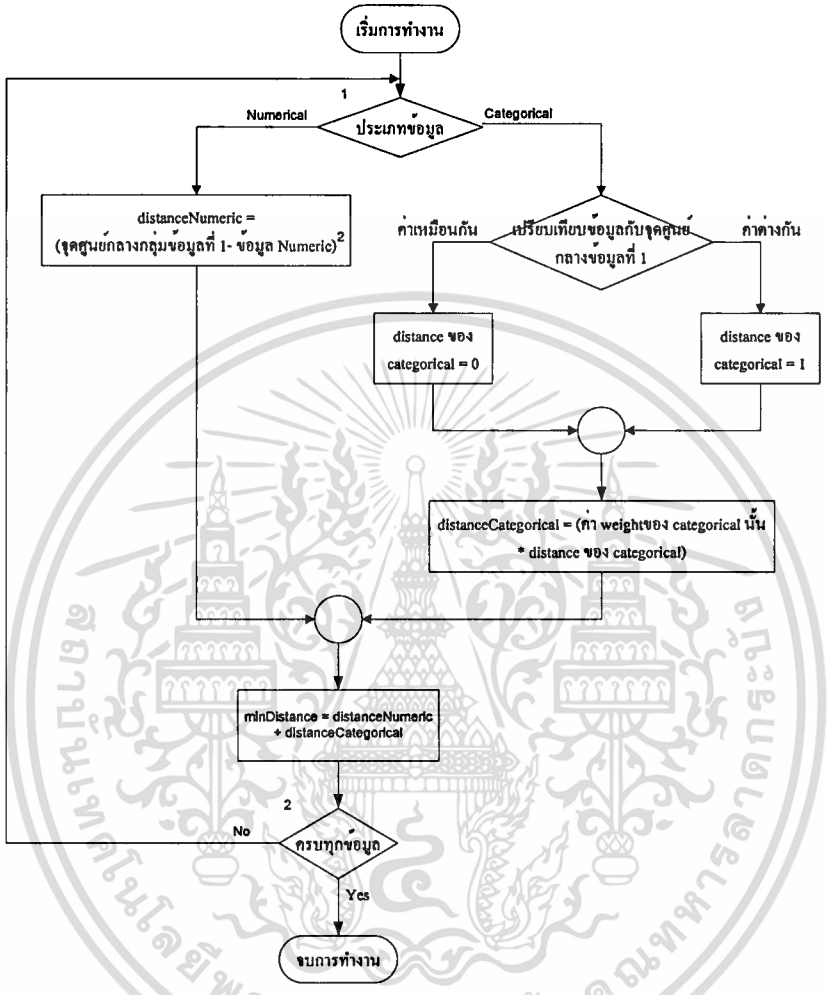


รูปที่ 3.11 ผังงานแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm



รูปที่ 3.12 (ต่อ) ผังงานแสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm

3.2.5 ขั้นตอนการคำนวณหาค่า Mindistance



รูปที่ 3.13 ผังงานแสดงขั้นตอนการคำนวณหาค่า Mindistance

รูปที่ 3.13 อธิบายขั้นตอนการคำนวณหาค่า Mindistance

1. ทำการตรวจสอบฟิลด์ข้อมูลว่าเป็นข้อมูลประเภทใด
   - ถ้าเป็นข้อมูลประเภท Categorical ให้นำค่าข้อมูลมาเปรียบเทียบกับข้อมูลที่เป็น Categorical ของจุดศูนย์กลางกลุ่มที่ 1
     - ถ้ามีค่าเหมือนกัน ค่าที่ได้จะเท่ากับศูนย์ นำค่าที่ได้จาการเปรียบเทียบคูณด้วยค่า weight ที่ได้กำหนดไว้ในขั้นตอน การปรับเปลี่ยนข้อมูล (Data Transformation) DistanceCategorical = (1 * ค่า weight ของฟิลด์)

- ถ้ามีค่าต่างกัน ค่าที่ได้จะเท่ากับหนึ่ง นำค่าที่ได้จากการเปรียบเทียบคูณด้วยค่า weight ที่ได้กำหนดไว้ในขั้นตอน การปรับเปลี่ยนข้อมูล (Data Transformation) DistanceCategorical = (0 * ค่า weight ของฟิลด์)

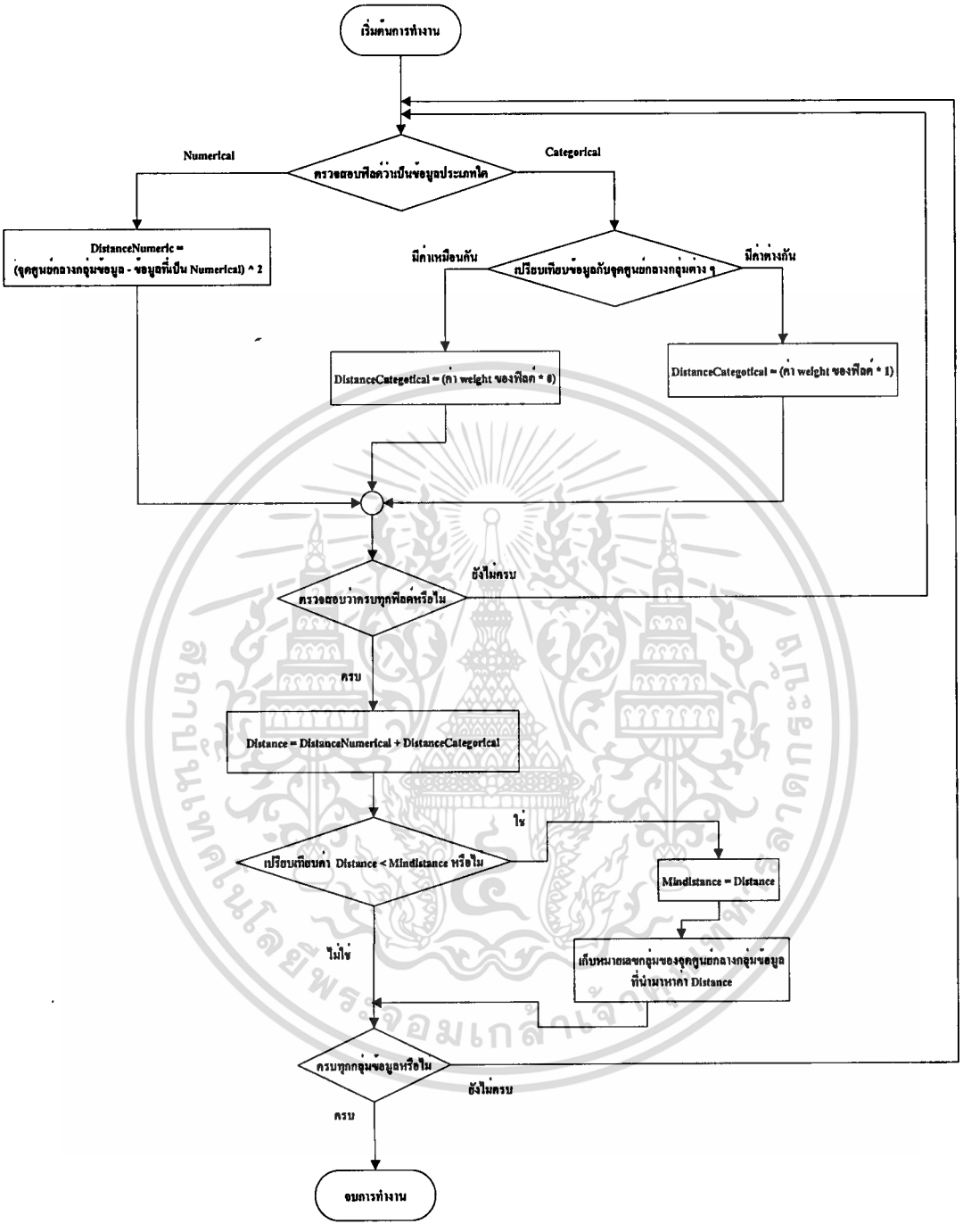
- ถ้าเป็นข้อมูลประเภท Numerical จะได้ DistanceNumerical = (ค่าจุดศูนย์กลางที่เป็น Numerical ของข้อมูลกลุ่มที่1- ข้อมูลที่เป็น Numerical)^2

2. นำค่าที่ได้จาก DistanceCategorical มารวมกับ DistanceNumerical จะได้ Mindistance = (DistanceCategorical + DistanceNumerical)
3. ทำการตรวจสอบว่าทำการคำนวณครบทุกฟิลด์ข้อมูลในเรคอร์ดแล้วหรือยัง ถ้ายังไม่ครบ กลับไปทำที่ข้อ 1

3.2.6 ขั้นตอนการคำนวณหาค่า Distance

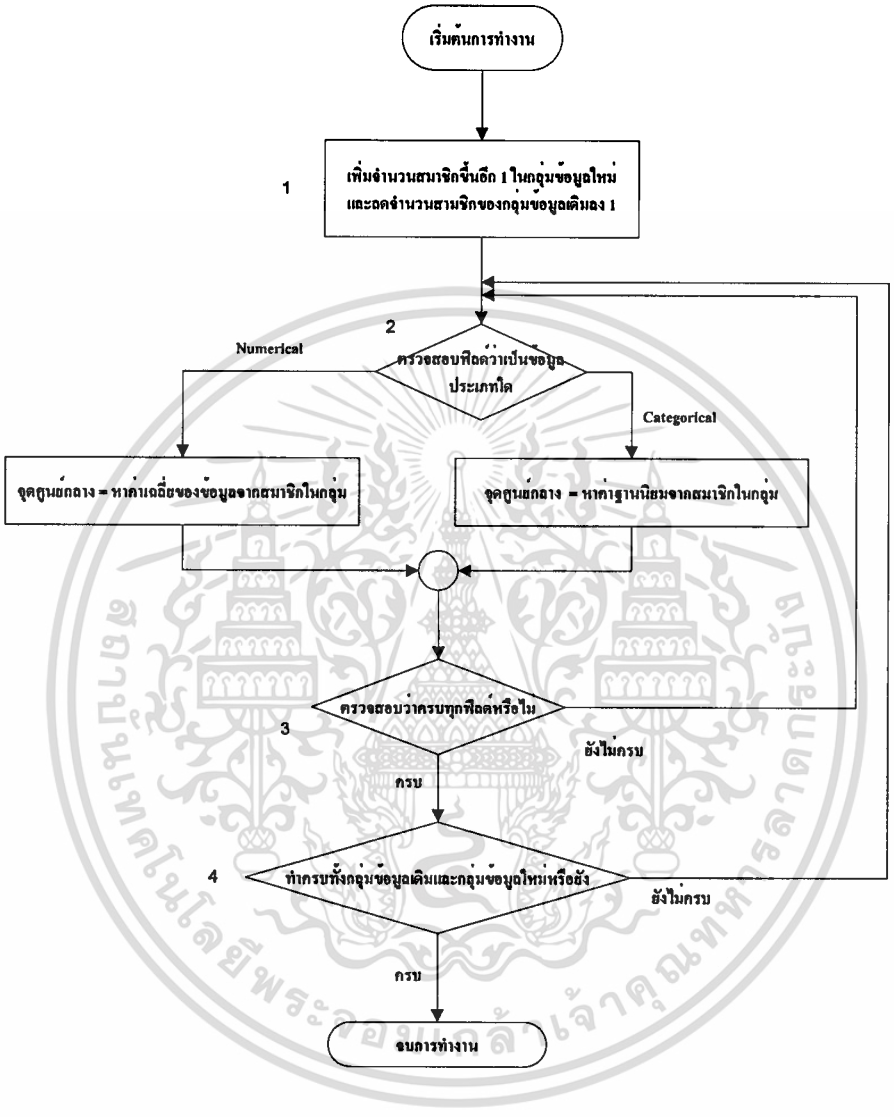
รูปที่ 3.14 อธิบายขั้นตอนการคำนวณหาค่า Distance

1. ทำการตรวจสอบฟิลด์ข้อมูลว่าเป็นข้อมูลประเภทใด
   - ถ้าเป็นข้อมูลประเภท Categorical ให้นำค่าข้อมูลมาเปรียบเทียบกับข้อมูลที่เป็น Categorical ของจุดศูนย์กลางกลุ่มนั้น ๆ
     - ถ้ามีค่าเหมือนกัน ค่าที่ได้จะเท่ากับศูนย์ นำค่าที่ได้จากการเปรียบเทียบคูณด้วยค่า weight ที่ได้กำหนดไว้ในขั้นตอน การปรับเปลี่ยนข้อมูล (Data Transformation) DistanceCategorical = (1 * ค่า weight ของฟิลด์)
     - ถ้ามีค่าต่างกัน ค่าที่ได้จะเท่ากับหนึ่ง นำค่าที่ได้จากการเปรียบเทียบคูณด้วยค่า weight ที่ได้กำหนดไว้ในขั้นตอน การปรับเปลี่ยนข้อมูล (Data Transformation) DistanceCategorical = (0 * ค่า weight ของฟิลด์)
   - ถ้าเป็นข้อมูลประเภท Numerical จะได้ DistanceNumerical = (ค่าจุดศูนย์กลางที่เป็น Numerical ของข้อมูลกลุ่มนั้น ๆ - ข้อมูลที่เป็น Numerical)^2
2. นำค่าที่ได้จาก DistanceCategorical มารวมกับ DistanceNumerical จะได้ minDistance = (DistanceCategorical + DistanceNumerical)
3. ทำการตรวจสอบว่าทำการคำนวณครบทุกฟิลด์ข้อมูลในเรคอร์ดแล้วหรือยัง ถ้ายังไม่ครบ กลับไปทำที่ข้อ 1



รูปที่ 3.14 ผังงานแสดงขั้นตอนการคำนวณหาค่า Distance

3.2.7 ขั้นตอนการคำนวณจุดศูนย์กลางกลุ่มใหม่



รูปที่ 3.15 ผังงานแสดงขั้นตอนคำนวณจุดศูนย์กลางกลุ่มใหม่

รูปที่ 3.15 อธิบายขั้นตอนการคำนวณจุดศูนย์กลางกลุ่มใหม่

1. เพิ่มจำนวนสมาชิกกลุ่มใหม่ขึ้น 1 ตัวและลดจำนวนสมาชิกในกลุ่มเดิมลง 1ตัว
2. ทำการตรวจสอบฟิลด์ข้อมูลว่าเป็นข้อมูลประเภทใด
   - ถ้าเป็นข้อมูลประเภท Categorical จะทำการหาค่าฐานนิยมจากสมาชิกในกลุ่มข้อมูล
   - ถ้าเป็นข้อมูลประเภท Numerical จะทำการหาค่าเฉลี่ยจากสมาชิกในกลุ่มข้อมูล

3. ตรวจสอบว่าทำครบทุกฟิลด์ข้อมูลหรือยัง ถ้ายังไม่ครบกลับไปทำข้อ 2 จนกว่าจะครบทุกฟิลด์ข้อมูล
4. ตรวจสอบว่าครบทั้งกลุ่มข้อมูลเดิม และกลุ่มข้อมูลใหม่หรือไม่ ถ้ายังไม่ครบกลับไปทำข้อที่ 1 จนกว่าจะครบทั้งสองกลุ่มข้อมูล

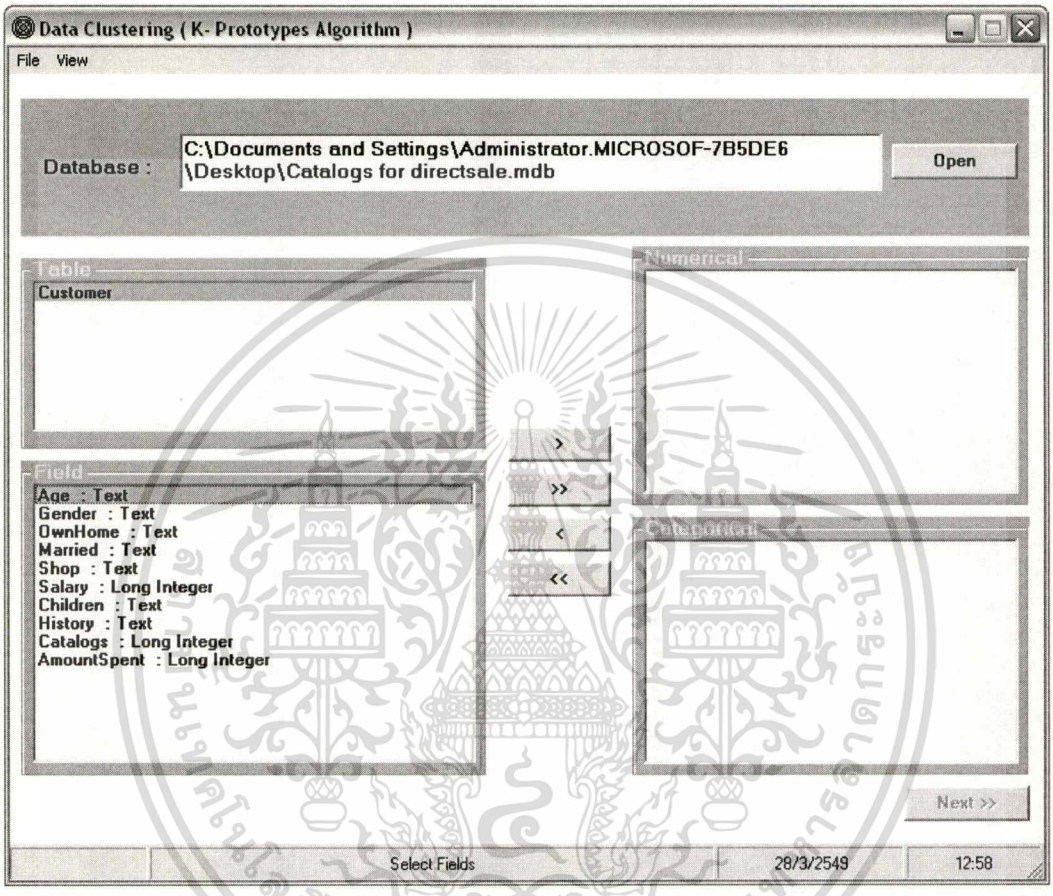
3.3 เครื่องมือที่ใช้ในการพัฒนาระบบ

โดยการพัฒนาเครื่องมือในการจัดกลุ่มข้อมูลโดยใช้ K-Prototypes Algorithm มีเครื่องมือที่ใช้ในการพัฒนาระบบดังนี้

- Software
  - ระบบปฏิบัติการ (Operating System) : Microsoft Windows XP
  - ฐานข้อมูล (Database) : Microsoft Access 2000, Microsoft Access XP
  - Development Language : Microsoft Visual Basic 6 Service Pack 5 เป็นโปรแกรมที่ใช้สร้างโปรแกรมประยุกต์สำหรับระบบปฏิบัติการ Windows สามารถทำงานร่วมกับ Microsoft Access ที่นำมาใช้เป็นฐานข้อมูล
  - โปรแกรมอื่นๆ : Microsoft Visio 2000 , Microsoft Excel, Notepad
- Hardware
  - ซีพียู (CPU) : อินเทล เพนเทียม (Intel Pentium 4) 1.50 GHz
  - แรม (RAM) : RDRAM ขนาด 512 MB
  - Hard Disk : ขนาด 40 GB
  - DVD Drive : สำหรับการ ติดตั้งโปรแกรมที่ใช้ในการทำงานในระบบ

3.4 การออกแบบหน้าจอการทำงานของระบบ

เมื่อเข้าสู่โปรแกรมจะปรากฏหน้าจอสำหรับเลือกฐานข้อมูลที่ต้องการ



รูปที่ 3.16 หน้าจอสำหรับเลือกฐานข้อมูลที่ต้องการนำมาจัดกลุ่ม

3.4.1 การเลือกข้อมูล (Data Selection)

ส่วนประกอบต่าง ๆ ของหน้าจอสำหรับเลือกฐานข้อมูลที่ใช้ในการจัดกลุ่มข้อมูล

- เลือกฐานข้อมูล, ตารางและ ฟิลด์
- แสดงฟิลด์และประเภทของข้อมูลที่ถูกเลือก แสดงฟิลด์และประเภทของข้อมูลที่ถูกเลือกโดยผู้ใช้

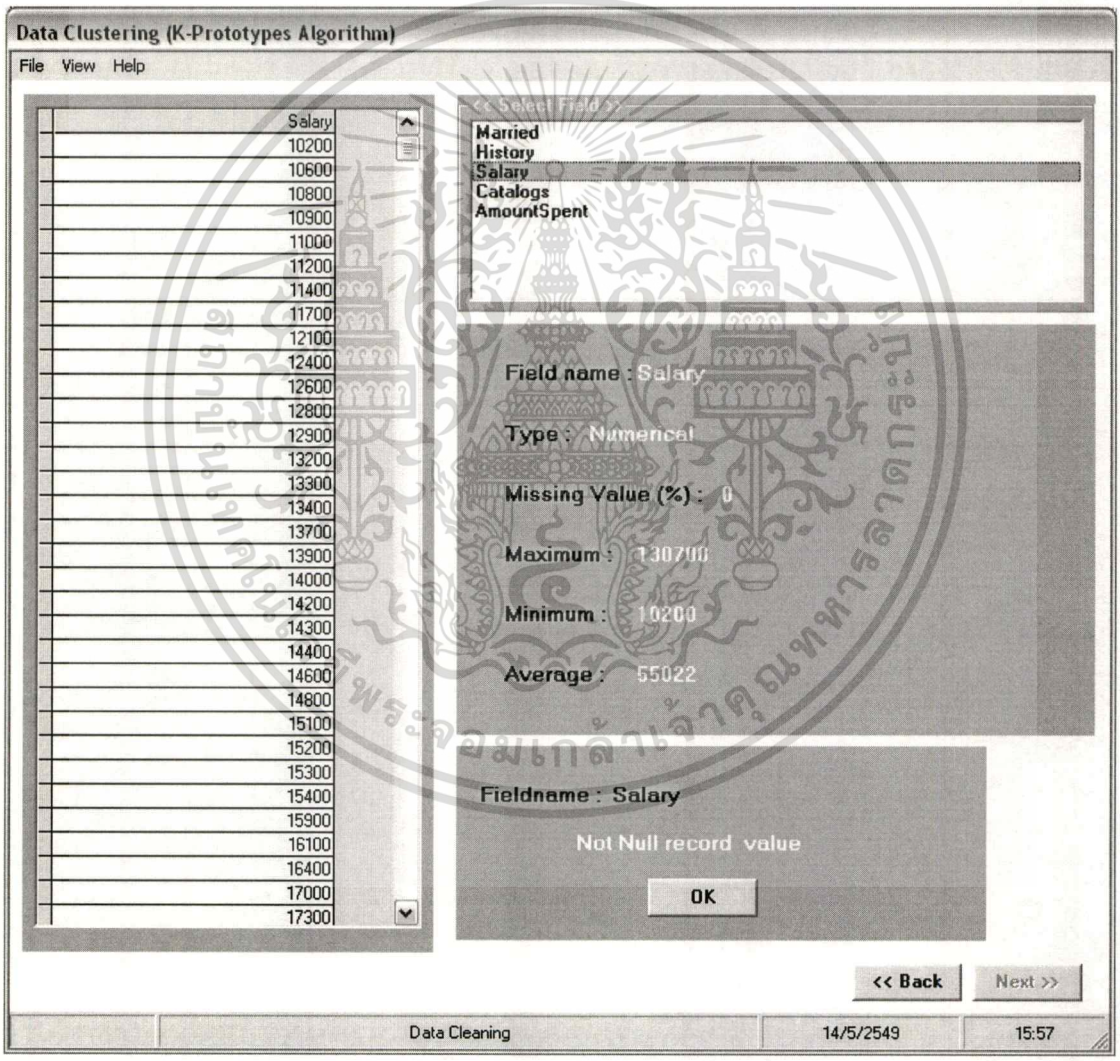
3.4.2 การทำความสะอาดข้อมูล (Data Cleaning)

การแก้ไขข้อมูลเพื่อจัดการกับเรคอร์ดต่าง ๆ ที่มีค่าว่าง (null) สามารถทำการแก้ไขข้อมูลได้ดังนี้

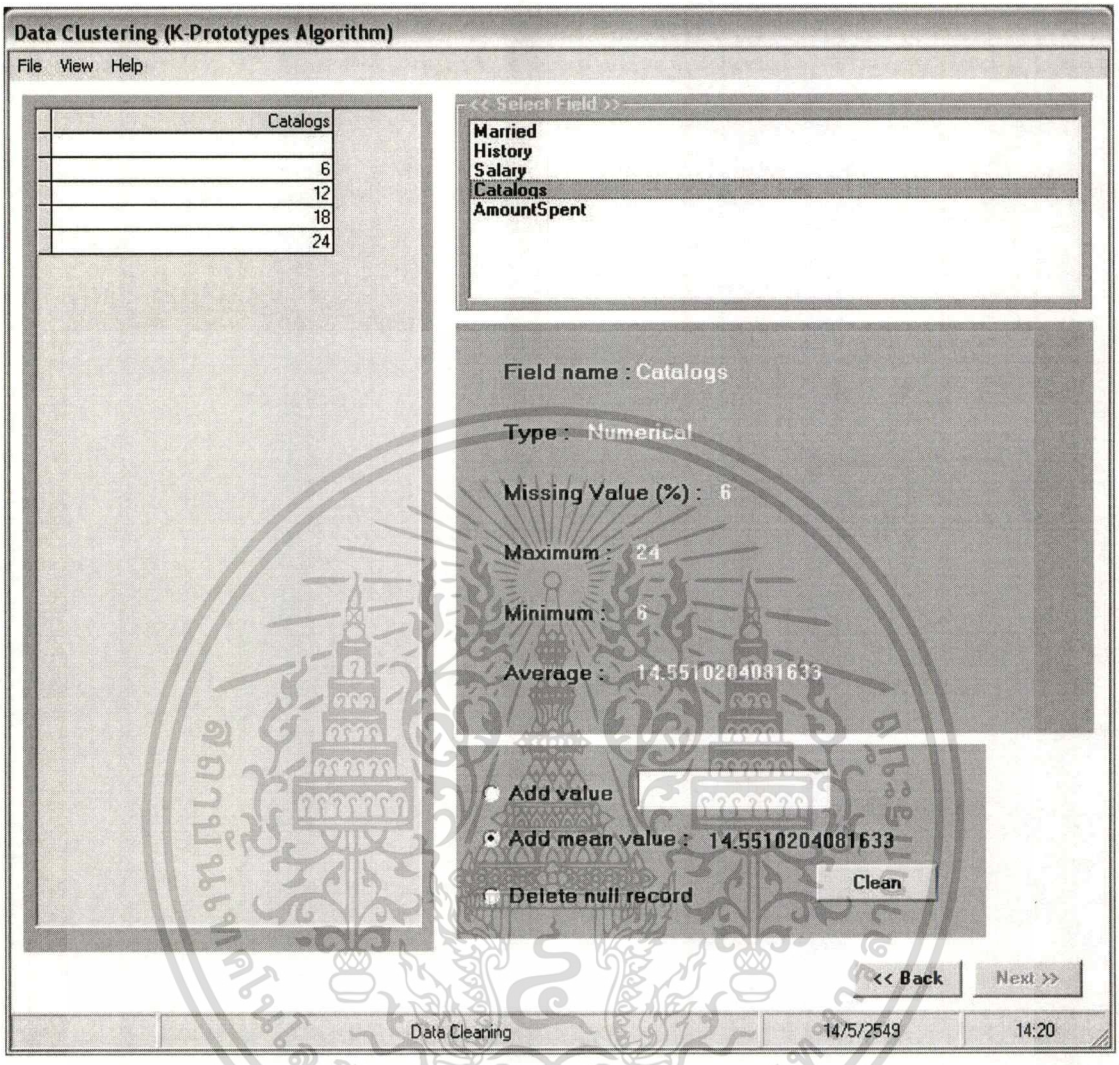
- สำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า null

แสดงรายละเอียดต่าง ๆ ในหน้าจอดังนี้

- ส่วนแสดงตัวอย่างข้อมูลในตาราง
- ชื่อฟิลด์ ,ชนิดของข้อมูล, จำนวนค่าว่าง, ค่าสูงสุด, ค่าต่ำสุดและ ค่าเฉลี่ย



รูปที่ 3.17 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า null

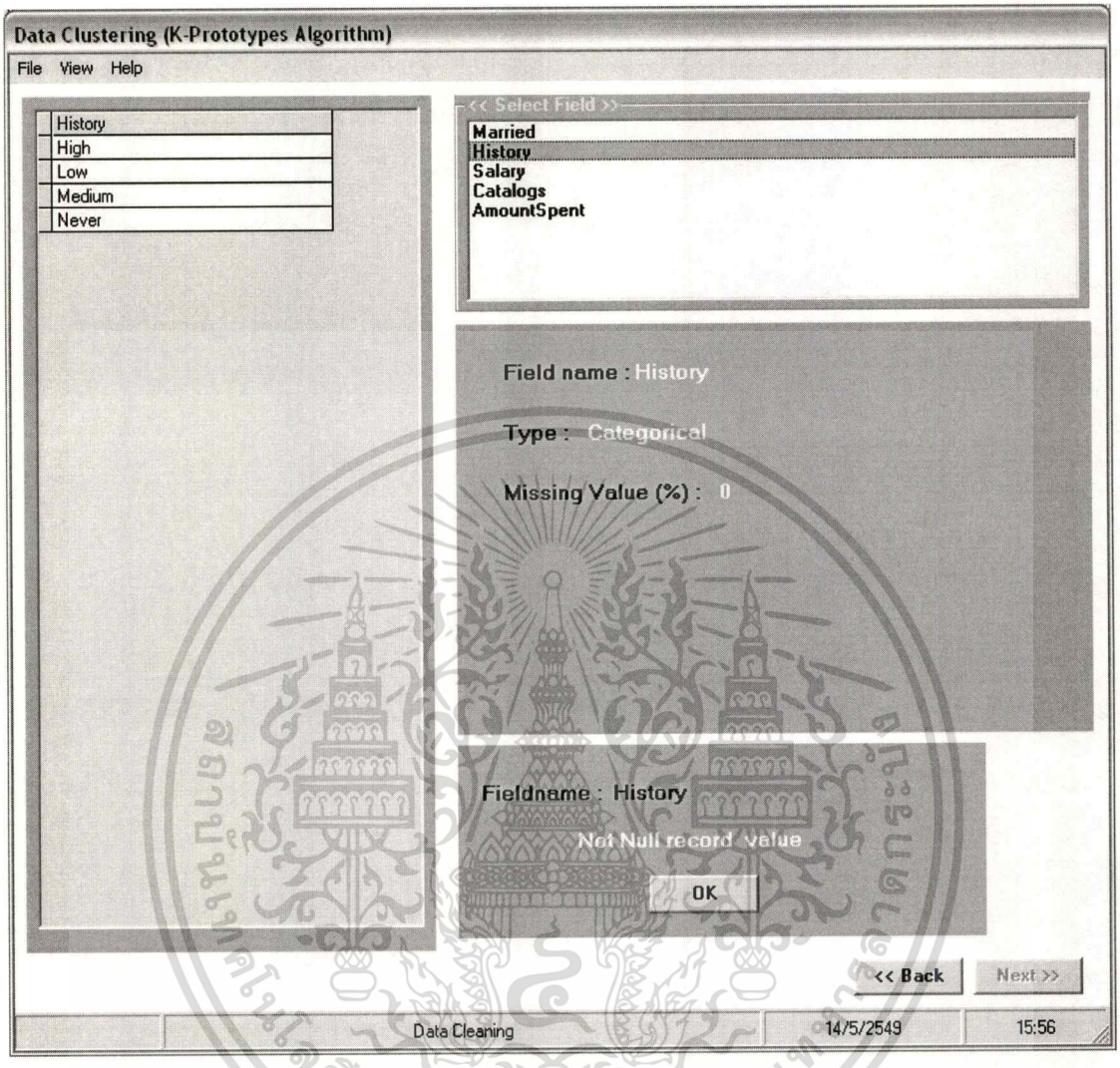


รูปที่ 3.18 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่มีค่า null

- สำหรับฟิลด์ที่เป็น Numerical ที่มีค่า null จะมีข้อมูลแสดงรายละเอียดต่าง ๆ ในหน้าจอดังนี้
    - ส่วนแสดงตัวอย่างข้อมูลในตาราง
    - ชื่อฟิลด์ ,ชนิดของข้อมูล, จำนวนค่าว่าง, ค่าสูงสุด, ค่าต่ำสุด และ ค่าเฉลี่ย

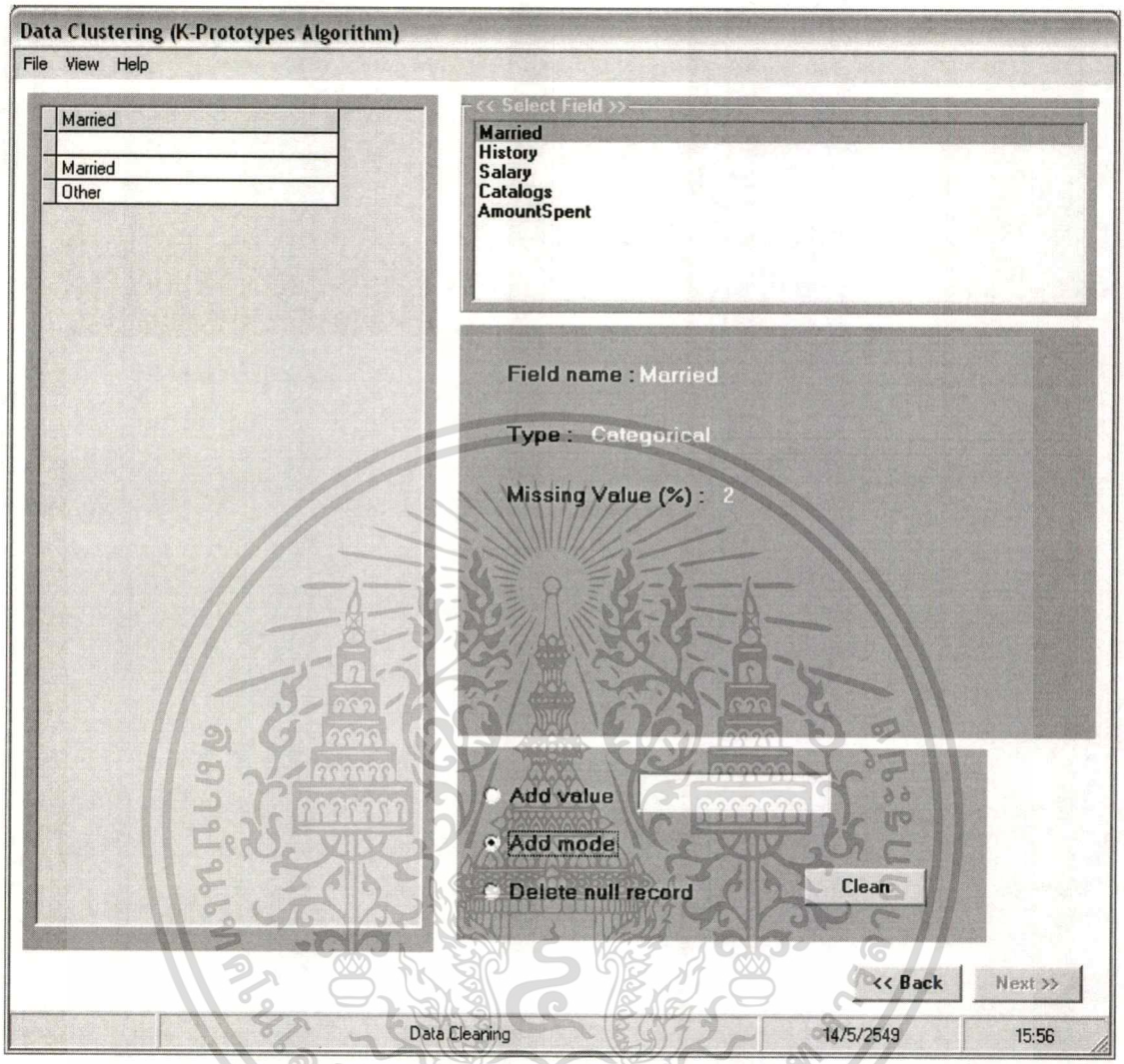
สามารถเลือกแก้ไขข้อมูลได้ดังนี้

- ใส่ค่าโดยที่ผู้ใช้กำหนดเอง
- ใส่ค่าค่าเฉลี่ยแทนข้อมูลที่เป็นค่า null
- ลบเรคอร์ดที่มีค่า null



รูปที่ 3.19 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่มีไม่มีค่า null

- สำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า null จะมีข้อมูลแสดงรายละเอียดต่าง ๆ ในหน้าจอดังนี้
  - ส่วนแสดงตัวอย่างข้อมูลในตาราง
  - ชื่อฟิลด์ ,ชนิดข้อมูลและจำนวนค่าว่าง



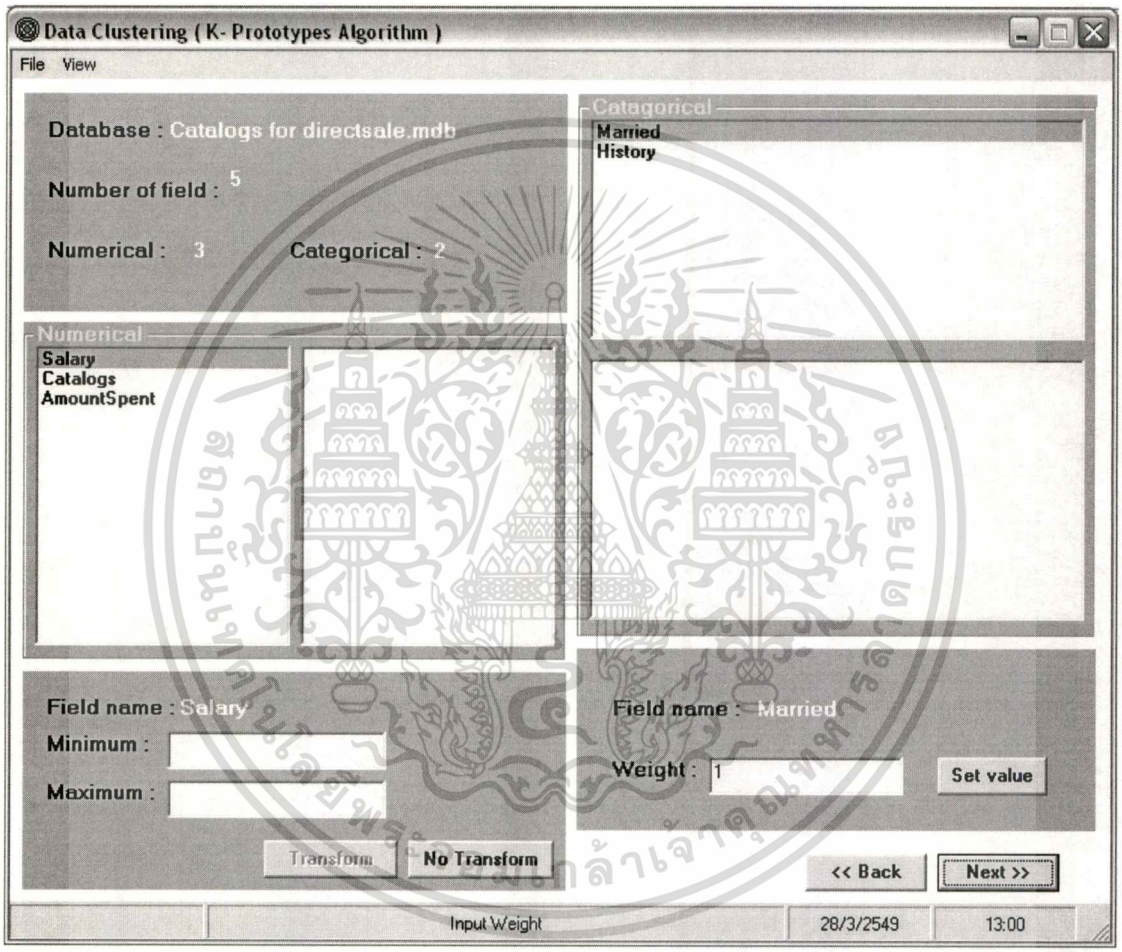
รูปที่ 3.20 แสดงหน้าจอการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่มีค่า null

- สำหรับฟิลด์ที่เป็น Categorical ที่มีค่า null จะมีข้อมูลแสดงรายละเอียดต่าง ๆ ในหน้าจอดังนี้
    - ส่วนแสดงตัวอย่างข้อมูลในตาราง
    - ชื่อฟิลด์ ,ชนิดข้อมูล และ จำนวนค่าว่าง

  การแก้ไขข้อมูลสามารถทำได้ดังนี้
    - ใส่ค่าโดยที่ผู้ใช้กำหนดเอง
    - ใส่ค่าฐานนิยมแทนข้อมูลที่เป็นค่า null
    - ลบเรคอร์ดที่มีค่า null

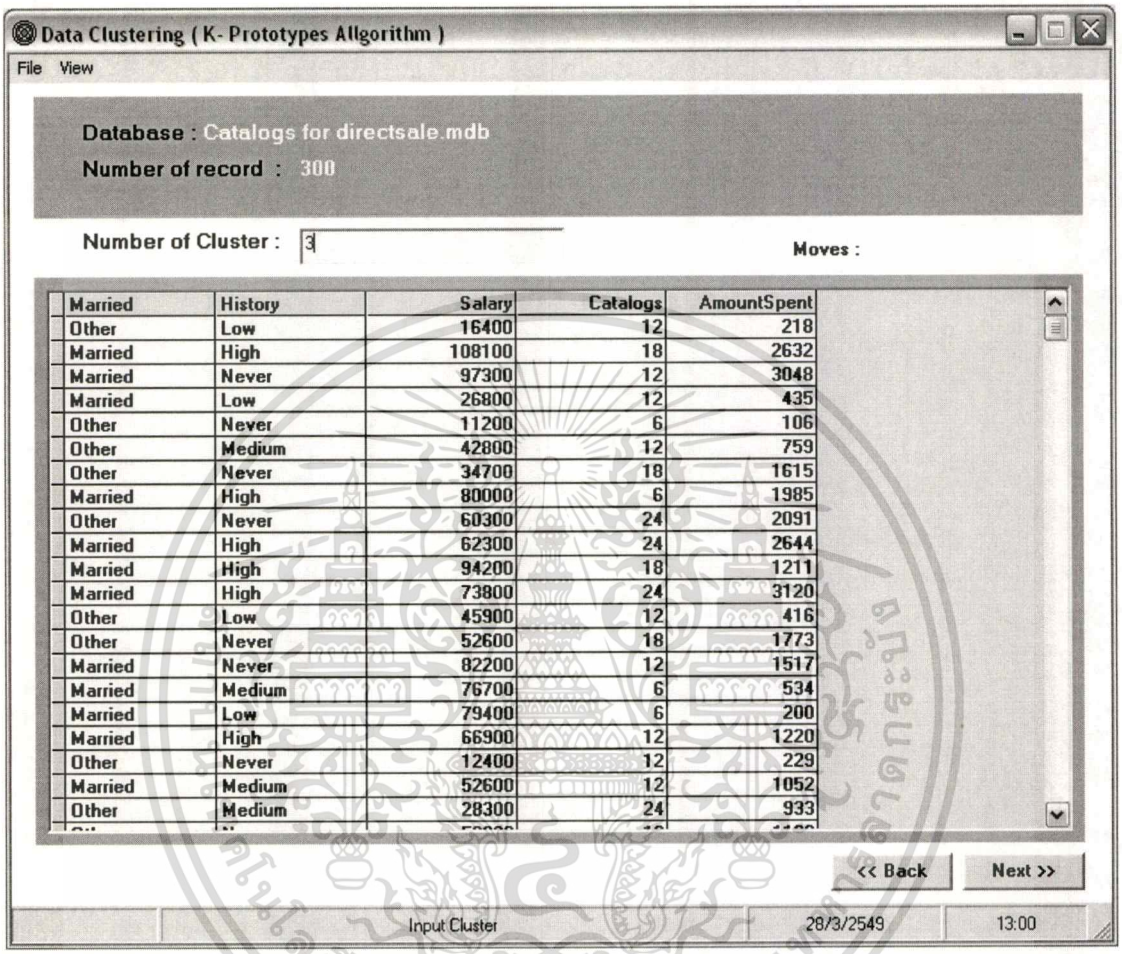
3.4.3 การปรับเปลี่ยนข้อมูล (Data Transformation)

ผู้ใช้สามารถทำการปรับเปลี่ยนข้อมูลของค่าที่เป็น Numerical ให้อยู่ในช่วง ๆ หนึ่งได้ โดยระบุค่า Min และ Max หรือไม่ต้องการปรับเปลี่ยนค่าก็ได้เช่นเดียวกัน สำหรับข้อมูลที่เป็น Categorical จะมีการกำหนดค่า weight เพื่อใช้ในการวัดความเหมือนกันของข้อมูล



รูปที่ 3.21 แสดงหน้าจอการปรับเปลี่ยนข้อมูลและใส่ค่า weight ให้กับข้อมูล

3.4.4 การกำหนดจำนวนกลุ่มข้อมูลที่ต้องการจัดกลุ่ม (Data Mining)

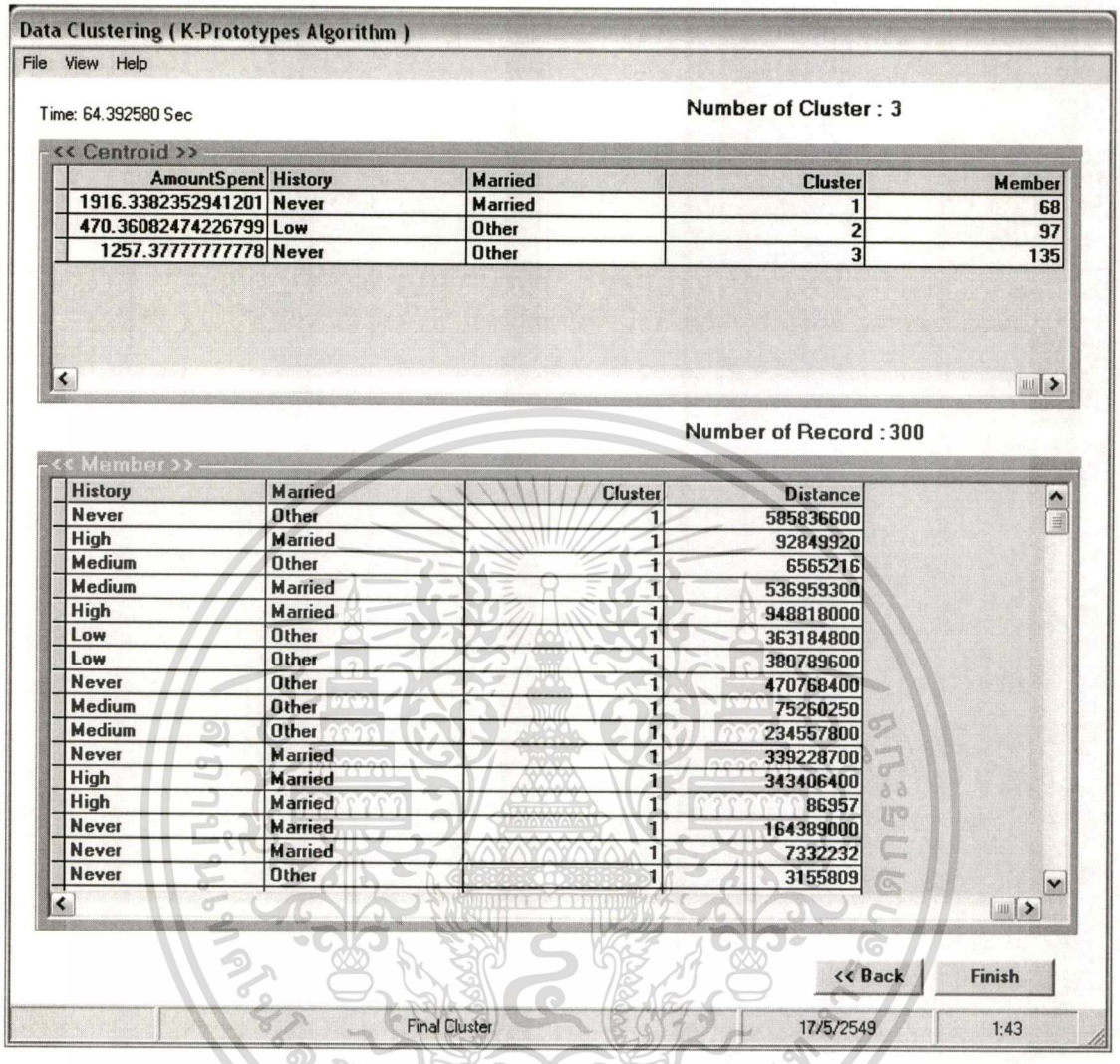


Married	History	Salary	Catalogs	AmountSpent
Other	Low	16400	12	218
Married	High	108100	18	2632
Married	Never	97300	12	3048
Married	Low	26800	12	435
Other	Never	11200	6	106
Other	Medium	42800	12	759
Other	Never	34700	18	1615
Married	High	80000	6	1985
Other	Never	60300	24	2091
Married	High	62300	24	2644
Married	High	94200	18	1211
Married	High	73800	24	3120
Other	Low	45900	12	416
Other	Never	52600	18	1773
Married	Never	82200	12	1517
Married	Medium	76700	6	534
Married	Low	79400	6	200
Married	High	66900	12	1220
Other	Never	12400	12	229
Married	Medium	52600	12	1052
Other	Medium	28300	24	933

รูปที่ 3.22 แสดงหน้าจอการกำหนดจำนวนสำหรับการจัดกลุ่มข้อมูล

3.4.5 การแสดงผลการจัดกลุ่มข้อมูล (Output)

- Centroid จะแสดงจุดศูนย์กลางของแต่ละกลุ่ม และบอกจำนวนสมาชิกที่อยู่ในกลุ่มนั้น ๆ
- Member จะแสดงสมาชิกที่อยู่ในกลุ่มใด ๆ และบอกถึงระยะห่างระหว่างสมาชิกกับจุดศูนย์กลางของกลุ่มนั้น ๆ



รูปที่ 3.23 หน้าจอแสดงผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูล

3.4.6 การส่งออกข้อมูล (Export)

เมื่อทำการจัดกลุ่มข้อมูลเรียบร้อยแล้ว สามารถที่จะส่งออกข้อมูลให้อยู่ในรูปแบบต่าง ๆ ได้ดังนี้

- เอกสารที่เป็นเท็กซ์ไฟล์ (.txt) จะแสดงข้อมูลที่เป็นจุดศูนย์กลางและสมาชิกในกลุ่มเป็นรูปแบบของเท็กซ์ไฟล์
- เอกสารที่เป็น Microsoft Excel (.xls)จะแสดงข้อมูลที่เป็นจุดศูนย์กลางและสมาชิกในกลุ่มเป็นรูปแบบของ Microsoft Excel และกราฟแสดงผล

ทำให้สะดวกต่อการนำผลลัพธ์ที่ได้ไปวิเคราะห์เพื่อใช้ในการสนับสนุนการตัดสินใจต่อไป

บทที่ 4
การทดลองใช้งานโปรแกรม

4.1 คู่มือและตัวอย่างการใช้งาน

การทำงานของระบบจัดกลุ่มข้อมูลลูกค้าที่พัฒนาขึ้น ประกอบด้วยส่วนหลัก 3 ส่วน คือ ส่วนนำข้อมูลเข้า (input), ส่วนประมวลผล (process), ส่วนแสดงผล (output)

4.1.1 ส่วนนำเข้าข้อมูล

ส่วนนำเข้าข้อมูลจะเป็นขั้นตอนในการนำข้อมูลเข้ามาเพื่อใช้ในการจัดกลุ่ม

1. ไฟล์ฐานข้อมูล ประกอบด้วยข้อมูลที่ต้องการจัดกลุ่ม โดยใช้ Microsoft Access 2000
2. ตารางและฟิลด์ข้อมูลเลือกจากฐานข้อมูลในข้อที่ 1
3. ค่า weight ของข้อมูลประเภท Categorical
4. จำนวนกลุ่มที่ต้องการจะจัดแบ่ง

4.1.2 ส่วนประมวลผล

การประมวลผลข้อมูลที่นำเข้ามาจากส่วนของการนำเข้าข้อมูลเพื่อทำการจัดกลุ่มข้อมูลตาม K-Prototypes Algorithm โดยขั้นตอนการประมวลผลมี ดังนี้

1. ส่วนการเตรียมข้อมูล เป็นขั้นตอนในการเตรียมข้อมูลเพื่อจะนำไปใช้ต่อในกระบวนการไมนิ่ง โดยมีขั้นตอนดังนี้
   - 1.1 เลือกฟิลด์ข้อมูลที่ต้องการนำมาจัดกลุ่ม
   - 1.2 ทำการแก้ไขข้อมูล (Data Cleaning) โดยจะตรวจสอบค่าที่หายไป (Missing Value) ถ้าตรวจพบก็จะมีการแก้ไข โดยผู้ใช้
   - 1.3 ทำการปรับเปลี่ยนข้อมูล (Data Transformation) ซึ่งในกระบวนการนี้จะทำหรือไม่ขึ้นอยู่กับผู้ใช้ต้องการ
2. ส่วนของการทำไมนิ่ง เป็นการจัดกลุ่มโดยใช้ K-Prototypes Algorithm
   - 2.1 Random จุดศูนย์กลางของกลุ่มตามจำนวนกลุ่มที่ต้องการจะแบ่ง เพื่อเป็นจุดศูนย์กลางของกลุ่มข้อมูล
   - 2.2 กำหนดค่า Mindistance โดยเปรียบเทียบกับจุดศูนย์กลางข้อมูลกลุ่มที่ 1
   - 2.3 คำนวณหาค่า Distance กับจุดศูนย์กลางของกลุ่มข้อมูลกลุ่มถัดไป

2.4. เปรียบเทียบว่าค่าที่คำนวณได้นั้นมีค่าน้อยกว่าค่า Mindistance หรือไม่ ถ้าค่า Mindistance น้อยกว่าค่า Distance ที่คำนวณได้ ดังนั้นให้ค่า Mindistance เท่ากับ ค่า Distance และเก็บเลขที่กลุ่มนั้น

2.5. ตรวจสอบว่าเปรียบเทียบกับจุดศูนย์กลางข้อมูลครบทุกกลุ่มหรือยัง ถ้ายังให้กลับไปทำข้อ 3 ซ้ำ

2.6. ตรวจสอบข้อมูลว่าต้องทำการย้ายกลุ่มหรือไม่ ถ้าต้องทำการย้ายกลุ่มให้ทำการคำนวณค่าจุดศูนย์กลางเดิมและจุดศูนย์กลางใหม่

2.7. ตรวจสอบข้อมูลว่าทำครบทุกข้อมูลหรือยัง ถ้ายังไม่ครบกลับไปทำซ้ำที่ข้อ 2.2-2.6

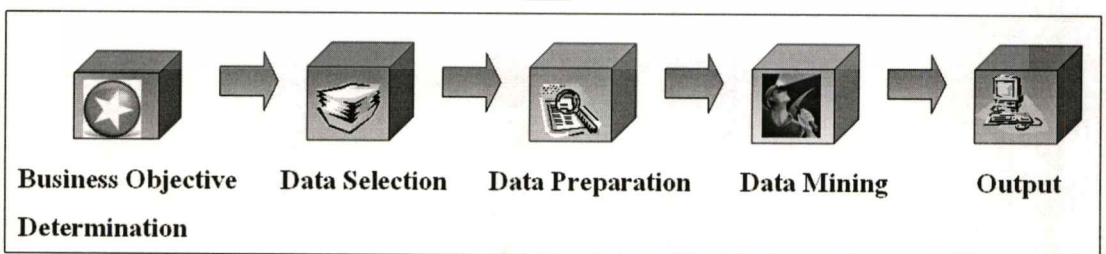
2.8. ทำซ้ำข้อที่ 2.2ใหม่ จนกระทั่งข้อมูลไม่มีการเปลี่ยนกลุ่ม

4.1.3 ส่วนแสดงผล

- Monitor ระบบจะแสดงผลการจัดกลุ่มทางหน้าจอ โดยแสดงรายละเอียดดังนี้
  - แสดงค่าจุดศูนย์กลางของกลุ่มข้อมูล และจำนวนสมาชิกในกลุ่ม
  - แสดงรายละเอียดของสมาชิกแต่ละตัวในกลุ่มข้อมูลและระยะห่างจากจุดศูนย์กลาง
- Export การส่งออกข้อมูล ทำให้สะดวกต่อการนำผลลัพธ์ที่ได้ไปวิเคราะห์เพื่อใช้ในการสนับสนุนการตัดสินใจต่อไป สามารถส่งออกข้อมูลในรูปแบบต่าง ๆ ดังนี้
  - Text Document หมายถึง เท็กซ์ไฟล์ (.txt)
  - Excel หมายถึง เอกสาร Microsoft Excel (.xls) และ การแสดงผลรูปแบบภาพชนิดกราฟ

4.2 ขั้นตอนการประยุกต์ใช้งานระบบ

การใช้งานระบบในการจัดกลุ่มลูกค้านั้นมีขั้นตอนต่าง ๆดังรูปที่ 4.1



รูปที่ 4.1 แสดงขั้นตอนการประยุกต์ใช้งานระบบ

จากรูปที่ 4.1 สามารถอธิบายการทำงานได้ดังนี้

- การกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objective Determination)

ปัจจุบันองค์กรทางธุรกิจ ส่วนใหญ่เริ่มหันมาให้ความสนใจในเรื่องการสร้างจุดขายในตลาดสินค้า ด้วยการกำหนดคุณลักษณะกลุ่มตลาดเป้าหมายสำหรับสินค้าและบริการที่ต้องการซึ่งมีความโดดเด่นเหนือกว่าสินค้าอื่น ๆ ซึ่งจะทำให้สามารถลดต้นทุนในการโฆษณาสินค้าและบริการเหล่านั้น ธุรกิจการขายตรง เป็นธุรกิจหนึ่งที่มีจำนวนมากขึ้นเรื่อย ๆ ในช่วงเวลาเพียงไม่กี่ปี ในเรื่องของการประชาสัมพันธ์สินค้าและบริการต่าง ๆ ของธุรกิจเหล่านี้นั้น ส่วนใหญ่แล้วจะส่งข้อมูลข่าวสารแก่ลูกค้าผ่านทางใบโฆษณาสินค้า (Catalog) ซึ่งบริษัทเหล่านี้จะต้องใช้เงินจำนวนมากในการโฆษณาและส่งใบโฆษณาสินค้าแก่ลูกค้า โดยที่ไม่ทราบว่าลูกค้ารายนั้น ๆ มีความสนใจที่จะซื้อสินค้าเหล่านั้นมากน้อยเพียงไร

บริษัทจำเป็นที่ต้องลดต้นทุนในการส่งใบโฆษณาสินค้า และจะจัดส่งให้กับลูกค้าที่คาดว่าจะซื้อสินค้าและบริการ เป็นการยากที่จะวิเคราะห์ข้อมูลเหล่านั้นจากฐานข้อมูลลูกค้าทั้งหมดที่มีอยู่ ดังนั้นจึงมีแนวคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้กับการแก้ปัญหานี้ โดยมีวัตถุประสงค์ในการจัดกลุ่มลูกค้าให้เหมาะสมต่อการจัดส่งใบโฆษณาสินค้าของบริษัท เพื่อนำมาใช้ในการสร้างแผนการตลาดแบบขายตรง และลดต้นทุนในการโฆษณาสินค้า เพื่อให้บริษัทมีกำไรมากขึ้น

- การคัดเลือกข้อมูล (Data Selection)
  ข้อมูลที่นำมาใช้เป็นข้อมูลจากฐานข้อมูลของบริษัทแห่งหนึ่งซึ่งทำธุรกิจการขายตรงเกี่ยวกับอุปกรณ์ระบบเสียง , คอมพิวเตอร์ และผลิตภัณฑ์อิเล็กโทรนิคต่าง ๆ โดยทำการบันทึกข้อมูลของลูกค้าลงในฐานข้อมูล Microsoft Access 2000 ชื่อ Catalogs for directsale.mdb และมีตารางชื่อ Customer ซึ่งได้ทำการคัดเลือกเฉพาะฟิลด์ที่ต้องการดังนี้

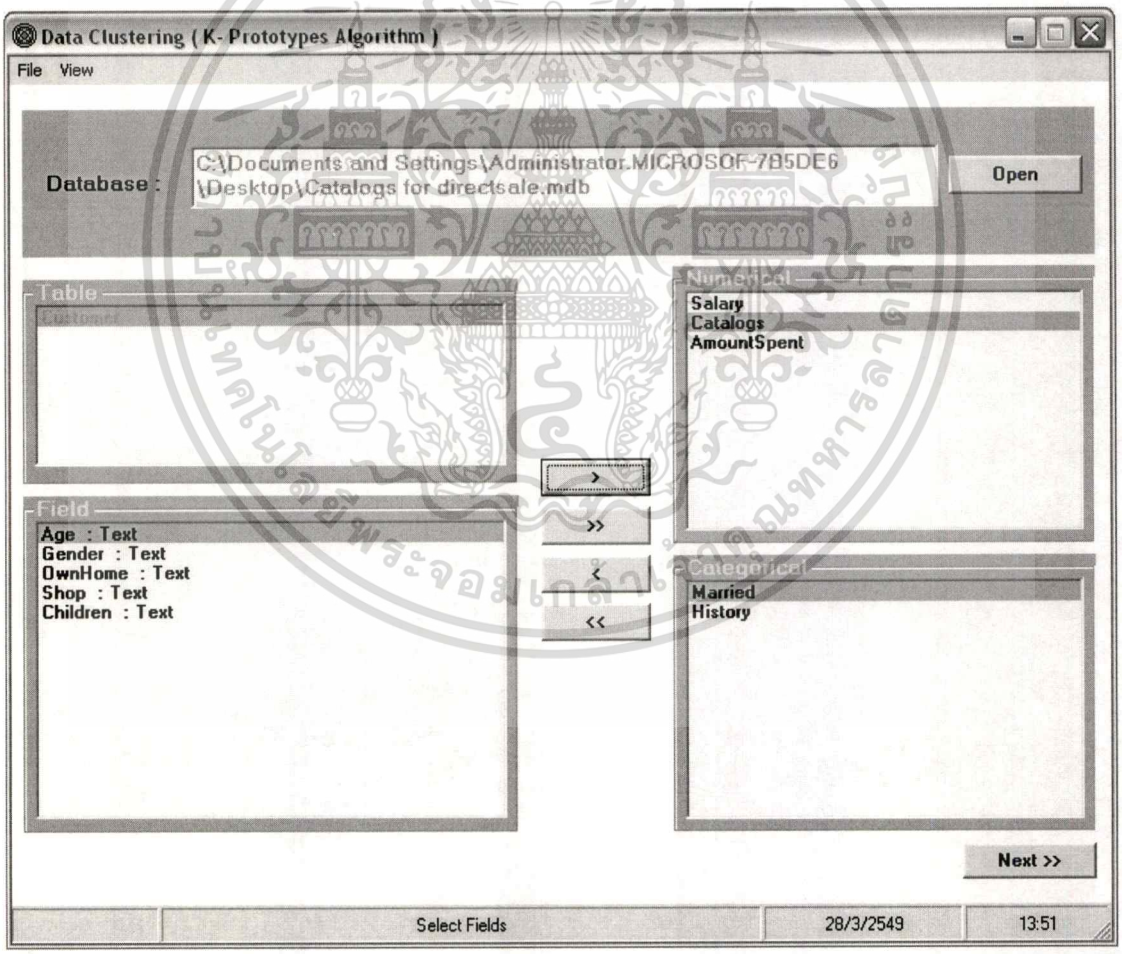
ตารางที่ 4.1 แสดงข้อมูลตาราง Customer ที่นำมาทำการคัดเลือกข้อมูลเพื่อจัดกลุ่ม

ชื่อฟิลด์ข้อมูล	**ประเภทของข้อมูล**
Married	Text(Categorical)
Salary	Single(Numerical)
History	Text(Categorical)
AmountSpent	Single(Numerical)
Catalogs	Single(Numerical)

จากตารางที่ 4.1 สามารถอธิบายความหมายของแต่ละฟิลด์ได้ดังนี้

- Married หมายถึง สถานภาพของลูกค้า
- Salary หมายถึง ผลรวมเงินเดือนทั้งปีของลูกค้าโดยถ้าลูกค้าสมรสแล้ว จะรวมเงินเดือนของทั้งลูกค้าและคู่สมรสด้วย
- History หมายถึง ระดับการซื้อสินค้าและบริการของลูกค้าในปีที่ผ่านมา
- AmountSpent หมายถึง ผลรวมจากการซื้อสินค้าและบริการของลูกค้าในปีนี้
- Catalogs หมายถึง จำนวนใบโฆษณาสินค้าที่บริษัทส่งถึงลูกค้าตลอดปี

หน้าจอการคัดเลือกข้อมูลของโปรแกรมจะแสดงได้ดังนี้

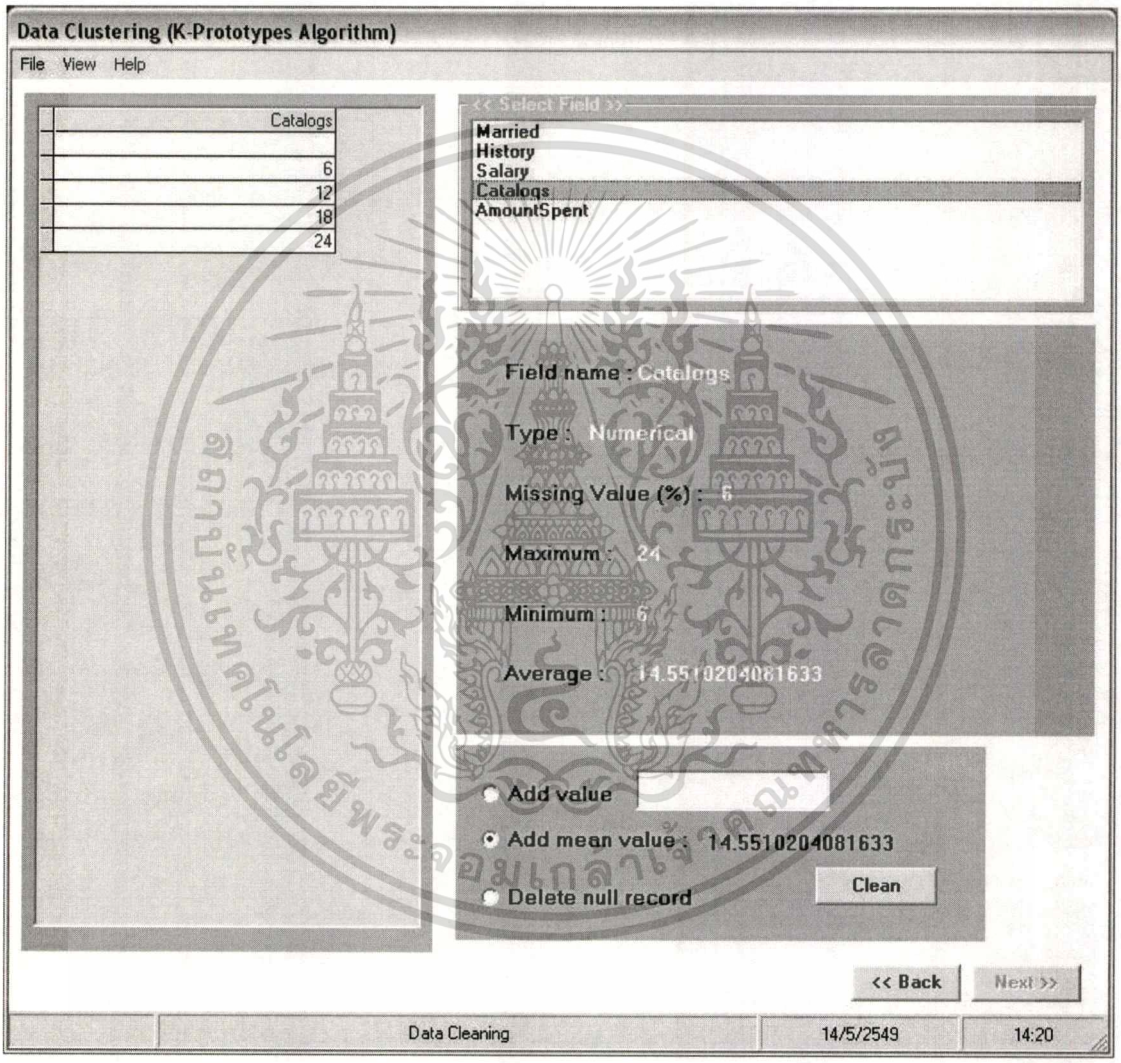


รูปที่ 4.2 หน้าจอเลือกตาราง, ฟิลด์ข้อมูลที่ต้องการจัดกลุ่ม(Data Selection)

- การเตรียมข้อมูล(Data Preparation) สำหรับการทำดาต้าไมนิ่ง

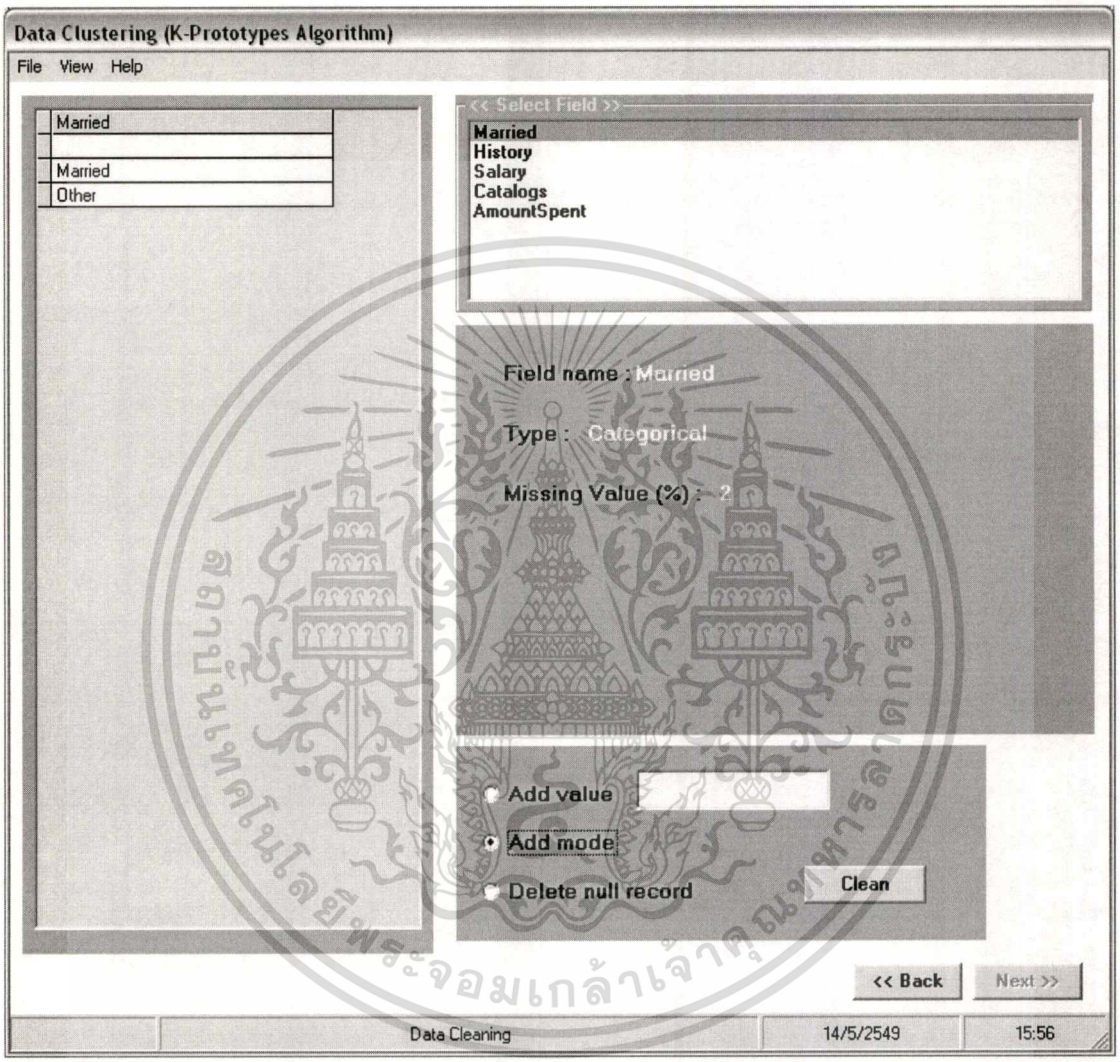
1. การทำความสะอาดข้อมูล(Data Cleaning) แก้ไขข้อมูลที่มีค่าว่าง

จะพบฟิลด์ที่มีค่าว่างคือฟิลด์ Catalogs และ Married โดยการแก้ไขข้อมูลสำหรับฟิลด์ Catalogs จะทำการแก้ไขโดยการใส่ค่า Mean ดังรูปที่ 4.3



รูปที่ 4.3 หน้าจอการทำความสะอาดข้อมูลสำหรับข้อมูลประเภท Numerical (Data Cleaning)

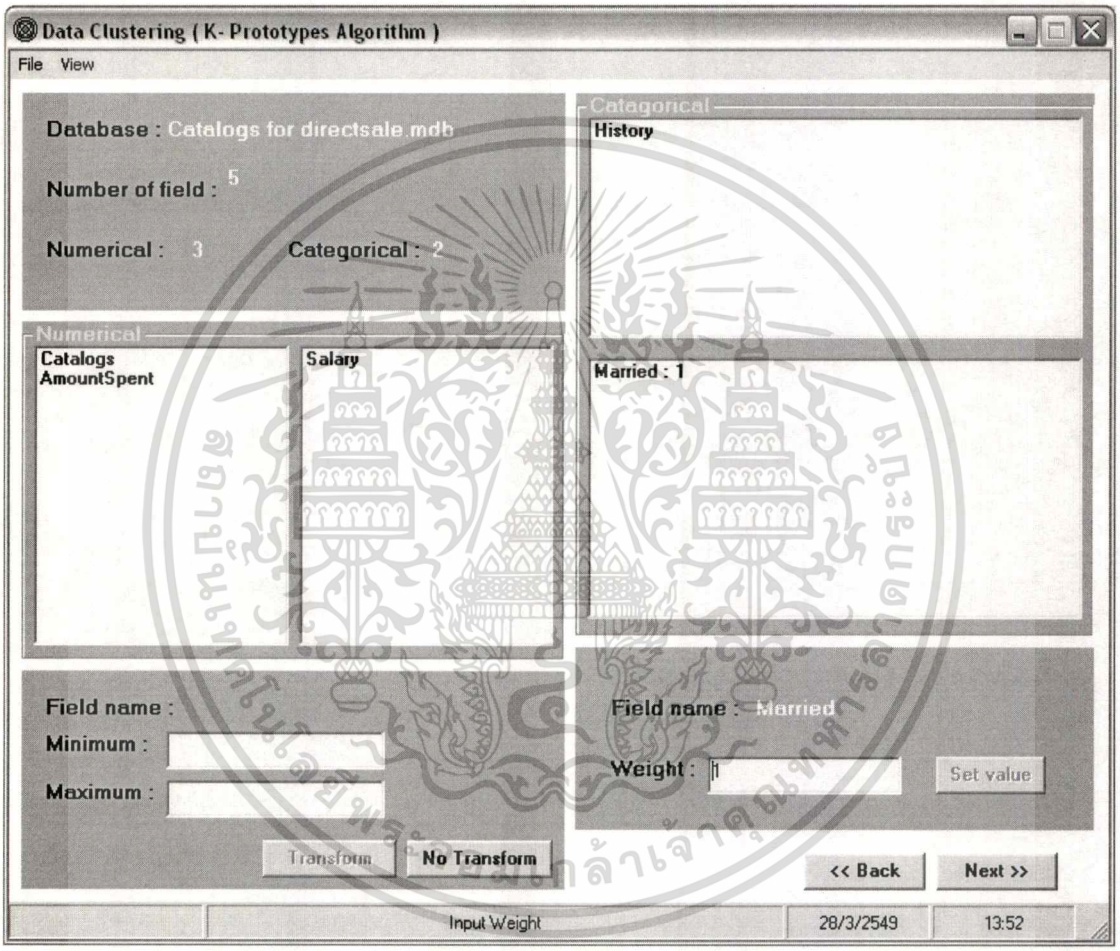
การแก้ไขข้อมูลสำหรับฟิลด์ Married จะทำการแก้ไขโดยการใส่ค่า Mode ส่วนฟิลด์อื่น ๆ ไม่พบค่าว่าง จึงไม่จำเป็นต้องทำการแก้ไขข้อมูล ดังรูปที่ 4.4



รูปที่ 4.4 หน้าจอการทำความสะอาดข้อมูลสำหรับข้อมูลประเภท Categorical (Data Cleaning)

2. การปรับเปลี่ยนข้อมูล (Data Transformation) การปรับเปลี่ยนข้อมูลให้อยู่ในช่วงใด ๆ

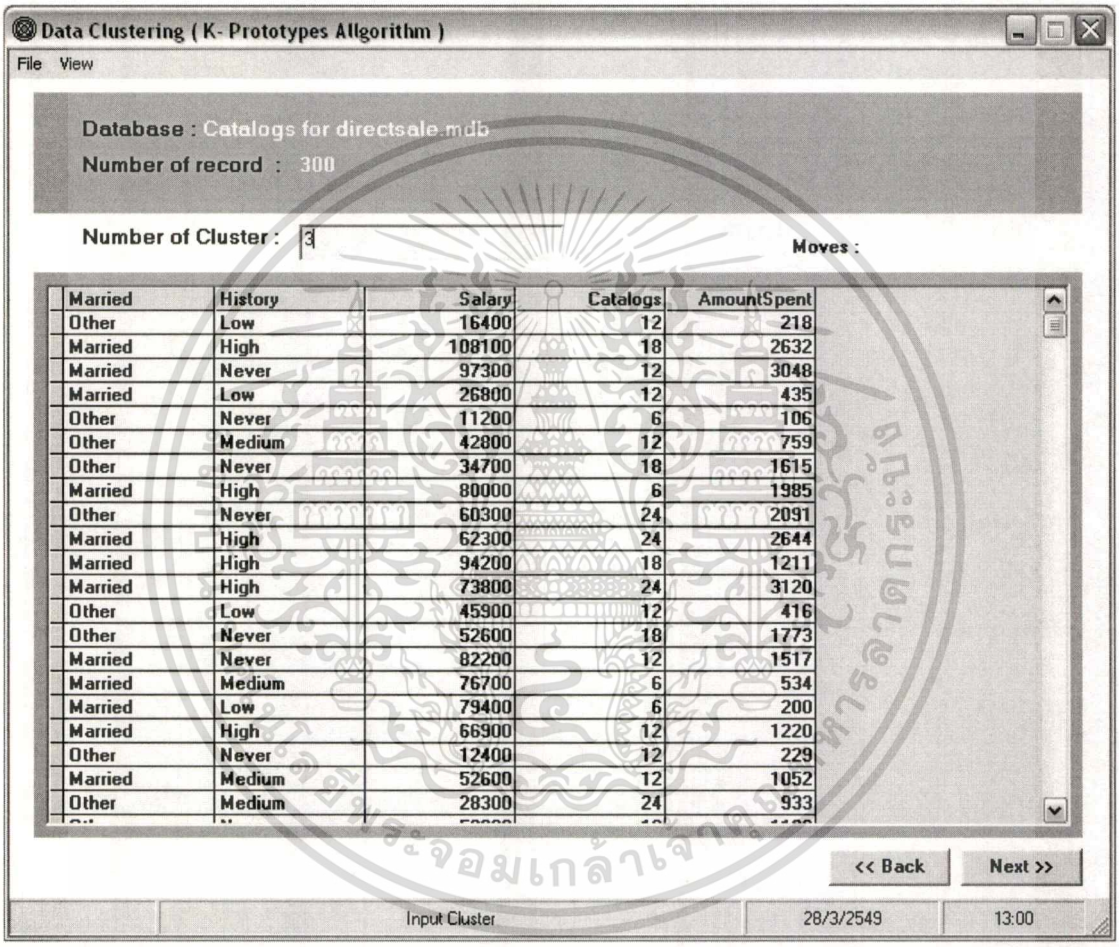
ในขั้นตอนนี้จะไม่ทำการ Transformation ฟิลด์ข้อมูลที่เป็น Numerical ซึ่งได้แก่ Salary, Catalogs และAmountSpent ส่วนฟิลด์ที่เป็น Categorical คือ History และ Married กำหนดให้มีค่า weight เป็น 1.0 ดังรูปที่ 4.5



รูปที่ 4.5 หน้าจอการปรับเปลี่ยนข้อมูล(Data Transformation)

- Data Mining การทำดาต้าไมนิ่งในการจัดกลุ่มข้อมูล

เมื่อทำการเตรียมข้อมูลสำหรับทำไมนิ่งเรียบร้อยแล้วนั้น ขั้นตอนนี้จะเป็นการกำหนดจำนวนกลุ่มข้อมูลที่ต้องการจัดแบ่งเป็น 3 กลุ่ม หลังจากที่ใส่จำนวนกลุ่มข้อมูลที่ต้องการทำการจัดกลุ่มแล้ว โปรแกรมจะทำการจัดกลุ่มข้อมูลโดยผ่าน K-Prototypes Algorithm ดังรูปที่ 4.6



Married	History	Salary	Catalogs	AmountSpent
Other	Low	16400	12	218
Married	High	108100	18	2632
Married	Never	97300	12	3048
Married	Low	26800	12	435
Other	Never	11200	6	106
Other	Medium	42800	12	759
Other	Never	34700	18	1615
Married	High	80000	6	1985
Other	Never	60300	24	2091
Married	High	62300	24	2644
Married	High	94200	18	1211
Married	High	73800	24	3120
Other	Low	45900	12	416
Other	Never	52600	18	1773
Married	Never	82200	12	1517
Married	Medium	76700	6	534
Married	Low	79400	6	200
Married	High	66900	12	1220
Other	Never	12400	12	229
Married	Medium	52600	12	1052
Other	Medium	28300	24	933

รูปที่ 4.6 หน้าจอการกำหนดจำนวนกลุ่มข้อมูล

- Output การแสดงผลข้อมูล

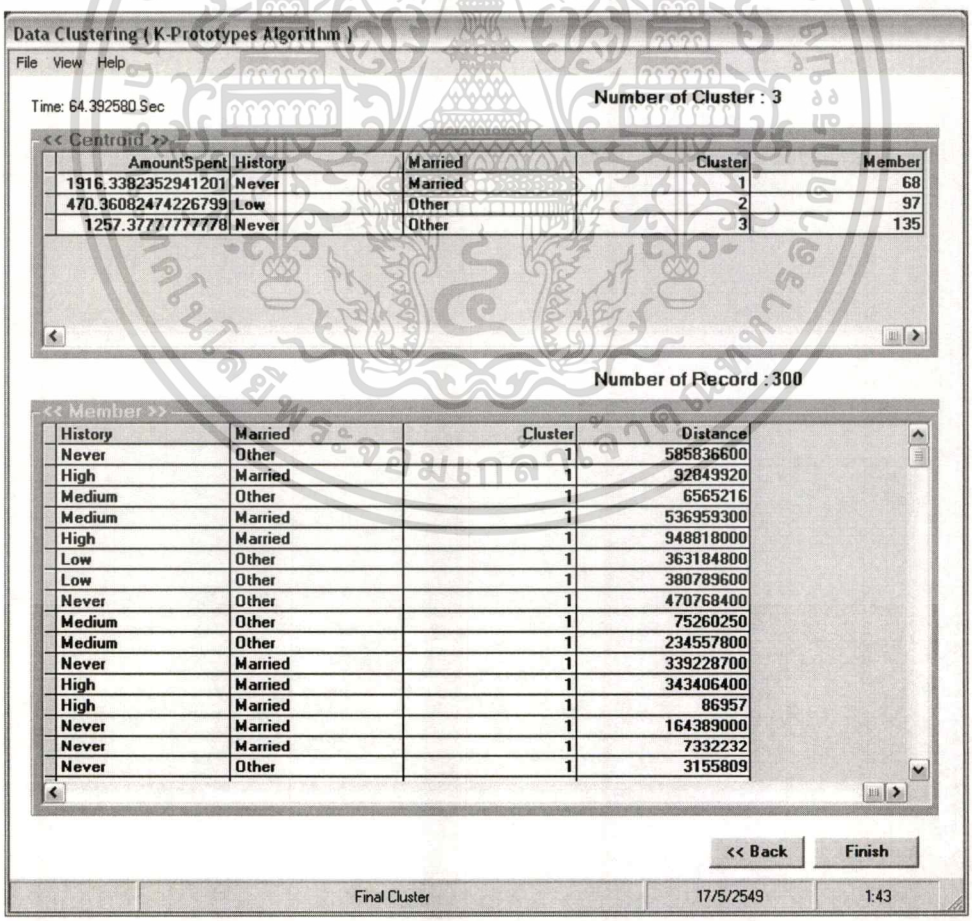
หน้าจอ จะเป็นส่วนของการแสดงผลเป็น 2 ส่วนด้วยกันคือ

1. Centroid หมายถึง จุดศูนย์กลางของกลุ่มข้อมูล ซึ่งได้แสดงข้อมูลที่จัดกลุ่มและจำนวนข้อมูลที่เป็นสมาชิกภายในกลุ่ม

2. Member หมายถึง สมาชิกของกลุ่มข้อมูลนั้น ๆ ซึ่งจะแสดงระยะห่างระหว่างตัวสมาชิกกับจุดศูนย์กลางของกลุ่มข้อมูล และแสดงให้เห็นว่าข้อมูลนั้นอยู่ในกลุ่มข้อมูลใด

- การส่งออกข้อมูล สามารถส่งออกข้อมูลเพื่อสร้างรายงานได้หลายรูปแบบดังนี้

1. เท็กซ์ไฟล์ บันทึกผลที่ได้จากการจัดกลุ่มข้อมูลให้อยู่ในรูปแบบของเท็กซ์ไฟล์ (. txt)
2. เอกสาร Microsoft Excel สามารถส่งออกผลลัพธ์ที่ได้จากการจัดกลุ่มข้อมูลในรูปแบบของเอกสาร Microsoft Excel (.xls) และกราฟแสดงผล



รูปที่ 4.6 หน้าจอแสดงผลการจัดกลุ่มข้อมูล(Output)

4.3 การวิเคราะห์ผลที่ได้จากการทำดาต้าไมนิ่งและการนำความรู้มาใช้

เมื่อได้ผลลัพธ์จากการทำดาต้าไมนิ่งแล้วจะต้องนำมาวิเคราะห์ต่อ โดยมีวัตถุประสงค์ในการจัดกลุ่มลูกค้าให้เหมาะสมต่อการจัดส่งใบโฆษณาสินค้าของบริษัท เพื่อนำมาใช้ในการสร้างแผนการตลาดแบบขายตรง และลูกค้าจะได้รับใบโฆษณาสินค้าตามจำนวนที่คาดว่าลูกค้านั้นจะสนใจ นำข้อมูลที่ได้จากฐานข้อมูลลูกค้ามาทำการจัดกลุ่ม วิเคราะห์ดูว่าควรที่จะมีการปรับเปลี่ยนเพิ่ม หรือลดการส่งใบโฆษณาสินค้าให้กับลูกค้ารายใดบ้าง ซึ่งทำการจัดกลุ่มออกเป็น 3 กลุ่ม จากจำนวนข้อมูลทั้งหมด 300 เรคอร์ด ประกอบด้วยฟิลด์ต่าง ๆ ดังนี้

- Married หมายถึง สถานภาพของลูกค้า
- Salary หมายถึง ผลรวมเงินเดือนทั้งปีของลูกค้าโดยถ้าลูกค้าสมรสแล้ว จะรวมเงินเดือนของทั้งลูกค้าและคู่สมรสด้วย
- History หมายถึง ระดับการซื้อสินค้าและบริการของลูกค้าในปีที่ผ่านมา
- AmountSpent หมายถึง ผลรวมจากการซื้อสินค้าและบริการของลูกค้าในปีนี้
- Catalogs หมายถึง จำนวนใบโฆษณาสินค้าที่บริษัทส่งถึงลูกค้าตลอดปี

จากการจัดกลุ่มโดยใช้อัลกอริทึม K-Prototypes ได้ผลลัพธ์ดังนี้

<u>กลุ่มที่ 1</u> มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 68 เรคอร์ด โดยลักษณะทั่วไปของกลุ่มนี้คือ ผลรวมเงินเดือนของลูกค้าตลอดทั้งปีเท่ากับ 97,180.88 ดอลลาร์สหรัฐ, จำนวนใบโฆษณาสินค้าที่บริษัทส่งถึงลูกค้าตลอดปีเท่ากับ 14.7 ครั้ง, ผลรวมจากการซื้อสินค้าและบริการของลูกค้าในปีนี้ 1,916.33 ดอลลาร์สหรัฐ, สถานภาพของลูกค้าคือ สมรสแล้ว, ลูกค้าไม่เคยซื้อสินค้าและบริการของบริษัทในปีที่ผ่านมา

<u>กลุ่มที่ 2</u> มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 97 เรคอร์ด โดยลักษณะทั่วไปของกลุ่มนี้คือ ผลรวมเงินเดือนของลูกค้าตลอดทั้งปีเท่ากับ 22,870.10 ดอลลาร์สหรัฐ, จำนวนใบโฆษณาสินค้าที่บริษัทส่งถึงลูกค้าตลอดปีเท่ากับ 14.2 ครั้ง, ผลรวมจากการซื้อสินค้าและบริการของลูกค้าในปีนี้เท่ากับ 470.36 ดอลลาร์สหรัฐ, สถานภาพของลูกค้าคือ อื่น ๆ , ลูกค้าเคยซื้อสินค้าและบริการของบริษัทในปีที่ผ่านมาซึ่งอยู่ในระดับต่ำ

<u>กลุ่มที่ 3</u> มีจำนวนข้อมูลทั้งหมดในกลุ่มนี้เท่ากับ 135 เรคอร์ด โดยลักษณะทั่วไปของกลุ่มนี้คือ ผลรวมเงินเดือนของลูกค้าตลอดทั้งปีเท่ากับ 56,923.80 ดอลลาร์สหรัฐ, จำนวนใบโฆษณาสินค้าที่บริษัทส่งถึงลูกค้าตลอดปีเท่ากับ 14.6 ครั้ง, ผลรวมจากการซื้อสินค้าและบริการของลูกค้าในปีนี้เท่ากับ 1,257.37 ดอลลาร์สหรัฐ, สถานภาพของลูกค้าคือ อื่น ๆ , ลูกค้าไม่เคยซื้อสินค้าและบริการของบริษัทในปีที่ผ่านมา

ผลจากการจัดกลุ่มทั้งสามกลุ่มดังที่กล่าวมาแล้ว จะพบว่ากลุ่มที่ 1 และ กลุ่มที่ 3 ได้รับใบโฆษณาสินค้าในจำนวนต่างกันเพียงเล็กน้อยและ ไม่เคยซื้อสินค้าและบริการของบริษัทมาก่อน ส่วนกลุ่มที่ 2 มีการซื้อสินค้าและบริการไม่มากนัก และมีประวัติการซื้อสินค้าและบริการอยู่ในระดับต่ำ กลุ่มที่ 1 มีการซื้อสินค้าและบริการมากที่สุด รองลงมาคือกลุ่มที่ 3 และกลุ่มที่ 2 ตามลำดับ ดังนั้นจากการจัดกลุ่ม สามารถนำผลลัพธ์เหล่านี้มาพิจารณาเพื่อปรับเพิ่มลดการส่งใบโฆษณาสินค้าให้แก่ลูกค้าที่คาดว่าจะซื้อสินค้าและบริการจริง ๆ ได้มากขึ้น ผลลัพธ์ที่ได้มานี้สามารถนำไปเป็นส่วนในการสนับสนุนการตัดสินใจของบริหารต่อไป

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

5.1 สรุปผลการศึกษา

จากการที่ได้ศึกษาทฤษฎีของดาต้าไมนิ่ง ทำให้ทราบได้ว่าดาต้าไมนิ่งเป็นกระบวนการที่ค้นหาข้อมูลที่เป็นประโยชน์จากฐานข้อมูลที่มีอยู่ ทำให้ได้รับสารสนเทศที่เป็นประโยชน์ และสามารถนำสารสนเทศนั้นไปช่วยสนับสนุนการตัดสินใจและการประยุกต์นำไปใช้งานกับธุรกิจต่างๆได้ การวิเคราะห์หาข้อมูลจำเพาะในจุดที่สินค้าหรือบริการขององค์กรหนึ่งมีความโดดเด่นเหนือ กว่าสินค้าและบริการอย่างเดียวกันที่มีอยู่ในตลาด ซึ่งข้อมูลที่มีอยู่ในระบบนั้นเปรียบเสมือนกับเหมืองขนาดใหญ่ที่จำเป็นที่จะต้องเสาะแสวงหาสินแร่ที่มีค่าให้พบ นั่นคือการหารูปแบบ นิสัย หรือรสนิยม และความชอบในการเลือกซื้อสินค้าของลูกค้า เพื่อเข้าใจถึงความต้องการเฉพาะของกลุ่มลูกค้าแต่ละกลุ่ม ซึ่งกระบวนการดังกล่าวจะเริ่มตั้งแต่ การกำหนดวัตถุประสงค์ของการทำดาต้าไมนิ่ง จากนั้นก็มีขั้นตอนการเตรียมข้อมูลมาวิเคราะห์ ซึ่งจะประกอบไปด้วยการคัดเลือกข้อมูล การทำความสะอาดข้อมูล และการแปลงข้อมูลให้เหมาะสม และการทำไมนิ่ง ผลลัพธ์ที่ได้จะก่อให้เกิดประโยชน์ในทางธุรกิจมากมายทีเดียว

เมื่อได้ศึกษาหลักการดังกล่าวแล้ว จึงได้มีแนวทางในการพัฒนาเครื่องมือในการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes ซึ่งเป็นดาต้าไมนิ่งในการจัดกลุ่มข้อมูลประเภทหนึ่งที่สามารถรองรับได้ทั้งข้อมูลประเภท Categorical และ Numerical และทำงานกับฐานข้อมูลที่เป็น Microsoft Access ได้

5.2 ข้อเสนอแนะ

1. ระบบที่พัฒนาขึ้นสามารถใช้ได้กับฐานข้อมูล Microsoft Access เท่านั้นไม่สามารถทำการวิเคราะห์กับฐานข้อมูลอื่น ๆ นอกเหนือจากนี้ได้ และสามารถเลือกได้เพียงตารางเดียวเท่านั้น

2. ระบบที่พัฒนาขึ้นอาจใช้เวลาในการประมวลผลเพื่อแสดงผลลัพธ์ต่างกัน โดยขึ้นอยู่กับสภาพแวดล้อมของหน่วยประมวลผล, จำนวนข้อมูลและจำนวนการเปลี่ยนกลุ่มของข้อมูล เป็นต้น

5.3 การประยุกต์ทดลองใช้งานโปรแกรมในรูปแบบต่าง ๆ

ข้อมูลต่อไปนี้เป็นการทดสอบการใช้งานระบบที่พัฒนาขึ้นในรูปแบบต่าง ๆ กัน โดยทำการเปรียบเทียบผลการจัดกลุ่มข้อมูลที่ได้จากระบบที่พัฒนาขึ้นกับผลการจัดกลุ่มข้อมูลโดยมนุษย์ เพื่อดูว่าผลที่ได้จากการทำงานของระบบที่พัฒนาขึ้นนั้น ง่ายต่อการทำความเข้าใจโดยผู้ใช้งานระบบหรือไม่ และเปรียบเทียบประสิทธิภาพการทำงานกับข้อมูลประเภทต่าง ๆ โดยการทดสอบการจัดกลุ่มข้อมูลโดยระบบที่พัฒนาขึ้น กับการจัดกลุ่มข้อมูลโดยมนุษย์นั้นแบ่งการทดสอบออกเป็น 3 ส่วนดังนี้

1. ทดสอบการจัดกลุ่มข้อมูลโดยใช้ข้อมูลประเภท Categorical
2. ทดสอบการจัดกลุ่มข้อมูลโดยใช้ข้อมูลประเภท Numerical
3. ทดสอบการจัดกลุ่มข้อมูลโดยใช้ข้อมูลแบบผสม คือ ข้อมูลประเภท Categorical และข้อมูลประเภท Numerical

การทดสอบการจัดกลุ่มข้อมูลจะทำการจัดกลุ่มข้อมูลออกเป็น 2 กลุ่มข้อมูล โดยมีข้อมูลจากตาราง ข้อมูลต่าง ๆ ดังนี้

1. ตาราง Actors มีปริมาณข้อมูลทั้งหมด 66 เรคอร์ด ประกอบด้วย 5 ฟิลด์ ดังนี้

(ที่มา : http://www.duxbury.com/awz/ , 2006)

ตารางที่ 5.1 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Actors

ชื่อฟิลด์	ประเภทของข้อมูล
DomesticGross	Numerical
ForeignGross	Numerical
Salary	Numerical
Name	Categorical
Gender	Categorical

2. ตาราง Bank มีปริมาณข้อมูลทั้งหมด 208 เรคอร์ด ประกอบด้วย 9 ฟิลด์ ดังนี้

(ที่มา : http://www.duxbury.com/awz/ , 2006)

ตารางที่ 5.2 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Bank

ชื่อฟิลด์	ประเภทของข้อมูล
Employee	Numerical
EducLev	Numerical
JobGrade	Numerical
YrHired	Numerical
YrBorn	Numerical
YrsPrior	Numerical
Salary	Numerical
Gender	Categorical
PCJob	Categorical

3. ตาราง SuperMKT มีปริมาณข้อมูล 99 เรคอร์ด ประกอบด้วย 9 ฟิลด์ ดังนี้ (ที่มา : http://www.duxbury.com/awz/ , 2006)

ตารางที่ 5.3 แสดงฟิลด์และประเภทของข้อมูลจากตาราง SuperMKT

ชื่อฟิลด์	ประเภทของข้อมูล
InitialWaiting	Numerical
Arrivals	Numerical
Departures	Numerical
EndWaiting	Numerical
Checkers	Numeraicl
TotalCustomers	Numcerical
Day	Categoarical
TimeInterval	Categorical
StartTime	Categorical

4. ตาราง GolfBall มีปริมาณข้อมูลทั้งหมด 149 เรคอร์ด ประกอบด้วย 3 ฟิลด์ ดังนี้ (ที่มา : http://www.duxbury.com/awz/ , 2006)

ตารางที่ 5.4 แสดงฟิลด์และประเภทของข้อมูลจากตาราง GolfBall

ชื่อฟิลด์	ประเภทของข้อมูล
Yards	Numerical
Brand	Categorical
Temp	Categorical

5. ตาราง Video มีปริมาณข้อมูลทั้งหมด128 เรคอร์ด ประกอบด้วย 8 ฟิลด์ ดังนี้

(ที่มา : http://www.duxbury.com/awz/ , 2006)

ตารางที่ 5.5 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Video

ชื่อฟิลด์	ประเภทของข้อมูล
Customer	Numcerical
Purchases	Numerical
DollarAmt	Numerical
State	Categorical
City	Categorical
Gender	Categorical
FirstChoice	Categorical
SecondChoice	Categorical

- การทดสอบการจัดกลุ่มข้อมูล 2 กลุ่มข้อมูล และใช้ข้อมูลประเภท Categorical จากตาราง ทั้ง 5 ตาราง โดยมีฟิลด์ข้อมูลดังนี้

1. ตาราง Actors มีปริมาณข้อมูลทั้งหมด 66 เรคอร์ด ประกอบด้วยCategorical 2 ฟิลด์ ดังนี้

ตารางที่ 5.6 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical ตาราง Actors

ชื่อฟิลด์	ประเภทของข้อมูล
Name	Categorical
Gender	Categorical

2. ตาราง Bank มีปริมาณข้อมูล 208 เรคอร์ด ประกอบด้วยCategorical 2 ฟิลด์ ดังนี้

ตารางที่ 5.7 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง Bank

ชื่อฟิลด์	ประเภทของข้อมูล
Gender	Categorical
PCJob	Categorical

3. ตาราง SuperMKT มีปริมาณข้อมูลทั้งหมด 99 เรคอร์ด ประกอบด้วยCategorical 3 ฟิลด์ ดังนี้

ตารางที่ 5.8 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง SuperMKT

ชื่อฟิลด์	ประเภทของข้อมูล
Day	Categorical
TimeInterval	Categorical
StartTime	Categorical

4. ตาราง GolfBall มีปริมาณข้อมูลทั้งหมด 149 เรคอร์ด ประกอบด้วย Categorical 2 ฟิลด์ ดังนี้

ตารางที่ 5.9 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง GolfBall

ชื่อฟิลด์	ประเภทของข้อมูล
Brand	Categorical
Temp	Categorical

5. ตาราง Video มีปริมาณข้อมูลทั้งหมด128 เรคอร์ด ประกอบด้วย Categorical 5 ฟิลด์ ดังนี้

ตารางที่ 5.10 แสดงฟิลด์ที่เป็นข้อมูลประเภทCategorical จากตาราง Video

ชื่อฟิลด์	ประเภทของข้อมูล
State	Categorical
City	Categorical
Gender	Categorical
FirstChoice	Categorical
SecondChoice	Categorical

ผลการทดสอบการจัดกลุ่มข้อมูลประเภท Categorical ดังตารางที่ 5.11

ตารางที่ 5.11 แสดงผลการจัดกลุ่มข้อมูลประเภทCategorical

ตาราง	ประเภทของข้อมูล (ฟิลด์)		เวลาที่ใช้ในการประมวลผล (วินาที)	สมาชิกในกลุ่มข้อมูล			
				ระบบ		มนุษย์	
	Categorical	Numerical		1	2	1	2
Actors	2	-	2.8281	18	48	16	50
Bank	2	-	18.9218	68	140	68	140
SuperMKT	3	-	12.3750	29	70	43	56
GolfBall	2	-	13.3007	47	102	61	88
Video	5	-	37.7031	48	80	42	86

จากตารางที่ 5.11 สามารถที่จะเปรียบเทียบค่าที่ตรงกันของข้อมูลประเภท Categorical ที่ทำการจัดกลุ่มโดยระบบและมนุษย์ ดังตารางที่ 5.12

ตารางที่ 5.12 แสดงค่าที่ match กันของข้อมูลประเภท Categorical

ตาราง	ค่าที่ match กันในการจัดกลุ่มข้อมูลโดยระบบกับมนุษย์ (record)		เปอร์เซ็นต์(%) ทั้งหมดของค่าที่ match กันในการจัดกลุ่มข้อมูล
	ข้อมูลกลุ่มที่1	ข้อมูลกลุ่มที่2	
Actors	6	37	65.10
Bank	68	140	100
SuperMKT	20	47	67.67
GolfBall	41	82	82.55
Video	35	73	84.37

จากตารางที่ 5.12 ค่าที่ match กันของข้อมูลที่นำมาทำการจัดกลุ่มนั้นจะมีค่าเฉลี่ยอยู่ที่ 73.87% โดยค่าที่ต่ำกว่าค่าเฉลี่ยมีอยู่ 2 ตารางคือ Actors มีค่าเท่ากับ 65.10% และ SuperMKT มีค่าเท่ากับ 67.67 % เนื่องจากค่าของฟิลด์บางฟิลด์ที่อยู่ในตารางทั้งสองนั้นเป็นค่าที่เป็นเอกลักษณ์ (Unique) เช่น ชื่อและนามสกุล เป็นต้น ข้อมูลแต่ละข้อมูลไม่สามารถบ่งบอกได้ว่าน่าจะอยู่ในกลุ่มใดมากกว่ากัน

- การทดสอบการจัดกลุ่มข้อมูล 2 กลุ่มข้อมูล และใช้ข้อมูลประเภท Numerical จากตาราง ทั้ง 5 ตารางข้างต้น โดยมีฟิลด์ข้อมูลดังนี้

1. ตาราง Actors มีปริมาณข้อมูลทั้งหมด 66 เรคอร์ด ประกอบด้วย 3 ฟิลด์ ดังนี้

ตารางที่ 5.13 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Actors

ชื่อฟิลด์	ประเภทของข้อมูล
DomesticGross	Numerical
ForeignGross	Numerical
Salary	Numerical

2. ตาราง Bank มีปริมาณข้อมูลทั้งหมด 208 เรคอร์ด ประกอบด้วย 7 ฟิลด์ ดังนี้

ตารางที่ 5.14 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Bank

ชื่อฟิลด์	ประเภทของข้อมูล
Employee	Numerical
EducLev	Numerical
JobGrade	Numerical
YrHired	Numerical
YrBorn	Numerical
YrsPrior	Numerical
Salary	Numerical

3. ตาราง SuperMKT มีปริมาณข้อมูล 99 เรคอร์ด ประกอบด้วย 6 ฟิลด์ ดังนี้

ตารางที่ 5.15 แสดงฟิลด์และประเภทของข้อมูลจากตาราง SuperMKT

ชื่อฟิลด์	ประเภทของข้อมูล
InitialWaiting	Numerical
Arrivals	Numerical
Departures	Numerical
EndWaiting	Numerical
Checkers	Numeraicl
TotalCustomers	Numcerical

4. ตาราง GolfBall มีปริมาณข้อมูลทั้งหมด 149 เรคอร์ด ประกอบด้วย 1 ฟิลด์ ดังนี้

ตารางที่ 5.16 แสดงฟิลด์และประเภทของข้อมูลจากตาราง GolfBall

ชื่อฟิลด์	ประเภทของข้อมูล
Yards	Numerical

5. ตาราง Video มีปริมาณข้อมูลทั้งหมด128 เรคอร์ด ประกอบด้วย 3 ฟิลด์ ดังนี้

ตารางที่ 5.17 แสดงฟิลด์และประเภทของข้อมูลจากตาราง Video

ชื่อฟิลด์	ประเภทของข้อมูล
Customer	Numcerical
Purchases	Numerical
DollarAmt	Numerical

ผลการทดสอบการจัดกลุ่มข้อมูลประเภท Numerical ดังตารางที่ 5.18

ตารางที่ 5.18 แสดงผลการจัดกลุ่มข้อมูลประเภทNumerical

ตาราง	ประเภทของข้อมูล (ฟิลด์)		เวลาที่ใช้ในการประมวลผล (วินาที)	สมาชิกในกลุ่มข้อมูล			
				ระบบ		มนุษย์	
	Categorical	Numerical		1	2	1	2
Actors	-	3	8.1796	24	42	31	35
Bank	-	7	122.6758	102	106	82	126
SuperMKT	-	6	24.6171	47	52	53	46
GolfBall	-	1	7.1718	68	81	68	81
Video	-	3	16.3437	49	79	65	63

จากตารางที่ 5.18 สามารถที่จะเปรียบเทียบค่าที่ตรงกันของข้อมูลประเภท Numerical ที่ทำการจัดกลุ่มโดยระบบและมนุษย์ ดังตารางที่ 5.19

ตารางที่ 5.19 แสดงค่าที่ match กันของข้อมูลประเภท Numerical

ตาราง	ค่าที่ match กันในการจัดกลุ่มข้อมูลโดยระบบกับมนุษย์ (record)		เปอร์เซ็นต์(%) ทั้งหมดของค่าที่ match กันในการจัดกลุ่มข้อมูล
	ข้อมูลกลุ่มที่1	ข้อมูลกลุ่มที่2	
Actors	20	31	77.27
Bank	75	100	84.13
SuperMKT	47	46	100
GolfBall	68	81	100
Video	49	63	87.50

จากตารางที่ 5.19 ค่าที่ match กัน ของข้อมูลที่นำมาทำการจัดกลุ่มนั้นจะมีค่าเฉลี่ยอยู่ที่ 89.78% โดยค่าที่ต่ำกว่าค่าเฉลี่ยมีอยู่ 2 ตารางคือ Actors มีค่าเท่ากับ 77.27% และ Bank มีค่าเท่ากับ 84.13% เนื่องจาก ค่าของฟิลด์บางฟิลด์ที่อยู่ในตารางทั้งสองนั้นเป็นค่าที่เป็นเลขทศนิยม ข้อมูลจะมีความละเอียดมาก ทำให้ช่วงในการจัดกลุ่มข้อมูลทำได้ยากขึ้น

- การทดสอบการจัดกลุ่มข้อมูล 2 กลุ่มข้อมูล และใช้ข้อมูลแบบผสมจากตาราง ทั้ง 5 ตาราง คือ ตารางที่5.1- 5.5 ผลการทดสอบการจัดกลุ่มข้อมูลแบบผสม ดังตารางที่ 5.20

ตารางที่ 5.20 แสดงผลการจัดกลุ่มข้อมูลแบบผสม

ตาราง	ประเภทของข้อมูล (ฟิลด์)		เวลาที่ใช้ในการประมวลผล (วินาที)	สมาชิกในกลุ่มข้อมูล			
				ระบบ		มนุษย์	
	Categorical	Numerical		1	2	1	2
Actors	2	3	10.3437	24	42	27	39
Bank	2	7	123.7500	102	106	82	126
SuperMKT	3	6	34.1250	47	52	50	49
GolfBall	2	1	20.5273	68	81	88	61
Video	5	3	34.3437	49	79	53	75

จากตารางที่ 5.20 สามารถที่จะเปรียบเทียบค่าที่ตรงกันของข้อมูลแบบผสม ที่ทำการจัดกลุ่มโดยระบบและมนุษย์ ดังตารางที่ 5.21

ตารางที่ 5.21 แสดงค่าที่ match กันของข้อมูลแบบผสม

ตาราง	ค่าที่ match กันในการจัดกลุ่มข้อมูลโดยระบบกับมนุษย์ (record)		เปอร์เซ็นต์(%) ทั้งหมดของค่าที่ match กันในการจัดกลุ่มข้อมูล
	ข้อมูลกลุ่มที่1	ข้อมูลกลุ่มที่2	
Actors	16	31	71.21
Bank	76	101	85.09
SuperMKT	47	49	96.96
GolfBall	65	56	81.20
Video	44	70	89.06

จากตารางที่ 5.21 ค่าที่ match กันของข้อมูลที่นำมาทำการจัดกลุ่มนั้นจะมีค่าเฉลี่ยอยู่ที่ 84.70% โดยค่าที่ต่ำกว่าค่าเฉลี่ยมีอยู่ 1 ตารางคือ Actors มีค่าเท่ากับ 71.21% เนื่องจาก ค่าของฟิลด์บางฟิลด์ที่อยู่ในตารางทั้งสองนั้นเป็นค่าที่เป็นเลขทศนิยมและค่าที่เป็นเอกลักษณ์(Unique) เช่น ชื่อและนามสกุล เป็นต้น ทำให้เป็นการยากในการที่จะจัดกลุ่มข้อมูลว่าควรที่จะอยู่ในกลุ่มใด เนื่องจากมีช่วงของข้อมูลที่ค่อนข้างละเอียด

จากตารางที่ 5.11-12 และตารางที่ 5.18-21 เป็นการทดสอบการจัดกลุ่มกับข้อมูลในประเภทต่าง ๆ ซึ่งผลที่ได้จากการทดสอบนั้น สามารถวิเคราะห์ได้ดังนี้

ผลการทดสอบด้วยข้อมูลที่เป็น Categorical นั้นใช้เวลาเฉลี่ยในการประมวลผลน้อยที่สุด เนื่องจากในการทำงานของระบบ จะใช้ K-Prototypes Algorithm ในการจัดกลุ่มข้อมูล โดยจะมีการเปรียบเทียบความเหมือนกันของข้อมูลที่เป็น Categorical ดังนั้นจึงสามารถทำงานได้เร็วกว่าการจัดกลุ่มข้อมูลกับข้อมูลประเภทอื่น ๆ

ผลการทดสอบด้วยข้อมูลที่เป็นNumerical นั้นให้ผลที่ match กับการจัดกลุ่มข้อมูลโดยมนุษย์มากที่สุด เนื่องจากการจัดกลุ่มข้อมูลโดยมนุษย์นั้นสามารถทำได้ง่ายกว่าการจัดกลุ่มข้อมูลที่เป็นแบบ Categorical เนื่องจากค่าที่เป็น Numerical นั้น จะมีช่วงของข้อมูล ทำให้ง่ายต่อการจัดกลุ่มข้อมูล

ผลการทดสอบด้วยข้อมูลแบบผสมนั้น ให้ผลที่เป็นค่ากลางระหว่างการทดสอบการจัดกลุ่มข้อมูลประเภท Categorical และNumerical เช่น เวลาในการประมวลผลข้อมูล, ค่าที่ match กันของข้อมูลที่ได้จากการจัดกลุ่มโดยระบบและมนุษย์ เป็นต้น ค่าเฉลี่ยของผลที่ได้จากการเปรียบเทียบผลของการจัดกลุ่มข้อมูลที่มีค่าตรงกันนั้น มีค่าเฉลี่ยของผลทีมีค่าตรงกันมากกว่า 80 เปอร์เซ็นต์ของผล

ที่ได้ทั้งหมด ดังนั้น ทำให้ทราบว่าผู้ใช้ระบบที่พัฒนาขึ้นสามารถที่จะเข้าใจผลที่ได้จากการจัดกลุ่มข้อมูลนั้น ๆ ได้โดยง่าย เนื่องจากผลที่ได้ใกล้เคียงกับความเป็นเหตุเป็นผลที่มนุษย์ใช้แยกแยะความเหมือนกันของกลุ่มข้อมูล

บรรณานุกรม

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์. 2544. "Micro Computer" **การใช้เทคนิคดาต้าไมน์นิ่งเพื่อพัฒนาคุณภาพการศึกษา คณะวิศวกรรมศาสตร์** . Vol.167

Albright, Winston and Zappe. 2003. **Data Analysis for Managers. [CD-ROM]**. Duxbury.

Daniel T. Larose. 2005. **Discovering Knowledge in data : An Introduction to Data Mining.** New Jersy : John Wiley & Sons, Inc.

Jain A.K., Murty M.N., and Flynn P.J. 1999. "Data Clustering : A Review." **ACM Computing Survey**. Vol.31 ,No.3.

Jiawei Han and Micheline Kamber . 2001. **Data Mining Concepts And Techniques**. Morgan Kaufmann Publisher.

Peter Cabena, [e.d.]. 1998. **Discovering Data Mining : From Concept to Implementation.** New Jersy : Prentice Hall PTR.

Richard J. Roiger and Micheal W. Geatz, 2003. **Data Mining a Tutorial – Based Primer.** United State of America : Addison Wesley.

Zhexue Huang. 1998. **Clustering large data sets with mixed numeric and categorical values.** Australia : CSIRO Mathematical and Information Sciences.

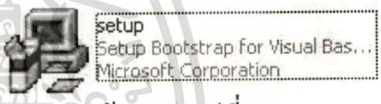
ภาคผนวก

ก.1 การติดตั้งโปรแกรม

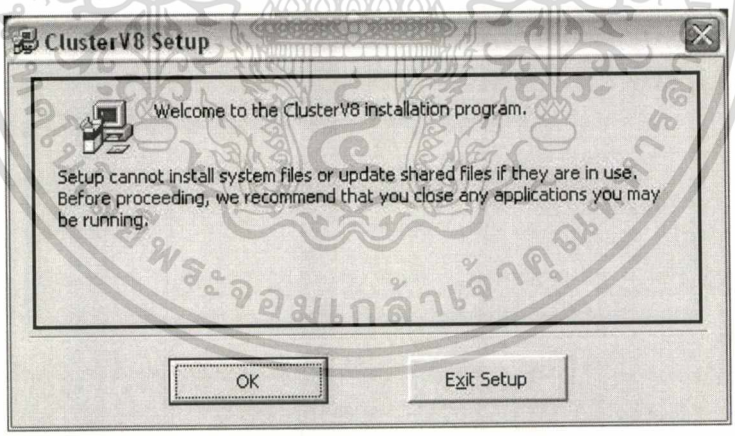
สำหรับการติดตั้งโปรแกรมเครื่องมือในการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Prototypes มีความต้องการของระบบขั้นต่ำที่มีรายละเอียดดังนี้

- ระบบปฏิบัติการ Windows 98/ 2000/ ME/ XP
- เครื่องคอมพิวเตอร์ที่ใช้หน่วยประมวลผลระดับ Pentium2 400 MHz เป็นอย่างน้อยขึ้นอยู่กับขนาดของฐานข้อมูล
- หน่วยความจำ 256 MBขึ้นไป
- พื้นที่ว่างอย่างน้อย 50 MB เพื่อเพียงพอสำหรับการประมวลผลของโปรแกรม
- ติดตั้ง Microsoft Access 2000 และ Microsoft Excel 2000

สำหรับการติดตั้งโปรแกรมให้ใส่แผ่น CD ติดตั้งโปรแกรม

1. เริ่มการติดตั้งโปรแกรมโดยทำการดับเบิ้ลคลิกที่ไฟล์ 

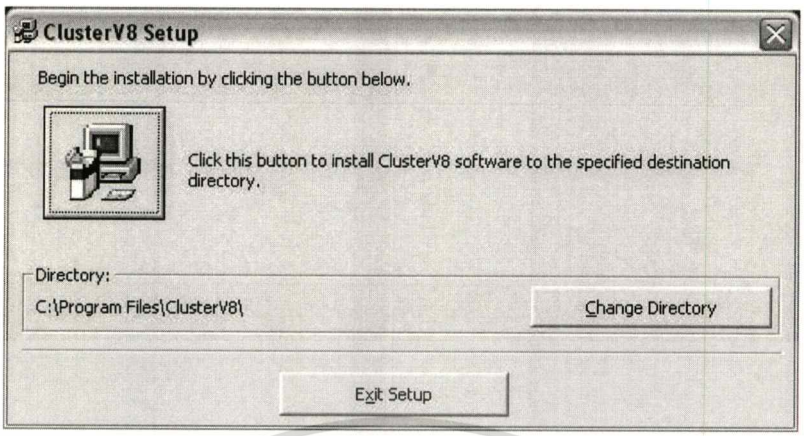
ที่อยู่ในไดร์ฟ CD-ROM เพื่อใช้ในการติดตั้งโปรแกรม จะปรากฏหน้าจอดังรูปที่ ก.1



รูปที่ ก.1 หน้าจอแรกเมื่อเข้าสู่การติดตั้งโปรแกรม

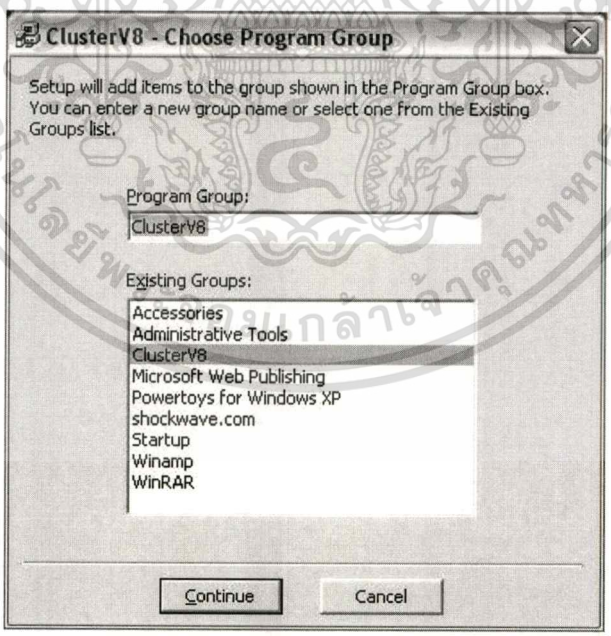
2. หลังจากคลิกที่ปุ่ม 

จะเข้าสู่หน้าจอการติดตั้งโปรแกรม ดังแสดงในรูปที่ ก.2



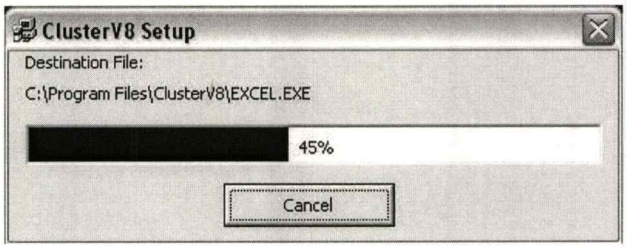
รูปที่ ก.2 หน้าจอที่สองของการติดตั้งโปรแกรม

3. เมื่อคลิกปุ่ม  ระบบจะทำการติดตั้งโดยจะเพิ่มกลุ่มของโปรแกรม ในProgram group box ดังแสดงในรูปที่ ก.3



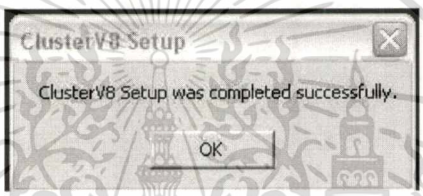
รูปที่ ก.3 หน้าจอแสดงกระบวนการติดตั้งระบบ

4. เมื่อคลิกที่ปุ่ม  ระบบจะทำการติดตั้งโปรแกรมดังรูปที่ ก.4



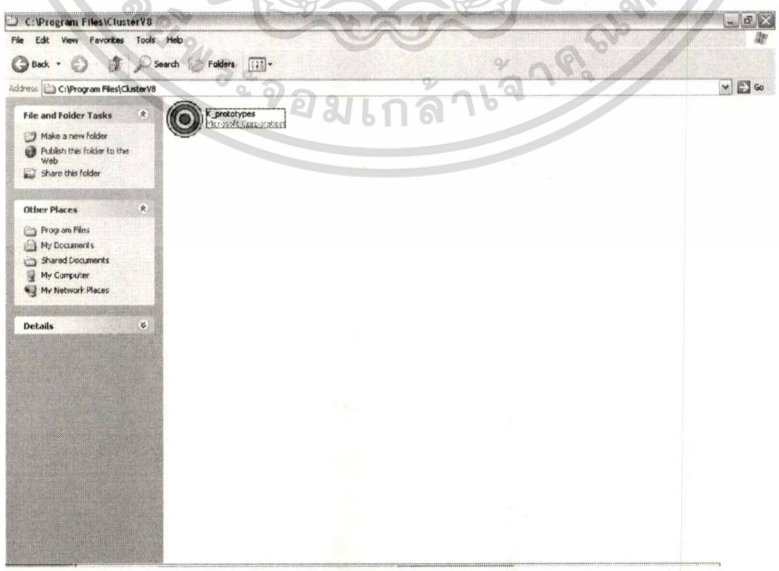
รูปที่ ก.4 หน้าจอแสดงความก้าวหน้าในการติดตั้งโปรแกรม

5. เมื่อโปรแกรมทำการติดตั้งเสร็จเรียบร้อยแล้วจะแสดงหน้าจอดังรูปที่ ก.5



รูปที่ ก.5 แสดงหน้าจอการติดตั้งโปรแกรมเสร็จสมบูรณ์

โปรแกรมจะถูกติดตั้งไว้ในไดเรกทอรี่ C:\Program Files\ClusterV8 ดังแสดงในรูปที่ ก.6



รูปที่ ก.6 แสดงไดเรกทอรี่ที่โปรแกรมได้ถูกติดตั้ง

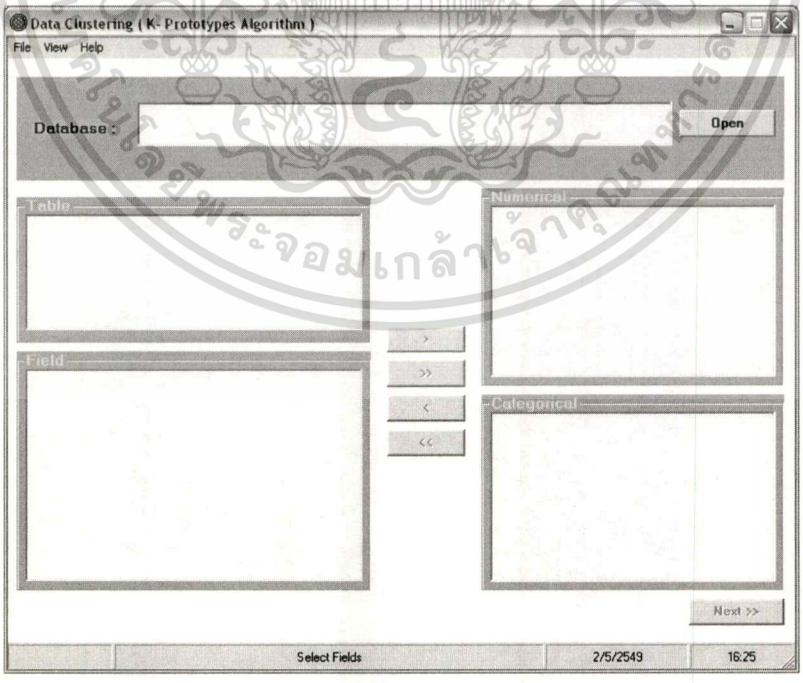
ก.2 การเข้าสู่โปรแกรม

การเข้าสู่โปรแกรม Data Clustering(K-Prototypes Algorithm) Program สามารถทำได้โดยคลิกเมาส์ที่ Start Menu เข้าไปยัง All programs จะพบกับโฟลเดอร์ ClusterV8 ดังแสดงในรูปที่ ก.7

รูปที่ ก.7 การเข้าสู่โปรแกรมโดยเลือกจาก Start Menu

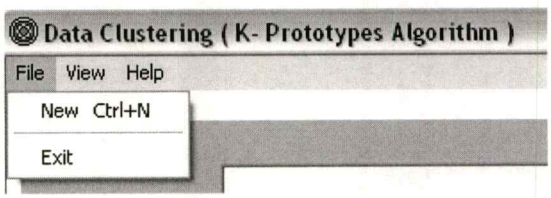
ก.3 การใช้งานโปรแกรม

เมื่อเข้าสู่โปรแกรมแล้วจะพบกับหน้าจอเริ่มต้นของโปรแกรมดังแสดงรูปที่ ก.8



รูปที่ ก.8 หน้าจอเริ่มต้นของโปรแกรม

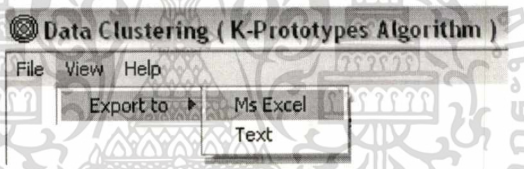
เลือกที่เมนู File จะปรากฏเมนูย่อยดังแสดงในรูปที่ ก.9



รูปที่ ก.9 หน้าจอแสดงผล เมนู File

- New Program – ทำหน้าที่ในการเริ่มโปรแกรมใหม่ (Restart)
- Exit – ทำการออกจากโปรแกรม

เลือกที่เมนู View จะปรากฏเมนูย่อยดังแสดงในรูปที่ ก.10

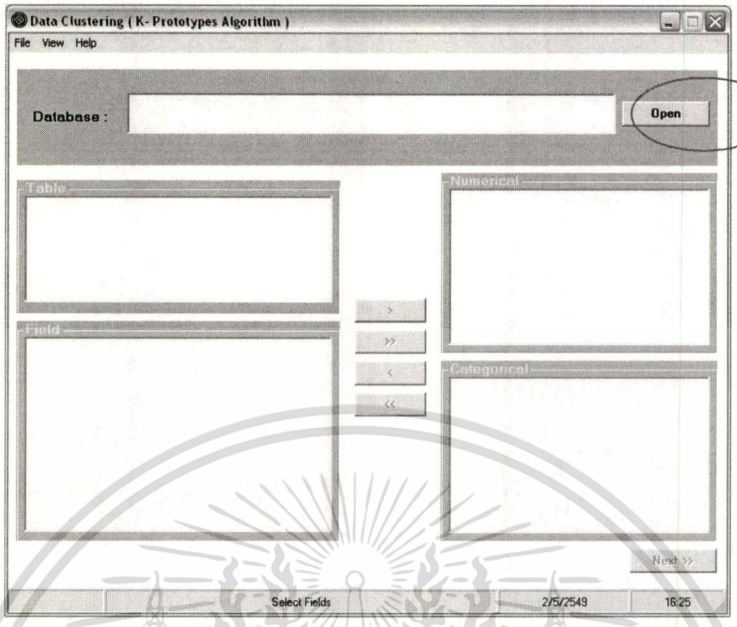


รูปที่ ก.10 หน้าจอแสดงผล เมนู View

- Export Excel – ทำหน้าที่ในการบันทึกผลการจัดกลุ่มข้อมูลออกมาเป็นไฟล์ประเภท Excel (.xls)
- Export Text – ทำหน้าที่ในการบันทึกผลการจัดกลุ่มข้อมูลออกมาเป็นไฟล์ประเภท Text (.txt)

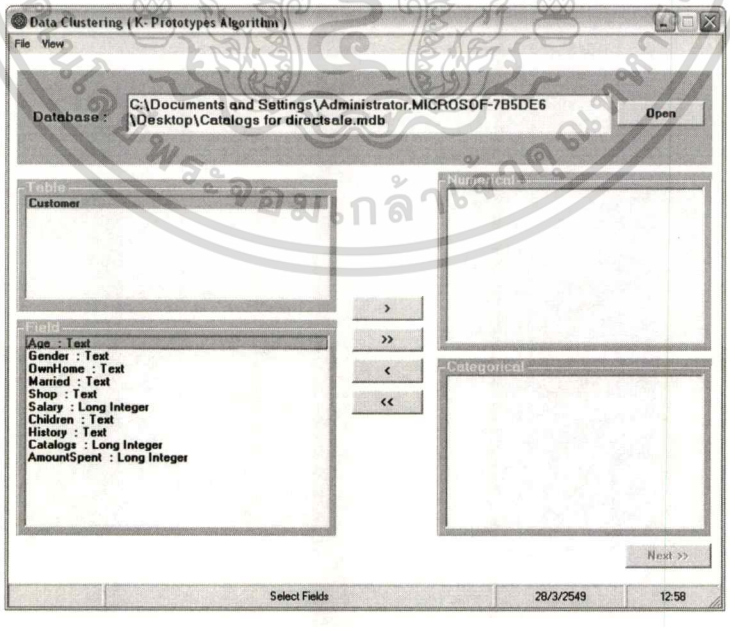
ก.4 การเลือกฐานข้อมูล (Data Selection)

ขั้นที่ 1 คลิกที่ปุ่ม Open เพื่อทำการเลือกฐานข้อมูล ดังแสดงในรูปที่ ก.11



รูปที่ ก.11 หน้าจอเลือกฐานข้อมูล

ขั้นที่ 2 หลักจากเลือกฐานข้อมูลเรียบร้อยแล้ว จะปรากฏหน้าจอให้เลือกตารางและฟิลด์ ดังรูปที่ ก.12



รูปที่ ก.12 หน้าจอเลือกตารางและฟิลด์

1) เลือกตารางที่ต้องการ

2) เลือกฟิลด์ข้อมูลจาก listbox จากหัวข้อเลือกฟิลด์ แล้วคลิกที่ปุ่มดังนี้

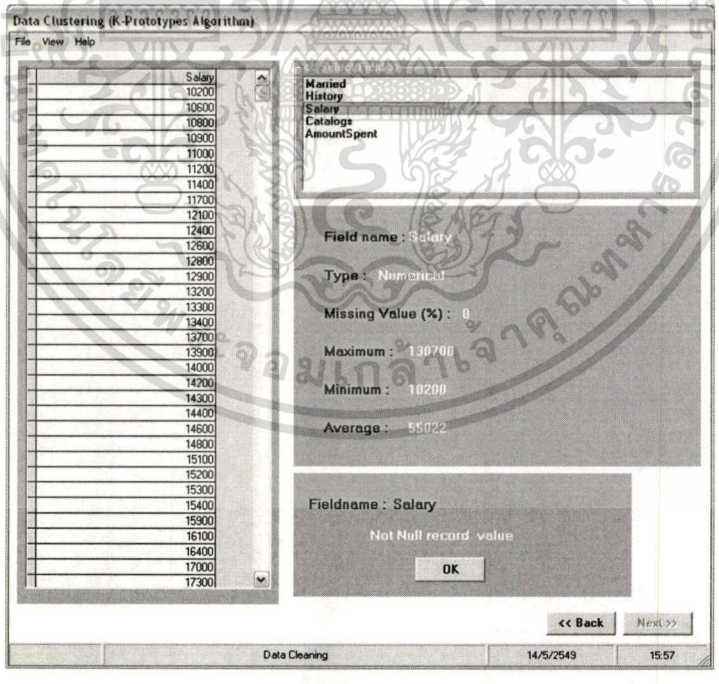
- [>] เพื่อทำการเลือกฟิลด์ข้อมูลที่ต้องการ
- [>>] เพื่อทำการเลือกฟิลด์ข้อมูลทั้งหมด
- [<] เพื่อทำการย้ายฟิลด์ข้อมูลกลับ
- [<<] เพื่อทำการย้ายฟิลด์ข้อมูลกลับทั้งหมด

3) คลิกที่ปุ่ม [Next >>] เพื่อทำขั้นตอนถัดไป

ก.5 การแก้ไขข้อมูล (Data Cleaning)

การแก้ไขข้อมูล ให้เลือกฟิลด์ที่ list box ซึ่งเมื่อเลือกแล้วหน้าจอจะแสดงผลต่างกัน ซึ่งมีทั้ง 4 กรณี ดังนี้

<u>กรณีที่ 1</u> ฟิลด์ที่เป็น Numerical ที่ไม่มีค่า Null ดังแสดงในรูปที่ ก.13



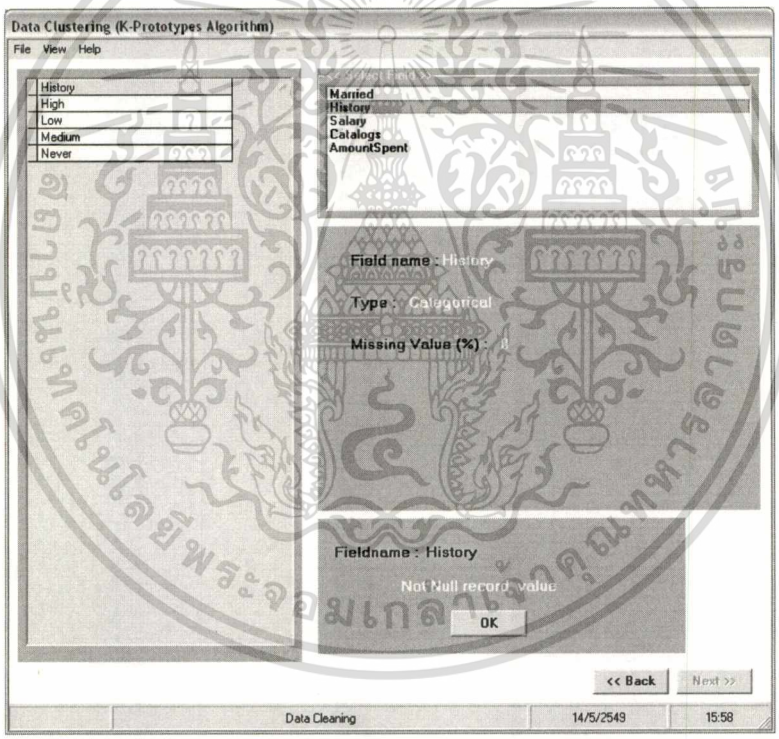
รูปที่ ก.13 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า Null

สำหรับฟิลด์ที่เป็น Numerical ที่ไม่มีค่า Null จะมีข้อมูลแสดงรายละเอียดต่างๆ ในหน้าจอดังนี้

- ส่วนแสดงตัวอย่างข้อมูลในตาราง
- ชื่อฟิลด์ - เป็นส่วนที่แสดงชื่อฟิลด์
- ชนิดข้อมูล - เป็นส่วนที่แสดงชนิดข้อมูล
- จำนวนค่าว่าง
- ค่า Minimum , ค่า Maximum , ค่า Average

ให้คลิกที่ปุ่ม OK หากฟิลด์นี้เป็นฟิลด์สุดท้ายในการแก้ไขข้อมูล ให้คลิกปุ่ม Next >> เพื่อทำขั้นตอนถัดไป หากไม่ใช่ฟิลด์สุดท้ายให้ดำเนินการกับฟิลด์อื่นๆแล้วแต่กรณี

กรณีที่ 2 ฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null ดังแสดงในรูปที่ ก.14



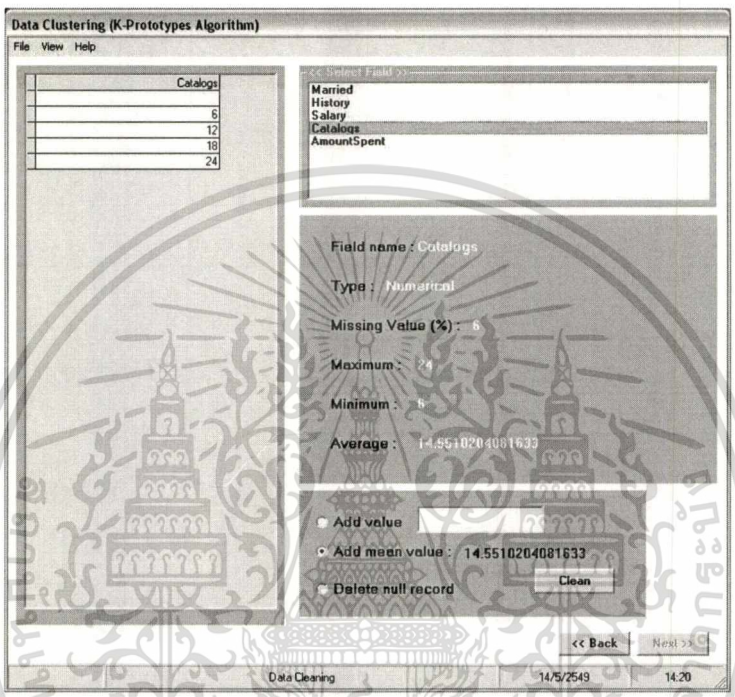
รูปที่ ก.14 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null

สำหรับฟิลด์ที่เป็น Categorical ที่ไม่มีค่า Null จะมีข้อมูลแสดงรายละเอียดต่างๆ ในหน้าจอดังนี้

- ส่วนแสดงตัวอย่างข้อมูลในตาราง
- ชื่อฟิลด์ - เป็นส่วนที่แสดงชื่อฟิลด์
- ชนิดข้อมูล - เป็นส่วนที่แสดงชนิดข้อมูล

ให้คลิกที่ปุ่ม OK หากฟิลด์นี้เป็นฟิลด์สุดท้ายในการแก้ไขข้อมูล ให้คลิกปุ่ม Next >> เพื่อทำขั้นตอนถัดไป หากไม่ใช่ฟิลด์สุดท้ายให้ดำเนินการกับฟิลด์อื่นๆแล้วแต่กรณี

<u>กรณีที่ 3</u> ฟิลด์ที่เป็น Numerical ที่มีค่า Null ดังแสดงในรูปที่ ก.15



รูปที่ ก.15 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Numerical ที่มีค่า Null

สำหรับฟิลด์ที่เป็น Numerical ที่มีค่า Null จะมีข้อมูลแสดงรายละเอียดต่างๆในหน้าจอดังนี้

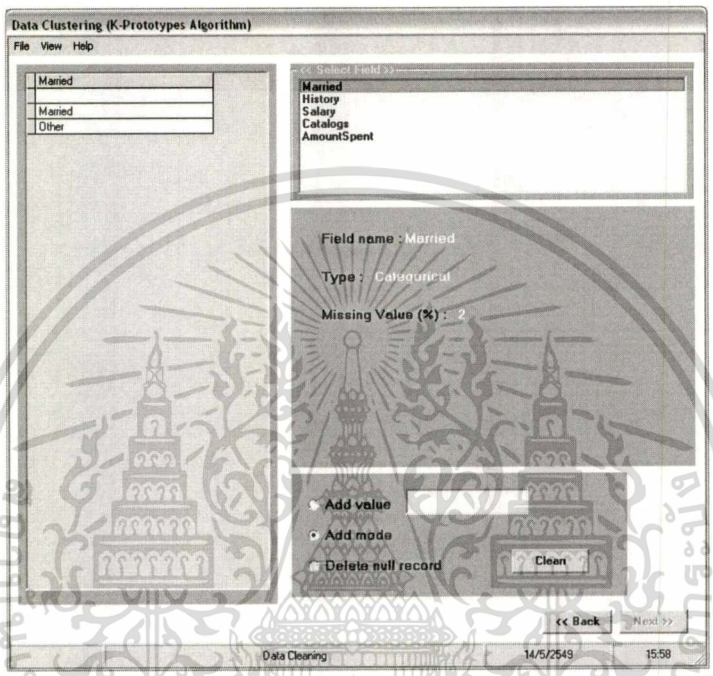
- ส่วนแสดงตัวอย่างข้อมูลในตาราง
- ชื่อฟิลด์ - เป็นส่วนที่แสดงชื่อฟิลด์
- ชนิดข้อมูล - เป็นส่วนที่แสดงชนิดข้อมูล
- จำนวนค่าว่าง
- ค่า Minimum , ค่า Maximum , ค่า Average

และมีตัวเลือกในการแก้ไขข้อมูลดังนี้

- ใส่ค่า Mean (ค่าเฉลี่ย)โดยจะใส่แทนข้อมูลที่เป็นค่า null
- ลบเรคอร์ดที่มีค่า null

ให้คลิกที่ปุ่ม Clean หากฟิลด์นี้เป็นฟิลด์สุดท้ายในการแก้ไขข้อมูล ให้คลิกปุ่ม Next >> เพื่อทำขั้นตอนถัดไป หากไม่ใช่ฟิลด์สุดท้ายให้ดำเนินการกับฟิลด์อื่นๆแล้วแต่กรณี

กรณีที่ 4 ฟิลด์ที่เป็น Categorical ที่มีค่า Null ดังแสดงในรูปที่ ก.16



รูปที่ ก.16 หน้าจอแสดงผลการแก้ไขข้อมูลสำหรับฟิลด์ที่เป็น Categorical ที่มีค่า Null

สำหรับฟิลด์ที่เป็น Categorical ที่มีค่า Null จะมีข้อมูลแสดงรายละเอียดต่างๆ ในหน้าจอดังนี้

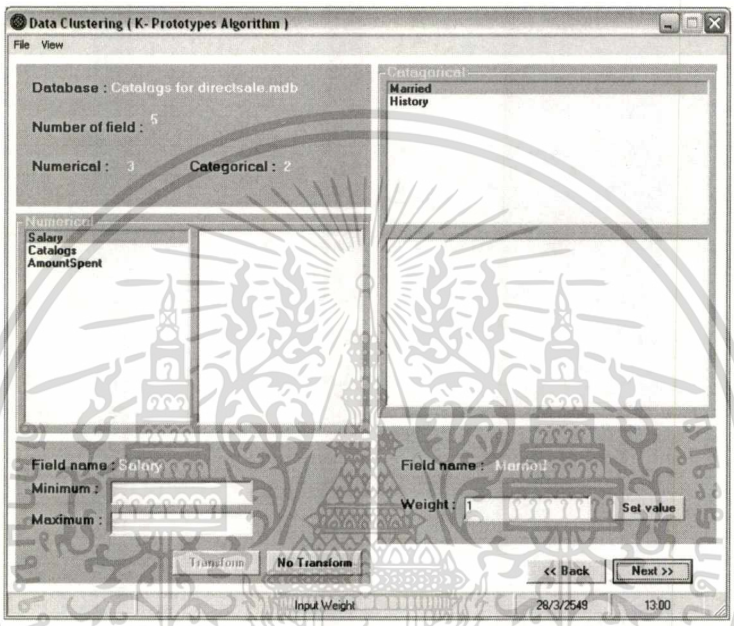
- ส่วนแสดงตัวอย่างข้อมูลในตาราง
- ชื่อฟิลด์ - เป็นส่วนที่แสดงชื่อฟิลด์
- ชนิดข้อมูล - เป็นส่วนที่แสดงชนิดข้อมูล
- จำนวนค่าว่าง

และมีตัวเลือกในการแก้ไขข้อมูลดังนี้

- ใส่ค่าโดยที่ผู้ใช้กำหนดเอง
- ใส่ค่า Mode(ฐานนิยม)โดยจะใส่แทนข้อมูลที่เป็นค่า null
- ลบเรคอร์ดที่มีค่า null

ให้คลิกที่ปุ่ม Clean หากฟิลด์นี้เป็นฟิลด์สุดท้ายในการแก้ไขข้อมูล ให้คลิกปุ่ม Next >> เพื่อทำขั้นตอนถัดไป หากไม่ใช่ฟิลด์สุดท้ายให้ดำเนินการกับฟิลด์อื่นๆแล้วแต่กรณี

ก.6 การปรับเปลี่ยนข้อมูล (Data Transformation)



รูปที่ ก.17 หน้าจอแสดงผลการปรับเปลี่ยนข้อมูล

ขั้นที่ 1 ข้อมูลประเภท Numerical ให้ผู้ใช้เลือกฟิลด์ที่เป็น Numerical ใน list box หลังจากนั้นจะมี 2 กรณีให้ผู้ใช้ได้เลือก ได้แก่

กรณีที่ 1 กำหนดให้มีการปรับเปลี่ยนข้อมูลของฟิลด์ที่เป็น Numerical มีค่าอยู่ในช่วงๆหนึ่ง

กำหนดช่วงในการปรับเปลี่ยนข้อมูลในช่อง Min ใน text box ซึ่งเป็นค่า Numerical ที่เป็นค่าน้อยที่สุดและช่อง Max ใน textbox ซึ่งเป็นค่า Numerical ที่เป็นค่าที่สูงที่สุดในฟิลด์ๆนั้น จากนั้นให้คลิกที่ปุ่ม Transform

กรณีที่ 2 ไม่ต้องการให้มีการปรับเปลี่ยนข้อมูล

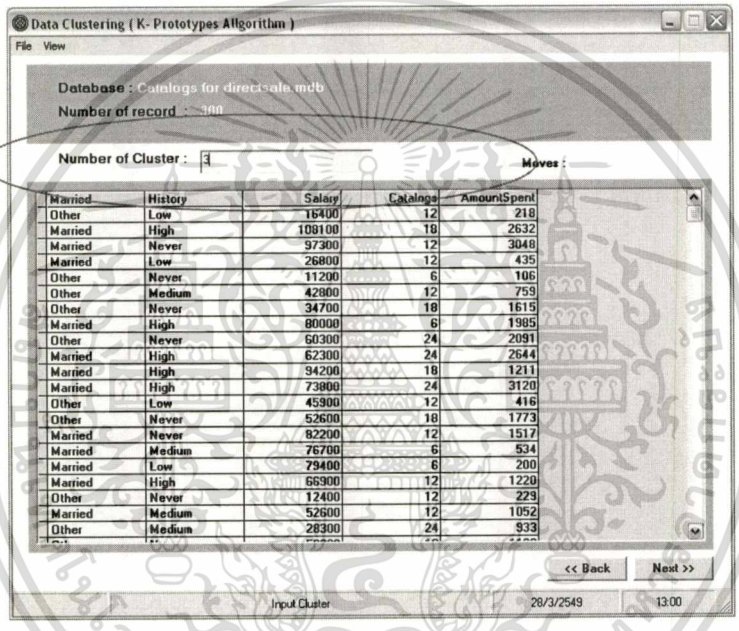
เมื่อไม่ต้องการปรับเปลี่ยนข้อมูล เพื่อให้ข้อมูลที่เป็น Numerical มีค่าเดียวกับค่าที่อยู่ในฐานข้อมูลตั้งแต่เริ่มต้น สามารถทำได้โดยคลิกที่ปุ่ม No Transform

ขั้นที่ 2 ข้อมูลประเภท Categorical ให้ผู้ใช้เลือกฟิลด์ที่เป็น Categorical ใน listbox จากนั้น ให้ใส่ค่าที่ต้องการกำหนด Weight ให้ข้อมูลที่ textbox จากนั้นคลิกที่ปุ่ม Set value

ขั้นที่ 3 เมื่อทำครบทุกฟิลด์แล้วให้คลิกที่ปุ่ม Next >> เพื่อทำในขั้นตอนถัดไป

ก.7 การกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่ม

ขั้นตอนในการกำหนดจำนวนกลุ่มให้ข้อมูลที่ต้องการจัดกลุ่ม ให้ผู้ใช้ใส่จำนวนที่ต้องการลงในช่อง text box แล้วให้คลิกที่ปุ่ม Next >> เพื่อให้ระบบทำการประมวลผลและแสดงผลลัพธ์

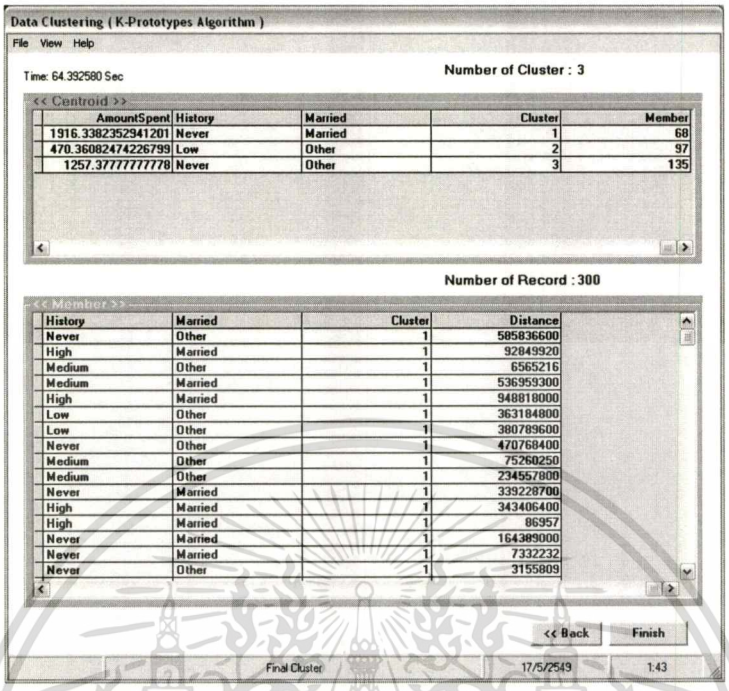


รูปที่ ก.18 หน้าจอแสดงผลการกำหนดจำนวนกลุ่มข้อมูล

ก.6 การแสดงผลการจัดกลุ่มข้อมูล

หลังจากที่ระบบได้ทำการวิเคราะห์และจัดกลุ่มข้อมูลให้เรียบร้อยแล้ว ก็จะมีการแสดงผล (Output) ซึ่งประกอบไปด้วย 2 ส่วนคือ

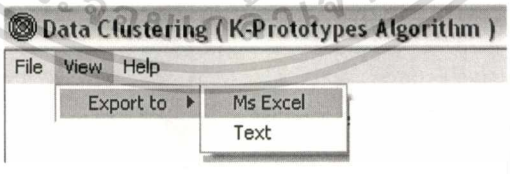
- Cluster Centroid จะแสดงจุดศูนย์กลางของแต่ละกลุ่ม และบอกจำนวนสมาชิกที่อยู่ในกลุ่มนั้นๆ
- Cluster Membership จะแสดงข้อมูลแต่ละเรคอร์ดที่อยู่ในกลุ่มใดๆ และบอกถึงระยะห่างระหว่างสมาชิกกับจุดศูนย์กลางของกลุ่มนั้นๆ



รูปที่ ก.19 หน้าจอแสดงผลการจัดกลุ่มข้อมูล

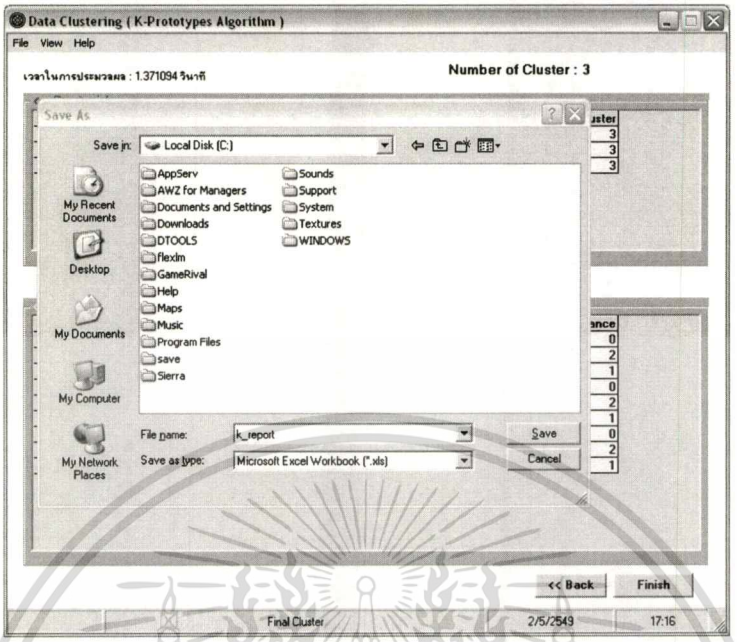
ก.7 การบันทึกผลการจัดกลุ่มข้อมูลออกมาเป็นไฟล์ประเภท Excel

1. เมื่อจัดกลุ่มข้อมูลจนเสร็จสมบูรณ์แล้ว เราสามารถที่จะบันทึกข้อมูลเพื่อสามารถเก็บไว้ดูผลลัพธ์ที่ทำการจัดกลุ่มข้อมูลได้ในรูปแบบของ Excel ทำได้โดย เลือกจากเมนู View > Export to > Ms Excel ดังแสดงในรูปที่ ก.20



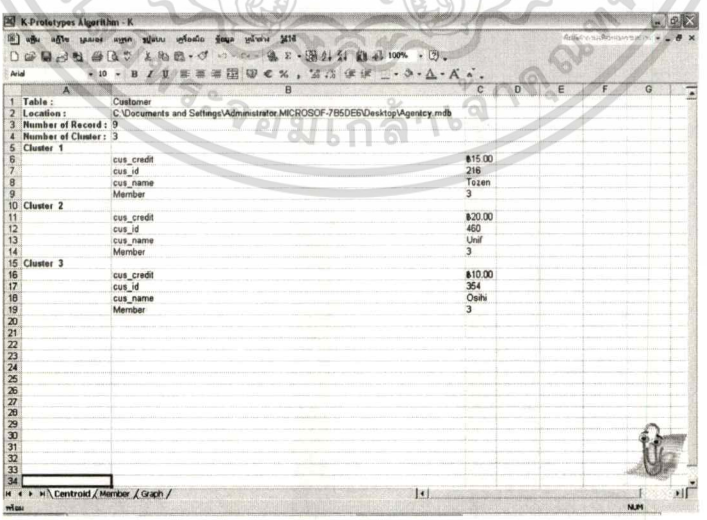
รูปที่ ก.20 หน้าจอแสดงผลโดยเลือกเมนู File > Export to > Ms Excel

2. ระบุตำแหน่งที่ต้องการในการบันทึกข้อมูล ดังแสดงในรูปที่ ก.21



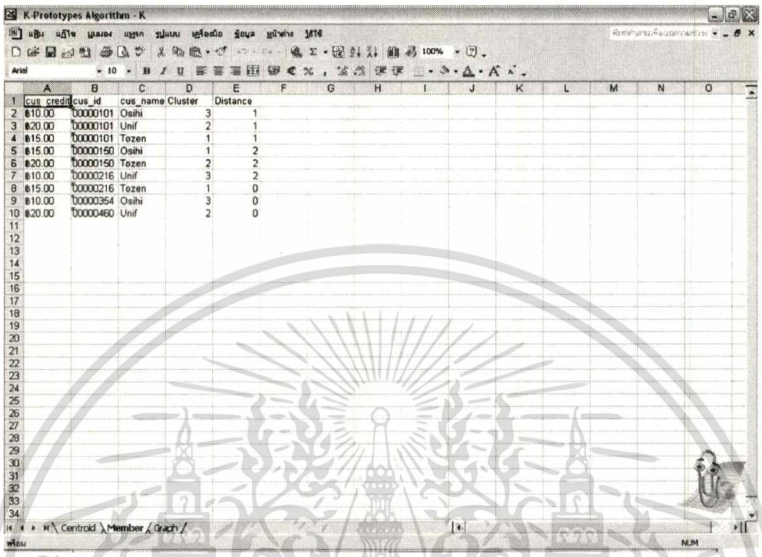
รูปที่ ก.21 หน้าจอแสดงผลการบันทึกข้อมูลแบบ Excel

3. เมื่อทำการบันทึกข้อมูลแล้ว เมื่อเข้าไปจะปรากฏผลจัดกลุ่มข้อมูลในรูปของ Excel ซึ่งเป็นแบ่ง 3 Worksheet ใน Worksheet ชื่อ Centroid จุดศูนย์กลางของกลุ่มข้อมูล ซึ่งได้แสดงข้อมูลที่จัดกลุ่มและ จำนวนข้อมูลที่เป็นสมาชิกภายในกลุ่ม ดังแสดงในรูปที่ ก.22



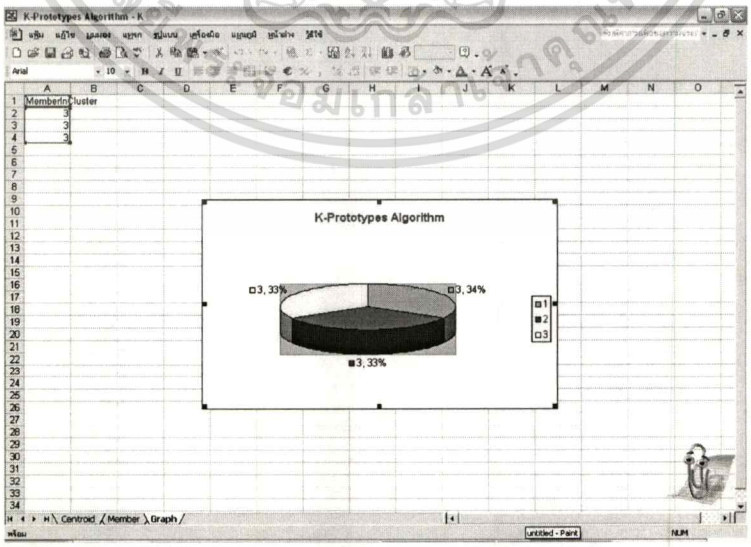
รูปที่ ก.22 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Centroid

Worksheetที่ 2 ชื่อ Member สมาชิกของกลุ่มข้อมูลนั้น ๆ ซึ่งจะแสดงระยะห่างระหว่างตัวสมาชิกกับจุดศูนย์กลางของกลุ่มข้อมูล และแสดงให้เห็นว่าข้อมูลนั้นอยู่ในกลุ่มข้อมูลใด



รูปที่ ก.23 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Member

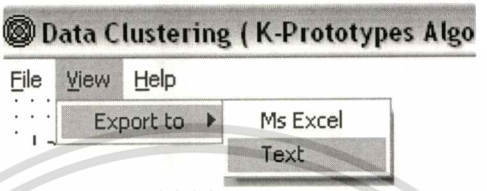
Worksheetที่ 3 ชื่อ Graph จะแสดงกราฟวงกลม 3 มิติ จำนวนสมาชิกในแต่ละกลุ่มข้อมูล เพื่อง่ายต่อความเข้าใจ ดังแสดงในรูปที่ ก.24



รูปที่ ก.24 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Excel จาก Worksheet ชื่อ Graph

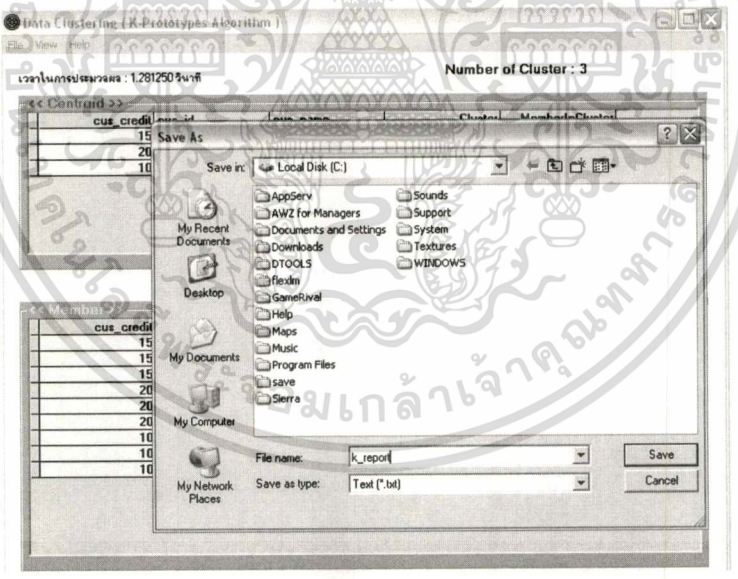
ก.8 การบันทึกผลการจัดกลุ่มข้อมูลออกมาเป็นไฟล์ประเภท Text

1. เมื่อจัดกลุ่มข้อมูลจนเสร็จสมบูรณ์แล้ว เราสามารถที่จะบันทึกข้อมูลเพื่อสามารถเก็บไว้ดูผลลัพธ์ที่ทำการจัดกลุ่มข้อมูลได้ในรูปแบบของ Text File สามารถทำได้โดย เลือกจากเมนู View > Export to > Text ดังแสดงในรูปที่ ก.25



รูปที่ ก.25 หน้าจอแสดงผล โดยเลือกเมนู File > Export To > Text

2. ระบุตำแหน่งที่ต้องการในการบันทึกข้อมูล ดังแสดงในรูปที่ ก.26



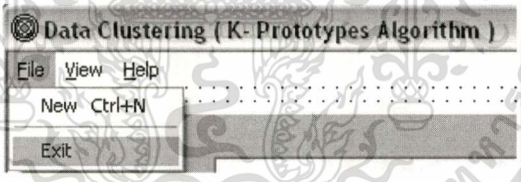
รูปที่ ก.26 หน้าจอแสดงผลการบันทึกข้อมูลแบบ Text

3. เมื่อทำการบันทึกข้อมูลแล้ว เมื่อเข้าไปจะปรากฏผลการจัดกลุ่มข้อมูลในรูปของ Text File ดังแสดงในรูปที่ ก.27

รูปที่ ก.27 หน้าจอแสดงผลข้อมูลเมื่อบันทึกข้อมูลแบบ Text

ก.11 การปิดโปรแกรม

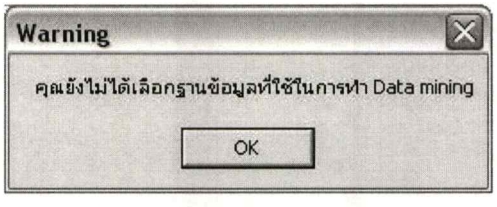
การปิดโปรแกรมสามารถปิดโปรแกรมดังกล่าวได้โดยเลือกจากเมนู File > Exit



รูปที่ ก.28 หน้าจอการปิดโปรแกรมโดยเลือกเมนู File > Exit

ก.12 Message Box

1. หน้าจอการเลือกฐานข้อมูล แสดงเมื่อไม่ได้ทำการเลือกฐานข้อมูล



รูปที่ ก.29 ข้อความเตือนเมื่อไม่ได้ทำการเลือกฐานข้อมูล

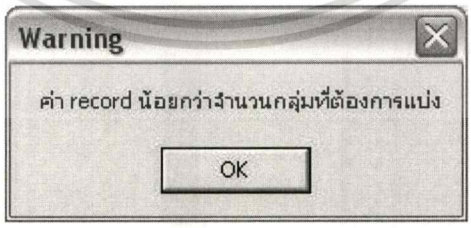
2. หน้าจอการทำความสะอาดข้อมูล เมื่อไม่ได้ใส่ค่าลงในช่องรับค่าสำหรับการทำความสะอาดข้อมูลที่เป็น Categorical

รูปที่ ก.30 ข้อความเตือนเมื่อไม่ได้ค่าลงในช่องรับข้อความ

3. หน้าจอการปรับเปลี่ยนข้อมูล เมื่อใส่ค่า weight ของข้อมูลประเภท Categorical ไม่ถูกต้อง

รูปที่ ก.31 ข้อความเตือนเมื่อใส่ค่า weight ไม่ถูกต้อง

4. หน้าจอการกำหนดกลุ่มข้อมูล โดยใส่ค่าจำนวนกลุ่มที่ต้องการแบ่งมากกว่าค่าจำนวนข้อมูล



รูปที่ ก.32 ข้อความเตือนเมื่อใส่ค่าจำนวนกลุ่มที่ต้องการแบ่งมากกว่าค่าจำนวนข้อมูล

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวสาลินี รุ่งเรือง
วันเดือนปีเกิด	25 ธันวาคม 2523
สถานที่เกิด	กรุงเทพมหานคร
ปริญญาตรี	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ คณะวิทยาศาสตร์และเทคโนโลยี สาขาวิทยาการคอมพิวเตอร์
ปีที่สำเร็จการศึกษา	2545