การบริหารงานลูกค้าสัมพันธ์โดยใช้การทำเหมืองข้อมูล
Customer Relationship Management with Data Mining

สมเพ็ชร จุลลาบุดดี Sompejch Junlabuddee *

บทคัดย่อ

ท่ามกลางการแข่งขันที่สูงขึ้นของธุรกิจและความต้องการของลูกค้าที่เปลี่ยนแปลงอย่างรวดเร็ว การบริหารงานลูกค้าสัมพันธ์(Customer Relationship Management : CRM)ได้เข้ามามีบทบาทสำคัญและมีความจำเป็นอย่างยิ่งต่อระบบธุรกิจในปัจจุบัน ทำให้มีปัญหาการจัดเก็บข้อมูลของลูกค้าเป็นจำนวนมากเพื่อใช้ศึกษาถึงพฤติกรรมของลูกค้า และวิธีการวิเคราะห์ข้อมูลที่มีอยู่มีความซับซ้อนไม่สามารถผลิตสารสนเทศที่นำไปใช้ในการสนับสนุนการตัดสินใจของธุรกิจเพื่อสนองต่อความต้องการของลูกค้าได้ การวิจัยนี้จึงได้นำเอาวิธีการขุดเจาะเหมืองข้อมูล มาใช้ในการวิเคราะห์ข้อมูล โดยใช้เทคนิคกระบวนการต้นไม้ตัดสินใจ (Decision Tree) วิเคราะห์แนวโน้มของผู้ถือบัตรเครดิต จากการทดลองทำให้ทราบถึงแนวโน้มของการมีบัตรเครดิต จากผลการวิจัยทำให้สามารถที่จะพัฒนาข้อเสนอแนะให้จูงใจและสร้างความสัมพันธ์กับลูกค้ากลุ่มต่างๆได้ต่อไป

Abstract

Due to higher business competitions and rapidly changing customers 'needs, the customer relationship management (CRM) has been playing its important role and it is strongly necessary for the current business systems. As a result, a large number of customer data have been stored for investigating customer behaviors. Also, current available data analyzing techniques are not yet able to analyze and refine data accurately. Such techniques cannot be implemented in supporting decision-making on the business for responding to their customer needs. Thus the data mining technique was implemented in analyzing data. In this case study the data mining by using the decision tree technique was implemented in analyzing trends of credit-card holders. From an experiment, trends of holding credit cards were known. Therefore, the research results could be used for developing recommendations to motivate and create relationships with customers in different groups in the future.

คำสำคัญ : การบริหารงานลูกค้าสัมพันธ์, การขุดเจาะเหมืองข้อมูล, กระบวนการต้นไม้ตัดสินใจ
Key word: Customer Relationship Management (CRM), Data Mining. Decision Tree

---

* อาจารย์ประจำกลุ่มวิชาการจัดการสารสนเทศและการสื่อสาร คณะมนุษยศาสตร์และสังคมศาสตร์มหาวิทยาลัยขอนแก่น

1.บทนำ

การที่ธุรกิจจะสามารถดำเนินอยู่ได้ จำเป็นอย่างยิ่งที่ต้องรับรู้และเข้าใจ และสนองตอบต่อความต้องการของลูกค้าได้อย่างชาญฉลาดและทันท่วงที รวมไปถึงยกระดับความต้องการของลูกค้าได้อย่างต่อเนื่อง นำมาซึ่งกลยุทธ์และวิธีการต่าง ๆ ที่จำเป็นต้องมีฐานความรู้ เพื่อใช้ในการสร้างกรอบการทำงาน ที่สนองตอบต่อกลยุทธ์ทางธุรกิจได้อย่างทันท่วงที และเทคโนโลยีที่สามารถสนับสนุนการตัดสินใจที่สำคัญๆ ในการใช้ประโยชน์จากฐานข้อมูลของลูกค้า เพื่อนำข้อมูลมาสร้างเป็นองค์ความรู้ โดยการศึกษาถึงรูปแบบธุรกรรมที่เหมือนกันในระบบของฐานข้อมูลขนาดใหญ่ ด้วยการทำเหมืองข้อมูลและนำข้อมูลที่วิเคราะห์ได้ไปประยุกต์ใช้ในการบริหารงานลูกค้าสัมพันธ์ต่อไป

การบริหารลูกค้าสัมพันธ์ (Customer Relationship Management : CRM)คือการสร้างความสัมพันธ์กับลูกค้า โดยการใช้เทคโนโลยีและการใช้บุคลากรอย่างมีหลักการ CRM ได้ถูกนำมาใช้มากยิ่งขึ้นเรื่อย ๆ เนื่องมาจากจำนวนคู่แข่งของธุรกิจแต่ละประเภทเพิ่มขึ้นสูงมาก การแข่งขันรุนแรงขึ้นในขณะที่จำนวนลูกค้ายังคงเท่าเดิม ธุรกิจจึงต้องพยายามสรรหาวิธีที่จะสร้างความพอใจให้แก่ลูกค้าอันจะนำไปสู่ความจงรักภักดีในที่สุด เป้าหมายของ CRM นั้นไม่ได้เน้นเพียงแค่การบริการ ลูกค้าเท่านั้น แต่ยังรวมถึงการเก็บข้อมูลพฤติกรรม ในการใช้จ่ายและความต้องการของลูกค้า จากนั้นจะนำข้อมูลเหล่านั้นมาวิเคราะห์และใช้ให้เกิดประโยชน์ โดยการใช้ เทคนิคของการทำเหมืองข้อมูล

การทำเหมืองข้อมูล(Data Mining-DM) คือกระบวนการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล (Database) แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากซึ่งจะช่วยในการเพิ่มมูลค่า (Value added)ให้กับข้อมูลที่องค์กรที่มีอยู่ โดยจะนำข้อมูลทั้งระดับปฏิบัติการ (Transaction data)ฐานข้อมูล (Data base) หรือคลังข้อมูล(Data warehouse) มาผ่านขบวนการทำงานที่เรียกว่า การสกัดข้อมูล (Extract data) จากฐานข้อมูลขนาดใหญ่ เพื่อให้ได้สารสนเทศ ที่เป็นประโยชน์ ที่เรายังไม่รู้ โดยเป็นสารสนเทศที่มีเหตุผล (Valid) และสามารถนำไปใช้ได้ ซึ่งเป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจ โดยใช้กระบวนการของต้นไม้ตัดสินใจ (Decision Tree) เข้ามาช่วย

งานวิจัยนี้ เริ่มต้นด้วย บทนำที่กล่าวถึงความสำคัญและแนวทางในการพัฒนากระบวนการบริหารงานลูกค้าสัมพันธ์(CRM)โดยใช้วิธีการการขุดเจาะเหมืองข้อมูล(Data Mining)เทคนิคต้นไม้ตัดสินใจ(Decision Tree) ส่วนที่ 2 กล่าวถึงแนวคิดและทฤษฎีที่เกี่ยวข้องได้แก่ การขุดเจาะเหมืองข้อมูล เทคนิคการจำแนกข้อมูล(Classification) ด้วยกระบวนการต้นไม้ตัดสินใจ(Decision Tree)และ อัลกอริธึมของต้นไม้ตัดสินใจ ส่วนที่ 3 วิธีดำเนินการวิจัย เริ่มตั้งแต่การจัดเตรียมข้อมูล การคัดเลือกและทดสอบเทคนิคที่ใช้ กระทั่งกระบวนการพัฒนาระบบงานด้วยวิชวลเบสิค(visul Basic) ส่วนที่ 4 กล่าวถึงผลการวิจัยและอภิปรายผลการทดลอง ส่วนที่ 5 ผลการวิจัย และส่วนที่ 6 ข้อเสนอแนะสำหรับการทำวิจัย

2.ทฤษฎีที่เกี่ยวข้อง

ทฤษฎีที่ใช้การทำวิจัยในครั้งนี้ ประกอบด้วย ความหมายของการบริหารงานลูกค้าสัมพันธ์ การขุดเจาะเหมืองข้อมูล เทคนิคการขุดเจาะเหมืองข้อมูล (data mining)โดยวิธีการต้นไม้ตัดสินใจ(Decision Tree) ที่เข้ามาช่วยให้การทำ CRM ได้สารสนเทศที่มีประโยชน์และถูกต้องมากขึ้น

2.1 Customer Relationship Management (CRM)

Customer Relationship Management (CRM) หรือ การบริหารลูกค้าสัมพันธ์ ซึ่งก็คือการสร้างความสัมพันธ์กับลูกค้า โดยการใช้เทคโนโลยีและการใช้บุคลากรอย่างมีหลักการ CRMได้ถูกนำมาใช้มากยิ่งขึ้นเรื่อยๆ เนื่องมาจากจำนวนคู่แข่งของธุรกิจแต่ละประเภทเพิ่มขึ้นสูงมาก การแข่งขันรุนแรงขึ้นในขณะที่จำนวนลูกค้ายังคงเท่าเดิมธุรกิจจึงต้องพยายามสรรหาวิธีที่จะสร้างความพอใจให้แก่ลูกค้าอันจะนำไปสู่ความจงรักภักดีในที่สุดเป้าหมายของ CRM นั้นไม่ได้เน้นเพียงแค่การบริการของลูกค้าเท่านั้น แต่ยังรวมถึงการเก็บข้อมูลพฤติกรรมในการใช้จ่าย และความต้องการของลูกค้า จากนั้นจะนำข้อมูลเหล่านั้นมาวิเคราะห์และใช้ให้เกิดประโยชน์ ในการพัฒนาผลิตภัณฑ์ หรือการบริการรวมไปถึงนโยบายในด้านการจัดการซึ่งเป้าหมาย สุดท้ายของการของพัฒนา CRM ก็คือการเปลี่ยนจากผู้บริโภคไปสู่การเป็นลูกค้าตลอดไป โดยเป้าหมายของ CRM แบ่งออกเป็นสามด้านหลัก ๆ(ดิสพงศ์ พรชนกนาถ,2543) ได้แก่

2.1.1 การได้มาซึ่งลูกค้าใหม่ (Acquire new customer) ทำอย่างไรจึงจะได้กลุ่มลูกค้าใหม่ ที่สร้างผลกำไรให้กับธุรกิจ

2.1.2 การรักษาลูกค้าเดิม (Retain old customer) ตามกฎ 20:80 ของพาเรโตกล่าวว่า ลูกค้าประมาณ 20% ที่ซื้อเราจะสร้างรายได้มูลค่าถึง 80 % ของยอดที่ขายได้ทั้งหมด การได้มาซึ่งลูกค้าใหม่ 1 คน จำเป็นต้องใช้ทรัพยากรจำนวนมหาศาล กว่าการรักษาลูกค้าเก่า 1 คนไว้ จึงเป็นเรื่องที่น่าสนใจว่าองค์กร จะรักษาลูกค้าเก่าไว้ได้อย่างไร และทำอย่างไรให้ลูกค้าเก่าเหล่านั้นมีความภักดีกับองค์กรเดิม

2.1.3 เพิ่มผลกำไร (Grow customer profitability) ทำอย่างไรที่องค์กรจะมีรายได้เพิ่มขึ้นจากลูกค้าเก่า ทำอย่างไรให้เกิดการซื้อเพิ่ม และไม่หนีหายไปไหน

2.2 ความหมายของการขุดเจาะเหมืองข้อมูล

Cabena , Hadjinian , Stadler , Verhees & Zanasi(1998) ได้แสดงความหมายของการขุดเจาะเหมืองข้อมูล(Data Mining) ว่าเป็น กระบวนการค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูลแต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากซึ่งจะช่วยในการเพิ่มมูลค่าให้กับข้อมูลที่องค์กรที่มีอยู่ โดยเป็นการทำงานร่วมกันระหว่างมนุษย์กับคอมพิวเตอร์ โดยมนุษย์จะทำหน้าที่ในการออกแบบฐานข้อมูล กำหนดจุดประสงค์ หรือเป้าหมายในการทำงานและคอมพิวเตอร์จะทำหน้าที่ในการคำนวณ หรือค้นหาความสัมพันธ์ของข้อมูลตามเป้าหมายที่ได้กำหนดไว้ โดยจะนำข้อมูลทั้งระดับปฏิบัติการ(Transaction data) ฐานข้อมูลหรือคลังข้อมูล (Data warehouse) มาผ่านขบวนการทำงานที่เรียกว่า การสกัดข้อมูล (Extract data) จากฐานข้อมูลขนาดใหญ่ เพื่อให้ได้สารสนเทศที่เป็นประโยชน์ที่เรายังไม่รู้ โดย

เป็นสารสนเทศที่มีเหตุผล และสามารถนำไปใช้ได้ ซึ่งเป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจ Data Mining เป็นโปรเซสที่สำคัญในการทำ Knowledge Discovery in Database ที่เราเรียกสั้นๆ ว่า KDD ส่วน Data Mining สามารถเรียกสั้น ๆ ว่า DM

2.2.1 วัฏจักรขั้นตอนการทำงานของ DM (Virtual cycle of data mining) ประกอบไปด้วย 4 ขั้นตอนหลัก ๆ ดังนี้

1) การระบุโอกาสทางธุรกิจหรือการระบุปัญหาที่เกิดขึ้นกับธุรกิจ คือ จะเป็นการระบุขอบเขตของข้อมูลที่จะนำมาทำการวิเคราะห์ เพื่อหาความได้เปรียบทางการตลาดหรือเพื่อนำมาแก้ปัญหา

2) ส่วนของ DM จะเป็นการนำเทคนิคของ DM ไปใช้การถ่ายทอดหรือทำการเปลี่ยนแปลงข้อมูลดิบให้อยู่ในรูปของข้อมูลที่สามารถนำไปปฏิบัติได้จริงในทางธุรกิจ

3) การปฏิบัติตามข้อมูล เราจะนำข้อมูลที่เป็นผลลัพธ์ของส่วน Data Mining นิ่งมาลองปฏิบัติจริงกับธุรกิจ

4) การวัดประสิทธิภาพจากผลลัพธ์ จะทำการวัดประสิทธิภาพของเทคนิคDM ที่จะนำมาใช้จากผลลัพธ์ ซึ่งสามารถตรวจสอบได้หลายทางอาจวัดจากส่วนแบ่งทางการตลาด , ปริมาณลูกค้า และวัดจากกำไรสุทธิที่ได้ เป็นต้น

จากทั้ง 4 ขั้นตอนที่กล่าวมาข้างต้น คือการนำ DM ไปใช้กับระบบธุรกิจ โดยแต่ละขั้นตอนจะพึ่งพาอาศัยกัน ผลลัพธ์จากขั้นตอนหนึ่งจะกลายมาเป็นอินพุทของขั้นตอนต่อไป ซึ่ง DM จะเปลี่ยนจากข้อมูลดิบเป็นข้อมูลประยุกต์ ที่สามารถนำมาใช้ประโยชน์ได้

2.3 วิธีการ การจัดหมวดหมู่ (Classification)

การจัดหมวดหมู่ถือว่าเป็นงานธรรมดาทั่วไปของดาต้าไมน์นิ่ง เพราะการทำความเข้าใจและการติดต่อสื่อสารต่างก็เกี่ยวข้องกับการแบ่งหมวดหมู่(Classifying),การจัดแยกประเภท(Categorizing) และการแบ่งแยกชนิด(Grading) งานในการแบ่งหมวดหมู่คือการบ่งบอกลักษณะ โดยการอธิบายจุดเด่นที่เป็นที่รู้จักดีในหมวดหมู่นั้น และ เทรนนิ่งเซต(Training Set)ของตัวอย่างในแต่ละหมวดหมู่ เช่น การจัดหมวดหมู่ของผู้ยื่นจองบัตรเครดิตเป็น ระดับต่ำ,กลาง,สูง การตัดสินใจโดยอาศัยหลักการแตกเป็นรูปทรี(Decision Tree) และ หลักการเมมโมรีเบสรีซันนิ่ง (Memory-Based Reasoning)เป็นเทคนิคที่เหมาะสมกับการจัดหมวดหมู่ ส่วนการวิเคราะห์การเชื่อมโยง(Link-Analysis)จะถูกนำมาใช้ประโยชน์สำหรับการจัดหมวดหมู่ในภาวะที่แน่นอนเท่านั้น (Alex Berson & Stephen J.Smith,1997)

2.4 เทคนิค Classification & Prediction

Classification เป็นกระบวนการสร้าง model จัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดให้ ตัวอย่างเช่น จัดกลุ่มนักเรียนว่า ดีมาก ปานกลาง ไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรือแบ่งประเภทของลูกค้าว่าเชื่อถือได้หรือไม่ได้ กระบวนการ Classification นี้แบ่งออกเป็น 3 ขั้นตอน (Alex Berson & Stephen J.Smith,1997)

Model Construction (learning) เป็นขั้นการสร้าง model โดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว (training data) ซึ่ง model ที่ได้อาจแสดงในรูปของแบบต้นไม้

(Decision Tree) เป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่ายมีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดจะแสดง attribute แต่ละกิ่งแสดงผลในการทดสอบและลีฟโหนดแสดงคลาสที่กำหนดไว้

Model Evaluate (Accuracy) เป็นขั้นของการประเมินความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (testing data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบจากคลาสที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง

Model Usage (Classification) เป็นโมเดลสำหรับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen data) โดยจะทำการกำหนดคลาส ให้กับ object ใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ

2.5 การสร้างดิซิชั่นทรี (Decision Tree Building)

Cabena , Hadjinian , Stadler , Verhees & Zanasi(1998) อธิบายว่าการสร้างดิซิชั่นทรีทำได้โดยใช้อัลกอริธึมที่อ้างอิงจาก Top-down Induction of Decision Trees (TDIDT) ใช้ induction เพราะความรู้ได้มาจากวิธีการเฉพาะจากกลุ่มข้อมูลซึ่งอยู่ในรูปแบบของ top-down กฎที่ถูกเลือกกฎแรกให้เป็นโหนดเดี่ยวบนสุด (root node) และจากนั้นก็จะมีการวนทำซ้ำในส่วนนั้นๆ ต่อไปการแบ่งแยกจะสิ้นสุดถ้าสมาชิก ทุกตัวของกลุ่มจัดอยู่ในประเภทเดียวกันแล้ว หรือไม่มีเกณฑ์ที่จะลงทางซ้ายสุดแล้ว ในบางอัลกอริธึมจะหยุดการแบ่งแยกโดยอ้างอิงถึงการ pre pruning ถ้าประเภทข้อมูลนั้น ได้รับการปรับปรุงแล้วดูเหมือนไม่มีความหมาย อัลกอริธึมอื่นๆ จะแทนที่ซับทรีที่ไม่มีความหมายโดยโหนดใบ หลังจากการประมวลผล ซึ่งเรียกว่า post pruning การนำข้อมูลมาสร้าง Decision Tree มีขั้นตอนพื้นฐานคือ

2.5.1 หาAttributeที่สำคัญที่สุดมาแบ่งข้อมูล โดย Attributeนี้จะถูกนำมาสร้างเป็นRoot node

2.5.2 นำค่าที่เป็นไปได้ใน Attribute ที่ถูกเลือกแตกออกมาเป็นกลุ่ม

2.5.3 แบ่งข้อมูลทั้งหมดตามกลุ่มที่แตกออกจาก Root node

2.5.4 นำข้อมูลแต่ละกลุ่มมาทำซ้ำขั้นตอนแรกคือหา Attribute ที่สำคัญที่สุด

2.6 อัลกอริธึมของดิซิชั่นทรี (Algorithm of Decision Tree)

การจัดกลุ่มข้อมูล หรือการพัฒนาต้นแบบแต่ละคลาส ในฐานข้อมูลให้อยู่บนพื้นฐานของลักษณะการนำเสนอ แบบการเทรนนิ่งข้อมูลกลุ่มของคลาส ซึ่งมีวิธีการจัดกลุ่มข้อมูลประกอบด้วยวิธีการดิซิชั่นทรี เช่น ID3, C4.5 เป็นต้น

2.6.1 อัลกอริธึม ID3 พัฒนาขึ้นมาในปี 1986 โดย Ross Quinlan เป็นดิซิชั่นทรี ที่สร้างอัลกอริธึมที่จัดกลุ่ม และตัดสินประเภทของ อ๊อปเจ็ค โดยการทดสอบค่าคุณสมบัติของอ๊อปเจ็คเหล่านั้น ซึ่งเป็นการสร้างทรีแบบบนลงล่าง (Top-down approach) ID3 Algorithm มีการสร้าง Tree ตามอัลกอริธึมดังนี้

1) Tree เริ่มต้นด้วย Node หนึ่ง Node แสดงถึงข้อมูลที่ใชTrain (Sample)

2) ถ้า Sample เป็น Class เดียวกัน Node ก็จะกลายเป็น leaf และมีชื่อตาม class

3) ถ้าไม่เช่นนั้น จะใช้การวัดแบบ Entropy-based ที่เรียกว่า Information gain เพื่อเลือก Attribute ที่เหมาะที่สุดที่จะเป็นตัวแยก Sample ออกเป็นแต่ละ Class ซึ่ง Attribute นี้จะกลายมาเป็น Attribute ที่ใช้ในการทดสอบหรือใช้ในการตัดสินใจที่ node

4) กิ่งของ Tree ถูกสร้างตามแต่ค่าของ Attribute และSample จะถูกแบ่งตามค่าของ Attribute นั้น

5) แล้วจะทำซ้ำกระบวนการเดิมโดยใช้ Sample จากแต่ละส่วน ที่ถูกแบ่งออกมากระบวนการทำซ้ำจะหยุดก็ต่อเมื่อเงื่อนไขเหล่านี้เป็นจริงSample ทั้งหมดอยู่ใน Class เดียวกัน, Sample ไม่มี Attribute เหลืออยู่อีก, ไม่มี Sample แล้ว

ต่อไปนี้จะเป็นการแสดงถึง การเลือก Attribute ที่มีความสำคัญที่สุดเพื่อใช้แบ่งข้อมูลโดยกำหนดให้

T แทน Training Set

S แทน Set ของข้อมูลใดๆ

P แทน ความน่าจะเป็นที่ Sample นั้นจะเป็นของ Class C_i

M แทน จำนวน Class

$$Info(S) = -\sum_{i=0}^{m} P_i \log_2(p_i)$$

และเมื่อนำสูตรนี้ไปประยุกต์ใช้กับ Training Set จะได้ Info(T) , $Info_x(T)$ เป็นการวัดค่าของ information เพื่อแบ่ง T โดยใช้ค่าที่เป็นไปได้ของ Attribute X

$$Info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \quad X \quad Info_x(T_i)$$

Gain (X) เป็นการวัดค่าของ information ที่ได้รับถ้าเลือก Attribute X

$$Gain\ (X) = Info\ (X) - Info_x\ (T)$$

2.6.2 อัลกอริธึม C4.5 อัลกอริธึมนี้เกิดขึ้นในปี 1993 โดย Quinlan เป็นการพัฒนาเพิ่มจากอัลกอริธึม ID 3 โดยอ้างอิงเทคนิค classification-decision เพื่อเรียกซ้ำสำหรับ data-set เพื่อใช้จัดการข้อมูลที่ ID3 ไม่สามารถจัดการได้ เช่น หลีกเลี่ยงข้อมูลที่ Over fitting Determining how deeply to grow a decision tree , Rule post-pruning, ลดการพรุนนิ่งผิดพลาด (error pruning), เลือก Attribute ที่เหมาะสมมาเป็นเครื่องมือในการคัดเลือก, ตรวจจับการเทรนนิ่งข้อมูลด้วยค่าที่ไม่ถูกต้อง, ตรวจจับ Attribute ที่เป็นข้อมูลต่อเนื่อง เช่น อุณหภูมิ, ตรวจจับแอททริบิวต์ด้วยค่าที่แตกต่างกัน, ปรับปรุงประสิทธิภาพในการประมวลผล

C4.5 จะสร้างดิซิชั่นทรีที่มีการจัดประเภทให้กลุ่มของข้อมูลได้เป็นแบบต่อเนื่องเพื่อแบ่งแยกข้อมูล โดยใช้วิธีการที่เรียกว่า Depth-first ซึ่งอัลกอริธึมนี้จะพิจารณาทดสอบความเป็นไปได้ในการแบ่งกลุ่มข้อมูลและเลือกมดสอบข้อมูลที่ดีที่สุด โดยในแต่ละ แอทริ-

บิวต์ ผลการทดสอบจะได้ตัวเลขมากมาย ค่าที่มีความแตกต่างจะถูกนำมาพิจารณา อัลกอริธึม C4.5 ได้เพิ่ม Feature ที่อัลกอริธึม ID3 ไม่มี

การหา Gain ratio criterion พัฒนาขึ้นเพื่อแก้ปัญหาของ Gain Criterion กรณีที่ Attribute มีค่าที่ unique การแบ่งข้อมูลโดยใช้ Attribute นี้อาจทำให้เกิด Subset จำนวนมากซึ่งแต่ละ Subset จะประกอบไปด้วยข้อมูลเพียง 1 record เท่านั้น ทำให้ $info_x(T) = 0$ ซึ่งจะมีผลให้ค่า information Gain ของ Attribute นี้มีค่าสูงมากและการแบ่งข้อมูลโดยใช้ Attribute นี้ไม่ก่อให้เกิดประโยชน์ใดๆ ต่อการทำนาย C4.5 แก้ไขโดยใช้ค่า Gain ratio ซึ่งคำนวณโดยใช้ split info(x) และ gain ratio(x) โดย split info(x) เป็นค่า information ที่ได้จากการแบ่ง T ออกเป็น n subset

$$\text{Split Info}(x) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2 \frac{|T_i|}{|T|}$$

gain ratio(x) เป็นการวัดว่า การแบ่งข้อมูลโดยใช้ Attribute นั้นก่อให้เกิดประโยชน์ต่อการทำนายหรือไม่

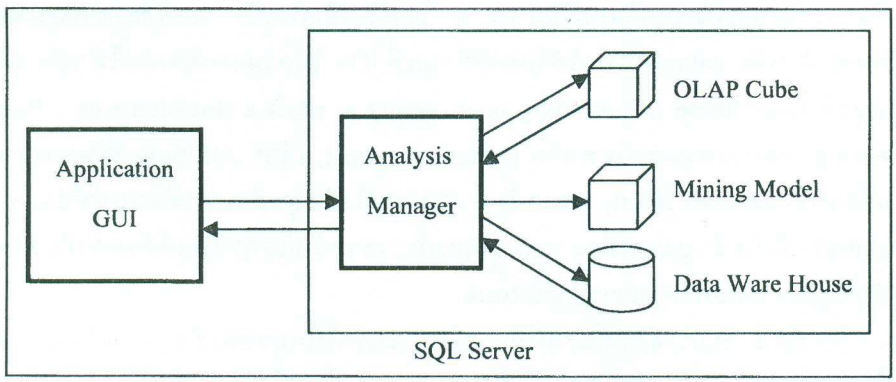
gain ratio(x) = gain(x)/split info(x)

ซึ่งการใช้ gain ratio ทำให้ tree ที่ได้มีขนาดเล็กกว่าการใช้ gain criterion

จากเทคนิคที่ได้ศึกษาทำให้เห็นว่าในกระบวนการทางธุรกิจหากต้องการจัดกลุ่มของลูกค้าด้วยเทคนิคที่เหมาะสมกับการจัดกลุ่มของข้อมูลโดยใช้กระบวนการทำเหมืองข้อมูลด้วยกระบวนการต้นไม้ตัดสินใจ โดยการดำเนินการวิจัยจะอยู่ที่หัวข้อถัด

3. วิธีดำเนินการวิจัย

บทนี้จะนำเสนอเกี่ยวกับวิธีดำเนินงานวิจัย ตั้งแต่ฐานข้อมูลที่ใช้วิธีการของ การขุดเจาะเหมืองข้อมูลที่ใช้เมื่อเปรียบเทียบกับชุดข้อมูลจริง รวมไปถึงการการสร้าง แอพพลิเคชั่นที่จะกลั่นกรองเอาข้อมูลที่ถูกต้องก่อนการนำไปใช้งานได้จริง



รูปที่ 1 แสดงการดำเนินการวิจัย

จากรูปที่ 1 การดำเนินการวิจัย เป็นการวิเคราะห์ข้อมูลจาก Data Warehouse โดยการสร้างโมเดลตามเทคนิคของ DM โดยใช้วิธีการ Decision Tree และสร้าง OLAP Cube เพื่อวิเคราะห์ข้อมูลและตรวจสอบความถูกต้องของโมเดลกับ Cube ที่สร้างขึ้นและเตรียมข้อมูลสำหรับนำไปแสดงผลยังแอพพลิเคชั่นที่สร้างขึ้น โดยดำเนินการตามขั้นตอนของ DM และ KDD ในการแยกความรู้จาก Data Warehouse ดังต่อไปนี้

3.1 การจัดเตรียมข้อมูลในการวิเคราะห์

ข้อมูลที่จะนำมาใช้ในการวิเคราะห์นี้ จะต้องจัดเตรียมข้อมูลให้อยู่ในรูปที่เหมาะสมและมีความถูกต้องก่อนนำมาใช้งาน โดยแต่ละแอทริบิวต์ต้องมีการทำความสะอาดข้อมูล (Cleaning) มีการปรับแต่งข้อมูลหรือตัดทิ้งข้อมูลที่ไม่ถูกต้องออกก่อน เช่นนข้อมูลที่มีค่าเป็น NULL หรือเป็นข้อมูลที่มีค่าไม่ถูกต้อง มีการแปลงข้อมูลให้เป็นข้อมูลในรูปที่ถูกต้อง ซึ่งถึงแม้ว่ากระบวนการนี้จะเป็นกระบวนการที่ใช้เวลานานในการจัดเตรียมข้อมูล แต่ก็นำมาซึ่งข้อมูลที่ถูกต้องเพื่อนำมาใช้ในการฝึกฝน(Training) ต่อไป ข้อมูลที่ใช้ในการวิเคราะห์เป็นข้อมูลจำลองของ MS SQL Server 2000 ที่ได้มีการรวบรวมเป็น Warehouse ไว้แล้วซึ่งข้อมูลที่เราเลือกมาทำการวิจัยในครั้งนี้ประกอบด้วย ตาราง Sale_Fact_1998 มีจำนวนข้อมูล 164,558 Record , customer มีจำนวนข้อมูล 10,281 Record , Product มีจำนวนข้อมูล 1,560 Record , Product_Class มีจำนวนข้อมูล 110 Record , Time_by_day มีจำนวนข้อมูล 730 Record , Store มีจำนวนข้อมูล 25 Record, Promotion มีจำนวนข้อมูล 1,864 Record

3.2 การสร้างโมเดล

สร้างโมเดลโดยใช้ไมโครซอฟต์เอสคิวแอลเซิร์ฟเวอร์ 2000 อนาไลซิสแมเนเจอร์เป็นชุดเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลต่างๆ โดยโครงสร้างการทำงานจะประกอบไปด้วย 2 ส่วน คือส่วน ดาต้าไมนิง ซึ่งใช้ในการสร้างดาต้าไมนิงโมเดลและส่วน โอแลป (OLAP Cube) ซึ่งเป็นเทคโนโลยีที่ช่วยให้การวิเคราะห์ข้อมูลในฐานข้อมูลได้อย่างมีประสิทธิภาพ โดยในการวิจัยในครั้งนี้ข้อมูลถูกจัดเก็บไว้ในรูปของฐานข้อมูลของ Microsoft Access ใช้เทคนิคต้นไม้ตัดสินใจในการจัดกลุ่มและวิเคราะห์แนวโน้มของข้อมูลโดยสร้างโมเดลตามการทดลองแบ่งออกเป็นสองส่วนดังนี้

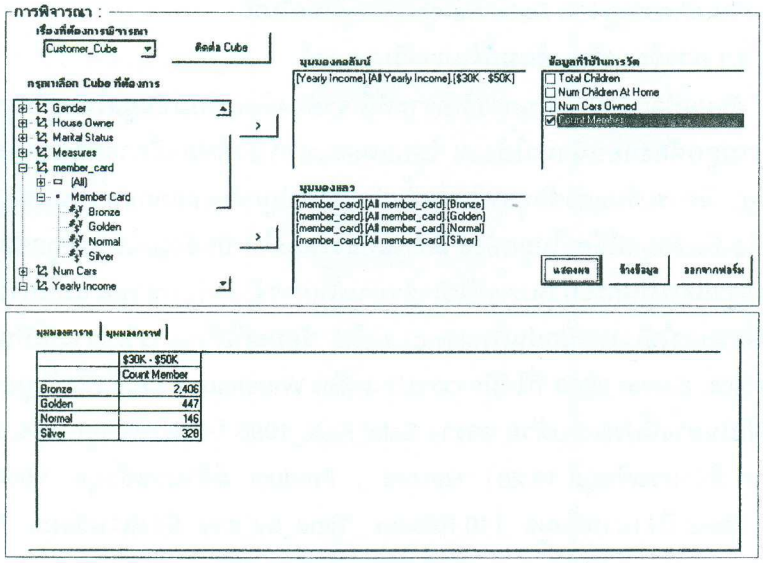
3.2.1 การทดลอง เพื่อหาแนวโน้มของการมีบัตรเครดิตของลูกค้า

การทดลองใช้ตารางของ customer ซึ่ง Attribute ที่ใช้ในการวิจัยครั้งนี้มีดังนี้Yearly Income, Marital Status, Education, Num cars Owned, Gender, House Owner, Member card ด้วยเหตุผลของการคัดเลือก Attribute นี้ เพราะคัดเลือกตามหลัก ปัจจัยที่มีผลต่อความเสี่ยงทางเครดิต ซึ่งประกอบด้วยปัจจัยหลัก 6 กลุ่มดังนี้ (สุนทรี สุนทราพรพล, 2539) (1) คุณสมบัติ(Character) สถานภาพสมรส (Marital Status), เพศ (Gender), การศึกษา (Education) (2) รายได้(Capacity) รายได้ต่อปี (Yearly Income) (3) เงินทุน (Capital) , รถที่มี(cars Owned) , บ้านที่มี (House Owner) (4) หลักประกัน (Collateral) (5) สภาวะแวดล้อม (Condition) (6) ประเทศ (Country)

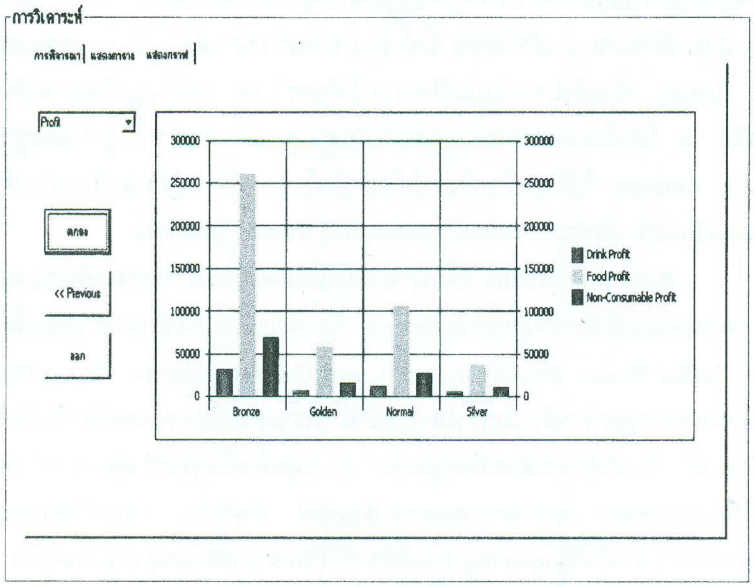
3.2.2 การแปรความหมายและนำความรู้ที่ได้มาใช้

ในส่วนนี้เป็นการสรุปความหมายของผลที่ได้ซึ่งจะเป็นข้อมูลความรู้ นำไปเป็นสารสนเทศในการตัดสินใจ โดยแสดงผลในรูปของประชากร ในกรณีที่สนใจ แสดงผลผ่าน

Application ด้วย Visual Basic โดยติดต่อยังCube ที่ได้สร้างเอาไว้แล้ว และสามารถตรวจสอบได้จาก cube ว่าผลลัพธ์โดยแอพลิเคชั่นที่สร้างถูกต้องตรงกันหรือไม่ และสามารถนำผลการวิเคราะห์ที่วิเคราะห์ได้แสดงผลไปยัง MS Excell เพื่อใช้ประกอบการตัดสินใจโดยแสดงในรูปของกราฟและข้อมูลที่ได้จากการวิเคราะห์



รูปที่ 2 แสดงแอพพลิเคชั่นที่สร้างขึ้น ทำการประมวลเงื่อนไขที่ต้องการแสดงในรูปของ



รูปที่ 3 แสดงแอพพลิเคชั่นที่สร้างขึ้นเพื่อแสดงผล ในรูปของกราฟ

จากการวิจัยได้ศึกษาเทคนิคและวิธีการต่าง ๆเพื่อประยุกต์เอาการขุดเจาะเหมืองข้อมูลมาใช้ โดยถูกต้องและแม่นยำพร้อมทั้งสร้างแอพพลิเคชั่นขึ้นมารองรับเพื่อให้ผู้ที่จะเข้ามาใช้ใช้งานได้ง่ายยิ่งขึ้น ส่วนผลของการทดลองและการประยุกต์ใช้งานจะอยู่ในหัวข้อถัดไป

4. ผลการวิจัยและอภิปรายผล

ผลการศึกษาที่ได้จากดำเนินงานวิจัย เป็นผลลัพธ์ที่ได้จากกระบวนการพัฒนาระบบ ซึ่งการวิจัยมุ่งการทดสอบการทำงานของแอพพลิเคชั่นที่สร้างขึ้นเพื่อการทำนายผลและเพื่อใช้ประกอบการพิจารณาตัดสินใจต่อไป

4.1 ผลการวิจัย

การทดลอง ทำการจัดแบ่งกลุ่มของลูกค้าออกตามประเภทของบัตรทั้ง 4 ประเภท เพื่อวิเคราะห์หาแนวโน้มในการทำบัตร โดยปัจจัยส่วนใหญ่ของการถือครองบัตรเครดิตจะอยู่ที่รายได้เป็นประการแรก ประการที่สองคือสถานะภาพการสมรส และได้แอพพลิเคชั่นโปรแกรมที่สามารถทำงานได้ถูกต้องตามกระบวนการของการทำเหมืองข้อมูล

4.2 การอภิปรายผล

ผลการทดลอง เพื่อตรวจสอบเงื่อนไขในทุก ๆ มุมมองที่จำเป็น ผลปรากฏว่า ปัจจัยแรกที่มีผลต่อการเลือกทำบัตรเครดิตคือ Yearly Income แบ่งออกตามช่วงของรายได้, Marital Status และ Num Car Owner โดยมีประชากรหนาแน่นที่บัตรแบบ Bronze เกือบทุกช่วงของรายได้ ยกเว้นช่วงเงินเดือนที่ $150K + จะเป็นบัตรแบบ Golden และ Silver ส่วนบัตรแบบ Normal คืออยู่ที่ช่วงรายได้ $10K - $30K

จากผลการทดลองทำให้ทราบถึงแนวโน้มของการมีบัตรโดยปัจจัยแรกที่มีผลกระทบต่อการทำบัตรคือเรื่องของเงินเดือน เพราะฉะนั้นการจะทำโปรโมชั่นใด ๆก็ควรที่จะดูระดับของราคาสินค้าและระดับเงินเดือนที่เหมาะสม เช่น บัตร Normal คืออยู่ที่ช่วงรายได้ $10K - $30K ฉะนั้นสินค้าที่จะส่งเสริมในประเภทของบัตรประเภทนี้ก็น่าจะอยู่ที่ราคาไม่น่าจะเกินรายได้ที่มีอยู่ เพราะด้วยปริมาณการซื้อและมูลค่ากำไรเฉลี่ยที่คำนวณได้ เรื่องต่อมาจะเป็นเรื่องของสถานะภาพการสมรสที่จะมีผลต่อการเลือกทำบัตรและมีการใช้จ่ายผ่านบัตรก็เป็นเรื่องที่น่าสนใจว่าคนที่ยังไม่แต่งงานจะมีการใช้จ่ายผ่านบัตรที่สูงกว่าคนที่แต่งงานแล้วการเลือกทำโปรโมชั่นก็น่าจะพิจารณาเงื่อนไขนี้ด้วยเช่นกัน

4.3 การเปรียบเทียบผลลัพธ์

ในการทดลองในครั้งนี้ได้มีการเปรียบเทียบผลลัพธ์ที่ได้จากการทดลอง โดยใช้วิธี CART และใช้เทคนิคแบบ Gini Index ผลปรากฏว่า node หรือ ข้อมูล ที่มีความสำคัญที่จะถูกนำมาใช้ในการพิจารณาเงื่อนไขของการทำบัตรคือ Yearly Income และความสำคัญอันดับรองลงมาสำหรับการพิจารณาในการทำบัตรเครดิตคือ Marital Status ซึ่งจากการเปรียบเทียบผลการทดลองปรากฏว่า ผลลัพธ์คือ node ที่ถูกเลือกมาเพื่อใช้ในการตัดสินใจในพิจารณาการทำบัตรเครดิตของลูกค้าถูกต้องตรงกับวิธีการที่ดำเนินการวิจัยอยู่

5.สรุปผลการวิจัย

ในการวิจัยในครั้งนี้ได้นำเอาวิธีการการขุดเจาะเหมืองข้อมูล และเทคนิคกระบวนการต้นไม้ตัดสินใจมาประยุกต์ใช้เพื่อหาแนวโน้มของการทำบัตร ทำให้ทราบปัจจัยที่มีผลต่อการทำบัตรคือเรื่องของเงินเดือน สถานะภาพการสมรส เพื่อสำหรับเป็นตัวแบบในการพิจารณาการทำบัตรของลูกค้าแต่ละประเภท และได้แอพพลิเคชั่นโปรแกรมที่สามารถวิเคราะห์ข้อมูลด้วยวิธีการของการทำเหมืองข้อมูลมีความถูกต้องและมีความยืดหยุ่น คือสามารถเพิ่มข้อมูลที่อาจจะเป็นผลกระทบต่อตัวโมเดลได้ตลอดเวลาซึ่งทำให้ค่าของข้อมูลสามารถปรับเปลี่ยนได้ตามข้อมูลจริง

6. ข้อเสนอแนะ

ยังมีความจำเป็นที่จะต้องทดสอบโมเดลเพื่อหาความถูกต้องและความแม่นยำของโมเดลที่สร้างขึ้นด้วยเครื่องมือหรืออัลกอริธึมเพิ่มเติม และในเรื่องของปริมาณของข้อมูล ถ้าข้อมูลที่ใช้ในการ train ชุดทดสอบมีปริมาณมากพอและมีทุกๆเหตุการณ์ที่จำเป็นต่อการจัดกลุ่มหรือตัดสินใจก็จะทำให้อัลกอริธึมส์สามารถที่จะแบ่งกลุ่มหรือจัดกลุ่มของข้อมูลได้อย่างถูกต้องและสุดท้ายเมื่อระยะเวลาและชุดข้อมูลที่เปลี่ยนไปก็จะมีผลกับโมเดลนี้ เพราะฉะนั้นควรที่จะปรับปรุงโมเดลอยู่อย่างสม่ำเสมอเพื่อให้ได้ข้อมูลสำหรับการตัดสินใจในการวิเคราะห์ที่ถูกต้องและทันสมัย

.....

บรรณานุกรม

ดิสพงศ์ พรชนกนาถ. (2543). **องค์ประกอบหลัก 8 ประการสร้าง CRM.** จาก http://www.ftpi.or.th/dwnld/pworld/pw43/43_8component.pdf.

ชูศักดิ์ เดชเกรียงไกร, นิทัศน์ คณะวรรณ และธีรพล แซ่ตั้ง . (2547). ***IRM-CRM การตลาดรุ่งมุ่งสัมพันธ์***. กรุงเทพฯ:ซีเอ็ดยูเคชั่น.

สถาบันวิจัยวิทยาศาสตร์และเทคโนโลยีแห่งประเทศไทย. 2546. **ประโยชน์ของการบริหารลูกค้าสัมพันธ์ (CRM).** จากhttp://www.tistr.or.th/crm/crm1.htm .

สุนทรี สุนทราพรพล . (2539). **การบริหารเครดิต**. พิมพ์ครั้งที่ 11. มหาวิทยาลัยกรุงเทพ ฯ:รามคำแหง.

Berson, A. & Smith, S. J. (1997). **Data Warehousing, Data Mining, & OLAP**. U.S.A. : McGraw-Hill.

Quinlan, J.R. (1993). **C4.5:PROGRAMS FOR MACHINE LEARNING**. San-Mateo, Calif. : Morgan Kaufmann.

Quinlan, J.R. (2003). **See5/C5.0 1.18**. from, http://www.rulequest.com/

see5-info.html.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). **Discovering Data Mining: From Concept to Implementation**. New Jersey : Prentice Hall.

Hui, S.C., Jha, G. (1993). **Application Data Mining for Customer Service Support**. from http://www.sciencedirect.com/

Soe-Tsyr Yuan, Wei-Lun Chang. (2001) .**Mixed-initiative Synthesized learning approach for web-base CRM**. from http://www.sciencedirect.com/

Surachai Wiwattanacharoenchai & Anongnart Srivihok . (n.d.). **Data Mining of Electronic Banking in Thailand : Usage Behavior Analysis by Using K-Means Algorithm**. Internet and. Inforamation Technology for Greater Mekong Bubregion(GMS) Business. from http://203.159.5.16/digital_gms/Proceedings/C55_SURACHAI_WIWATTANACHARO.pdf

YongSeog Kim. (2002). **Toward a successful CRM : variable selection sampling and Ensemble**. from http://www.sciencedirect.com/