การค้นหาผู้มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มโดยใช้ดีซีชั่นทรี

Classification of Value Added Tax Taxpayers by Using Decision Tree

โดย

นางสาวกรรณิกา สุดกังวล

รหัส 41067154

อาจารย์ที่ปรึกษา

ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 1 ปีการศึกษา 2543

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การค้นหาผู้มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มโดยใช้ ดีซีชั่นทรี
นักศึกษา	นางสาวกรรณิกา สุดกังวล
อาจารย์ที่ปรึกษา	ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2543

บทคัดย่อ

กฎเกณฑ์ของการคัดเลือกรายผู้ประกอบการที่ยื่นแบบแสดงรายการภาษีมูลค่าเพิ่ม (ภ.พ.30) เพื่อส่งตรวจปฏิบัติการที่มีอยู่ในปัจจุบัน ยังไม่สามารถค้นพบผู้ที่จะหลีกเลี่ยงภาษีได้ โดยกฎเกณฑ์ดังกล่าวนั้นตั้งขึ้นมาจากความเชื่อ หรือสมมติฐานของคนกลุ่มใดกลุ่มหนึ่ง ซึ่งใช้ประสพการณ์เป็นตัวกำหนด อีกทั้งกฎเกณฑ์ที่กำหนดขึ้นมายังไม่ครอบคลุมทุกกรณี จึงทำให้ยังมีผู้ประกอบการบางรายสามารถหลบเลี่ยงกฎเกณฑ์ดังกล่าวได้ จึงได้มีการนำเทคนิคของดาต้าไมนิ่ง (Data Mining) ที่มีลักษณะการทำงานแบบดีซีชั่นทรี(Decision Tree) มาใช้ โดยเลือกใช้อัลกอริทึ่มของ ID3 เข้ามาช่วยเพื่อเพิ่มประสิทธิภาพในการคัดเลือกรายผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มให้ได้ผลดียิ่งขึ้น

Title	Classification of Value Added Tax Taxpayer By Using Decision Tree
Student	Miss Kannikar Sudkangvon
Advisor	Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2000

Abstract

Presently, the auditing criteria of selecting a value added tax taxpayer , which decided by experienced person , are not efficient for finding taxpayers who are high possibility of fault claim. This project proposes ID3 algorithm , which is a algorithm to generate a decision tree , for finding the auditting criteriea. The proposed technique is more efficient for finding taxpayers who are high possibility of fault claim.

กิตติกรรมประกาศ

ในการพัฒนาระบบงานนี้สามารถดำเนินการลุล่วงมาได้ด้วยดี เพราะได้รับการสนับสนุนจากหลายๆ ท่าน ข้าพเจ้าจึงใคร่ขอขอบพระคุณ

1. ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษา ที่ให้ความรู้ แนะนำหนังสือ และให้คำปรึกษา แนะนำแนวทางแก้ไขปัญหา
2. คุณพ่อคุณแม่ที่ให้กำเนิด
3. หน่วยส่งเสริมประสิทธิภาพ กรมสรรพากร ที่อำนวยความสะดวกเกี่ยวกับรายละเอียดของภาษีมูลค่าเพิ่ม
4. ขอขอบพระคุณคณาจารย์ทุกท่านที่ประสิทธิประสาทวิชาให้

กรรณิกา สุดกังวล

สารบัญ

หน้า

หน้า

สารบัญภาพ

หน้า

ภาพที่

สารบัญตาราง

หน้า

ตารางที่

บทที่1
บทนำ

1.1 ความสำคัญและเหตุผลในการศึกษา

ในโลกปัจจุบันเทคโนโลยีได้พัฒนาก้าวไกลไปมาก ทั้งเทคโนโลยียังเข้ามามีบทบาทเป็นผลทำให้ธุรกิจของโลกเราเข้าสู่ยุคโลกาภิวัฒน์(gobolization) โดยมีการแข่งขันกันทั้งภายในและภายนอกประเทศ จึงมีผู้ประกอบกิจการต่างๆเกิดขึ้นอย่างมากมาย ซึ่งในบรรดาผู้ที่ประกอบกิจการเหล่านี้ ก็มีผู้ที่ประกอบธุรกิจที่ต้องชำระภาษีมูลค่าเพิ่มเป็นจำนวนมากโดยจะเห็นได้จากการยื่นแบบชำระภาษีมูลค่าเพิ่ม (ภ.พ.30) ซึ่งหากผู้ประกอบการมีความซื่อตรงในการยื่นแบบชำระภาษีมูลค่าเพิ่มตามจริงที่ได้ประกอบกิจการแล้วก็จะไม่เกิดกรณีของการตรวจสอบกิจการของผู้ประกอบการ แต่ในความเป็นจริงมีผู้ประกอบการจำนวนมากที่พยายามหาทางหลีกเลี่ยงการชำระภาษีมูลค่าเพิ่ม ดังนั้นกรมสรรพากรจึงต้องทำการค้นหาผู้ประกอบการที่มีแนวโน้มของการหลีกเลี่ยงภาษีจากข้อมูลของแบบแสดงรายการภาษีมูลค่าเพิ่ม (ภ.พ.30) โดยทำการคัดเลือกรายผู้ประกอบการที่ยื่นแบบภาษีมูลค่าเพิ่มเพื่อดำเนินการส่งตรวจสอบภาษี แต่ในปัจจุบันหลักเกณฑ์การคัดเลือกรายดังกล่าวยังขาดความรัดกุม จึงเป็นหนทางให้ผู้ประกอบการยื่นแบบภาษีมูลค่าเพิ่มโดยมิชอบด้วยกฎหมาย รวมทั้งยังหลีกพ้นจากเงื่อนไขของระบบการตรวจสอบและมีอัตราเสี่ยงต่อการเลี่ยงภาษีมูลค่าเพิ่มด้วย จึงต้องหาวิธีที่จะทำให้การคัดเลือกรายของผู้ประกอบการที่มีความเสี่ยงต่อการเลี่ยงภาษีมูลค่าเพิ่มจากฐานข้อมูลการยื่นแบบภาษีมูลค่าเพิ่ม (ภ.พ.30) มีประสิทธิภาพที่สุดจึงได้นำเทคนิคของดาต้าไมนิ่ง(Data Mining) เข้ามาช่วยในการคัดเลือกข้อมูลที่จะสนับสนุนการตรวจสอบผู้ประกอบการที่มีความเสี่ยงต่อการเลี่ยงภาษีให้ได้ผลยิ่งขึ้น

1.2 วัตถุประสงค์ของการศึกษา

เพื่อนำเอาเทคนิคของดาต้าไมนิ่ง(Data Mining) ที่ศึกษามาประยุกต์ใช้ในกระบวนการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจสอบภาษีมูลค่าเพิ่ม เพื่อที่จะให้ได้ผู้ประกอบการรายที่มีอัตราเสี่ยงต่อการหลีกเลี่ยงการชำระภาษีมูลค่าเพิ่มจริงๆ โดยการใช้ทรัพยากรอย่างเหมาะสมและมีประสิทธิภาพ เพื่อทำให้กระบวนการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจสอบให้ได้ประสิทธิภาพและ รัดกุมมากที่สุด

1.3 ขอบเขตของการศึกษา

โครงการนี้เป็นการศึกษาถึงการนำเอาเทคนิคของดาต้าไมนิ่ง(Data Mining) มาปรับใช้ในกระบวนการคัดเลือกรายผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มเพื่อส่งตรวจสอบภาษีมูลค่าเพิ่ม โดยใช้ข้อมูลจากภายในหน่วยงานของกรมสรรพากรเพื่อหาผู้ประกอบการกลุ่มเป้าหมายที่ตอบสนองกับกระบวนการมากที่สุด ซึ่งเป็นการพิจารณาที่พฤติกรรมการยื่นแบบภาษีมูลค่าเพิ่ม (ภ.พ.30) ของผู้ประกอบการ , ระยะเวลาการจดทะเบียนเป็นผู้ประกอบการภาษีมูลค่าเพิ่มของผู้ประกอบกิจการ (ภ.พ.01) และ ประวัติของผลการตรวจสอบผู้ประกอบการ

1.4 หลักการที่เกี่ยวข้องในการพัฒนาระบบงาน

1.4.1 เทคนิคของดาต้าไมนิ่ง(Data Mining)

1.4.2 หลักการของดีซีชั่นทรี(Decision Tree)โดยใช้อัลกอริทึ่ม ID3

1.5 องค์ประกอบของการพัฒนาระบบงาน

1.5.1 คอมพิวเตอร์ NT ที่ประกอบด้วย RAM ขนาด 130 MB

1.5.2 Visual Basic

1.5.3 ระบบปฏิบัติการ DOS

1.5.4 Windows 98

1.5.5 Visual FoxPro6

1.6 ขั้นตอนการศึกษา

เพื่อให้บรรลุถึงวัตถุประสงค์ที่กำหนดไว้ภายใต้ขอบเขตของการศึกษาจึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1.6.1 ศึกษาแนวคิดและทฤษฎีที่เกี่ยวกับดาต้าไมนิ่ง(Data Mining)

1.6.2 ศึกษาและวิเคราะห์ในขั้นต้นถึงปัจจัยที่มีผลต่อการหลีกเลี่ยงภาษี

1.6.3 เก็บรวบรวมข้อมูลที่ต้องการใช้ในการดำเนินงาน

1.6.4 ทำความสะอาดข้อมูลและทำการแปลงข้อมูล

1.6.5 จัดทำโมเดลสำหรับคัดเลือกผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษี

1.7 ประโยชน์ที่คาดว่าจะได้รับจากโครงการ

1.7.1 เพื่อเพิ่มประสิทธิภาพของการคัดเลือกรายผู้ประกอบการในการสุ่มตรวจภาษีมูลค่า

เพิ่มให้ดียิ่งขึ้น

1.7.2 เพื่อให้ขบวนการของการทำดาต้าไมนิ่ง(Data Mining) หารูปแบบความสัมพันธ์ระหว่างข้อมูลด้วยเงื่อนไขต่างๆ ที่คนไม่สามารถคาดการณ์ได้จากข้อมูลเก่าๆ

1.7.3 เพื่อการใช้ทรัพยากรอย่างมีประสิทธิภาพยิ่งขึ้น

ในบทนี้เป็นการกล่าวถึงวัตถุประสงค์ และขอบเขตของการทำงานในเบื้องต้นของระบบนี้เท่านั้น ส่วนในบทถัดไปจะกล่าวถึงสภาพแวดล้อมของขององค์กรที่จะพัฒนาระบบและความรู้เบื้องต้นเกี่ยวกับงานขององค์กรที่จะพัฒนาระบบต่อไป

บทที่ 2
สภาพแวดล้อมขององค์กร

2.1 ความเป็นมา

กรมสรรพากรเป็นหน่วยงานของรัฐที่มีหน้าที่จัดเก็บภาษีอากร ภาษีที่จัดเก็บอยู่หลายภาษี เช่น ภาษีเงินได้บุคคลธรรมดา ภาษีนิติบุคคล ภาษีมูลค่าเพิ่ม ภาษีปิโตรเลียม เป็นต้น แต่ในโครงการพัฒนาระบบงานนี้จะของกล่าวในส่วนของภาษีมูลค่าเพิ่ม ซึ่งเป็นภาษีที่กรมสรรพากรจัดเก็บจากผู้ประกอบกิจการที่ต้องชำระภาษีมูลค่าเพิ่ม โดยหากผู้ประกอบการดำเนินการยื่นแบบแสดงรายการภาษีมูลค่าเพิ่มตามความเป็นจริงแล้วการเก็บภาษีก็จะดำเนินไปโดยไม่มีข้อติดขัด หากแต่ในความเป็นจริงแล้วมีผู้ประกอบการบางรายที่ยื่นแบบแสดงรายการภาษีมูลค่าเพิ่มโดยมีข้อมูลที่ผิดไปจากความเป็นจริง ซึ่งอาจจะเกิดจากความผิดพลาดทั้งโดยตั้งใจและมิได้ตั้งใจ กรมสรรพากรจึงต้องดำเนินการตรวจสอบข้อมูลตามแบบแสดงรายการภาษีมูลค่าเพิ่มที่ผู้ประกอบการยื่นไว้ โดยให้ฝ่ายตรวจปฏิบัติการเป็นหน่วยงานที่ตรวจสอบ ซึ่งหน่วยงานตรวจปฏิบัติการได้ข้อมูลของผู้ที่จะทำการตรวจปฏิบัติการมาจากสำนักมาตรฐาน ซึ่งกฎเกณฑ์ของการคัดเลือกรายผู้ประกอบการที่ยื่นแบบแสดงรายการภาษีมูลค่าเพิ่ม (ภ.พ.30) เพื่อส่งตรวจปฏิบัติการที่สำนักมาตรฐานใช้อยู่ในปัจจุบัน ยังไม่สามารถค้นพบผู้ที่จะหลีกเลี่ยงภาษีได้ โดยกฎเกณฑ์ดังกล่าวนั้นตั้งขึ้นมาจากความเชื่อ หรือสมมติฐานของคนกลุ่มใดกลุ่มหนึ่ง ซึ่งใช้ประสพการณ์เป็นตัวกำหนด อีกทั้งกฎเกณฑ์ที่กำหนดขึ้นมายังไม่ครอบคลุมทุกกรณี จึงทำให้ยังมีผู้ประกอบการบางรายสามารถหลบเลี่ยงกฎเกณฑ์ดังกล่าว หรือ ยื่นแบบแสดงรายการภาษีมูลค่าเพิ่มโดยมิชอบด้วยกฎหมายได้ จึงได้มีแนวคิดที่จะนำเอาเทคนิคใหม่ๆ เข้ามาใช้ซึ่งในปัจจุบันพบว่ามีเทคโนโลยีเกิดขึ้นมามากมาย ซึ่งดาต้าไมนิ่ง (Data Mining) ก็เป็นเครื่องมือหนึ่งในนั้นที่มีประสิทธิภาพในการค้นหาความหมายที่แอบแฝงอยู่ในฐานข้อมูลต่างๆ จึงได้เลือกนำเทคนิคของดาต้าไมนิ่ง(Data Mining) ที่มีลักษณะการทำงานแบบดีซีชั่นทรี (Decision Tree) มาใช้ โดยเลือกใช้อัลกอริทึ่มของ ID3 เข้ามาช่วยในการคัดเลือกรายของผู้ประกอบการที่มีความเสี่ยงต่อการเลี่ยงภาษีมูลค่าเพิ่มจากฐานข้อมูลการยื่นแบบภาษีมูลค่าเพิ่ม (ภพ.30)เพื่อเพิ่มประสิทธิภาพในการคัดเลือกรายผู้ประกอบการ ที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มให้ได้ผลดียิ่งขึ้น

กระบวนการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจสอบภาษีมูลค่าเพิ่มนั้นเป็นกระบวนการอย่างหนึ่งในหลายๆ กระบวนการของระบบการตรวจสอบภาษีมูลค่าเพิ่มซึ่งมีความสำคัญอย่าง

มาก การคัดเลือกรายจะช่วยเพิ่มประสิทธิภาพในกระบวนการตรวจสอบได้มากขึ้น แต่เราจะความมั่นใจได้อย่างไรว่ากระบวนการคัดเลือกรายที่กรมสรรพากรทำอยู่นั้นมีประสิทธิภาพและมีความถูกต้องรัดกุมเพียงพอแล้ว เพื่อให้บรรลุจุดประสงค์ในการคัดเลือกรายให้มีประสิทธิภาพและคุ้มค่ามากที่สุด จึงได้นำเอาเทคนิคของดาต้าไมนิ่ง(Data Mining) เข้ามาใช้โดยพิจารณาจากสิ่งแวดล้อมทั้งภายใน และ ภายนอกองค์กรเองโดยใช้ทรัพยากรที่มีอยู่แล้วภายในองค์กร

2.2 ความรู้พื้นฐานของระบบภาษีมูลค่าเพิ่ม

เมื่อจะดำเนินการเกี่ยวกับภาษีมูลค่าเพิ่มแล้วในขั้นแรกจึงควรจะต้องทำการเรียนรู้ถึงคำศัพท์พื้นฐานในระบบภาษีมูลค่าเพิ่มก่อนซึ่งมีดังนี้

2.2.1 ภาษีมูลค่าเพิ่ม (Value Added Tax) คือภาษีที่นำมาใช้จัดเก็บแทนภาษีการค้าที่ได้ยกเลิกไป โดยจัดเก็บจากผู้ประกอบการที่ขายสินค้าหรือบริการทุกประเภทเฉพาะมูลค่าส่วนที่เพิ่มขึ้นในแต่ละขั้นตอนของการขายหรือการให้บริการเท่านั้น [5]

2.2.2 ภาษีขาย หมายถึง ภาษีที่ผู้ประกอบการเรียกเก็บในขณะที่ขายสินค้าหรือบริการ [5]

2.2.3 ภาษีซื้อ หมายถึง ภาษีที่ผู้ประกอบการได้เสียไปในขณะที่ซื้อสินค้าหรือบริการต่างๆ รวมทั้งสินค้าทุนด้วย เพื่อใช้ในกิจการของตนเอง [5]

2.2.4 ผู้ประกอบการ หมายถึง บุคคลซึ่งขายสินค้า หรือ ให้บริการในทางธุรกิจ ไม่ว่าการกระทำดังกล่าวจะได้รับประโยชน์ ค่าตอบแทน หรือไม่ก็ตาม [7]

2.2.5 ผู้ที่มีหน้าที่ชำระภาษีมูลค่าเพิ่มได้แก่ผู้ประกอบการที่เป็นผู้ขายสินค้าทั้งที่เป็นผู้ผลิต ผู้ขายส่ง ,ผู้ขายปลีก หรือผู้ให้บริการ ผู้ส่งออก ตลอดจนผู้นำเข้า ไม่ว่าจะเป็น บุคคลธรรมดา ห้างหุ้นส่วนจำกัด บริษัท รัฐวิสาหกิจ หรือองค์กรใดที่มีรายได้เกินกว่า 1,200,000 บาท [7]

2.2.6 กำหนดของการยื่นชำระแบบภาษีมูลค่าเพิ่ม โดยภาษีมูลค่าเพิ่มยื่นชำระเป็นรายเดือนปกติภาษีมูลค่าเพิ่มสำหรับเดือนภาษีใดกำหนดชำระภายในเดือนถัดไปไม่เกินวันที่ 15 ที่สรรพากรเขตหรือสรรพากรอำเภอ ส่วนกรณีนำเข้าและส่งออกจากต่างประเทศให้ชำระในวันที่มีการชำระอากรขาเข้า [7]

2.3 ขั้นตอนของการรับแบบภาษีมูลค่าเพิ่ม

เมื่อผู้ประกอบการประกอบกิจการที่ต้องชำระภาษีมูลค่าเพิ่มแล้ว หากถึงกำหนดเวลาใน

การยื่นแบบก็ต้องไปยื่นแบบภาษีมูลค่าเพิ่ม (ภ.พ.30) ตามที่กฎหมายกำหนด โดยมีขั้นตอนดังนี้[5]

2.3.1 ผู้ประกอบการภาษีมูลค่าเพิ่มยื่นแบบ ภ.พ.30 เพื่อชำระภาษีมูลค่าเพิ่มที่สรรพากรอำเภอ หรือสรรพากรเขต

2.3.2 เจ้าหน้าที่สรรพากรอำเภอหรือสรรพากรเขตทำการรับแบบตรวจสอบความถูกต้องของแบบ ภ.พ.30 เสร็จแล้วก็ทำการบันทึกเข้าเครื่องออกใบเสร็จ (POS) จากนั้นก็นำใบเสร็จคืนกลับให้ผู้ประกอบการ

2.3.3 เจ้าหน้าที่สรรพากรอำเภอหรือสรรพากรเขตจัดแฟ้มนำส่งแบบให้กับสำนักงานสรรพากรพื้นที่หรือสำนักงานสรรพากรภาคต้นสังกัด

2.3.4 ฝ่ายบริหารของสำนักงานสรรพากรพื้นที่หรือสรรพากรภาคนั้นๆเมื่อรับแบบเสร็จก็จัดส่งแบบดังกล่าวให้กับฝ่ายบริหารงานกรรมวิธีของสรรพากรพื้นที่เพื่อบันทึกแบบภ.พ.30 ส่วนสรรพากรภาค ฝ่ายบริหารจะจัดส่งให้ฝ่ายประมวลผลของภาค

2.3.5 ดำเนินการบันทึกแบบภ.พ.30 ซึ่งข้อมูลที่ถูกบันทึกจะถูกส่งทางออนไลท์(On-line) เพื่อไปเก็บที่สำนักเทคโนโลยีสารสนเทศ เพื่อให้ดำเนินการประมวลผลข้อมูลต่อ

2.3.6 เมื่อบันทึกแบบเสร็จฝ่ายกรรมวิธี หรือฝ่ายประมวลผล ก็จะจัดส่งแบบภ.พ.30 ที่บันทึกแล้วนั้นคืนให้ฝ่ายบริหารเพื่อดำเนินการจัดเก็บต่อไป

2.4 ขั้นตอนในการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจสอบในปัจจุบัน

จากขั้นตอนการรับแบบดังที่ได้กล่าวไว้แล้วข้างต้นนั้น เมื่อสำนักงานสรรพากรภาค 4 ประมวลผลข้อมูลแบบแสดงรายการภาษีมูลค่าเพิ่มที่ได้รับแล้ว ก็จะออกเป็นรายงานของผู้ประกอบการที่ยื่นแบบแสดงรายการผิดพลาดออกมาเพื่อส่งให้ฝ่ายกรรมวิธีของสำนักงานสรรพากรพื้นที่สำนักงานสรรพากรจังหวัด ดำเนินการตรวจรายงานดังกล่าวกับแบบแสดงรายการภาษีมูลค่าเพิ่ม แต่ตามความเป็นจริงแล้วมีผู้ประกอบการเพียงส่วนน้อยที่จะยื่นแบบตามความเป็นจริง จึงได้มีการคัดเลือกรายผู้ประกอบการขึ้นมาเพื่อสุ่มตรวจสอบโดยมีขั้นตอนของการคัดเลือกรายเพื่อตรวจปฏิบัติการภาษีมูลค่าเพิ่มมีดังต่อไปนี้[7]

2.4.1 สำนักมาตรฐานทำการกำหนดหลักเกณฑ์ต่างๆที่จะใช้เพื่อทำการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจปฏิบัติการภาษีมูลค่าเพิ่ม แล้วดำเนินการส่งหลักเกณฑ์ดังกล่าวให้กับสำนักเทคโนโลยีสารสนเทศเพื่อดำเนินการ

2.4.2 สำนักเทคโนโลยีสารสนเทศจัดทำโปรแกรมเพื่อคัดเลือกรายผู้ประกอบการจากข้อมูลภาษีมูลค่าเพิ่ม(ภ.พ.30) ตามหลักเกณฑ์ที่สำนักมาตรฐานกรรมวิธีกำหนด แล้วส่งข้อมูลผลลัพธ์ที่ได้ให้กับสำนักมาตรฐานต่อไป

2.4.3 เมื่อสำนักมาตรฐานได้รับข้อมูลจากสำนักเทคโนโลยีสารสนเทศแล้ว จะดำเนินการจัดส่งข้อมูลดังกล่าวกระจายให้กับสำนักงานสรรพากรพื้นที่หรือสำนักงานสรรพากรจังหวัดเพื่อดำเนินการ

2.4.4 เมื่อสำนักงานสรรพากรพื้นที่หรือสำนักงานสรรพากรจังหวัด ได้รับรายชื่อผู้ประกอบการจากสำนักมาตรฐานแล้วก็จะส่งรายนั้นให้กับฝ่ายตรวจปฏิบัติการเพื่อดำเนินการตรวจสอบต่อไป

2.5 หลักเกณฑ์ในการคัดเลือกรายผู้ประกอบการเพื่อส่งตรวจสอบ

เมื่อพิจารณาถึงการรับแบบภาษีมูลค่าเพิ่มที่ได้กล่าวมาแล้วในข้างต้นพบว่าผู้ประกอบการที่ยื่นแบบภาษีมูลค่าเพิ่มมีเป็นจำนวนมาก จึงทำให้ไม่สามารถที่จะทำการตรวจสอบผู้ประกอบการได้ทั้งหมดทุกราย และ ทุกแบบแสดงรายการภาษีมูลค่าเพิ่ม จึงต้องมีการกำหนดหลักเกณฑ์ในการคัดเลือกรายผู้ประกอบการที่เสียภาษีมูลค่าเพิ่มบางรายขึ้นมาสุ่มตรวจ ซึ่งสำนักมาตรฐานจะกำหนดหลักเกณฑ์ของการคัดเลือกรายผู้ประกอบการเพื่อการตรวจปฏิบัติการภาษีมูลค่าเพิ่มเป็นคราวๆไป

ในการคัดเลือกรายครั้งหนึ่งๆนั้น หลักเกณฑ์ต่างๆอาจจะไม่ได้ถูกใช้ทั้งหมด แต่ขึ้นอยู่กับประเภทของกิจการที่ต้องการสุ่มตรวจว่าต้องการตรวจกิจการใด ซึ่งมีหลักเกณฑ์การคัดเลือกรายสำหรับผู้ประกอบการซื้อมาขายไปเพื่อตรวจปฏิบัติการภาษีมูลค่าเพิ่มตามตารางที่2.1 ดังนี้[6]

ลำดับที่	หลักเกณฑ์	คะแนน
1.	อัตราส่วนภาษีซื้อ > ภาษีขาย	2
2.	ระยะเวลาที่จดทะเบียนมาแล้ว > 1ปี	1
3.	ระยะที่มียอดเครดิตยกไปติดต่อกัน > 4 ถึง 8 ครั้ง	1
4.	ยอดซื้อ รวม 12 เดือนภาษี > 10 ล้วน	2
5.	ยอดขายรวม 12 เดือนภาษี > 10 ล้วน	2
6.	จำนวนครั้งที่ขอคืนเป็นเงินสด > 4 ถึง 8 ครั้ง	1
7.	ขอคืนรวม 12 เดือนภาษีล่าสุดเป็นเงิน > 240,000 บาท	1
8.	ยื่นแบบขอคืนสำหรับเดือนภาษีเดียวกัน > 3 ครั้ง	1

ตารางที่ 2.1 หลักเกณฑ์การคัดเลือกรายผู้ประกอบการซื้อมา-ขายไปเพื่อตรวจปฏิบัติการ

จากตารางจะเห็นได้ว่าการคัดเลือกข้อมูลเพื่อส่งตรวจจะคัดเลือกตามหลักเกณฑ์ทั้ง 8 ข้อ คือ

2.5.1 ดูว่าอัตราภาษีซื้อมากกว่าภาษีขาย หรือไม่ ถ้ามากกว่าก็จะได้ 2 คะแนน

2.5.2 ดูว่าระยะเวลาที่กิจการได้จดทะเบียน มีมากว่า 1 ปีหรือไม่ ถ้ามากกว่า 1 ปีจะได้ 1 คะแนน

2.5.3 ดูว่าระยะที่มียอดเครดิตยกไปติดต่อกันมีมากกว่า 4-8 ครั้งหรือไม่ หากมากกว่าจะได้ 1 คะแนน

2.5.4 ดูว่ายอดซื้อรวม 12 เดือนภาษีว่ามีค่ามากกว่า 10 ล้าน หรือไม่ ถ้ามากกว่าก็ให้ 2 คะแนน

2.5.5 ดูว่ายอดขายรวม 12 เดือนภาษีว่ามีค่ามากกว่า 10 ล้าน หรือไม่ ถ้ามากกว่าก็ให้ 2 คะแนน

2.5.6 ดูว่าจำนวนครั้งที่ขอคืนเป็นเงินสด (ภายใน 12 เดือนภาษี) มากกว่า 4 - 8 ครั้งหรือไม่ ถ้ามากกว่าจะได้คะแนน 1 คะแนน

2.5.7 ดูว่าจำนวนเงินที่ขอคืนรวม 12 เดือนภาษีล่าสุดเป็นเงินมากกว่า 240,000 บาทหรือไม่ ถ้ามากว่า จะได้คะแนน 1 คะแนน

2.5.8 ดูว่ายื่นแบบขอคืนสำหรับเดือนภาษีเดียวกันมากกว่า 3 ครั้งหรือไม่ ถ้ามากกว่า 3 ครั้ง จะได้คะแนน 1 คะแนน

เมื่อได้คะแนนมาแล้วก็เรียงคะแนนของผู้ประกอบการแต่ละรายจากมากไปหาน้อย และเลือกออกมาเท่ากับจำนวนที่ประมาณการให้แต่ละหน่วยงานดำเนินการ

2.6 ปัญหาที่พบจากการคัดเลือกรายเพื่อการตรวจสอบภาษีมูลค่าเพิ่มในปัจจุบัน

จากข้อมูลผู้ประกอบการรายที่ได้รับมาจากสำนักมาตรฐานนั้นเมื่อดำเนินการตรวจปฏิบัติการภาษีมูลค่าเพิ่มตามขั้นตอนที่ได้กล่าวมาในข้างต้นแล้วนั้นพบปัญหาดังนี้

2.6.1 จำนวนรายที่ส่งตรวจสอบมีมากจนเกินกำลังของเจ้าหน้าที่ทำให้งานล้นมือ

2.6.2 ผู้ประกอบการรายเดิมถูกตรวจซ้ำแล้วซ้ำอีกทำให้เกิดความไม่พอใจ

2.6.3 จำนวนรายของผู้ประกอบการที่ใช้หลักเกณฑ์ดังกล่าวคัดเลือกมามีจำนวนของผู้ไม่เลี่ยงภาษีมากกว่าผู้ที่เลี่ยงภาษีทำให้สูญเสียทรัพยากรไปอย่างมากในการดำเนินการ

2.6.4 เนื่องจากหลักเกณฑ์ที่ใช้ในการคัดเลือกรายเพื่อสุ่มตรวจนั้น จะเห็นว่าเป็นหลักเกณฑ์ที่เกิดจากความคิดหรือความรู้สึกของบุคคลกลุ่มใดกลุ่มหนึ่งเท่านั้น จึงไม่

สามารถที่จะยืนยันได้ว่าหลักเกณฑ์ดังกล่าวนั้นถูกต้องหรือไม่

ซึ่งหากนำเอาดาต้าไมนิ่ง(Data Mining) มาใช้แล้วจะช่วยในการแก้ปัญหาที่เกิดขึ้นได้บางข้อเท่านั้น โดยปัญหาที่สามารถช่วยแก้ไขได้ก็คือ

- สามารถใช้ประวัติตามข้อมูลเก่าสร้างความสัมพันธ์เพื่อเป็นหลักเกณฑ์ในการคัดเลือกรายเพื่อตรวจสอบ ทำให้หลักเกณฑ์ที่ได้มีความน่าเชื่อถือขึ้น
- สามารถคัดเลือกเฉพาะรายที่มีโอกาสเลี่ยงภาษีจริงๆ โดยใช้หลักเกณฑ์ที่ได้จากการไมนิ่ง (mining) ในการคัดเลือกรายเพื่อส่งตรวจ ทำให้สามารถใช้ทรัพยากรได้อย่างมีประสิทธิภาพขึ้น

2.7 วัตถุประสงค์ของการนำดาต้าไมนิ่ง(Data Mining) เข้ามาช่วยในกระบวนการตรวจสอบภาษีมูลค่าเพิ่ม

จากที่ได้เคยทราบมาในเบื้องต้นแล้วว่าดาต้าไมนิ่ง(Data Mining)นั้นเป็นการค้นหาความหมายที่แอบแฝงอยู่ในฐานข้อมูลและความสัมพันธ์ของข้อมูลทั้งหมดที่ยังไม่เคยมีใครรู้มาก่อนเพื่อให้ได้สารสนเทศที่ถูกต้อง ใช้งานได้และมีประโยชน์มาช่วยสนับสนุนการตัดสิน ด้วยคุณสมบัตินี้เองจึงเป็นเหตุจูงใจที่จะนำเอาดาต้าไมนิ่ง(Data Mining) มาใช้กับกระบวนการในการคัดเลือกรายผู้ประกอบการที่ยื่นแบบภาษีมูลค่าเพิ่ม (ภ.พ.30) โดยมีวัตถุประสงค์ดังนี้

2.7.1 เพื่อให้ขบวนการของการทำดาต้าไมนิ่ง(Data Mining) หารูปแบบความสัมพันธ์ระหว่างข้อมูล ด้วยเงื่อนไขต่างๆ ที่คนไม่สามารถคาดการณ์ได้

2.7.2 เพื่อเพิ่มประสิทธิภาพของการคัดเลือกรายผู้ประกอบการในการสุ่มตรวจภาษีมูลค่าเพิ่ม

2.7.3 เพื่อลดจำนวนรายของการส่งตรวจสอบผู้ประกอบการที่มีอัตราเสี่ยงต่อการหลีกเลี่ยงการชำระภาษีมูลค่าเพิ่ม

จากที่กล่าวมาในบทนี้จะทำให้มองเห็นถึงปัญหาขององค์กรที่มีความจำเป็นจะต้องนำเอาเทคนิคอย่างใดอย่างหนึ่งเข้ามาช่วยให้การทำงานขององค์กรดีขึ้นโดยได้ทำการเลือกเอาทฤษฎีของดาต้าไมนิ่งเข้ามาช่วยเพื่อที่จะทำให้การทำงานขององค์กรดีขึ้น โดยทฤษฎีและส่วนรายละเอียดของดาต้าไมนิ่งจะได้กล่าวในต่อไป

บทที่ 3
ดาต้าไมนิ่ง และ ทฤษฎีที่เกี่ยวข้อง

เมื่อพูดถึงเครื่องมือที่ทันสมัยและกำลังเป็นที่รู้จักในปัจจุบันนี้คงหลีกหนีไม่พ้นเครื่องมือที่เรียกว่าดาต้าไมนิ่ง(Data Mining) เพราะเป็นเครื่องมือที่มีประสิทธิภาพในการค้นหาสารสนเทศที่มีประโยชน์ออกมาจากฐานข้อมูลอันมหาศาลในยุคโลกาภิวัฒน์(gobolization) นี้ แต่ทั้งนี้ก็ยังมีคนอีกมากมายที่ยังไม่รู้ว่าดาต้าไมนิ่ง(Data Mining) คืออะไร และจะมีประโยชน์อย่างไรต่อพวกเขาเหล่านั้น ดังนั้นในบทนี้ก็จะกล่าวให้ทราบเกี่ยวกับคำว่าดาต้าไมนิ่ง(Data Mining)ว่าคืออะไรและสามารถนำไปใช้ประโยชน์ได้อย่างไร

3.1 กำเนิดของดาต้าไมนิ่ง

ดาต้าไมนิ่ง(Data Mining) นั้นเป็นการหาความหมายที่แอบแฝงอยู่ในฐานข้อมูลและความสัมพันธ์ของข้อมูลทั้งหมดที่ยังไม่เคยมีใครรู้มาก่อน เพื่อให้ได้สารสนเทศที่มีประโยชน์มาช่วยสนับสนุนการตัดสินใจ ซึ่งในปัจจุบันความสนใจเรื่องดาต้าไมนิ่ง(Data Mining) เพิ่มขึ้นจากเดิมโดยเกิดจากอิทธิพล 2 อย่าง คือ [1]

3.1.1 ความจำเป็นที่ต้องเอาไมนิ่ง(Mining) มาใช้(Drivers) เพื่อค้นหาพฤติกรรมของผู้บริโภค เพื่อค้นหากลยุทธ์เชิงธุรกิจที่จะใช้ในการแข่งขันขึ้น เพื่อหาส่วนแบ่งทางการตลาดการเพิ่มขึ้น และ เพื่อค้นหาสารสนเทศ(information) ที่เก็บในฐานข้อมูล (DataBase) ให้ได้

3.1.2 การมุ่งที่จะทำให้เกิดเป็นจริงขึ้นมา (Enablers) โดยเป็นการพัฒนาด้านเทคนิคชั้นสูงเกี่ยวกับการวิจัยเรื่องการเรียนรู้ของเครื่องจักรกล, ฐานข้อมูล และเทคโนโลยีเสมือนจริง

3.2 สาเหตุที่ทำให้มีดาต้าไมนิ่ง

เมื่อการวิจัยและเทคโนโลยีสารสนเทศมีความก้าวหน้าขึ้น จึงทำให้เกิดการพัฒนาเรื่องของการตัดสินใจทางธุรกิจ โดยสิ่งแวดล้อมที่มีผลต่อการเปลี่ยนแปลงของธุรกิจซึ่งมีดังนี้: [1]

3.2.1 รูปแบบพฤติกรรมของผู้บริโภค(Customer behavior patterns) พฤติกรรมของผู้บริโภคมีการเปลี่ยนแปลงไป ซึ่งจะเกิดจากการปรับเปลี่ยนตามสภาพแวดล้อม

หรือ สภาพของเศรษฐกิจ อีกทั้งลูกค้ายังมีความต้องการและมีการรับรู้ข่าวสารได้รวดเร็วขึ้น จึงทำให้ผู้บริโภคซื้อสินค้าได้รวดเร็วตามที่ต้องการ

3.2.2 สภาวะตลาดที่อิ่มตัว(Market saturation) เกิดขึ้นจากปริมาณของการเสนอขายสินค้ามีมากเกินกว่าความต้องการของผู้บริโภคเป็นผลให้การขยายส่วนแบ่งทางการตลาดมีโอกาสน้อยมาก ทำให้คู่แข่งขันต้องปรับตัวเองเพื่อสนองตอบความต้องการของลูกค้าให้ได้มากที่สุด

3.2.3 การเพิ่มขึ้นของสินค้า (Increased commoditization) สินค้าและบริการหลายชนิดมีการเพิ่มที่แตกต่างจากสายของสินค้าชนิดนั้นๆ เพราะว่ามีตลาดกลุ่มใหม่ๆ เกิดขึ้นตลอดเวลา เนื่องจากว่าพฤติกรรมของผู้บริโภค แต่ละกลุ่มมีความชอบที่แตกต่างกัน ทำให้เกิดพฤติกรรมการบริโภคแบบใหม่ซึ่งหลบซ่อนอยู่

3.2.4 แนวโน้มทางการตลาดภายใต้สภาวะแรงกดดัน(Traditional marketing approaches under pressure) เกิดจากจำนวนข่าวสารที่เพิ่มมากขึ้นแต่การดำเนินการเกี่ยวกับฐานข้อมูล เริ่มไม่มีประสิทธิภาพขณะที่ความต้องการของลูกค้าเพิ่มมากขึ้น

3.2.5 เวลา(Time To Market) ถูกให้ความสำคัญเพิ่มมากขึ้นเพราะว่าคู่แข่งขันในตลาดมีมากขึ้น องค์กรแต่ละองค์กรจึงพยายามหาทางที่จะทำให้สินค้าไปสู่ผู้บริโภคเร็วที่สุด ซึ่งหากว่าใครดำเนินการก่อนย่อมได้ผลประโยชน์มากกว่า

3.2.6 วงจรชีวิตของผลิตภัณฑ์สั้นลง (Shorter product life cycle) เนื่องจากผลิตภัณฑ์ต่างๆ สามารถหาซื้อจากตลาดได้อย่างรวดเร็วและพฤติกรรมของผู้บริโภคมีการเปลี่ยนแปลงอย่างรวดเร็วมากทำให้อายุของสินค้ามีช่วงชีวิตที่สั้นผล เป็นผลให้ผู้ใช้ต้องซื้อสินค้าชนิดนั้นๆ บ่อยขึ้นและคนขายก็มีเวลาในการทำกำไรน้อยลง

3.2.7 การเพิ่มขึ้นของคู่แข่งและความเสี่ยงของธุรกิจ(Increased competition and business risks) มีมากขึ้นโดยเมื่อพิจารณาแล้วพบว่าแนวโน้มของสินค้าจะมีการซื้อขายกันทั่วโลกโดยผ่านทางอินเตอร์เน็ตทำให้ยากที่จะเก็บข้อมูลต่างๆ อีกทั้งแนวโน้มการเปลี่ยนแปลงของลูกค้าก็เป็นไปอย่างรวดเร็วทำให้ธุรกิจเกิดความเสี่ยงอย่างไม่เคยเป็นมาก่อน

3.3 ความหมายของดาต้าไมนิ่ง

ความหมายของดาต้าไมนิ่ง(Data Mining) นั้นเป็นเรื่องยากสำหรับการทำความเข้าใจ ซึ่งไม่ได้มีคำจำกัดความเพียงหนึ่งเดียว แต่ยังพบว่าคำจำกัดความของดาต้าไมนิ่ง(Data Mining) มีมากมายจากหลายแหล่ง ซึ่งพบว่ามีความหมายในทำนองเดียวกัน โดยเป็นที่รับรองทางสากลโดย ทั่วๆ

ไปมีคำจำกัดความว่า

ดาต้าไมนิ่ง(Data Mining) คือกระบวนการของการกลั่นกรองสารสนเทศที่ไม่มีใครรู้มาก่อน โดยสารสนเทศที่ได้ต้องถูกต้อง และเอาไปใช้งานได้ โดยมีการเคลื่อนไหวของข่าวสารจากฐานข้อมูลขนาดใหญ่ และ ใช้ข่าวสารเพื่อไปทำการตัดสินใจทางธุรกิจ [1]

ดาต้าไมนิ่ง(Data Mining) คือ กระบวนการสืบค้นข้อมูลสำคัญอันเป็นประโยชน์ต่อการดำเนินธุรกิจออกจากกองข้อมูลขนาดมหาศาลที่จัดเก็บอยู่ในแหล่งข้อมูลของแต่ละองค์กร [4]

ดาต้าไมนิ่ง(Data Mining) คือกระบวนการในการสืบค้นหารูปแบบความสัมพันธ์ที่มีประโยชน์ของข้อมูลในฐานข้อมูลขนาดใหญ่ [3]

ซึ่งจากคำจำกัดความของดาต้าไมนิ่ง(Data Mining) ดังกล่าวข้างต้นได้ให้ความเข้าใจที่ลึกซึ้งถึงจุดสำคัญของดาต้าไมนิ่ง(Data Mining) และช่วยขยายความแตกต่างทางพื้นฐานระหว่างดาต้าไมนิ่ง(Data Mining) และแนวโน้มของการวิเคราะห์ข้อมูล ดาต้าไมนิ่ง(Data Mining) มีจุดมุ่งหมายที่จะค้นพบสิ่งที่ต้องการจากข้อมูลที่มี โดยปราศจากกฎเกณฑ์ข้อสมมติ ซึ่งมี 3 สิ่งที่สำคัญคือ

สิ่งแรก ข่าวสารที่ค้นพบต้องไม่มีใครรู้มาก่อน และเป็นความจริงที่ไม่เคยมีใครนำมาเป็นสมมติฐานมาก่อนด้วย ดังนั้นสิ่งที่ซ่อนเร้นอยู่ภายในข้อมูลจึงมีค่ามากตัวอย่างเช่นเรื่องของเบียร์กับผ้าอ้อม ถูกค้นพบว่ามีการเกี่ยวพันกันอย่างเหนียวแน่น โดยจะพบว่าในสุดสัปดาห์ผู้ซื้อที่เป็นผู้ชายต้องกักตุนของจำเป็นของเด็กและในเวลาเดียวกันก็จะซื้อของส่วนตัวที่พวกเค้าต้องการด้วย

สิ่งที่สอง ข่าวสารที่ได้ต้องใช้ได้และมีความถูกต้องซึ่งต้องมีการตรวจสอบที่เพียงพอเพราะข้อมูลมากมายที่รวบรวมมานั้นมาจากแหล่งข้อมูลหลายแหล่ง ซึ่งอาจจะพบบางสิ่งที่น่าสนใจโดยจะทำให้แก้ไขได้ทันเวลาตัวอย่างเช่น ความผิดพลาด จากการนำข้อมูลรายการซื้อของลูกค้ากับจำนวนของรายการสินค้าทั้งหมดที่มีอยู่มารวมกันทำให้ข้อมูลไม่ถูกต้อง

สิ่งที่สาม สารสนเทศที่ได้ต้องสามารถเคลื่อนไหวได้ นั่นคือต้องถ่ายทอดสิ่งที่เป็นประโยชน์สำหรับธุรกิจได้ ดังเช่น ในกรณีของร้านขายบ่ลึก ผลของการวิเคราะห์การจัดวางเบียร์และผ้าอ้อมไว้ด้วยกันในร้านทำให้แน่ใจได้ว่าสินค้าทั้ง 2 รายการไม่ต้องลดราคาลงในเวลาเดียวกัน.

3.4 โอเปอเรชั่นของดาต้าไมนิ่ง

เมื่อพูดถึงโอเปอเรชั่นในการทำงานของดาต้าไมนิ่ง(Data Mining) แล้วจะพบว่ามี 4 โอเปอเรชั่นดังนี้

3.4.1 Predictive Modeling เป็นโมเดลที่ใช้ในการทำนายพฤติกรรมของมนุษย์ ซึ่งมีอยู่ 2 เทคนิคคือ

3.4.1.1 Classification เป็นการแบ่งกลุ่มโดยมีกลุ่มอยู่แล้ว แล้วใส่ข้อมูลเข้าไป เช่น ใช้เทคนิคนี้ตรวจดูบัตรเครดิตว่าคนนี้จะโกงหรือเปล่า ต้องมีกำหนดขนาด(scale) ว่าดี , ไม่ดี , ปานกลาง ใช้สำหรับผู้ที่จะมาของกู้หรือพวกประกันภัยดูว่าลักษณะลูกค้าประเภทไหนมีแนวโน้มว่าจะย้ายไปทำกับบริษัทอื่น

3.4.1.2 Value predictive ใช้เพื่อทำนายค่าของเหตุการณ์ในอนาคต เช่น ทำนายค่าหุ้นเป็นต้น

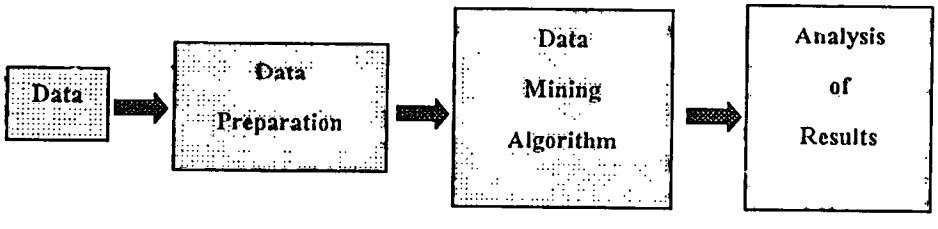
3.4.2 Database Segmentation เป็นการแบ่งกลุ่มของฐานข้อมูลซึ่งสมาชิกในแต่ละกลุ่มจะมีค่าความเหมือนกันอยู่ จากนั้นก็เอาข้อมูลนั้นมาวิเคราะห์หาความหมาย เช่น เรามีเรคคอร์ดอยู่ 2000 เรคคอร์ด เมื่อผ่านอัลกอริทึ่มนี้ก็จะทำการแบ่งกลุ่มของเรคคอร์ดออกเป็นประเภทต่างๆ โดยที่ยังไม่เคยมีกลุ่มอยู่เลย

3.4.3 Link Analysis เป็นการวิเคราะห์ความสัมพันธ์ระหว่างข้อมูล เช่น สินค้าในซุปเปอร์มาเก็ต คนซื้อเสื้อเชิ้ตมักจะซื้อไทด้วยหรือเปล่า

3.4.4 Deviation Detection เป็นความพยายามหาสิ่งที่แปลกปลอมออกจากกลุ่มของมัน ส่วนมากอาศัยการ Plot กราฟ แล้วดูว่าจุดมันมีการกระจายออกหรือไม่

3.5 ขั้นตอนในการทำดาต้าไมนิ่ง

ขั้นตอนในการทำดาต้าไมนิ่ง(Data Mining) นั้นมีมากมายหลายขั้นตอน ซึ่งไมนิ่งเป็นเพียงขั้นตอนหนึ่งเท่านั้น โดยที่ขั้นตอนของการทำดาต้าไมนิ่ง(Data Mining) นั้นเป็นขั้นตอนในการสร้างโมเดลจากกลุ่มของข้อมูล (data set) เพื่อสร้างรูปแบบและความเกี่ยวข้องกันของกลุ่มข้อมูลเพื่อใช้ในการทำนายบนข้อมูลนั้นๆ กระบวนการของดาต้าไมนิ่ง(Data Mining) เป็นขั้นตอนที่มีการย้อนกลับไปกลับมาได้ตลอดเวลาและแต่ละขั้นตอนยังต้องอาศัยเวลาในการทำงานแตกต่างกันไปด้วย แต่เมื่อพิจารณาอย่างคร่าวๆแล้วสามารถที่จะแยกขั้นตอนของการทำดาต้าไมนิ่ง (Data Mining) ได้ ดังแสดงให้เห็นในภาพที่3.1 โดยมีรายละเอียดของแต่ละขั้นตอนดังนี้



ภาพที่ 3.1 ขั้นตอนการทำ Data Mining[1]

3.5.1 Data Preperation เมื่อมีการนำเอาข้อมูลซึ่งอาจจะได้มาจากฐานข้อมูล คลังข้อมูลหรือ ไฟล์ข้อมูลเข้ามาสู่กระบวนการ เริ่มแรกต้องทำการกำหนดวัตถุประสงค์ (Objective) เสียก่อนว่าต้องการทำอะไร ซึ่งวัตถุประสงค์นี้ต้องมีความชัดเจนเพราะว่าถ้าไม่มีความชัดเจนหรือไม่รู้ว่าวัตถุประสงค์ของการทำคืออะไรแล้วก็ไม่รู้จะทำไปทำไมผลลัพท์ที่ได้ก็จะไม่มีความหมาย จึงต้องมีการกำหนดปัญหาขอบเขตของปัญหาที่ต้องการขึ้นมาจากนั้นก็ทำการวิเคราะห์ถึงตัวปัญหา แล้วจึงทำความเข้าเกี่ยวกับปัญหาซึ่งมีความ สำคัญเป็นอันดับแรก แล้วจึงค่อยเริ่มทำในขั้นตอนอื่นๆต่อไปแต่ไม่ได้หมายความว่าทุกปัญหาจะสามารถแก้ได้ด้วยเทคนิคดาต้าไมนิ่ง (Data Mining) ซึ่งอาจจะนำมาสู่ความเข้าใจที่ผิดได้ ดังนั้นการกำหนดปัญหาไม่ถูกต้องย่อมยากที่จะนำไปสู่ความสำเร็จในการแก้ปัญหา เมื่อรู้วัตถุประสงค์การทำแล้วก็มาทำการจัดเตรียมข้อมูลก่อนโดยในขั้นตอนนี้ประกอบด้วยขั้นย่อยๆ 3 ขั้นคือ

3.5.1.1 การคัดเลือกข้อมูล (Selection Data) เมื่อกำหนดหนดวัตถุประสงค์ได้แล้วก็ต้อง คัดเลือกว่าจะใช้ข้อมูลอะไร จากส่วนไหนเพื่อที่จะทำให้บรรลุถึงวัตถุประสงค์ที่ต้องการ ซึ่งข้อมูลที่ได้อาจได้มาจากฐานข้อมูลของการปฏิบัติงาน จากแฟ้มข้อมูล หรือจากฐานข้อมูลหลายๆ แห่ง โดยจะต้องทำให้ข้อมูลเหล่านั้นอยู่ในรูปแบบเดียวกันเสียก่อน ซึ่งการเลือกข้อมูลนั้นจะแตกต่างกันไปตามวัตถุประสงค์ของแต่ละธุรกิจที่ได้มีการกำหนดไว้แล้วตั้งแต่ต้น โดยตัวแปรที่ถูกเลือกมาใช้แต่ละตัวนั้นจะต้องทำความเข้าใจด้วยว่าแต่ละตัวหมายความว่าอะไร อีกทั้งยังต้องมีคำอธิบายเกี่ยวกับตัวแปร ชนิดของข้อมูล และ ค่าที่เป็นไปได้ของแต่ละตัวแปรด้วย ซึ่งตัวแปรมีอยู่ 2 ชนิดคือ

- ตัวแปรแบบหมวดหมู่(Categorical) แบ่งเป็น
  - Nomonal Variable เป็นตัวแปรอ้างถึงที่ค่าของข้อมูลไม่มีลำดับ เช่น สถานะการสมรส (โสด หย่า แต่งงาน) เป็นต้น
  - Ordinal Variable เป็นตัวแปรอ้างถึงที่ค่าของข้อมูลมีลำดับ เช่น ระดับของลูกค้า (ดี ปานกลาง ไม่ดี) เป็นต้น
- ตัวแปรแบบควอนติเตทีฟ(Quantitative) แบ่งเป็น
  - Continuous เป็นค่าที่ต่อเนื่อง เช่น 10.2 1.3 เป็นต้น
  - Discrete เป็นค่าที่ไม่ต่อเนื่อง เช่น 1 2 4 เป็นต้น

3.5.1.2 การเตรียมข้อมูล (Preprocessed Data) เมื่อคัดเลือกข้อมูลที่ต้องการใช้ทำงานได้แล้วก็เอาข้อมูลที่คัดเลือกมานั้น มาทำความสะอาด โดยการตรวจสอบข้อมูลเหล่านั้นดูว่า มีความถูกต้องหรือไม่ ซึ่งความผิดพลาดของข้อมูลอาจจะเกิดขึ้นได้ขณะที่มีการรวบรวมข้อมูลจากฐานข้อมูลหลายๆแหล่ง ซึ่งหากมีข้อมูลขาดหายไปก็ตัดส่วนนั้นทิ้งไป หรือ หาค่าเฉลี่ยเพื่อใส่ข้อมูลลงไป ซึ่งนักวิเคราะห์ที่ดีควรจะนึกถึงข้อนี้ด้วย

3.5.1.3 การแปลงข้อมูล (Transform Data) ขั้นตอนนี้เป็นการพิจารณาถึงข้อมูลที่จะนำมาใช้ว่าเป็นข้อมูลประเภทใด เพราะบางอัลกอริทึมที่เราใช้จะรับข้อมูลประเภทหนึ่งแต่ข้อมูลที่เข้ามาอาจจะเป็นข้อมูลอีกประเภทหนึ่งก็ได้ จึงจำเป็นต้องมีการแปลงให้เหมาะสมกับอัลกอริทึมที่จะใช้ เช่น Neural network รับข้อมูลเป็นตัวเลขแต่หากข้อมูลที่เข้ามาเป็นข้อมูลแบบหมวดหมู่(Categorical) ก็ต้องทำการแปลงให้เป็นตัวเลขเสียก่อน โดยการแปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปใช้ในอัลกอริทึมได้นั้นมีอยู่หลายแบบ เช่น

- ข้อมูลแบบหมวดหมู่(Categorical) อาจจะแบ่งค่าของตัวแปรให้เป็นช่วงๆ เช่น การแปลงเงินเดือน

  0 ----- 1 แสน

  1 แสน ----- 5 แสน

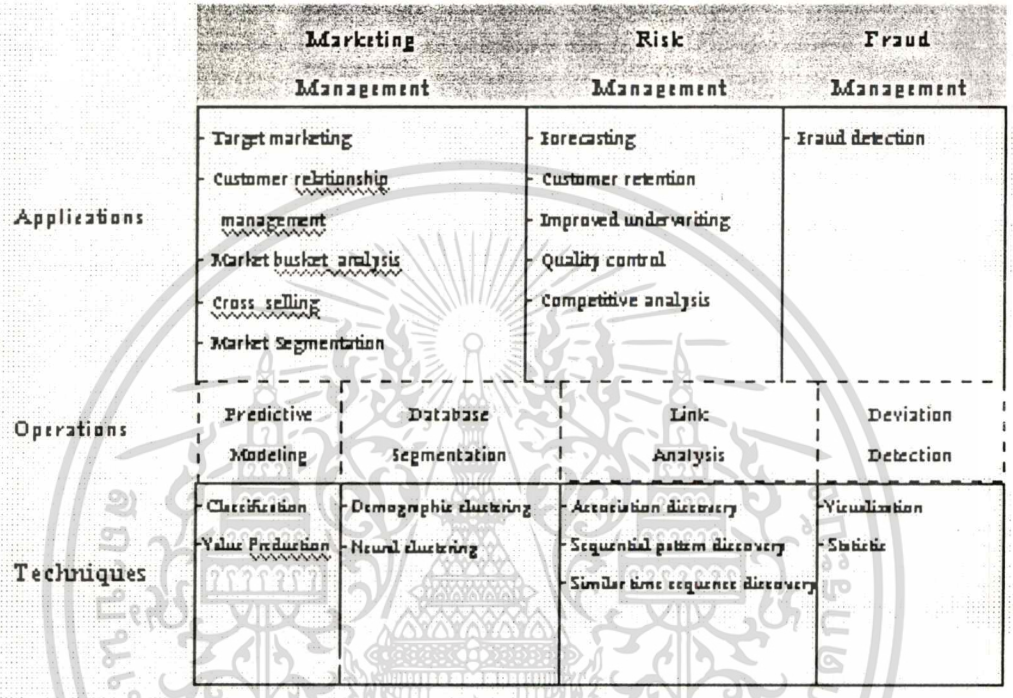
  5 แสน ----- 1 ล้าน

  1 ล้าน ----- 10 ล้าน

- วันออฟเอ็นโคดดิ้ง (One of N coding) เป็นเทคนิคของการแปลงข้อมูลแบบหมวดหมู่(Categorical) ให้เป็นข้อมูลที่เป็นตัวเลข เช่น ชนิดของรถ Ford ก็แปลงเป็น 100 เป็นต้น
- เทคนิคการสเกล (Scaling Technique) เป็นลักษณะ linear mapping คือ เป็นค่าเส้นตรงที่มีสโลป(slope) คงที่

3.5.2 อัลกอริทึมของ Data mining หลังจากที่ทำการเตรียมข้อมูลเป็นที่เรียบร้อยแล้ว จากนั้นก็เป็นขั้นตอนของการเลือกใช้อัลกอริทึมที่เหมาะสม โดยการเลือกใช้อัลกอริทึมนั้นจะขึ้นอยู่กับลักษณะของปัญหาและลักษณะของข้อมูล โดยอาจจะเลือกมากกว่า 1 อัลกอริทึมก็ได้ เพื่อที่จะได้นำผลลัพธ์ที่ได้มาเปรียบเทียบกัน โดยการทำงานหลักๆของ Data Mining มีอยู่ 4 Operation ได้แก่ แบบจำลองการทำนาย

(Predictive Modeling) การแบ่งส่วนฐานข้อมูล (Database Segmentation) การวิเคราะห์ความสัมพันธ์ (Link Analysis) และการตรวจสอบค่าเบี่ยงเบน (Deviation detection) ซึ่งหากจะเลือกนำเทคนิคแบบใดของ Data Mining ไปใช้จะต้องพิจารณาให้เหมาะสมกับลักษณะของงานด้วยดังภาพที่ 3.2



ภาพที่3.2 Data Mining Application และOperation และ Techniques ที่สนับสนุน [1]

เมื่อทำการเลือกอัลกอริทึมได้แล้วก็นำเอาอัลกอริทึมนั้นมาประมวลผลกับข้อมูลที่ได้เตรียมไว้ ซึ่งจากการทำ mining ก็จะได้ความรู้หรือสารสนเทศที่มีประโยชน์ออกมา

3.5.3 การวิเคราะห์ผลลัพธ์ (Analysis of Results) เป็นการวิเคราะห์ผลลัพธ์ที่ได้ออกมา เพราะว่าผลลัพธ์ที่ได้อาจจะไม่สามารถนำไปใช้ตรงๆได้ ซึ่งเป็นการตีความหมายจากผลลัพธ์ซึ่งอาจจะอยู่ในรูปของตัวเลขก็ได้ เหตุที่ต้องมีการตีความหมายเพราะว่า ผลลัพธ์ที่ได้ในบางครั้งอาจจะมีความผิดพลาดหากไม่มีการตีความหมายเสียก่อนจะทำให้นำผลลัพธ์ที่ผิดๆไปใช้จะทำให้เกิดมีข้อผิดพลาดไปด้วย จากนั้นขั้นตอนนี้เป็นการนำเอาแบบจำลองที่ได้ไปทดสอบกับข้อมูลชุดอื่นที่ไม่ใช่เป็นชุดข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง เพื่อที่จะนำเอาผลลัพธ์ที่ได้มาเปรียบเทียบกับผลตามแบบจำลองว่ามีความแม่นยำหรือไม่และยอมรับได้หรือไม่ ซึ่งถ้าไม่

สามารถยอมรับได้ ก็แก้ไข โดยการเพิ่มจำนวนของข้อมูลให้มากขึ้นอีก หรือเปลี่ยนไปใช้อัลกอริทึ่มอื่นแทน ดังนั้นกระบวนการทำ Data Mining นี้จึงเป็นกระบวนการที่ต้องมีการตรวจสอบและทำซ้ำอยู่ตลอดเวลาเนื่องจากสิ่งแวดล้อมมีการเปลี่ยนแปลงอยู่เสมอ

จากที่ได้กล่าวมาแล้วข้างต้นว่าโอเปอเรชั่นของ Data Mining มีมากมาย แต่ในโครงการฉบับนี้จะขอนำเฉพาะโอเปอเรชั่นของ Predictive modeling มาใช้เพื่อทำนายผู้ที่มีแนวโน้มในการหลีกเลี่ยงภาษีมูลค่าเพิ่มเท่านั้นซึ่งจะได้กล่าวถึงในหัวข้อถัดไป

3.6 โอเปอเรชั่น Predictive Modeling [1]

Predictive Modeling นั้นได้กล่าวไว้แล้วในข้างต้นว่าเป็นรูปแบบการทำนายพฤติกรรมของมนุษย์ ซึ่งหลักเกณฑ์การคัดเลือกรายผู้ประกอบการเพื่อสุ่มตรวจสอบภาษีมูลค่าเพิ่มของกรมสรรพากรนั้น ก็อาศัยปัจจัยต่างๆที่คาดว่าจะมีผลให้ผู้ประกอบการมีแนวโน้มที่จะหลีกเลี่ยงภาษีเป็นเกณฑ์ในการพิจารณาคัดเลือกรายเพื่อสุ่มตรวจด้วยเหมือนกัน เพียงแต่ปัจจัยดังกล่าวที่นำมาเป็นเกณฑ์ในการพิจารณานั้น อาจจะยังไม่ครอบคลุมถึงความสัมพันธ์ที่แฝงอยู่ในข้อมูลของแบบแสดงรายการภาษีมูลค่าเพิ่มดังกล่าวได้ทั้งหมด โดยปัจจัยแต่ละตัวที่นำมาเป็นเกณฑ์ในการพิจารณานั้นก็จะสามารถค้นหาผู้ประกอบการที่มีแนวโน้มว่าจะหลีกเลี่ยงภาษีได้ แต่ยังไม่สามารถค้นหาได้ทั้งหมด ทำให้ผู้ประกอบการบางรายก็ยังสามารถเล็ดรอดจากเงื่อนไขของการคัดเลือกรายเพื่อสุ่มตรวจภาษีมูลค่าเพิ่ม (ภ.พ.30) ได้

ใน Data Mining เราใช้รูปแบบของการทำนายเพื่อที่จะวิเคราะห์ฐานข้อมูลที่มีอยู่ในปัจจุบันเพื่อที่จะได้กำหนดคุณลักษณะบางอย่างเกี่ยวกับข้อมูลเหล่านั้น ซึ่งข้อมูลนั้นจะต้องมีความสมบูรณ์ และทำการสังเกตุโดยการใช้หลักเหตุผล โดยมีคำตอบของทางแก้ไขไว้เรียบร้อยแล้วดังนั้นอัลกอริทึ่มที่ทำงานแบบนี้จึงถูกเรียกว่าการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) คือต้องมีการกำหนดรูปแบบของอินพุตและเอ้าท์พุตมาก่อนจากข้อมูลตัวอย่าง โดยทางกายภาพแล้วรูปแบบนี้สามารถทำโดยใช้กฎของ IF THEN และใช้ความสามารถของ SQL หรือ การเขียนด้วยภาษา C ก็ได้

โมเดลของการพัฒนาแบ่งการพัฒนาเป็น 2 เฟส คือ เฟสของการฝึกหัด(training) หรือการเรียนรู้ (Learning)และ เฟสของการทดสอบ (testing) หรือใช้งานจริง โดยที่ Training เป็นการสร้างแบบจำลองใหม่โดยการใช้ข้อมูลเก่าๆ และ Testing เป็นการการทดสอบแบบจำลองที่สร้างขึ้นมาใหม่นั้น โดยข้อมูลที่จะนำมาทำการทดสอบนั้นต้องไม่ใช่ข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง จึงจะให้ผลที่มีความถูกต้องและมีประสิทธิภาพ ซึ่งจะเห็นได้ว่า Training เป็นการทำงาน

กับข้อมูลส่วนใหญ่ของข้อมูลที่ใช้ทั้งหมด ขณะที่ Testing ทำงานกับเปอร์เซ็นส่วนน้อยของข้อมูลที่นำมาใช้

แบบจำลองการทำนาย(Predictive modeling) มีความเหมาะสมกับอุตสาหกรรมหลายๆ อุตสาหกรรม โดยเป็นแอพพลิเคชั่นทางธุรกิจที่สนับสนุนในเรื่องรักษาลูกค้า , การอนุมัติการให้เครดิต และ เป้าหมายทางการตลาด ซึ่ง Predictive Modeling มีเทคนิค 2 อย่าง คือ Classification และ Value prediction ซึ่ง Classification เป็นแบบจำลองการทำนายที่ถูกใช้เพื่อที่จะกำหนดคลาสเฉพาะสำหรับแต่ละเรคคอร์ดในฐานข้อมูล ซึ่งคลาสต้องเป็นเซ็ตของความเป็นไปได้ที่มีค่าแน่นอน โดยมีการกำหนดค่าของคลาสไว้ก่อนล่วงหน้าแล้ว ส่วน Value prediction เป็นแบบจำลองการทำนายที่ถูกใช้เพื่อที่จะกำหนดค่าตัวเลขที่ต่อเนื่อง ซึ่งมีความสัมพันธ์กับเรคคอร์ดในฐานข้อมูล ถึงแม้ว่าทั้ง 2 อย่างจะมีวัตถุประสงค์เหมือนกัน คือเพื่อที่จะทำการศึกษาเรื่องการคาดเดาเกี่ยวกับตัวแปรที่น่าสนใจบางอย่าง แต่ทั้ง 2 เทคนิคนี้ก็มีความแตกต่างกันตรงลักษณะของตัวแปรที่ใช้ในการทำนาย

จากที่ได้กล่าวมาในข้างต้นจะเห็นได้ว่า Classification ก็เป็นเทคนิคที่สามารถนำมาใช้ในการค้นหาผู้ที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มได้อย่างหนึ่ง โดยจะพบว่า Classification นั้นมีเทคนิคที่สามารถนำมาใช้งานได้อยู่ 2 เทคนิคคือ Tree Induction และ Neural Induction ทั้ง 2 เทคนิคนี้อยู่บนพื้นฐานของ Supervised Learning ซึ่งแต่ละเทคนิคก็ยังมีอัลกอริทึ่มที่จะนำมาใช้งานได้อีกมากมายอย่างเช่นเทคนิคแบบ Neural มีอัลกอริทึ่ม การแพร่ย้อนกลับ (Back Propagation) ซึ่งเป็นอัลกอริทึ่มที่ใช้สำหรับการ Training หลายๆ layer (Multi layer) เพื่อที่จะปรับระดับค่ากลางของความผิดพลาดให้น้อยลง ซึ่งแสดงให้เห็นถึงรูปแบบสถาปัตยกรรมของโหนด และ การเชื่อมต่อน้ำหนัก ส่วน Tree เป็นการสร้างแบบจำลองซึ่งแสดงให้เห็นได้ทั้งแบบการตัดสินใจแบบต้นไม้ (Decision Tree) หรือแบบกฎของ IF THEN โดยในที่นี้จะนำเสนอเทคนิคของ Decistion Tree มาใช้ในการค้นหาผู้มีแนวโน้มในการหลีกเลี่ยงภาษีมูลค่าเพิ่ม ซึ่งจะได้กล่าวในรายละเอียดในบทถัดไป

บทที่ 4
เทคนิคและอัลกอริทึ่มของ Decision Tree

ตามที่เคยได้กล่าวมาแล้วว่า เทคนิคที่ใช้ในการค้นหาผู้มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่มนั้นมีอยู่หลายเทคนิคที่จะนำไปใช้ในการทำงานซึ่งดีซีชั่นทรี(Decision Tree) เองก็เป็นอีกเทคนิคหนึ่งที่สามารถนำไปใช้ไป โดยมีรายละเอียดของคุณลักษณะดังจะกล่าวต่อไป

4.1 เทคนิคที่ใช้ในการคัดเลือกรายผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่ม [1]

ตามที่ได้ทราบกันอยู่แล้วว่าเทคนิคของดาต้าไมนิ่ง(Data mining) มีอยู่มากมายแต่ในที่นี้จะกล่าวถึงในส่วนของการแบ่งระดับ(Classification) ซึ่งเป็นการทำนายว่าสิ่งที่เราสนใจจะอยู่ในประเภทไหน หรือกลุ่มไหน โดยจะเห็นว่า การแบ่งระดับ(Classification) เป็นการจำแนกคลาสของเรคคอร์ดในฐานข้อมูล ซึ่งคลาสต้องเป็นเซ็ตของความเป็นไปได้ที่มีค่าแน่นอนโดยมีการกำหนดค่าของคลาสไว้ก่อนล่วงหน้าแล้ว โดยเป็นโมเดลที่ใช้ในการทำนาย ซึ่งใช้ข้อมูลในอดีตเพื่อสร้างโมเดลโดยเฟสในการสร้างโมเดล มี 2 เฟต คือ

4.1.1 เฟตของการเทรนนิ่ง(Training Phase) คือเฟสที่ใช้ในการสร้างโมเดล โดยสร้างจากการใช้ข้อมูลในอดีต

4.1.2 เฟตของการทดสอบ(Test Phase) คือเฟสของการตรวจสอบว่าโมเดลที่ได้ ดีหรือไม่ซึ่งข้อมูลที่ใช้ในการทดสอบต้องไม่ใช้ข้อมูลที่ใช้ในการสร้างโมเดล

ดังนั้นกระบวนการทำงานนี้จึงถูกเรียกว่าการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) คือต้องมีการกำหนดรูปแบบของอินพุตและเอ้าท์พุตมาก่อนจากข้อมูลตัวอย่าง โดย Classification มีเทคนิคที่ใช้ในการทำงานอยู่หลายเทคนิค แต่ในที่นี้จะขอกล่าวถึงรายละเอียดในส่วนของเทคนิคทรีอินดักชั่น(Tree Induction) เท่านั้น เทคนิคนี้อยู่บนพื้นฐานของการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) ซึ่งเป็นการสร้างแบบจำลองซึ่งแสดงให้เห็นได้ทั้งแบบดีซีชั่นทรี (Decision Tree) หรือแบบกฎของ IF THEN ซึ่งดีซีชั่นทรี(Decision tree) เป็นการจำแนกค่าของแอททริบิวเพื่อจัดเรียงเป็นต้นไม้(Tree) จากโหนดที่อยู่บนสุดของต้นไม้ (root node) ถึง โหนดที่อยู่ล่างสุดของต้นไม้ (leaf node) โดยแต่ละโหนดในต้นไม้จะทำการแตกกิ่งออกตามค่าต่างๆที่เป็นไปได้ของแอททริบิวนั้น เพื่อสร้างลำดับถัดไปของต้นไม้ กระบวนการนี้จะถูกทำซ้ำในต้นไม้ย่อยๆ จนกว่าจะถึงโหนดสุดท้ายของต้นไม้

4.2 ความหมายของดีซีชั่นทรี(Decision Trees)

ดีซีชั่นทรี(decision tree) คือ ต้นไม้(tree) ซึ่งในแต่ละกิ่งของโหนดแสดงให้เห็นถึงทางเลือกระหว่างจำนวนของทางเลือก และ แต่ละโหนดที่เป็นใบ(leaf node) แสดงให้เห็นถึงการจำแนกพวกหรือการตัดสินใจ ดีซีชั่นทรี(Decision tree) มีอัลกอริทึ่มที่ใช้ในการทำหลายอัลกอริทึ่ม แต่ในโครงการนี้จะนำเสนอ อัลกอริทึ่มของ ID3 โดย ID3 เป็นอัลกอริทึ่มพื้นฐานของดีซีชั่นทรี (Decision Trees) ซึ่งเป็นการเรียนรู้แบบมีเป้าหมาย (Supervised Learning) โดยการสร้างเป็น กฎในการแบ่งระดับ(classification rules) ในรูปแบบของดีซีชั่นทรี(decision tree) มันใช้เซ็ตของการเทรนนิ่ง(training set) เป็นอินพุตและทำการสร้างดีซีชั่นทรี(decision tree) โดยการแบ่งส่วนของ เซ็ตของการเทรนนิ่ง(training set) แอททริบิวจะถูกเลือกใช้เพื่อที่จะเป็นตัวแบ่งเซ็ตของข้อมูล และ tree จะถูกสร้างตามค่าย่อยๆของแอททริบิวนั้น ทำเช่นนี้จนกระทั่งสมาชิกทุกตัวมีคลาสเดียวกัน

ฟังก์ชั่นฮิวริสติค(heuristic function) ถูกใช้เพื่อเลือกแอททริบิวที่ดีที่สุดเพื่อแตกข้อมูลของเซ็ตที่ในในการเทรนนิ่ง(training set) ID3 เป็นอัลกอริทึ่มแบบกรีดี้(greedy algorithm) และ การเลือกแอททริบิวที่ไม่ดีจะมีผลกระทบต่อผลลัพท์ตอนสุดท้าย ในตอนแรก ID3 นั้นควบคุมจำนวนของค่าที่ไม่ต่อเนื่องที่มีจำนวนน้อยๆเท่านั้น แต่ต่อมาภายหลังได้มีการปรับปรุงให้สามารถควบคุมได้ทั้งค่าที่เป็นลำดับและแอททริบิวที่มีค่าต่อเนื่อง และอัลกอริทึ่ม ID3 ยังได้ผนวกความสามารถในการควบคุมเรื่องของสิ่งรบกวน(noise) ด้วย [10]

ตัวอย่างของการนำไปใช้ เช่น เราอาจจะมีดีซีชั่นทรี(decision tree) เพื่อช่วยกรมสรรพากรในการคัดเลือกรายผู้ประกอบการที่มีแนวโน้มในการหลีกเลี่ยงภาษีขึ้นมาตรวจสอบ หรือ ช่วยบริษัทการเงินในการตัดสินใจว่าควรจะให้กู้หรือไม่ โดยมีหลายอัลกอริทึ่มที่จะนำไปสร้างต้นไม้(implement tree) และ ID3 ก็เป็นหนึ่งในหลายๆอัลกอริทึ่มเหล่านั้น

ID3 เป็นอัลกอริทึ่มพื้นฐานโดยที่ ตัวอย่าง*(examples)* คือเซ็ตของข้อมูลที่ใช้ในการเรียนรู้ (training Exsample) แอททริบิวเป้าหมาย(*Target_atrributes*)คือ แอททริบิวที่ใช้ค่าของมันในการทำนายผลในต้นไม้(tree) และ แอททริบิว(*Attributes*)คือ แอททริบิวอื่นๆที่ใช้ในการสร้างโหนดในต้นไม้(Tree) และ ไม่ใช่ แอททริบิวเป้าหมาย(*Target_atrributes*) ซึ่งมีลักษณะของ อัลกอริทึ่มดังนี้ [8]

ID3(*Examples*, *Target_attribute*, *Attributes*)

Create a *Root* node for the tree

If all *Examples* are positive, Return the single-node tree *Root*, with label = +

If all *Examples* are negative, Return the single-node tree *Root*, with label = -

If *Attributes* is empty, Return the single-node tree *Root*, with label = most common value of *Target_attribute* in *Examples*

Otherwise **Begin**

A ::= the attribute from *Attributes* that best classifies *Examples* according to the Information Gain

The decision attribute for *Root* ::= A

For each possible value, v_i, of A,

Add a new tree branch below *Root*, corresponding to the test A = v_i

Let $Examples_{vi}$ be the subset of *Examples* that have value v_i for A

If $Examples_{vi}$ is empty

Then below this new branch add a leaf node with label = most common value of *Target_attribute* in *Examples*

Else below this new branch add the subtree

ID3($Examples_{vi}$, *Target_attribute*, *Attributes* - {A})

End

ซึ่ง ID3 เป็นอัลกอริทึ่มที่จะสร้างดีซีชั่นทรี(decision tree) โดยจะมี ตัวอย่าง(Examples) เป็นข้อมูลที่จะเข้าไปทดสอบเพื่อหาแอทริบิว(Attribute) มาสร้างเป็นโหนดในดีซีชั่นทรี(decision tree) โดย ID3 มีขั้นตอนต่างๆดังนี้

4.2.1 ทำการทดสอบว่า ตัวอย่าง(examples) ทุกตัวที่รับเข้ามามีค่าของ แอททริบิวเป้าหมาย(target attribute)เป็นบวกทั้งหมดหรือไม่ ถ้าใช่ก็จะคืนโหนดที่มีค่าเป็นบวกให้ แล้วก็จบการทำงาน แต่ถ้าไม่ใช่ก็ทำข้อต่อไป

4.2.2 ทำการทดสอบว่า ตัวอย่าง(examples) ทุกตัวที่รับเข้ามามีค่าของ แอททริบิวเป้าหมาย(target attribute)เป็นลบทั้งหมดหรือไม่ ถ้าใช่ก็จะคืนโหนดที่มีค่าเป็นลบให้แล้วก็จบการทำงาน แต่ถ้าไม่ใช่ก็ทำข้อต่อไป

4.2.3 ทำการทดสอบว่า แอททริบิว อื่นๆที่ไม่ใช่ แอททริบิวเป้าหมาย(target attribute)มีค่าหรือไม่ถ้าไม่มีค่าก็จบการทำงาน แต่ถ้ายังมีค่าอยู่ก็ทำกระบวนการดังนี้

4.2.3.1 หาค่า อินฟอร์เมชั่นเกรน(information gain)ของทุกๆ แอททริบิว ที่ไม่ใช่แอททริบิวเป้าหมาย(target attribute)ซึ่ง แอททริบิว ไหนมีค่า gain สูงสุด จะได้ถูกกำหนดเป็น โหนดบนสุด(root node) ตามสมการดังนี้

$$Gain(S,A) \equiv Entropy(S) - \sum_{\upsilon \in Values(A)} \frac{|S_\upsilon|}{|S|} Entropy(S_\upsilon) \quad \text{_____(1)}$$

โดยที่ Values(A) คือ ค่าที่เป็นไปได้ทั้งหมดของ แอททริบิว A

S_υ คือ ตัวอย่างทั้งหมดของค่าที่เป็นไปได้หนึ่งๆ ของ แอททริบิว A

Entropy(S) คือ เอ็นโทรปรี(entropy) ของการรวบรวม S แบบเดิม

$Entropy(S_\upsilon)$ คือ ค่าที่ถูกคาดหวังของ เอ็นโทรปรี(entropy) หลังจาก S ถูกแบ่งด้วยการใช้ แอททริบิว A

Gain(S,A)คือ ค่า อินฟอร์เมชั่นเกรน(information gain)ของ A

จากสมการ (1) สามารถหาค่าเอ็นโทรปรี(entropy) ได้จากสมการ (2)

$$Entropy(S) = - p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus \quad \text{_____(2)}$$

โดยที่ $p_\oplus$ คือ อัตราส่วนของตัวอย่างที่เป็นบวกใน S

$p_\ominus$ คือ อัตราส่วนของตัวอย่างที่เป็นลบใน S

ถ้าค่า แอททริบิวเป้าหมาย(target attribute) สามารถมีได้ c ค่าแล้ว เอ็นโทรปรี(entropy) ของ S จะเป็นดังนี้

$$Entropy(S) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \quad \text{_____(3)}$$

โดยที่ P_I คือความสัมพันธ์ของ S ที่คลาส I จะเห็นได้ว่าถ้าแอททริบิวเป้าหมาย (taget attribute) สามารถรับค่าได้ c ค่าแล้ว เอ็นโทรปรี(entropy) ก็จะเท่ากับ $\log_2 c$

4.2.3.2 กำหนดให้ A เท่ากับ แอททริบิว ที่มีค่า อินฟอร์เมชั่นเกรน(information gain) สูงสุด แล้วให้ A เป็นโหนดบนสุด(root node)

4.2.3.3 ทำการวนลูป ค่าต่างๆทั้งหมดที่เป็นไปได้ของ A โดยให้ υ แทนค่าของ A ตัวที่ 1 ถึง i ตามขั้นตอนข้างล่างนี้ จนกว่าจะหมดแล้วทำการเลือก ตัวอย่าง(examples) ของ แอททริบิว A ที่มีค่าเท่ากับ υ_i มา

4.2.4 ทำตั้งแต่ 1. ซ้ำจนกว่าจะสร้างต้นไม้เสร็จโดยทำการส่งค่าพารามิเตอร์ต่างๆ เข้าไปดังนี้

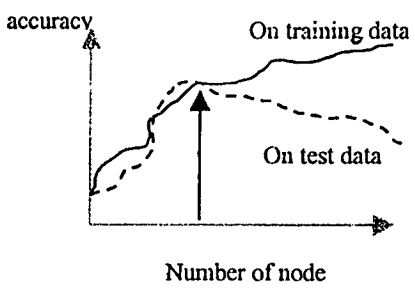
4.2.4.1 ค่าตัวอย่าง(example)ทั้งหมดที่เลือกมา

4.2.4.2 ค่าแอททริบิวเป้าหมาย(Target attribute)

4.2.4.3 ค่าของ แอททริบิว ที่ลบเอา แอททริบิว A ออก

สรุปขั้นตอนการทำงานโดย เริ่มจากการให้อัลกอริทึ่มทำงานบนเซ็ตของเรคคอร์ดที่ใช้ในการ เทรนนิ่ง(training) ซึ่งในที่นี้คือ S แล้วทำการตรวจสอบเงื่อนไขว่าถ้าเซ็ตของเรคคอร์ดทั้งหมด (all instances) ใน S เป็นคลาส C ก็จะทำการสร้างโหนด *C* แล้วหยุด แต่ถ้ามีคลาสอื่นปนอยู่ด้วยก็ต้องทำการเลือกแอททริบิว A และสร้างโหนดการตัดสินใจ(decision node) ขึ้นมา แล้วทำการแบ่งส่วนเรคคอร์ดที่ใช้ในการเทรนนิ่ง(training)ใน *S* ตามค่า *V* ของแอททริบิว A จากนั้นก็ทำซ้ำไปซ้ำมาในแต่ละเซ็ตย่อยของ Sv ต่อไป

ในการเทรนนิ่ง(training) จะทำการแตกกิ่งไปเรื่อยๆ จนถึงข้อมูลตัวสุดท้าย โดยที่ต้นไม้ (tree)จะทำการแตกกิ่งไปเรื่อยๆ แต่ในความเป็นจริงแล้วไม่ได้ง่ายอย่างที่คิดซึ่งจะพบกับอุปสรรคมากมายในระหว่างการสร้างต้นไม้(Tree) โดยเมื่อยิ่งถ้าต้องการทำให้ความถูกต้องยิ่งสูงขึ้นก็ต้องใช้ข้อมูลมากๆในกรณีของการเทรนนิ่งข้อมูล(training data) แต่ถ้าเป็นกรณีของการทดสอบข้อมูล (test data) ความถูกต้องจะเพิ่มจนถึงจุดๆหนึ่งเท่านั้นแล้วถ้ามีการแตกกิ่งไปเรื่อยๆจะทำให้ความถูกต้องยิ่งลดลง ซึ่งจุดที่ความถูกต้องถึงจุดสูงสุดของการ ทดสอบข้อมูล(test data) นี้เรียกว่า จุดโอเวอร์ฟิต(overfit) ดังภาพที่4.1 จึงเป็นผลให้เกิดการ pruning tree (การแต่งกิ่ง หรือ การตัดกิ่ง) และ สาเหตุอีกอย่างที่ทำให้เกิด โอเวอร์ฟิต(overfit)อีกอย่างคือ เมื่อข้อมูลที่ใช้มีสิ่งรบกวน (noise) อยู่ หรือ เมื่อข้อมูลที่ใช้ในการ เทรนนิ่ง(training) มีจำนวนน้อยเกินกว่าที่จะสร้าง tree เพื่อแสดงให้เห็นถึงคำตอบที่ถูกต้อง ซึ่งสิ่งเหล่านี้สามารถทำให้อัลกอริทึ่มนี้ผลิต tree ที่ทำให้เกิดการโอเวอร์ฟิต (overfit) ได้ โดยเมื่อแสดงการ Plot กราฟดังภาพที่3 เส้นกราฟของการ เทรนนิ่ง(training) กับ เส้นกราฟของการทดสอบ(testing) ตัดกันพอดี ซึ่งณ.จุดที่เกิดการ โอเวอร์ฟิต(overfit)ของข้อมูลนั้นจะมีผลทำให้ความถูกต้องของผลที่ได้ลดลง ดังนั้นทางที่ดีจึงควรใช้ข้อมูลจำนวนมากๆในการทำเทรนนิ่ง(training) เพื่อที่จะทำให้ได้โมเดลที่ครอบคลุมมากที่สุด และ ใช้ข้อมูลจำนวนน้อยๆเพื่อทำทดสอบ(testing)



ภาพที่4.1 แสดงถึงจุด Overfit

ข้อดีของทรีอินดักชั่น(Tree Induction)

- มีประสิทธิภาพทางด้านเวลาในการประมวลผลในช่วงหลังจากการทำแบ่งระดับ (Classification) แล้ว
- การแตกต้นไม้(tree) มีลักษณะที่เข้าใจได้ง่ายเมื่อสร้างเป็นกฎ ถ้าเทียบกับนูรอลอินดักชั่น (neural induction) ที่บางครั้งก็ไม่สามารถรู้ได้เลยว่าให้คำตอบมาได้อย่างไร
- สามารถบอกได้ว่าปัจจัยใดที่มีอิทธิพลต่อการทำนายมากที่สุด

ข้อเสียของทรีอินดักชั่น(Tree Induction)

- การสร้างต้นไม้(tree) เป็นเรื่องยากที่จะการคำนวนและมีความซับซ้อนมาก
- ยิ่งข้อมูลน้อยๆ ความผิดพลาดจะสูงขึ้น ซึ่งจะเห็นได้จากว่าเมื่อทำการแตกต้นไม้ (tree) ไปเรื่อยๆจำนวนของข้อมูลจะลดลงเรื่อยๆซึ่งข้อมูลน้อยๆจะเป็นตัวแทนทางสถิติได้ยาก โดยยิ่งทำให้มี level มากขึ้นความน่าเชื่อถือยิ่งน้อยลง

4.3 การหลีกเลี่ยงการเกิด โอเวอร์ฟิต(overfit)ในข้อมูล

การหลีกเลี่ยงการเกิด โอเวอร์ฟิต(overfit)สามารถทำได้ 2 แบบคือ

แบบที่ 1 หยุดต้นไม้(tree) ก่อนที่ ต้นไม้(tree) จะแตกไปเรื่อยๆจนเกิดโอเวอร์ฟิต (overfit)คือ ในระหว่างการแตกต้นไม้(tree) ก็ทำการทดสอบด้วยว่ามันเกิด โอเวอร์ฟิต(overfit)หรือเปล่าไปด้วย ถ้าผลที่ได้จากการทดสอบทำให้ความถูกต้องแม่นยำลดลงก็หยุดทันที

แบบที่ 2 วิธีการโพสพรุนทรี(Post-Prune Tree) คือทำการแตกต้นไม้(tree) จนหมดก่อนแล้วค่อยมาตัดกิ่งต้นไม้(tree)ในภายหลัง

จากทั้งหมดที่ได้กล่าวมาจึงพอสรุปได้ว่า ID3 อยู่ในกลุ่มของอัลกอริที่มการเรียนรู้ดีซีชั่นทรี (decision tree) โดยใช้ทฤษฎีข่าวสาร(information theory) เพื่อตัดสินใจ ซึ่งแชร์แอททริบิวโดยการรวบรวมเรคคอร์ดเพื่อแตกข้อมูลต่อไป แอททริบิวที่ถูกเลือกในวิธีนี้ถูกทำซ้ำไปซ้ำมาจนกระทั่ง ดีซีชั่นทรี(decision tree) สมบูรณ์ นั่นคือ การจำแนกทุกๆอินพุตท์ที่เข้ามา ถ้าข้อมูลมีสิ่งรบกวน (noise) จะมีผลทำให้เรคคอร์ดอาจจะถูกจำแนกผิดพลาด อาจจะเป็นไปได้ที่จะทำการแต่งดีซีชั่นทรี (decision tree) เพื่อที่จะลดความผิดพลาดในการจำแนกข้อมูลที่มีสิ่งรบกวน(noisy) อยู่ได้ จากที่

กล่าวมาทั้งหมดเป็นเพียงทฤษฎีของอัลกอริทึ่มนี้ แต่บางท่านอาจจะยังมองไม่เห็นถึงประโยชน์เท่าไรนั้น แต่ในบทถัดไปจะเป็นการนำทฤษฎีที่ได้ศึกษามาไปใช้ประโยชน์ได้กับงานจริง

บทที่ 5
การเตรียมข้อมูล

เพื่อเป็นการเพิ่มประสิทธิภาพของการคัดเลือกรายผู้ประกอบการที่มีแนวโน้ม ว่าจะเลี่ยงภาษีมูลค่าเพิ่ม จึงได้มีการศึกษาแนวทางที่จะสามารถคัดเลือกรายผู้ประกอบการเพื่อตรวจสอบที่ตรงเป้าหมายที่สุด ในการนำทฤษฎีเกี่ยวกับคาต้าไมนิ่งมาประยุกต์ใช้โดยมีวัตถุประสงค์เพื่อหารูปแบบของการตัดสินใจสำหรับการค้นหาและคัดเลือกรายผู้ประกอบการภาษีมูลค่าเพิ่มที่มีแนวโน้มว่าจะเลี่ยงภาษีจากฐานข้อมูลภาษีมูลค่าเพิ่ม แต่ข้อมูลในฐานข้อมูลเดิม ไม่สามารถนำมาดำเนินการได้ทันที ต้องมีกระบวนการต่างๆ สำหรับจัดเตรียมข้อมูลให้อยู่ในรูปที่เหมาะสมก่อนนำมาใช้งาน ซึ่งจะได้กล่าวต่อไปนี้

5.1 แหล่งที่มาของข้อมูล

ข้อมูลที่ต้องนำมาใช้ในการดำเนินการทั้งหมด รวบรวมได้จากแฟ้มข้อมูลดังนี้

5.1.1 ข้อมูลภาษีมูลค่าเพิ่ม (ภ.พ.30) แฟ้มข้อมูล PP30.dbf จากสำนักเทคโนโลยีสารสนเทศ ซึ่งจัดเก็บรายละเอียดของข้อมูลตามแบบ ภ.พ.30 ที่ผู้ประกอบการยื่นเพื่อแสดงรายการต่างๆ ดังมีโครงสร้างตามที่แสดงไว้ในตารางที่ 5.1

5.1.2 ข้อมูลการจดทะเบียนภาษีมูลค่าเพิ่ม (ภ.พ.01) จากแฟ้มข้อมูล REGPP01.dbf จากสำนักเทคโนโลยีสารสนเทศ ซึ่งจัดเก็บรายละเอียดของผู้ประกอบการเมื่อผู้ประกอบการยื่นแบบ ภ.พ.01 เพื่อจดทะเบียนเป็นผู้ประกอบการภาษีมูลค่าเพิ่ม ดังมีโครงสร้างตามที่แสดงไว้ในตารางที่ 5.2

5.1.3 ข้อมูลการตรวจสอบภาษีมูลค่าเพิ่ม จากแฟ้มข้อมูล Auditde.dbf จากหน่วยส่งเสริมประสิทธิภาพดังมีโครงสร้างตามที่แสดงไว้ในตารางที่ 5.3

5.1.4 ข้อมูลเชื่อมโยงรหัสผลการตรวจ จากแฟ้มข้อมูล Audits.dbf จากหน่วยส่งเสริมประสิทธิภาพ ดังมีโครงสร้างตามที่แสดงไว้ในตารางที่ 5.4

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
1	BATOFFCO	Numeric	7	รหัสสำนักงานนำส่งชุดข้อมูล
2	BATDAT	Character	9	วันที่นำส่งชุดข้อมูล
3	BATNO	Numeric	4	เลขที่ชุดข้อมูล
4	BATSEQNO	Numeric	3	ลำดับที่ในชุดข้อมูล
5	TIN	Numeric	10	เลขประจำตัวผู้เสียภาษีอากร
6	BRANO	Numeric	4	เลขที่สาขา
7	ADDNO	Character	10	ที่อยู่
8	POSCOD	Numeric	5	รหัสไปรษณีย์
9	VATMON	Numeric	2	เดือนภาษี
10	VATYEA	Numeric	4	ปีภาษี
11	PAYAMO	Numeric	15	จำนวนเงินที่ชำระ
12	PAYDAT	Character	9	วันที่ชำระ
13	ICRVATCC	Numeric	1	รหัสยอดขายแจ้งไว้ขาด/ยอดซื้อแจ้งไว้เกิน
14	SLEAMO	Numeric	15	ยอดขายในเดือนนี้
15	SLEEXPAM	Numeric	15	ยอดขายที่เสียภาษีในอัตราร้อยละ 0
16	SLEEXEAM	Numeric	15	ยอดขายที่ได้รับยกเว้นภาษี
17	VATSLEAM	Numeric	15	ยอดขายที่ต้องเสียภาษี
18	SLETAXAM	Numeric	15	ภาษีขายเดือนนี้
19	DCRVATCC	Numeric	1	รหัสยอดซื้อแจ้งไว้ขาด/ยอดขายแจ้งไว้เกิน
20	PURAMO	Numeric	15	ยอดซื้อเดือนนี้
21	PURTAXAM	Numeric	15	ยอดภาษีซื้อของเดือนนี้
22	PABTAXAM	Numeric	15	ภาษีที่ต้องชำระ
23	REBTAXAM	Numeric	15	ภาษีที่ชำระเดือนนี้

ตารางที่ 5.1 โครงสร้างข้อมูลภาษีมูลค่าเพิ่ม (PP30.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
24	OLDFWDAM	Numeric	15	ภาษีที่ชำระเกินยกมาจากเดือนก่อน
25	TOTPABTA	Numeric	15	ยอดรวมภาษีที่ต้องชำระ
26	TOTREBTA	Numeric	15	ยอดรวมภาษีที่ชำระไว้เกิน
27	SURAMO	Numeric	15	จำนวนเงินเพิ่ม
28	PENAMO	Numeric	15	จำนวนเงินเบี้ยปรับ
29	GRAPABTA	Numeric	15	รวมภาษีเงินเพิ่มเบี้ยปรับที่ต้องชำระ
30	GRAREBTA	Numeric	15	รวมภาษีที่ชำระเกินหลังคำนวนเงินเพิ่มและเบี้ยปรับแล้ว
31	FWDCOD	Numeric	1	รหัสขอคืนภาษีเป็นเงินสด
32	SIGCOD	Numeric	1	รหัสการลงชื่อผู้เสียภาษี
33	INSCOD	Numeric	1	รหัสการผ่อนชำระภาษีเป็นงวด

ตารางที่ 5.1 (ต่อ) โครงสร้างข้อมูลภาษีมูลค่าเพิ่ม (PP30.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
1	TIN	Numeric	10	เลขประจำตัวผู้เสียภาษีอากร
2	BRANO	Numeric	4	เลขที่สาขา
3	TITCOD	Numeric	5	รหัสรายละเอียด
4	BRANAM	Character	79	ชื่อสาขา
5	ADDNO	Character	10	เลขที่
6	SOINAM	Character	30	ซอย
7	MOONO	Numeric	2	หมู่ที่
8	THNNAM	Character	30	ถนน

ตารางที่ 5.2 โครงสร้างข้อมูลการจดทะเบียนภาษีมูลค่าเพิ่ม (regpp01.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
9	TAMNAM	Character	30	ตำบล
10	AMPCOD	Numeric	7	รหัสอำเภอ
11	POSCOD	Numeric	5	รหัสไปรษณีย์
12	TELNO	Numeric	9	เบอร์โทรศัพท์
13	FORRECDA	Character	9	วันที่ยื่นขอจดทะเบียน ภ.พ.01
14	VATREGDAT	Character	9	วันที่ได้รับการอนุมัติให้เป็นผู้ประกอบการภาษีมูลค่าเพิ่ม

ตารางที่ 5.2 (ต่อ) โครงสร้างข้อมูลการจดทะเบียนภาษีมูลค่าเพิ่ม (regpp01.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
1	TIN	Numeric	10	เลขประจำตัวผู้เสียภาษี
2	BRANCH	Character	4	สาขาที่
3	STATUS	Character	1	สถานภาพ 1 = บุคคลธรรมดา 2 = นิติบุคคล
4	VATCOD	Character	1	ผู้ประกอบการจดทะเบียน VAT (1 = 1.5 , 2 = 7/10)
5	RANGAUD	Character	1	ประเภทการตรวจ (1 = คืน , 2 = ทั่วไป , 3 = เฉพาะประเด็น , 4 = สอบยัน , 5 = ประเมินสถานะ)

ตารางที่ 5.3 โครงสร้างประวัติผลการตรวจสอบภาษีมูลค่าเพิ่ม (Auditde.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
6	PP30TYP	Character	1	การยื่นแบบภ.พ.30 1 = ขอคืนเงินสด 2 = ขอคืนเครดิต 3 = ชำระ)
7	RESULT	Character	2	ผลการตรวจ(Audits.dbf)
8	DMY_30	Character	8	วันที่ยื่นแบบภ.พ.30
9	MM_YY	Character	5	เดือน/ปีภาษี
10	MM_YY_EX	Character	3	เดือน/ปีภาษี (กรณียื่นเพิ่มเติม)
11	TAX_30	Numeric	15	ภาษีที่ขอคืน/ชำระ(pp30typ 1,2 = ขอคืน , 3 = ชำระ)
12	TAX_PRA	Numeric	15	ภาษีที่ประเมินหรือชำระเพิ่ม
13	TAX_INT	Numeric	15	เบี้ยปรับ
14	TAX_ADD	Numeric	15	เงินเพิ่ม
15	TAX_ALL	Numeric	15	ยอดรวม
16	TAX_REFOK	Numeric	15	ภาษีที่คืน
17	TAX_REFN	Numeric	15	ภาษีที่ไม่คืน
18	TAX_LAW	Numeric	15	ปรับอาญา
19	PRANO	Character	23	เลขที่ใบประเมิน/ในเสร็จ
20	CIT50	Numeric	15	การวิเคราะห์แบบภ.ง.ด50

ตารางที่ 5.3 (ต่อ) โครงสร้างประวัติผลการตรวจสอบภาษีมูลค่าเพิ่ม (Auditde.dbf)

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
1	Result	Numeric	2	ผลการตรวจ
2	Def_result	Character	15	รายละเอียดผลการตรวจ

ตารางที่ 5.4 โครงสร้างข้อมูลการเชื่อมโยงรหัสผลการตรวจ (Audits.dbf)

5.2 การคัดเลือกข้อมูล

การคัดเลือกแอททริบิวเป็นส่วนที่ที่มีความสำคัญมาก ทั้งนี้เพราะถ้าเลือกแอททริบิวที่ไม่มีความสำคัญหรือไม่เหมาะสมก็จะมีผลต่อโมเดลที่จะได้ ดังนั้นในการคัดเลือกจึงจำเป็นต้องอ้างถึงหลักการเดิมที่ใช้ในการตรวจสอบ เพื่อเลือกแอททริบิวที่มีแนวโน้มว่าจะเกี่ยวข้องมาใช้ในการเทรนนิ่ง หลักการนี้ได้อธิบายไว้แล้วในตอนต้น ผลลัพท์ที่ได้คือแฟ้มข้อมูล data_audit โดยข้อมูลในแฟ้ม data_audit ได้มาจากแฟ้มต่างๆ ดังนี้

5.2.1 จากแฟ้มข้อมูล ภ.พ.30 มี 12 แอททริบิว คือ PP30TYP, SLEAMO , PURAMO , TOTPABTA , TOTREBTA , OLDFWDAM , SURAMO , PENAMO , PAYAMO , PAYDAT , VATMON , VATYEA

5.2.2 จากแฟ้มข้อมูล ภ.พ.01 มี 1 แอททริบิวคือ VATREGDAT

5.2.3 จากแฟ้มข้อมูลการตรวจสอบ มี 2 แอททริบิวคือ RESULT , TIN

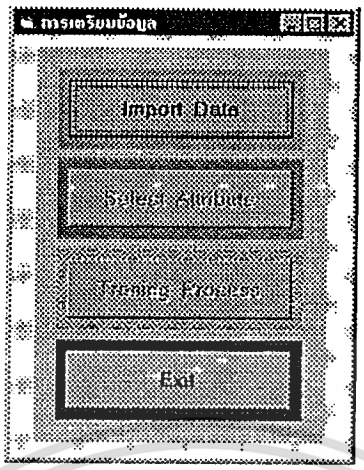
ซึ่งการเชื่อมโยงข้อมูล (Join) นั้นใช้คำสั่ง SQL ดังนี้

```
INSERT INTO data_audit ( TIN, VATMON, VATYEA, PP30TYP, SLEAMO, PURAMO, TOTPABTA, TOTREBTA, OLDFWDAM, SURAMO, PENAMO, PAYAMO, PAYDAT, VATREGDAT, RESULT )
SELECT auditde.TIN, pp30.VATMON, pp30.VATYEA, auditde.PP30TYP, pp30.SLEAMO, pp30.PURAMO, pp30.TOTPABTA, pp30.TOTREBTA, pp30.OLDFWDAM, pp30.SURAMO, pp30.PENAMO, pp30.PAYAMO, pp30.PAYDAT, regpp01.VATREGDAT, auditde.RESULT
FROM regpp01 INNER JOIN (pp30 INNER JOIN (audits INNER JOIN auditde ON audits.RESULT = auditde.RESULT) ON pp30.TIN = auditde.TIN) ON regpp01.TIN = auditde.TIN
WHERE (((pp30.VATMON)=Val(Left$([auditde].[MM_YY],2))) AND ((pp30.VATYEA)=Val('25' & Right$([auditde].[MM_YY],2))-543))
```

จากคำสั่ง SQL จะให้เลือกข้อมูล 15 แอททริบิวจากตารางทั้ง 3 ตารางเชื่อมโยงกันด้วยเลขประจำตัวผู้เสียภาษี และ ผลการตรวจสอบ โดยจะคัดเลือกมาเฉพาะที่ ค่าในแอททริบิวเดือนภาษีของตาราง pp30 เท่ากับ ค่าของตัวแรกในแอททริบิวเดือนปีภาษีของ ตาราง Auditde ก็จะได้ข้อมูลตามโครงสร้างที่แสดงไว้ในตารางที่ 5.5

คอลัมน์ที่	ชื่อคอลัมน์	ประเภท	ขนาด	คำอธิบาย
1	TIN	Numeric	10	เลขประจำตัวผู้เสียภาษี
2	VATMON	Numeric	2	เดือนภาษี
3	VATYEA	Numeric	4	ปีภาษี
4	pp30typ	Character	1	การยื่นแบบภ.พ.30 1 = ขอคืนเงินสด 2 = ขอคืนเครดิต 3 = ชำระ
5	SLEAMO	Numeric	15	ยอดขายในเดือนนี้
6	PURAMO	Numeric	15	ยอดซื้อในเดือนนี้
7	TOTPABTA	Numeric	15	ยอดรวมภาษีที่ต้องชำระ
8	TOTREBTA	Numeric	15	ยอดรวมภาษีที่ชำระไว้เกิน
9	SURAMO	Numeric	15	จำนวนเงินเพิ่ม
10	PENAMO	Numeric	15	จำนวนเงินเบี้ยปรับ
11	OLDFWDAM	Numeric	15	ภาษีที่ชำระเกินยกมาจากเดือนก่อน
12	PAYAMO	Numeric	1	จำนวนเงินที่ชำระ
13	PAYDAT	Character	9	วันที่ชำระ
14	VATREGDAT	Character	9	วันที่ได้รับการอนุมัติให้เป็นผู้ประกอบการภาษีมูลค่าเพิ่ม
15	RESULT	Character	2	ผลการตรวจสอบ

ตารางที่ 5.5 โครงสร้างตารางข้อมูล Data_Audit



ภาพที่ 5.1 การนำข้อมูลเข้าสู่ระบบ

5.3 การทำความสะอาดข้อมูล

เป็นกระบวนการที่อยู่ในส่วนการเตรียมข้อมูล ซึ่งเมื่อเรียกใช้ Import Data ตามภาพที่ 5.1 ก็จะรวมการทำความสะอาดข้อมูลหรือการคัดข้อมูลที่ไม่ถูกต้องออกด้วยเพื่อให้ได้เฉพาะข้อมูลที่ดีเข้าไปใช้ในการเทรนนิ่ง โดยดำเนินการคัดทิ้งข้อมูลที่ไม่ถูกต้อง จนไม่สามารถปรับแต่งได้ตามเงื่อนไขดังนี้

- TIN ที่มีค่าเท่ากับ 0 หรือ Null หรือ มีขนาดไม่ครบ 10 หลัก
- VATMON ที่มีค่าเท่ากับ 0 หรือ Null หรือ มีค่าไม่อยู่ในช่วง 1-12
- VATYEA ที่มีค่าเท่ากับ 0 หรือ Null หรือ มีขนาดไม่ครบ 4 หลัก
- RESULT ที่มีค่าเป็น Null หรือ มีค่าไม่เท่ากับ 0 หรือ 1

5.4 การแปลงข้อมูล

5.4.1 การแปลงข้อมูลในแอททริบิว Paydat

- แปลงค่าเดือนที่เป็นตัวอักษรให้เป็นค่าตัวเลข เช่น 22-Sep-90 แปลงเป็น 22-09-90
- ทำการแปลงข้อมูลตามเงื่อนไขดังนี้

5.4.1.1 ถ้าเดือนภาษีเป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เท่ากับเดือนภาษีของแอททริบิว vatmon และ ปีที่ชำระภาษีของแอททริบิว paydat เท่ากับปีภาษีของแอททริบิว vatyea แล้วแสดง

ว่าผู้ประกอบการรายนั้นยื่นแบบตรงตามกำหนดเวลาให้แทนค่า "1" ลงในแอททริบิว paydat

5.4.1.2 ถ้าเดือนภาษีเป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เป็นเดือน 1 และ วันที่ชำระภาษีของแอททริบิว paydat ไม่เกินวันที่ 15 แล้วแสดงว่าผู้ประกอบการรายนั้นยื่นแบบตรงตามกำหนดเวลา ให้แทนค่า "1" ลงในแอททริบิว paydat

5.4.1.3 ถ้าเดือนภาษีเป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เป็นเดือน 1 และ วันที่ชำระภาษีของแอททริบิว paydat เกินวันที่ 15 แล้วแสดงว่าผู้ประกอบการรายนั้นยื่นแบบเกินกำหนดเวลา ให้แทนค่า "2" ลงในแอททริบิว paydat

5.4.1.4 ถ้าเดือนภาษีไม่เป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เท่ากับเดือนภาษีของแอททริบิว vatmon และ ปีที่ชำระภาษีของแอททริบิว paydat เท่ากับปีภาษีของแอททริบิว vatyea แล้วแสดงว่าผู้ประกอบการรายนั้นยื่นแบบตรงตามกำหนดเวลาให้แทนค่า "1" ลงในแอททริบิว paydat

5.4.1.5 ถ้าเดือนภาษีไม่เป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เท่ากับเดือนภาษีบวกอีก 1 และ ปีที่ชำระภาษีของแอททริบิว paydat เท่ากับ ปีภาษีของแอททริบิว vatyea และ วันที่ชำระภาษีของแอททริบิว paydat ไม่เกินวันที่ 15 แล้ว แสดงว่าผู้ประกอบการรายนั้นยื่นแบบตรงตามกำหนดเวลา ให้แทนค่า "1" ลงในแอททริบิว paydat

5.4.1.6 ถ้าเดือนภาษีไม่เป็นเดือน 12 และ เดือนที่ชำระภาษีของแอททริบิว paydat เท่ากับเดือนภาษีบวกอีก 1 และ ปีที่ชำระภาษีของแอททริบิว paydat เท่ากับ ปีภาษีของแอททริบิว vatyea และ วันที่ชำระภาษีของแอททริบิว paydat เกินวันที่ 15 แล้วแสดงว่าผู้ประกอบการรายนั้นยื่นแบบเกินกำหนดเวลา ให้แทนค่า "2" ลงในแอททริบิว paydat

5.4.2 การแปลงข้อมูลในแอททริบิว Vatregdat

- แปลงค่าเดือนที่เป็นตัวอักษรให้เป็นค่าตัวเลข เช่น 22-Sep-90 แปลงเป็น 22-09-90

- แปลงค่า Vatregdat ให้เป็นจำนวนวันที่จดทะเบียนด้วยการนำไปลบออกจากวันที่ปัจจุบัน

5.4.3 การแปลงข้อมูลในแอททริบิวที่ต้องจัดกลุ่ม ข้อมูลบางแอททริบิวมีค่าต่อเนื่องเช่น ข้อมูลที่เป็นจำนวนเงิน ซึ่งต้องแปลงให้อยู่ในรูปของกลุ่มข้อมูล แต่ในการจัดกลุ่มไม่ควรกำหนดช่วงของกลุ่มข้อมูลเอง ควรจะจัดกลุ่มตามค่าของข้อมูลจริงๆ ด้วยการใช้เครื่องมือช่วยสำหรับระบบนี้ได้ใช้ MATLAB ช่วยในการสร้างฮีทโทรแกรม (histrogram) เพื่อจัดกลุ่มข้อมูลก่อนที่จะแปลงข้อมูล ซึ่งผลของฮีทโทรแกรมได้แสดงไว้ในภาพที่ 5.2 ถึงภาพที่ 5.12 รายละเอียดแยกตามแอททริบิวดังนี้

5.4.3.1 ข้อมูลยอดขายเดือนนี้ sleamo สามารถจัดกลุ่มข้อมูลได้ 5 กลุ่ม คือ

กลุ่มที่ 1 คือ มียอดขายน้อยกว่าหรือเท่ากับ 10,000 (<= 10,000)

กลุ่มที่ 2 คือ มียอดขาย 10,001 ถึง 100,000 (10,001 – 100,000)

กลุ่มที่ 3 คือ มียอดขาย 100,001 ถึง 1,000,000(100,001 – 1,000,000

กลุ่มที่ 4 คือ มียอดขาย 1,000,001 ถึง 10,000,000(1,000,001-10,000,000)

กลุ่มที่ 5 คือ กลุ่มที่มียอดขายมากกว่า 10,000,000 ขึ้นไป (> 10,000,000)

5.4.3.2 ข้อมูลยอดซื้อเดือนนี้ puramo สามารถจัดกลุ่มได้ 5 กลุ่ม คือ

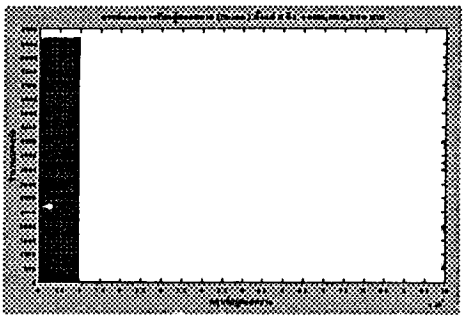
กลุ่มที่ 1 คือ มียอดซื้อน้อยกว่าหรือเท่ากับ 10,000 (<= 10,000)

กลุ่มที่ 2 คือ มียอดซื้อ 10,001 ถึง 100,000 (10,001 – 100,000)

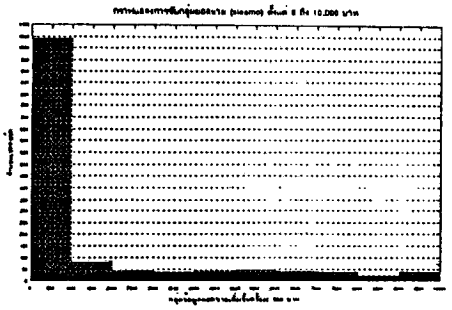
กลุ่มที่ 3 คือ มียอดซื้อ 100,001 ถึง 1,000,000 (100,001 – 1,000,000)

กลุ่มที่ 4 คือ มียอดซื้อ 1,000,001 ถึง 10,000,000 (1,000,001- 10,000,000)

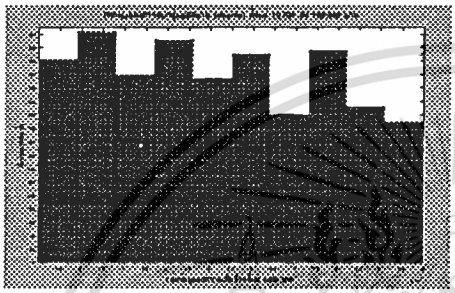
กลุ่มที่ 5 คือ มียอดซื้อมากกว่า 10,000,000 (> 10,000,000)



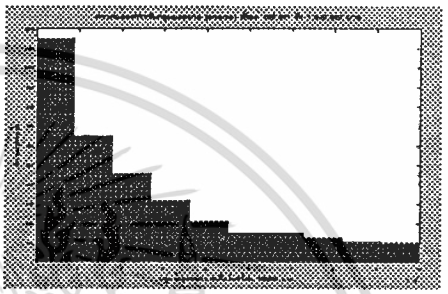
ภาพแสดงการจัดกลุ่มข้อมูลยอดขาย



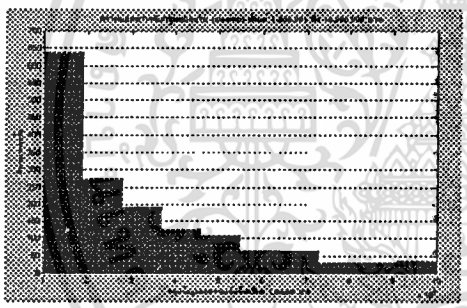
กลุ่มที่ 1



กลุ่มที่ 2



กลุ่มที่ 3



กลุ่มที่ 4

กลุ่มที่ 5

ภาพที่ 5.2 แสดงการจัดกลุ่มยอดขาย

5.4.3.3 ข้อมูลยอดรวมภาษีที่ต้องชำระ (totpabta) สามารถจัดได้ 5 กลุ่มคือ

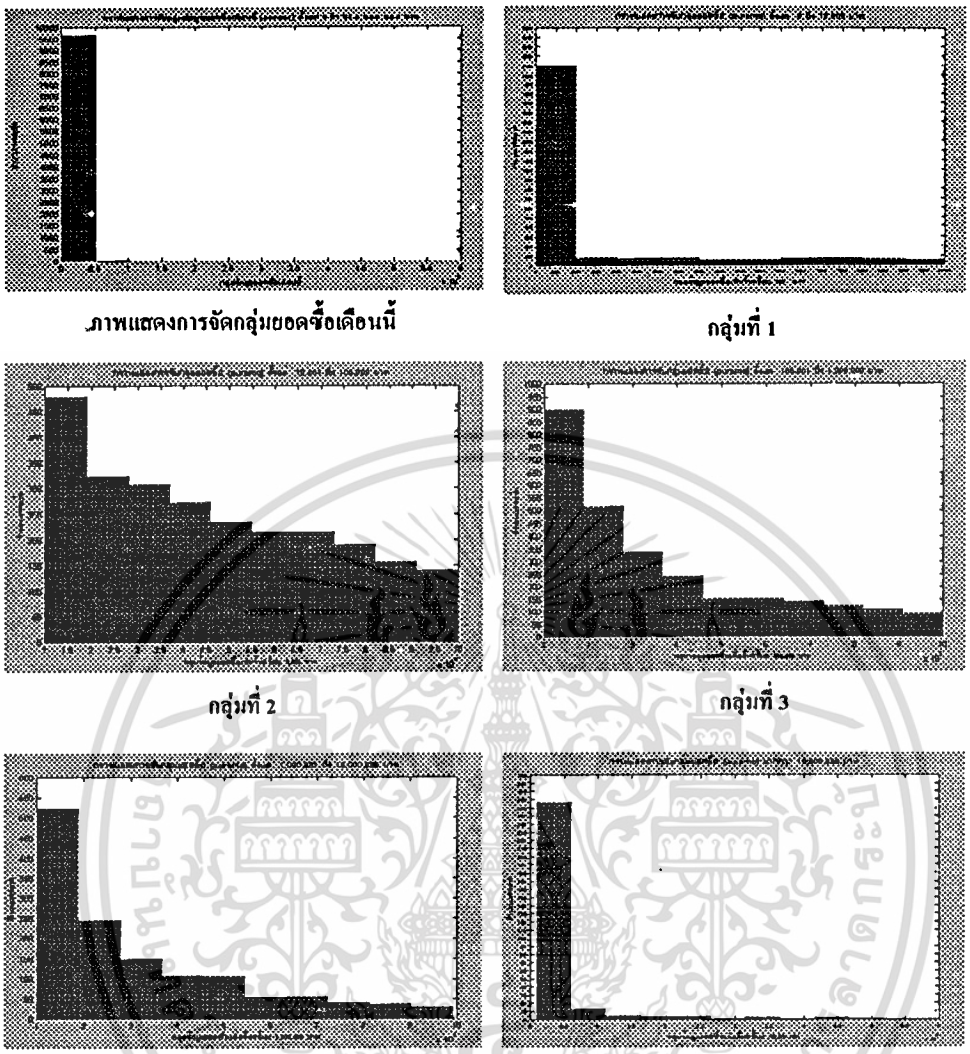
กลุ่มที่ 1 คือ มียอดรวมภาษีที่ต้องชำระน้อยกว่าหรือเท่ากับ 5,000

กลุ่มที่ 2 คือ มียอดรวมภาษีที่ต้องชำระ 5,001 ถึง 10,000

กลุ่มที่ 3 คือ มียอดรวมภาษีที่ต้องชำระ 10,001 ถึง 50,000

กลุ่มที่ 4 คือมียอดรวมภาษีที่ต้องชำระ 50,001 ถึง 100,000

กลุ่มที่ 5 คือ มียอดรวมภาษีที่ต้องชำระมากกว่า 100,000 (> 100,000)



ภาพที่ 5.3 แสดงการจัดกลุ่มข้อมูลยอดซื้อ

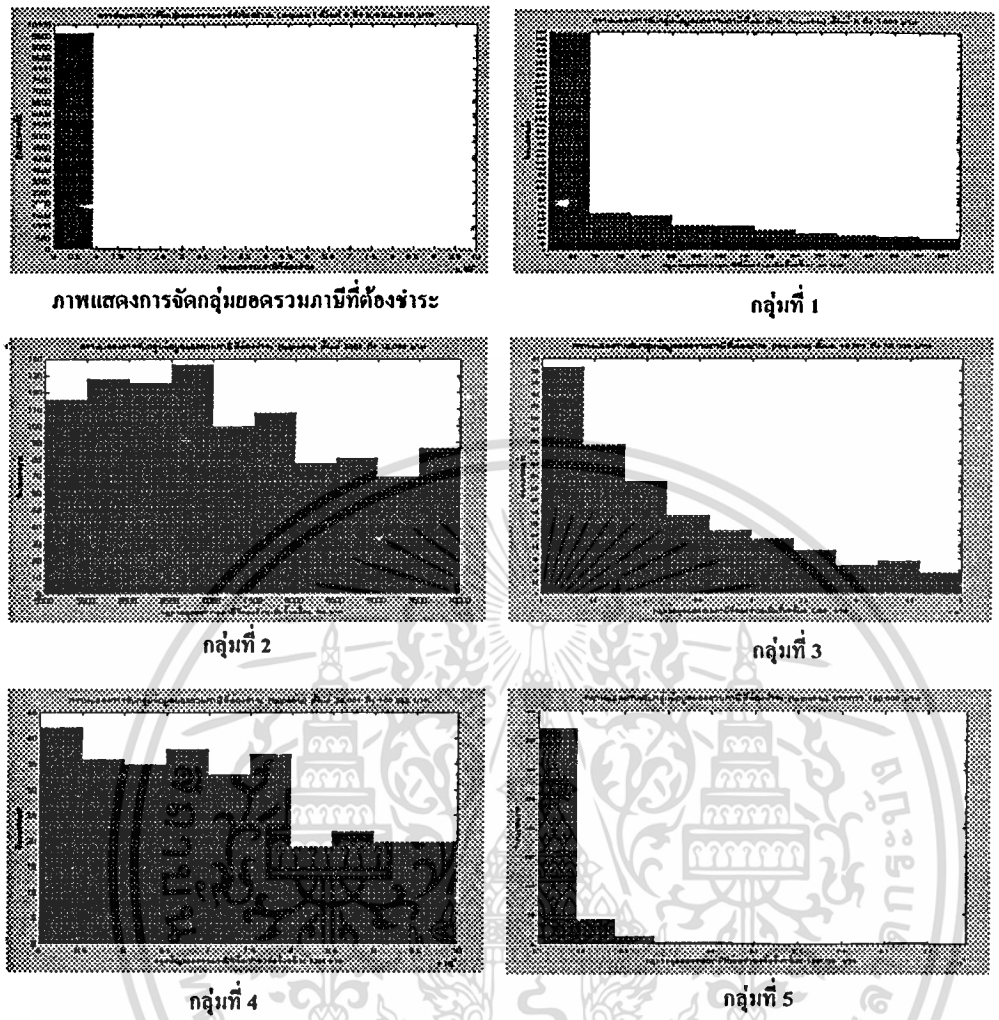
5.4.3.4 ข้อมูลยอดรวมภาษีที่ชำระไว้เกิน (totrebta) สามารถจัดกลุ่มได้ 4 กลุ่ม

กลุ่มที่ 1 คือ มียอดภาษีที่ชำระไว้เกินน้อยกว่าหรือเท่ากับ 5,000

กลุ่มที่ 2 คือ มียอดภาษีที่ชำระไว้เกิน 5,001 ถึง 50,000

กลุ่มที่ 3 คือ มียอดภาษีที่ชำระไว้เกิน 50,001 ถึง 500,000

กลุ่มที่ 4 คือ มียอดภาษีที่ชำระไว้เกินมากกว่า 500,000



ภาพที่ 5.4 แสดงการจัดกลุ่มยอดรวมภาษีที่ต้องชำระ

5.4.3.5 ข้อมูลยอดรวมภาษีที่ชำระไว้เกินยกมาจากเดือนก่อน (oldfwdam) จัดได้ 4 กลุ่มคือ

กลุ่มที่ 1 คือ มียอดชำระเกินยกมาจากเดือนก่อน <= 5,000

กลุ่มที่ 2 คือ มียอดชำระเกินยกมาจากเดือนก่อน 5,001 ถึง 50,000

กลุ่มที่ 3 คือ มียอดชำระเกินยกมาจากเดือนก่อน 50,001 ถึง 500,000

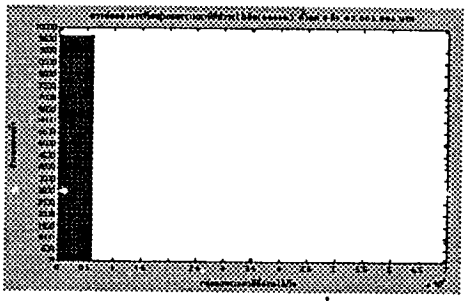
กลุ่มที่ 4 คือ มียอดชำระเกินยกมาจากเดือนก่อนมากกว่า 500,000

5.4.3.6 ข้อมูลยอดเงินเพิ่ม (suramo) จัดได้ 3 กลุ่ม

กลุ่มที่ 1 คือ มียอดเงินเพิ่มน้อยกว่าหรือเท่ากับ 500 (<= 500)

กลุ่มที่ 2 คือ มียอดเงินเพิ่ม 501 ถึง 2,500 (501 – 2,500)

กลุ่มที่ 3 คือ มียอดเงินเพิ่มมากกว่า 2,500 (> 2,500)



ภาพแสดงการจัดกลุ่มยอดรวมภาษีที่ชำระไว้เกิน

กลุ่มที่ 1

กลุ่มที่ 2

กลุ่มที่ 3

กลุ่มที่ 4

ภาพที่ 5.5 แสดงการจัดกลุ่มข้อมูลยอดภาษีที่ชำระไว้เกิน

5.4.3.7 ข้อมูลยอดเบี้ยปรับ (penamo) จัดได้ 3 กลุ่ม

กลุ่มที่ 1 คือ มียอดเบี้ยปรับน้อยกว่าหรือเท่ากับ 500 (<= 500)

กลุ่มที่ 2 คือ มียอดเบี้ยปรับ 501 ถึง 5,000 (501 – 5,000)

กลุ่มที่ 3 คือ มียอดเบี้ยปรับ มากกว่า 5,000 (> 5,000)

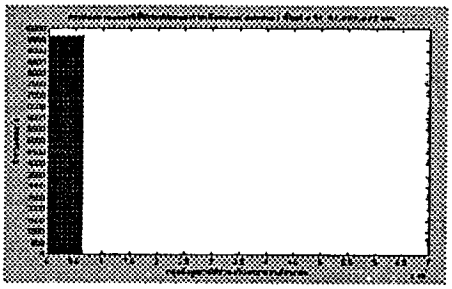
5.4.3.8 ข้อมูลยอดจำนวนเงินที่ชำระ (payamo) จัด 4 กลุ่ม

กลุ่มที่ 1 คือ มียอดจำนวนเงินที่ชำระน้อยกว่าหรือเท่ากับ 1,000

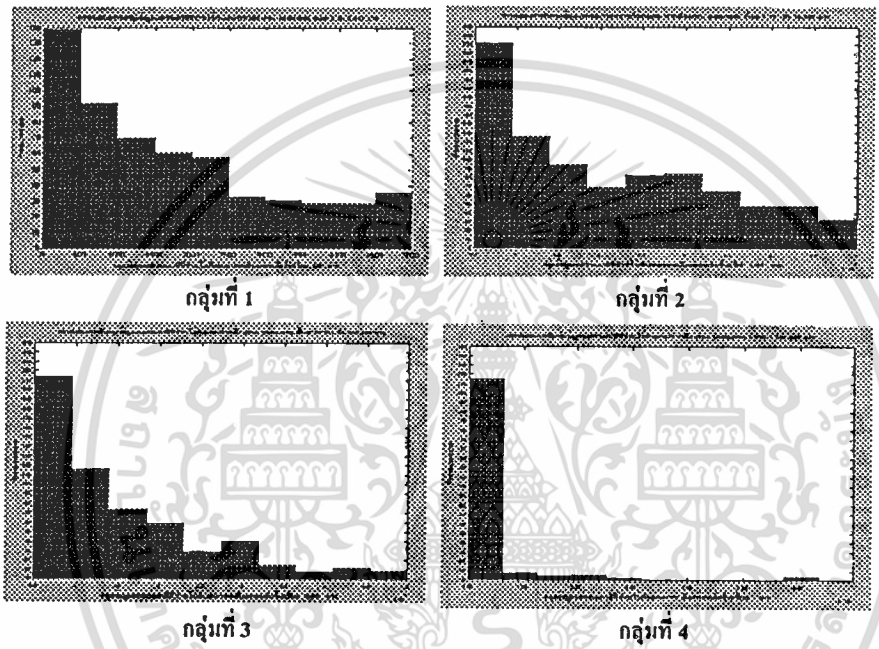
กลุ่มที่ 2 คือ มียอดจำนวนเงินที่ชำระ 1,001 ถึง 10,000

กลุ่มที่ 3 คือ มียอดจำนวนเงินที่ชำระ 10,001 ถึง 100,000

กลุ่มที่ 4 คือ มียอดจำนวนเงินที่ชำระมากกว่า 100,000 (> 100,000)



ภาพแสดงการจัดกลุ่มยอดรวมภาษีที่ชำระเกินยกมาจากเดือนก่อน



ภาพที่ 5.6 แสดงการจัดกลุ่มยอดรวมภาษีที่ชำระเกินยกมาจากเดือนก่อน

5.4.3.9 ข้อมูลวันที่ยื่นแบบ (paydat) จัดได้ 2 กลุ่ม คือ

กลุ่มที่ 1 คือ มีวันที่ยื่นแบบเท่ากับ 1

กลุ่มที่ 2 คือ กลุ่มที่มีวันที่ยื่นแบบเท่ากับ 2

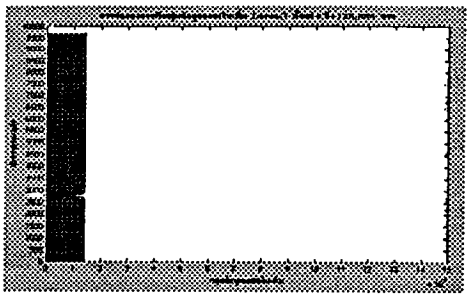
5.4.3.10 ข้อมูลจำนวนวันที่จดทะเบียนเป็นผู้ประกอบการภาษีมูลค่าเพิ่ม (vatregdat) จัดได้ 4 กลุ่ม คือ

กลุ่มที่ 1 คือ มีระยะเวลาน้อยกว่าหรือเท่ากับ 1000 วัน

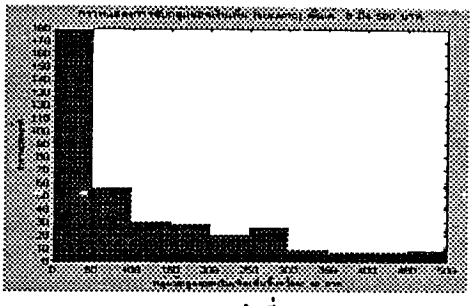
กลุ่มที่ 2 คือ มีระยะเวลา ตั้งแต่ 1001 ถึง 2000 วัน

กลุ่มที่ 3 คือ มีระยะเวลา ตั้งแต่ 2001 ถึง 3000 วัน

กลุ่มที่ 4 คือ มีระยะเวลา มากกว่า 3000 วัน



ภาพแสดงการจัดกลุ่มข้อมูลยอดเงินเพิ่ม



กลุ่มที่ 1

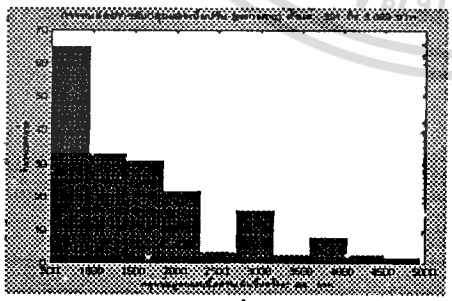
กลุ่มที่ 2

กลุ่มที่ 3

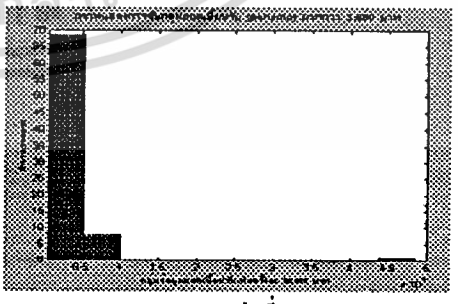
ภาพที่ 5.7 แสดงการจัดกลุ่มยอดเงินเพิ่ม

ภาพแสดงการจัดกลุ่มข้อมูลยอดเบี้ยปรับ

กลุ่มที่ 1

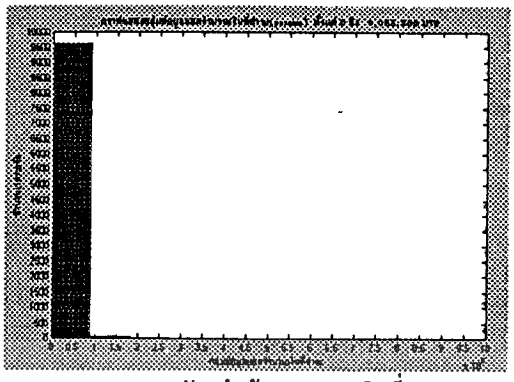


กลุ่มที่ 2

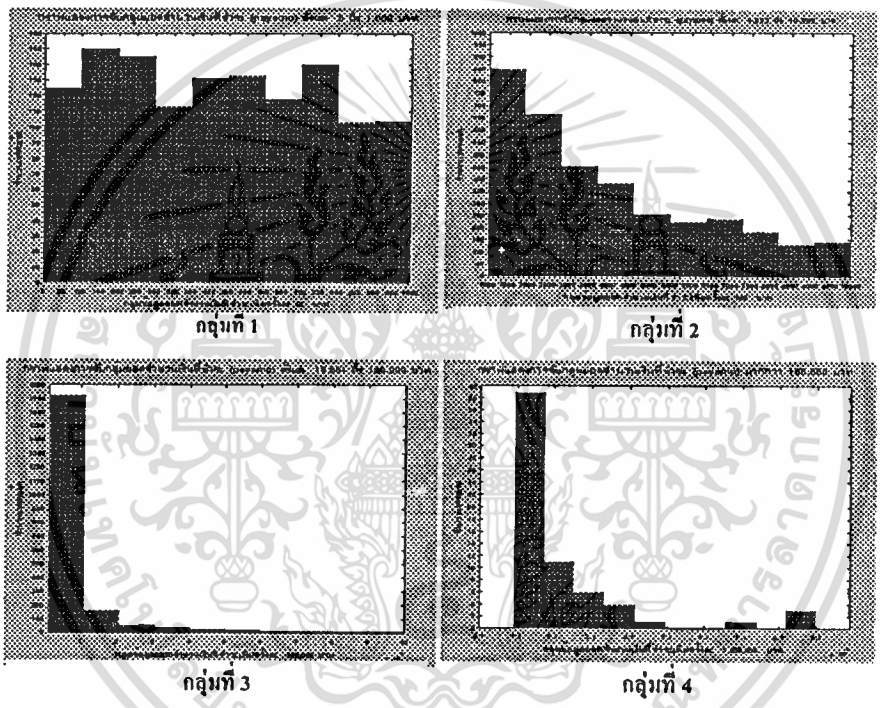


กลุ่มที่ 3

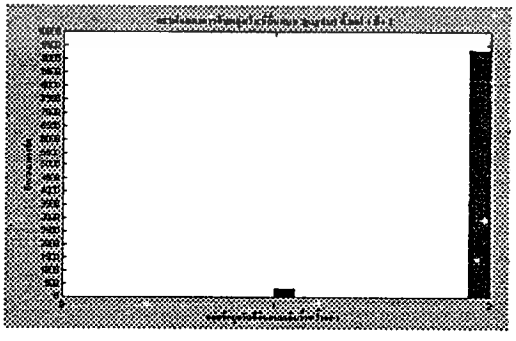
ภาพที่ 5.8 แสดงการจัดกลุ่มข้อมูลยอดเบี้ยปรับ



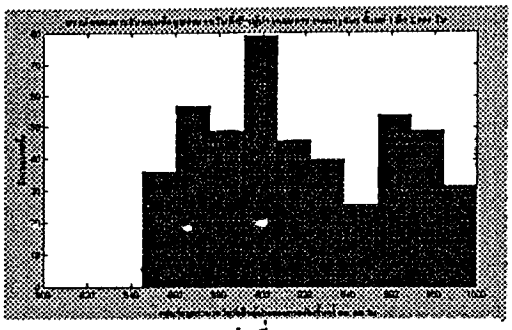
ภาพแสดงการจัดกลุ่มข้อมูลจำนวนเงินที่ชำระ



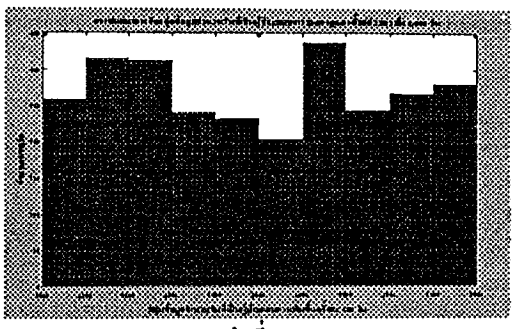
ภาพที่ 5.9 แสดงการจัดกลุ่มยอดจำนวนเงินที่ชำระ



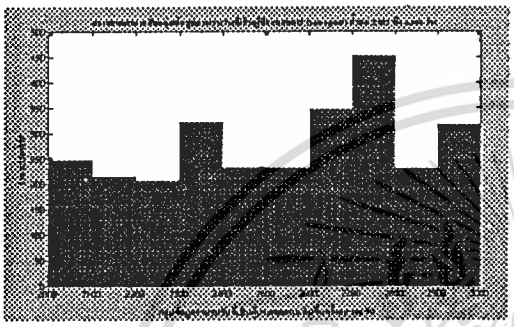
ภาพที่ 5.10 แสดงการจัดกลุ่มข้อมูลวันที่ยื่นแบบ



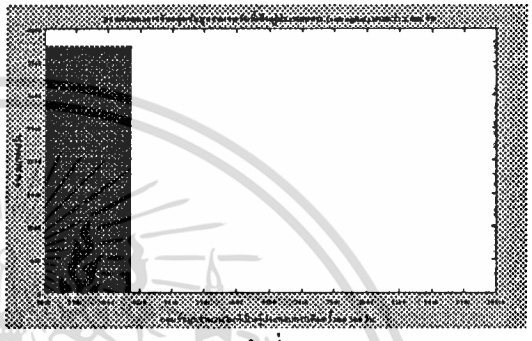
กลุ่มที่ 1



กลุ่มที่ 2



กลุ่มที่ 3



กลุ่มที่ 4

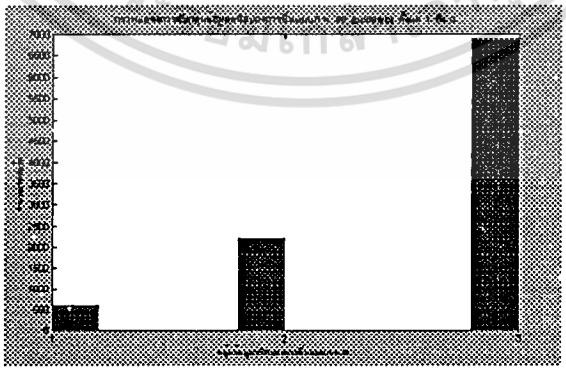
ภาพที่ 5.11 แสดงการจัดกลุ่มจำนวนวันที่จดทะเบียนเป็นผู้ประกอบการภาษีมูลค่าเพิ่ม

5.4.3.11 ข้อมูลชนิดของการยื่นแบบภ.พ. 30 (pp30typ) จัดได้ 3 กลุ่ม

กลุ่มที่ 1 คือ กลุ่มการขอคืนเงินสด

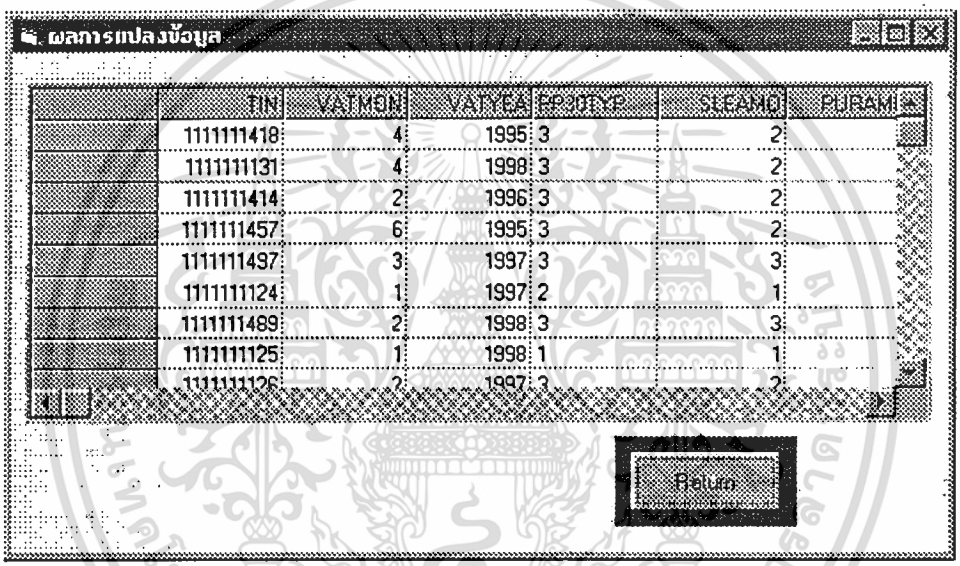
กลุ่มที่ 2 คือ กลุ่มขอคืนเป็นเครดิต

กลุ่มที่ 3 คือกลุ่มชำระภาษี



ภาพที่ 5.12 แสดงการจัดกลุ่มข้อมูลชนิดของการยื่นแบบ

เมื่อจัดกลุ่มในแต่ละแอททริบิวได้แล้ว ก็แปลงข้อมูลในแต่ละแอททริบิวให้ตรงกับค่ากลุ่ม แล้วคัดเลือกข้อมูล โดยแอททริบิวที่คัดเลือกมา มีทั้งหมด 15 แอททริบิว โดยแอททริบิว TIN , VATMON , VATYEA เป็นแอททริบิวสำหรับเชื่อมโยงเรคคอร์ด ส่วนแอททริบิวที่เหลือ 12 แอททริบิว ที่จะใช้ในการเทรนนิ่ง(Training) คือ PP30TYP, SLEAMO , PURAMO , TOTPABTA , TOTREBTA , OLDFWDAM , SURAMO , PENAMO , PAYAMO , PAYDAT, VATREGDAT, RESULT กระบวนการทั้งหมดที่กล่าวมาใน 5.2 ถึง 5.4 ดำเนินการได้ด้วยการคลิกไอคอน Import ในภาพที่ 5.1 และจะได้ผลลัพธ์ ดังในภาพที่ 5.15



ภาพที่ 5.13 แสดงตารางที่ได้จากการ Import ข้อมูล

จากกระบวนการในการเตรียมข้อมูลที่กล่าวมา จะเห็นว่าเป็นกระบวนการที่ใช้เวลาในการดำเนินการนานที่สุด เพื่อที่ให้ได้ข้อมูลที่สมบูรณ์และมีความถูกต้องมากที่สุด จบจากกระบวนการนี้ต่อไปก็เป็นการนำข้อมูลที่ได้ไปดำเนินการในส่วนของการเทรนนิ่ง ซึ่งจะได้กล่าวในบทถัดไป

บทที่6
การเทรนนิ่งและการทดสอบ

บทที่ผ่านมาเป็นการนำเสนอกระบวนการในการจัดเตรียมข้อมูลเพื่อให้ได้ข้อมูลที่มีคุณภาพซึ่งเมื่อจัดเตรียมข้อมูลเรียบร้อยแล้วต่อไปก็เป็นกระบวนการของการนำเอาข้อมูลที่ได้เตรียมไว้แล้วนั้นมาทำการเทรนนิ่งเพื่อสร้างโมเดลสำหรับการค้นหาผู้มีแนวโน้มที่จะหลีกเลี่ยงภาษี โดยในบทนี้จะได้กล่าวถึงกระบวนการทำงานเทรนนิ่งและนำเอาผลที่ได้จากการเทรนนิ่งนั้นไปทำการทดสอบต่อไป

6.1 การเทรนนิ่ง

การเทรนนิ่ง เป็นกระบวนการนำข้อมูลที่ได้จากการเตรียมข้อมูลเข้าสู่การประมวลผลด้วยอัลกอลิทึ่มของดาต้าไมนิ่ง สำหรับระบบนี้จะใช้อัลกอริทึ่มในการทำงานคือ ID3 และแฟ้มข้อมูลที่จำเป็นต้องใช้มี 2 แฟ้ม โดยมีรายละเอียด คือ

6.1.1 แฟ้มข้อมูลที่มีชนิด(extension) เป็น .tag ซึ่งจะเก็บชื่อแอททริบิวของข้อมูล ที่จะใช้เทรนนิ่ง โดยมีแอททริบิวที่เป็นเป้าหมาย(target) อยู่สุดท้าย เช่น

ชื่อคอลัมน์	รายละเอียด
pp30typ	การยื่นแบบภ.พ.30 (1 = ขอคืนเงินสด , 2 = ขอคืนเครดิต , 3 = ชำระ)
SLEAMO	ยอดขายในเดือนนี้
PURAMO	ยอดซื้อในเดือนนี้
TOTPABTA	ยอดรวมภาษีที่ต้องชำระ
TOTREBTA	ยอดรวมภาษีที่ชำระไว้เกิน
SURAMO	จำนวนเงินเพิ่ม
PENAMO	จำนวนเงินเบี้ยปรับ

ตารางที่ 6.1 แสดงข้อมูลในแฟ้ม Exam

ชื่อคอลัมน์	รายละเอียด
OLDFWDAM	ภาษีที่ชำระเกินยกมาจากเดือนก่อน
PAYAMO	จำนวนเงินที่ชำระ
PAYDAT	วันที่ชำระ
VATREGDAT	วันที่ได้รับการอนุมัติให้เป็นผู้ประกอบการภาษีมูลค่าเพิ่ม
RESULT	ผลการตรวจสอบ

ตารางที่ 6.1 (ต่อ) แสดงข้อมูลในแฟ้ม Exam

6.1.2 แฟ้มข้อมูลที่มีชนิด(extension) เป็น .dat ซึ่งจะเก็บข้อมูลที่จะเทรนนิ่งเป็นข้อมูลที่ผ่านการเตรียมข้อมูลแล้ว เช่น

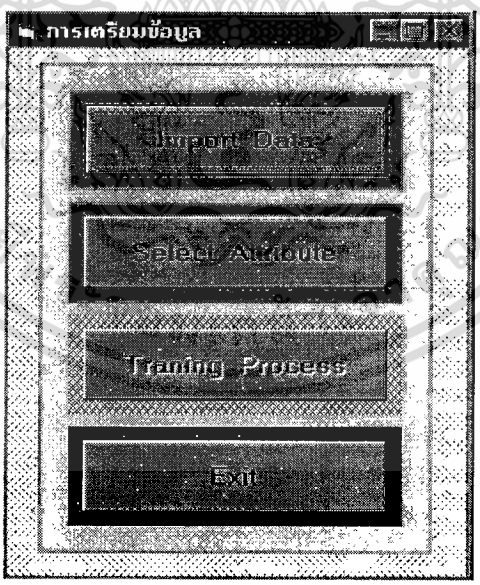
ข้อมูล :
1 1 1 1 1 1 1 1 1 2 4 0
1 1 1 1 1 1 1 1 1 2 4 0
1 1 2 1 1 1 1 1 1 1 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 1 1 1 1 1 2 4 1
1 1 2 1 2 1 1 1 1 2 4 1
1 1 2 1 2 1 1 1 1 2 4 1

ตารางที่ 6.2 แสดงลักษณะข้อมูลที่เข้าเทรนนิ่ง

ข้อมูล :
1 1 2 1 2 1 1 1 1 2 4 1
1 1 3 1 2 1 1 1 1 2 4 0
1 1 3 1 2 1 1 1 1 2 4 0
1 1 3 1 2 1 1 1 1 2 4 0
1 1 3 1 2 1 1 1 1 2 4 0
1 1 3 1 2 1 1 1 1 2 4 0

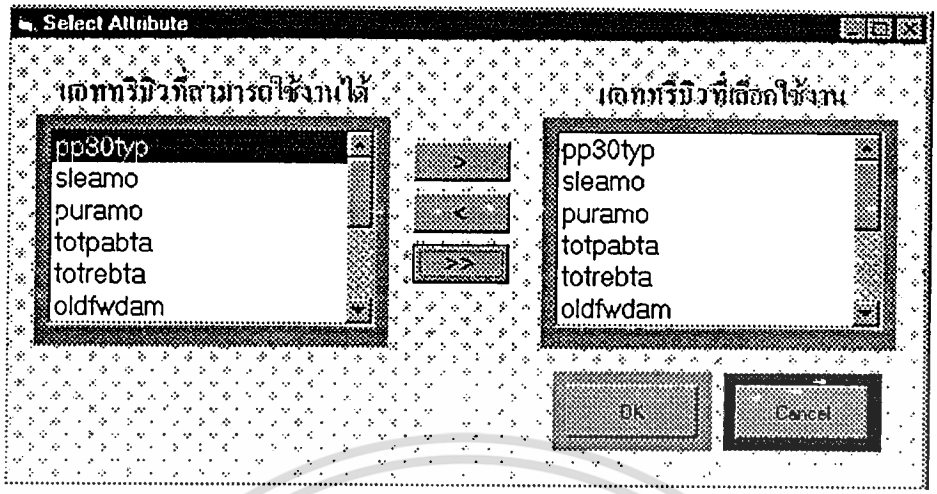
ตารางที่ 6.2 (ต่อ) แสดงลักษณะข้อมูลที่เข้าเทรนนิ่ง

จากตัวอย่างที่แสดงในข้อ 6.1.1 และข้อ 6.1.2 เป็นลักษณะของข้อมูลตัวอย่างที่จะใช้ในการเทรนนิ่ง ซึ่งเมื่อทำการ Import Data แล้วสถานะของ Select Attribute ก็จะถูกเปลี่ยนให้สามารถเรียกใช้งานได้ ตามภาพที่6.1



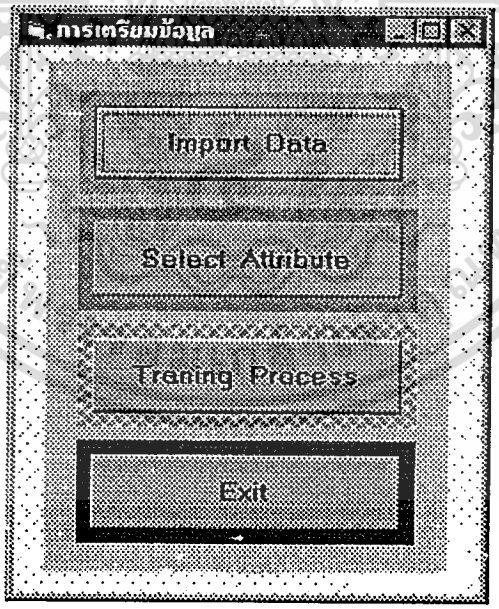
ภาพที่6.1 หน้าจอแสดงการเปลี่ยนสถานะของปุ่ม Select Attribute

จากภาพเมื่อคลิกปุ่ม Select Attribute แล้วจะเข้าสู่หน้าจอการคัดเลือกแอททริบิวที่ต้องการใช้งานดังภาพที่ 6.2



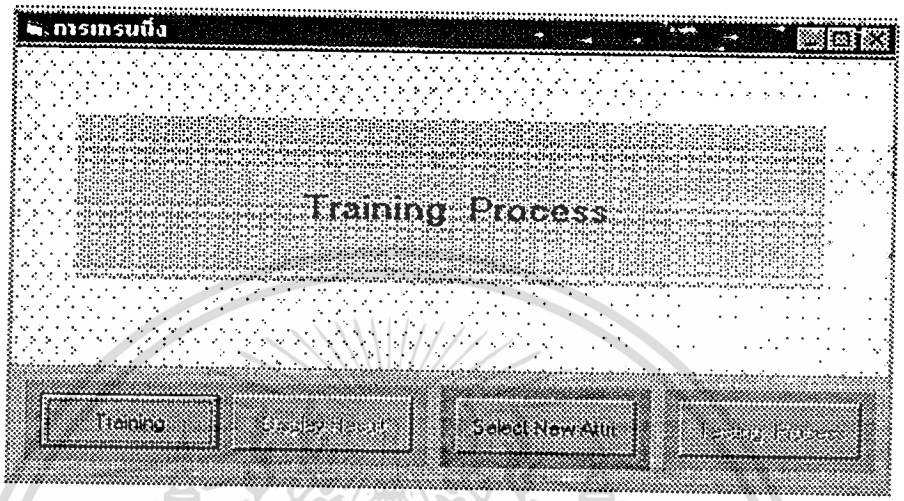
ภาพที่6.2 แสดงหน้าจอการคัดเลือกแอททริบิว

จากภาพทำการเลือกแอททริบิวทั้งหมดจากนั้นก็คลิกปุ่ม OK ก็จะทำการโหลดข้อมูลเหล่านั้นเข้าสู่ระบบและจะปรากฏภาพที่ 6.3



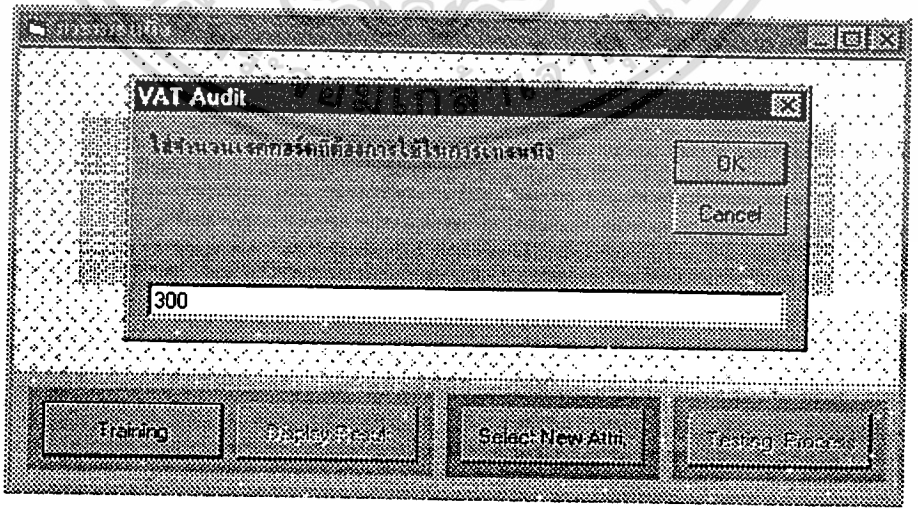
ภาพที่6.3 หน้าจอแสดงการเปลี่ยนสถานะของปุ่ม Training Process

จากภาพจะพบว่าจะมีปุ่มที่สามารถใช้งานได้เพิ่มขึ้นมาอีกปุ่มหนึ่งคือปุ่มของการเทรนนิ่ง จากนั้นก็ทำการเทรนนิ่งข้อมูลโดยการคลิกที่ปุ่ม Training Process ก็จะเข้าสู่หน้าจอตามภาพที่ 6.4



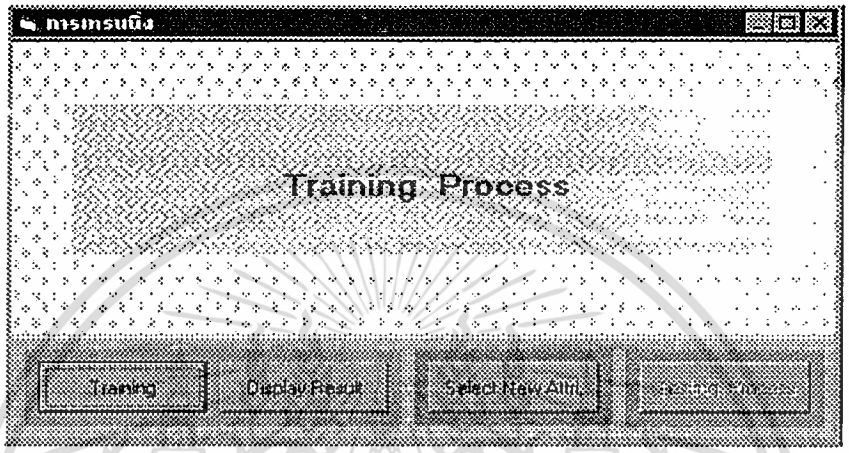
ภาพที่6.4 แสดงหน้าจอกระบวนการเทรนนิ่งข้อมูล

จากภาพจะเห็นว่ามีปุ่มอยู่ 4 ปุ่มแต่สามารถใช้งานได้เพียง 2 ปุ่มเท่านั้นคือ ปุ่มการเทรนนิ่ง และ ปุ่มของการคัดเลือกแอททริบิวใหม่เมื่อต้องการเปลี่ยนแอททริบิวที่จะใช้ในการเทรนนิ่ง โดยเมื่อคลิกปุ่มเทรนนิ่งแล้วจะปรากฏตามจอภาพที่6.5



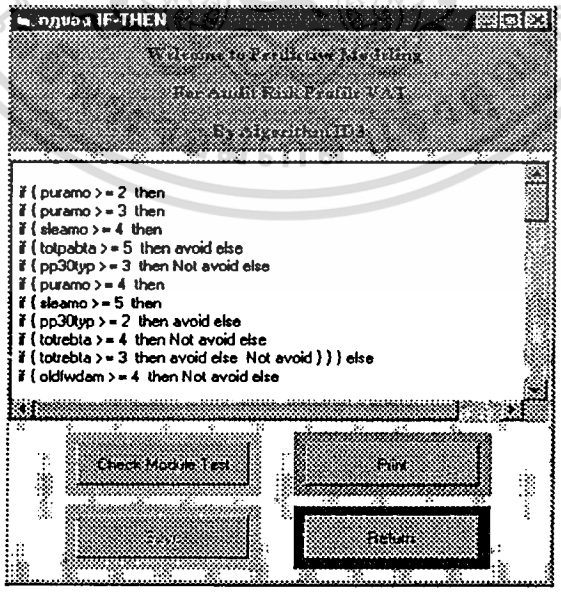
ภาพที่6.5 แสดงหน้าจอรับจำนวนเรคคอร์ดที่จะใช้ในการเทรนนิ่ง

จากภาพเป็นหน้าจอการรับจำนวนข้อมูลที่ต้องการจะใช้ในการเทรนนิ่ง ซึ่งเมื่อใส่จำนวนข้อมูลเสร็จแล้วก็กดปุ่ม OK ระบบจะก็ดำเนินการเทรนนิ่งข้อมูล โดยขณะที่ดำเนินการเทรนนิ่งข้อมูลอยู่นั้นจะเห็นว่าหน้าจอจะไม่ทำงาน แต่เมื่อทำการเทรนนิ่งเรียบร้อยแล้วหน้าจอจะกลับมาทำงาน อีกครั้งดังภาพที่ 6.6



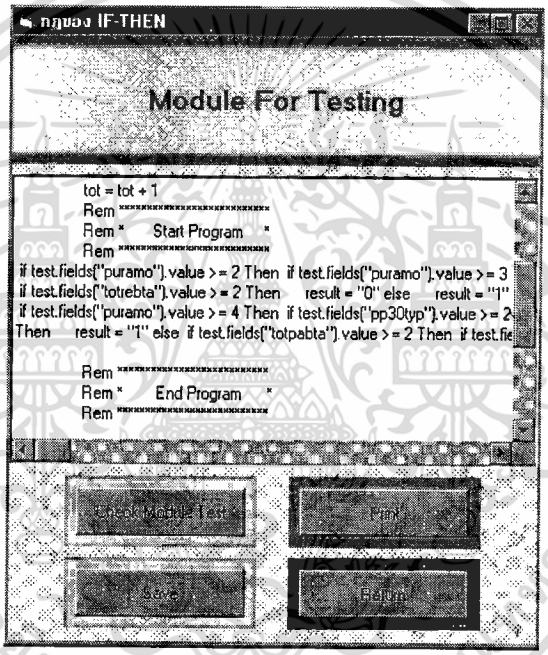
ภาพที่6.6 แสดงหน้าจอเมื่อทำการเทรนนิ่งเรียบร้อยแล้ว

จากภาพจะเห็นว่ามีปุ่มที่สามารถใช้งานได้เพิ่มขึ้นมาอีก 1 ปุ่มคือปุ่มของการแสดงผลลัพท์ที่ได้จากการเทรนนิ่งซึ่งแยกออกเป็น 2 ส่วน โดยส่วนแรกคือกฎของ IF-THEN ซึ่งเมื่อคลิกที่ปุ่มแสดงผลลัพท์แล้วจะปรากฏหน้าจอตามภาพที่ 6.7



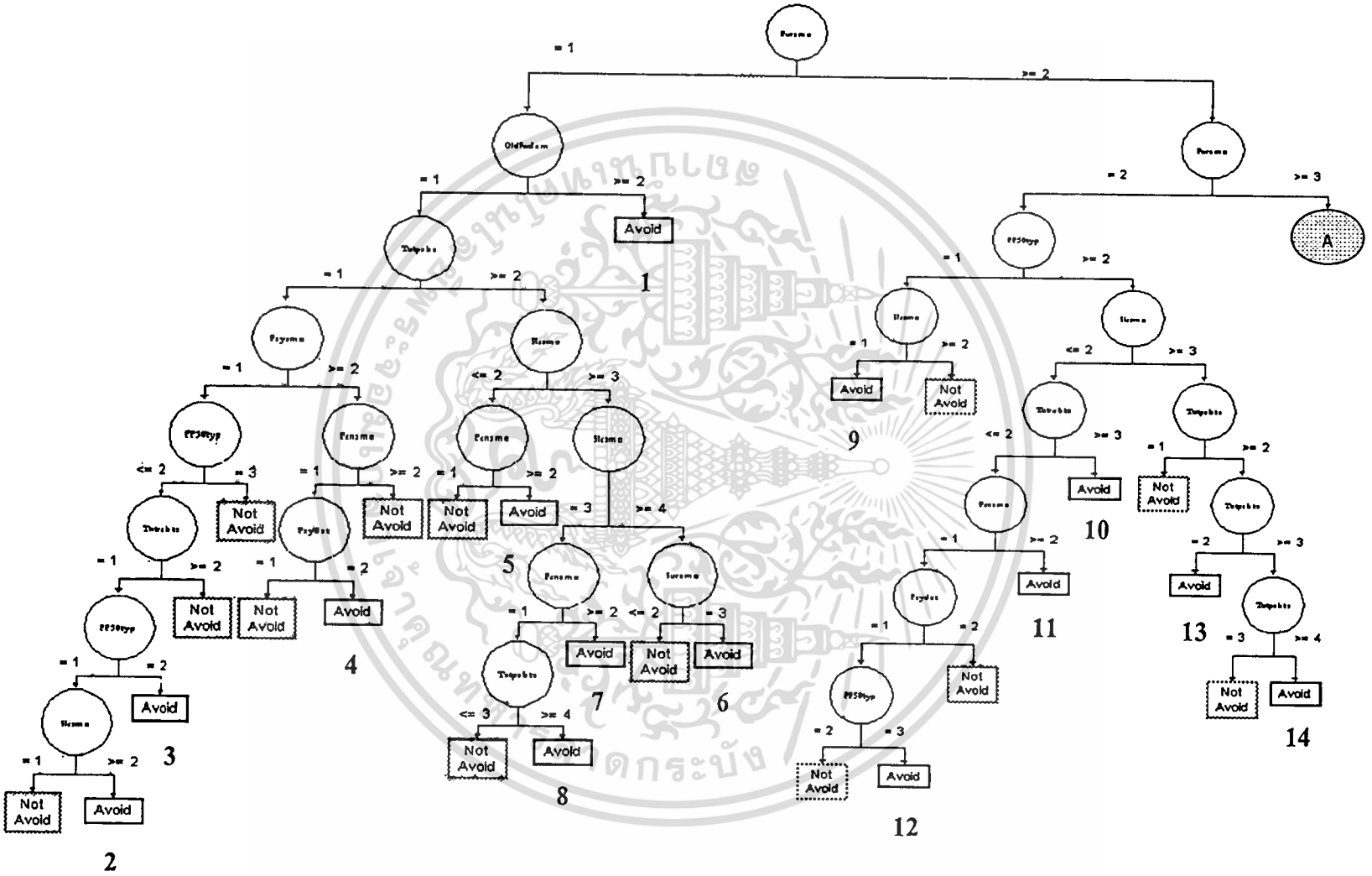
ภาพที่6.7 แสดงหน้าจอผลลัพท์ที่ได้จากการเทรนนิ่งในรูปของกฎ IF-THEN

จากภาพแสดงให้เห็นว่ามีปุ่มที่จะทำงานอยู่ 4 ปุ่มแต่ปุ่มที่สามารถทำงานได้มีเพียง 3 ปุ่มเท่านั้น จากนั้นก็ทำการสั่งพิมพ์กฎ IF-THEN ที่ได้ออกมาโดยกดปุ่มการสั่งพิมพ์ จากผลที่ได้ในรูปของกฎ if-then สามารถเปลี่ยนให้อยู่ในรูปของดีซีชั่นทรี(Decision Tree) เพื่อง่ายและสะดวกในการวิเคราะห์ใช้งาน ซึ่งดีซีชั่นทรี (Decision Tree) ที่ได้มีรูปดังแสดงในภาพที่ 6.9 ส่วนผลลัพท์ที่ได้จากการเทรนนิ่งอีกอย่างหนึ่งสามารถแสดงได้โดยกดปุ่ม Check Module Test ตามภาพที่6.7 จะปรากฏโมดูลที่จะใช้ในการทดสอบความถูกต้องของโมเดลที่ได้ดังภาพที่ 6.8

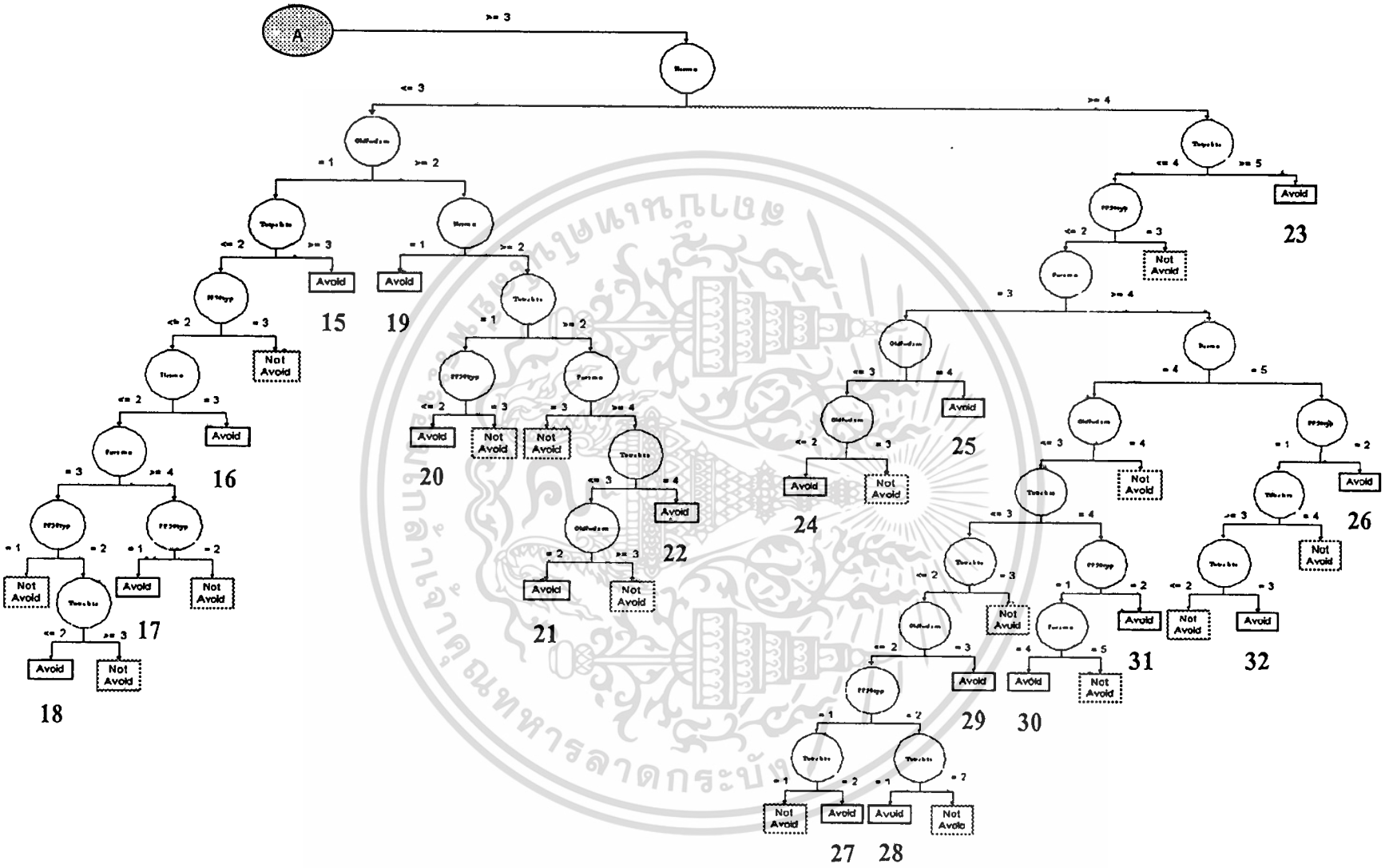


ภาพที่6.8 แสดงหน้าจอโมดูลที่ได้จากการเทรนนิ่ง

จากภาพจะเห็นว่ามีปุ่มที่สามารถใช้งานได้เพิ่มขึ้นมาอีก 1 ปุ่มคือปุ่มจัดเก็บข้อมูล โดยเมื่อทำการตรวจสอบความถูกต้องของโมดูลแล้ว จากนั้นก็ทำการจับเก็บการเปลี่ยนแปลงโดยกดปุ่ม Save แล้วก็กดปุ่ม Return เพื่อกลับมาสู่หน้าจอกระบวนการเทรนนิ่ง



ภาพที่ 6.9 แสดงการสร้างต้นไม้ (Tree) จากเงื่อนไขที่ได้จากการเทรนนิ่ง



ภาพที่ 6.9 (ต่อ) แสดงการสร้างต้นไม้ (Tree) จากเงื่อนไขที่ได้จากการเทรนนิ่ง

จากภาพที่ 6.9 และ ภาพที่ 6.9 (ต่อ) เป็นดีซีชั่นทรี(Decision Tree) ซึ่งรูปแบบที่ได้นั้นยังไม่สามารถที่จะนำไปใช้งานได้ทันทีเพราะผู้ใช้ไม่สามารถที่จะเข้าใจได้จึงต้องมีกระบวนการวิเคราะห์ผลลัพท์ที่ได้ออกมาจากการเทรนนิ่งเพื่อสามารถที่จะให้ผู้ใช้สามารถใช้งานได้ต่อไป

6.2 การวิเคราะห์ผล

ผลลัพท์ที่ได้จากการเทรนนิ่งตามที่กล่าวในหัวข้อ 6.7 นั้นเป็นกฎเกณฑ์ในการปฏิบัติจริงได้ แต่โมเดลที่ได้อยู่ในรูปของกฎ if-then ไม่สะดวกในการนำไปปฏิบัติจึงต้องมีการวิเคราะห์กฎ if-then และเปลี่ยนให้อยู่ในรูปแบบที่สามารถใช้ปฏิบัติได้สะดวก โดยเมื่อพิจารณาดีซีชั่นทรี (Decision Tree) ที่ได้สามารถวิเคราะห์ความหมายของแต่ละกิ่งของต้นไม้ได้ โดยอธิบายเป็นเงื่อนไขที่ผู้เสียภาษีรายนั้นจะมีโอกาสหลีกเลี่ยงภาษีได้ 32 เงื่อนไข ดังนี้

1. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาท และยอดชำระเกินยกมาจากเดือนก่อนมีค่าระหว่าง 5,001 บาท ถึง 50,000 บาท หรือ
2. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาท และยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และยอดรวมภาษีที่ต้องชำระน้อยกว่าหรือเท่ากับ 5,000 บาท และยอดจำนวนเงินที่ชำระน้อยกว่าหรือเท่ากับ 1,000 บาท และ เป็นแบบขอคืนเงินสด และยอดภาษีที่ชำระไว้เกินน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดขายเดือนนี้มีค่าตั้งแต่10,001 บาทขึ้นไป หรือ
3. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดจำนวนเงินที่ชำระน้อยกว่าหรือเท่ากับ 1,000 บาท และ เป็นแบบขอคืนเครดิต และ ยอดภาษีที่ชำระไว้เกินน้อยกว่าหรือเท่ากับ 5,000 บาท หรือ
4. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดจำนวนเงินที่ชำระมีค่าตั้งแต่ 1,001 บาทขึ้นไป และ ยอดเบี้ยปรับน้อยกว่าหรือเท่ากับ 500 บาทและ วันที่ชำระภาษีนั้นมีการชำระเกินกำหนดเวลา หรือ
5. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าตั้งแต่ 5,001 บาทขึ้นไป และ ยอดขายเดือนนี้มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และ ยอดเบี้ยปรับมีค่าตั้งแต่ 501 บาทขึ้นไป หรือ

6. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าตั้งแต่ 5,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดเงินเพิ่มมากกว่า 2,500 บาทขึ้นไป หรือ
7. ถ้ายอดซื้อในเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าตั้งแต่ 50,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาท และ ยอดเบี้ยปรับน้อยกว่าหรือเท่ากับ 500 บาท หรือ
8. ถ้ายอดซื้อในเดือนนี้ น้อยกว่าหรือเท่ากับ 10,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระ มีค่าตั้งแต่ 50,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาท และ ยอดเบี้ยปรับ น้อยกว่าหรือเท่ากับ 500 บาท หรือ
9. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบขอคืนเงินสดและยอดขายเดือนนี้น้อยกว่าหรือเท่ากับ 10,000 บาท หรือ
10. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบขอคืนเครดิตหรือเป็นแบบชำระ และยอดขายเดือนนี้มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่าตั้งแต่ 50,001 บาทขึ้นไป หรือ
11. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบขอคืนเครดิตหรือเป็นแบบชำระ และยอดขายเดือนนี้ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท และ ยอดเบี้ยปรับมีค่าตั้งแต่ 501 บาทขึ้นไป หรือ
12. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบที่มีเงินชำระ และยอดขายเดือนนี้มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท และ ยอดเบี้ยปรับมีค่าน้อยกว่าหรือเท่ากับ 500 บาท และ วันที่ชำระภาษีตรงตามกำหนดเวลาการชำระ หรือ
13. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบขอคืนเครดิตหรือเป็นแบบที่มีเงินชำระ และยอดขายเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดรวมภาษีที่ต้องชำระ มีค่าตั้งแต่ 5,001 ถึง 10,000 บาท หรือ
14. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 100,000 บาทและเป็นแบบขอคืนเครดิตหรือเป็นแบบที่มีเงินชำระ และยอดขายเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ

ยอดรวมภาษีที่ต้องชำระมีค่าตั้งแต่ 50,001 ถึง 100,000 บาท หรือ

15. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าน้อยกว่าหรือเท่ากับ 1,000,000 บาท และ ยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่ามากกว่าหรือเท่ากับ 10,001 บาทขึ้นไป หรือ
16. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาทและ ยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 10,000 บาทและเป็นแบบขอคืนเงินสดหรือเป็นแบบเครดิต หรือ
17. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาทและ ยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 10,000 บาทและเป็นแบบขอคืนเป็นเงินสด หรือ
18. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาท และ ยอดขายเดือนนี้ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาทและ ยอดชำระเกินยกมาจากเดือนก่อนน้อยกว่าหรือเท่ากับ 5,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 10,000 บาทและเป็นแบบขอคืนเป็นเครดิต และ ยอดภาษีที่ชำระไว้เกิน มีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท หรือ
19. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าน้อยกว่าหรือเท่ากับ 10,000 บาท และยอดชำระเกินยกมาจากเดือนก่อน มีค่าตั้งแต่ 5,001 บาทขึ้นไป หรือ
20. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 10,001 ถึง 1,000,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนมีค่าตั้งแต่ 5,001 บาทขึ้นไป และ ยอดภาษีที่ชำระไว้เกิน น้อยกว่าหรือเท่ากับ 5,000 บาท และเป็นแบบขอคืนเงินสดหรือเป็นแบบขอคืนเครดิต หรือ
21. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้มีค่าตั้งแต่ 10,001 ถึง 1,000,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนมีค่าตั้งแต่ 5,001 ถึง 50,000 บาท และ ยอดภาษีที่ชำระไว้เกิน น้อยกว่าหรือเท่ากับ 500,000 บาท หรือ
22. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่

10,001 ถึง 1,000,000 บาทและยอดชำระเกินยกมาจากเดือนก่อนมีค่าตั้งแต่ 5,001 บาท ขึ้นไป และ ยอดภาษีที่ชำระไว้เกิน มากกว่า 500,000 บาท หรือ

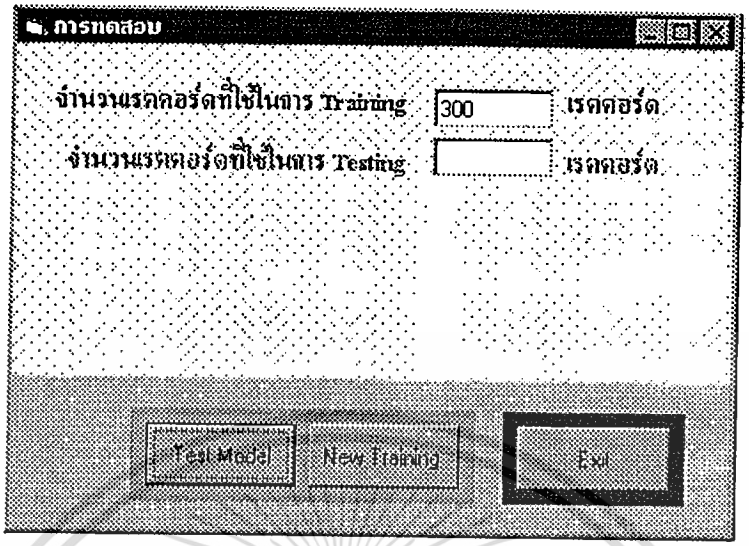
23. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดรวมภาษีที่ต้องชำระมีค่ามากกว่า 100,000 บาท หรือ
24. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาท และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสดหรือเป็นแบบขอคืนเครดิต และยอดชำระเกินยกมาจากเดือนก่อนมีค่าน้อยกว่าหรือเท่ากับ50,000 บาท หรือ
25. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 100,001 ถึง 1,000,000 บาท และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสดหรือเป็นแบบขอคืนเครดิต และยอดชำระเกินยกมาจากเดือนก่อนมีค่ามากกว่า 500,000 บาท หรือ
26. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้มีค่ามากกว่า 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และ เป็นแบบขอคืนเครดิต หรือ
27. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสด และยอดชำระเกินยกมาจากเดือนก่อนมีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท และ ยอดภาษีที่ชำระไว้เกินมีค่าตั้งแต่ 5,001 ถึง 50,000 บาท หรือ
28. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระมีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเครดิต และยอดชำระเกินยกมาจากเดือนก่อนมีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่าน้อยกว่าหรือเท่ากับ 5,000 บาท หรือ
29. ถ้ายอดซื้อในเดือนนี้มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสด หรือ เป็นแบบขอคืนเครดิต และยอดชำระเกินยกมาจากเดือนก่อน มีค่าตั้งแต่ 50,001 ถึง 500,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่าน้อยกว่าหรือเท่ากับ 50,000 บาท หรือ

30. ถ้ายอดซื้อในเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสด และยอดชำระเกินยกมาจากเดือนก่อน มีค่าน้อยกว่าหรือเท่ากับ 500,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่ามากกว่า 500,000 บาท หรือ
31. ถ้ายอดซื้อในเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่าตั้งแต่ 1,000,001 ถึง 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเครดิต และยอดชำระเกินยกมาจากเดือนก่อน มีค่าน้อยกว่าหรือเท่ากับ 500,000 บาท และ ยอดภาษีที่ชำระไว้เกิน มีค่ามากกว่า 500,000 บาท หรือ
32. ถ้ายอดซื้อในเดือนนี้ มีค่าตั้งแต่ 1,000,001 บาทขึ้นไป และ ยอดขายเดือนนี้ มีค่ามากกว่า 10,000,000 บาท และ ยอดรวมภาษีที่ต้องชำระ มีค่าน้อยกว่าหรือเท่ากับ 100,000 บาท และเป็นแบบขอคืนเงินสด และ ยอดภาษีที่ชำระไว้เกิน มีค่าตั้งแต่ 50,001 ถึง 500,000 บาท หรือ

อย่างไรก็ตามแม้ว่าจะได้โมเดลแล้วแต่ก็ยังไม่สามารถบอกได้ว่าโมเดลที่ได้มาจะเหมาะสมและมีความน่าเชื่อถือมากพอที่จะนำไปใช้ปฏิบัติหรือไม่ จึงต้องมีการดำเนินการทดสอบโมเดลที่ได้ก่อนซึ่งจะกล่าวในหัวข้อถัดไป

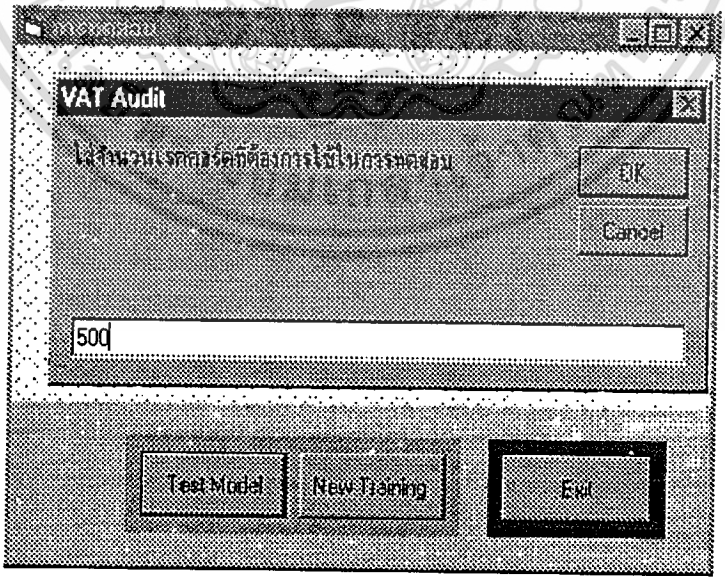
6.3 การทดสอบ

การทดสอบ เป็นการนำโมเดลที่ได้มาทำการตรวจสอบว่ามีความน่าเชื่อมากแค่ไหน และมีความน่าเชื่อถือเพียงพอที่จะนำไปใช้ปฏิบัติหรือไม่ โดยการตรวจสอบความน่าเชื่อถือของโมเดลนั้นก็ทำโดยการนำข้อมูลที่ไม่ใช่ข้อมูลที่นำมาใช้ในการสร้างโมเดลมาทดสอบและข้อมูลที่นำมาทดสอบนั้นจะต้องมีผลการตรวจสอบมาแล้ว เพื่อมาเช็คว่าผลการตรวจสอบกับผลของการใช้โมเดลในการทำนายว่ามีความน่าเชื่อถือเพียงใด โดยนำเอาโมดูลที่ได้ไปทำการคอมไพล์แล้วนำไปทดสอบกับข้อมูลชุดใหม่ เพื่อเปรียบเทียบและนับจำนวนรายที่ผลการตรวจตรงกับผลการใช้โมเดล โดยเมื่อทำการเรียกใช้โปรแกรมสำหรับทดสอบความถูกต้องของรูปแบบที่ได้จะปรากฏตามภาพที่ 6.10



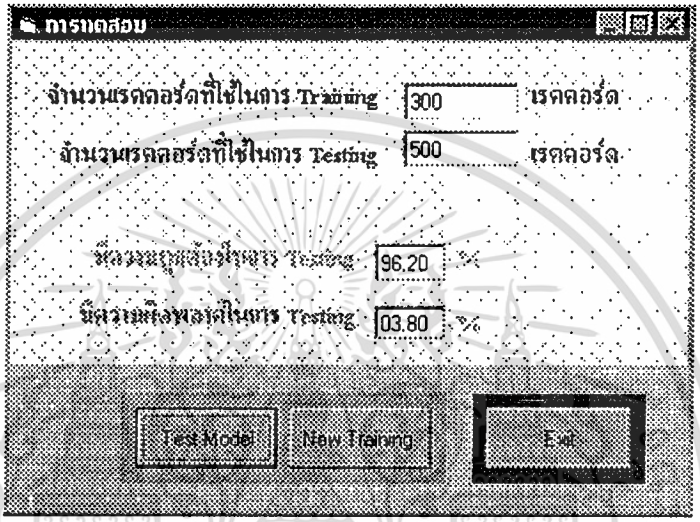
ภาพที่6.10 แสดงหน้าจอกระบวนการทดสอบรูปแบบที่ได้จากการเทรนนิ่ง

จากภาพจะปรากฏจำนวนเรคคอร์ดที่เราใช้ในการเทรนนิ่งซึ่งในที่นี้ใช้ 300 เรคคอร์ดในการเทรนนิ่ง ซึ่งจากภาพจะเห็นว่ามีปุ่มที่สามารถใช้งานได้อยู่ 3 ปุ่มซึ่งเมื่อต้องการจะทดสอบรูปแบบที่ได้จากการเทรนนิ่งก็ให้กดปุ่ม Test Model ก็จะปรากฏดังจอภาพที่6.11



ภาพที่6.11 แสดงหน้าจอให้ใส่จำนวนเรคคอร์ดที่ต้องการทดสอบ

จากภาพทำการใส่จำนวนเรคคอร์ดที่จะใช้ในการเทรนนิ่งลงไป ซึ่งในที่นี้จะใช้ข้อมูลในการทดสอบรูปแบบที่ได้จำนวน 500 เรคคอร์ด เมื่อใส่จำนวนเรคคอร์ดเรียบร้อยแล้วก็ให้คลิกปุ่ม OK ระบบก็จะทำการทดสอบรูปแบบที่ได้จากการเทรนนิ่งกับข้อมูลที่ใส่เข้าไปทดสอบซึ่งผลที่ได้จากการทดสอบจะปรากฏตามจอภาพที่6.12



ภาพที่6.12 แสดงหน้าจอผลที่ได้จากการทดสอบ

จากการทดสอบตามภาพที่ 6.12 จะเห็นว่าการพัฒนาระบบได้ใช้ข้อมูลในการเทรนนิ่งทั้งหมด 300 ราย และมีข้อมูลที่ใช้ในการทดสอบทั้งหมด 500 ราย ซึ่งผลที่ได้จากการทดสอบ ปรากฏว่าผลการตรวจสอบตรงกับผลการทำนายของโมเดลคิดเป็น 96.20 เปอร์เซ็นต์ของข้อมูลทั้งหมดที่ทดสอบซึ่งเป็นเปอร์เซ็นต์ที่สูงพอสมควร และ เมื่อลองนำโมเดลที่ได้ย้อนกลับไปดำเนินการทดสอบกับข้อมูล ที่ใช้ในการเทรนนิ่งแล้วปรากฏว่าผลของความน่าเชื่อถือเป็น 100 เปอร์เซ็นซึ่งแสดงให้เห็นได้ว่าโมเดลที่ได้จากการเทรนนิ่งมีความน่าเชื่อถือพอที่จะสามารถนำไปใช้ในการคัดเลือกรายผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีได้

อย่างไรก็ตามสามารถที่จะทำให้การทำงานดีขึ้นโดยทำการแต่งกิ่งต้นไม้ตามภาพที่ 6.9 และภาพที่ 6.9 (ต่อ) ในกิ่งที่ 27 28 18 12 และ 2 หลังจากนั้นเมื่อทำการทดสอบต้นไม้ที่ได้ทำการแต่งกิ่งแล้วพบว่า มีเปอร์เซ็นต์ของความถูกต้อง 90.80 เปอร์เซ็นต์ แตกต่างกับตอนที่ยังไม่ได้แต่งกิ่งต้นไม้

เพียง 0.80 เปอร์เซ็นต์ ซึ่งความน่าเชื่อถือไม่แตกต่างกันมากนักและเมื่อเทียบกับผลของการตรวจโดยใช้เงื่อนไขของการคัดเลือกรายแบบในปัจจุบัน ซึ่งเป็นข้อมูลชุดเดียวกับข้อมูลที่ทดสอบกับเงื่อนไขที่ได้จากการเทรนนิ่งพบว่ามีเปอร์เซ็นของความถูกต้องเพียง 34.90 เปอร์เซ็นต์ แสดงให้เห็นว่าโมเดลที่ได้จากการเทรนนิ่งมีประสิทธิภาพในการคัดเลือกรายผู้ประกอบการที่หลีกเลี่ยงภาษีมูลค่าเพิ่มได้ดีกว่าแบบปัจจุบัน

บทที่ 7
บทสรุป

โครงการพัฒนาระบบนี้เป็นโครงการที่จัดทำขึ้นมาเพื่อนำเสนอให้เห็นถึงประโยชน์ของการนำทฤษฎีของดาต้าไมนิ่งมาใช้เพิ่มประสิทธิภาพในการค้นหาผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่ม โดยเงื่อนไขในการคัดเลือกรายในปัจจุบันไม่สามารถที่จะค้นหาผู้ประกอบการที่มีแนวโน้มที่จะหลีกเลี่ยงภาษีที่ได้ดีเท่าที่ควร เพราะเงื่อนไขในปัจจุบันที่ใช้อยู่เกิดจากความคิดของบุคคลกลุ่มหนึ่งเท่านั้นซึ่งไม่สามารถยืนยันได้ว่าเงื่อนไขดังกล่าวถูกต้องหรือไม่ โดยในโครงการนี้ได้นำเอาเทคนิคของ Classification เข้ามาใช้ให้เกิดประโยชน์

7.1 สรุปหลักการที่ใช้ในระบบระบบ

เทคนิคที่จะใช้ในการเลือกรายของผู้ประกอบการที่มีแนวโน้มในการหลีกเลี่ยงภาษีมูลค่าเพิ่มเพื่อสุ่มตรวจของระบบนี้ได้นำเทคนิคของการ Classification ในส่วนของ decision tree มาใช้ในการพัฒนาระบบ ซึ่งเป็นเทคนิคของการสร้างรูปแบบการทำนายโดยอาศัยประสบการการเรียนรู้หรือความสนใจในสิ่งบางอย่างที่จะเป็นหลักสำคัญในการตัดสินใจ โดยอัลกอริทึ่มที่เลือกใช้คืออัลกอริทึ่มของ ID3 ซึ่งเป็นกระบวนการเรียนรู้แบบมีรูปแบบของอินพุตและเอ้าท์พุตไว้ก่อนแล้วจากข้อมูลตัวอย่างซึ่งเป็นข้อมูลเก่าๆ จากนั้นจึงนำเอาข้อมูลที่ได้มาเพื่อที่จะเลือกหา attribute ที่มีความสำคัญเพียงพอที่จะสามารถคัดเลือกรายผู้ประกอบการที่หลีกเลี่ยงภาษีออกมาได้ จากนั้นก็นำเอาแอททริบิวที่เราสนใจไปดำเนินการค้นหาว่าแอททริบิวไหนมีความสำคัญมากที่สุด โดยการใช้วิธีการของ information gain เพื่อเลือก attribute ที่มีค่ามากที่สุดเป็น attribute สำหรับทดสอบเพื่อทำการสร้างดีซีชั่นทรี (Decision Tree) ให้สมบูรณ์ เมื่อได้ต้นไม้(Tree)หรือโมเดล แล้วก็ต้องนำเอาโมเดลที่ได้ไปทำการทดสอบกับข้อมูลอื่นๆที่ไม่ใช่ข้อมูลที่ใช้ในการสร้างโมเดล เพื่อทดสอบโมเดลที่ได้ว่ามีความถูกต้องหรือไม่ด้วย

7.2 สรุปกระบวนการในการทำงาน

โครงการนี้เริ่มต้นจากการค้นหาแอททริบิวที่มีความสำคัญเพียงพอที่จะระบุได้ว่าผู้ประกอบการรายไหนมีแนวโน้มที่จะหลีกเลี่ยงภาษีมูลค่าเพิ่ม โดยจากการวิเคราะห์แอททริบิวที่เกี่ยวข้องทำให้คัดเลือกแอททริบิวที่คาดว่าจะเป็นปัจจัยที่มีความสำคัญขึ้นมา 12 แอททริบิวซึ่งมีทั้ง

แอททริบิวที่เป็นตัวเลขที่ต่อเนื่อง แอททริบิวที่เป็นตัวเลขที่ไม่ต่อเนื่อง และ บางแอททริบิวก็ยังเป็นข้อความ จึงต้องนำข้อมูลเหล่านั้นมาทำการแปลงรูปแบบให้เหมาะสมที่จะทำงานได้โดยทำการแปลงข้อมูลต่างๆที่ได้มาให้เป็นตัวเลขทั้งหมด ซึ่งถ้าเป็นตัวเลขที่ต่อเนื่องก็จะต้องทำการจัดกลุ่มข้อมูลเสียก่อนโดยใช้แอพพลิเคชั่น MATLAB เข้ามาช่วยสร้างฮีทโทรแกรมเพื่อที่จะกำหนดได้ว่าจะทำการแบ่งข้อมูลเหล่านั้นออกเป็นกี่กลุ่ม เมื่อได้จำนวนกลุ่มเรียบร้อยแล้วก็ดำเนินการสร้างโปรแกรมขึ้นมาโดยในโปรแกรมเมื่อทำการนำเข้าข้อมูลจะมีการทำความสะอาดข้อมูลเหล่านั้นเพราะถ้าหากข้อมูลเป็นขยะเข้าไปก็จะไม่มีประโยชน์พร้อมทั้งทำการแปลงข้อมูลให้อยู่ในรูปที่ได้แบ่งกลุ่มไว้แล้ว จากนั้นถึงจะนำข้อมูลที่ได้ไปทำการเทรนนิ่ง ผลที่ได้ก็จะออกมาเป็นโมเดลที่เป็นกฎ IF-THEN ซึ่งยังไม่สามารถใช้งานได้ทันทีจึงต้องมีการวิเคราะห์ผลที่ได้ออกมาให้เป็นรูปแบบที่มนุษย์สามารถที่จะอ่านได้เข้าใจจากนั้นจึงนำผลที่ได้ไปทดสอบกับข้อมูลอื่นที่ไม่ใช่ข้อมูลที่ใช้ในการสร้างโมเดลอีกครั้งเพื่อเป็นการทดสอบความถูกต้องของโมเดลที่ได้

7.3 สรุปผลการทดสอบ

เมื่อนำเอาผลที่ได้จากการเทรนนิ่งไปทำการทดสอบกับข้อมูลจำนวน 500 เรคคอร์ดแล้วพบว่ามีความถูกต้องอยู่ในระดับที่สูงเพียงพอที่หากจะนำเอาเทคนิคดังกล่าวข้างต้นมาใช้กับการสุ่มหาข้อมูลของผู้ประกอบการที่มีอัตราเสี่ยงต่อการหลีกเลี่ยงภาษีมูลค่าเพิ่มก็น่าจะทำให้ได้การคัดเลือกรายขึ้นมาสุ่มตรวจมีประสิทธิภาพและรัดกุมยิ่งขึ้นเพราะว่าเปอร์เซ็นต์ของความถูกต้องในการคัดเลือกรายจากเงื่อนไขของโมเดลที่ได้จากการเทรนนิ่งมีเปอร์เซ็นต์ที่สูงกว่าเมื่อใช้หลักเกณฑ์ในปัจจุบันคัดเลือกรายจากข้อมูลเดียวกัน

7.4 ข้อเสนอแนะ

ระบบนี้สามารถที่จะลดขั้นตอนการทำงานลงได้ ด้วยการใช้ข้อมูลจาก Datawarehouse เพราะว่าจะเป็นข้อมูลที่ได้รับ การทำความสะอาดมาเรียบร้อยแล้วทำให้สามารถลดขั้นตอนในการทำความสะอาดข้อมูลลงได้ ทั้งยังให้ความมั่นใจในความถูกต้องเพิ่มขึ้นด้วย

เอกสารอ้างอิง

[1] Peter Cabena, et.al. 1998. **Discovering Data Mining: From Concept to Implementation.** New Jersey. Prentice Hall PTR.

[2] Dr.Dobb's Journal. 1996. **Algorithm Alley.** [Online]. Avialable : http://www.ddj.com/ftp/1996/1996.06

[3] จิราภรณ์ แจ่มชัดใจ. 2540. **"บริหารการตลาดด้วย Data Mining."** สารเนคเทค. เดือนกันยายน-ตุลาคม 2540 : 17-22.

[4] สุรพล ศรีบุญทรง. 2541. **"Data Mining เครื่องมือการตลาดยุคดิจิตอล."** IT.SOFT. มกราคม 2541 : 116-123.

[5] กรมสรรพากร. ตุลาคม 2539. **ภาษีมูลค่าเพิ่ม.** กรุงเทพฯ : กรมสรรพากร.

[6] สำนักมาตรฐานกรรมวิธี กรมสรรพากร. ธันวาคม 2540. **Audit Selection Criteria.** กรุงเทพฯ : กรมสรรพากร.

[7] สำนักกฎหมาย กรมสรรพากร. ธันวาคม 2540. **ประมวลกฎหมาย.** กรุงเทพฯ : กรมสรรพากร.

[8] Tom M. Mitchell. 1997. **MACHINE LEARNING.** United of America. The McGraw-Hill Companies Inc.

[9] Helge Grenager Solheim. 1996. **ID3.** [Online]. Avialable : http://www.recursive-partitioning.com/Classification_Trees/more36.html.

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาวกรรณิกา สุดกังวล
สถานที่เกิด	จังหวัดกรุงเทพมหานคร
การศึกษา	บริหารธุรกิจบัณฑิต (คอมพิวเตอร์) มหาวิทยาลัยรังสิต
ปัจจุบัน	รับราชการตำแหน่ง นักวิชาการคอมพิวเตอร์ 5
	สังกัดสำนักเทคโนโลยีสารสนเทศ กรมสรรพากร