ขั้นตอนวิธีคัดเลือกหลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน

นาวสาวชรีย์วรรณ สิริศรีสัมฤทธิ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคณนา ภาควิชาคณิตศาสตร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2550

ASSOCIATION RULE SELECTION ALGORITHM BASESD ON WEAK SUPPORT

Miss Chareewan Sirisrisumrit

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computational Science
Department of Mathematics
Faculty of Science
Chulalongkorn University
Academic Year 2007

หัวข้อวิทยานิพนธ์	ขั้นตอนวิธีคัดเลือกหลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน
โดย	นางสาวชรีย์วรรณ สิริศรีสัมฤทธิ์
สาขาวิชา	วิทยาการคณนา
อาจารย์ที่ปรึกษา	ผู้ช่วยศาตราจารย์ ดร. กรุง สินอภิรมย์สราญ

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิทยาศาสตร์

(ศาสตราจารย์ ดร. สุพจน์ หารหนองบัว)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(อาจารย์ ดร. จารุโลจน์ จงสถิตย์วัฒนา)

..... อาจารย์ที่ปรึกษา

(ผู้ช่วยศาตราจารย์ ดร. กรุง สินอภิรมย์สราญ)

..... กรรมการ

(อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล)

ชรีย์วรรณ สิริศรีสัมฤทธิ์ : ขั้นตอนวิธีคัดเลือกหลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน. (ASSOCIATION RULE SELECTION ALGORITHM BASED ON WEAK SUPPORT) อ. ที่ปรึกษา : ผศ. ดร. กรุง สินอภิรมย์สราญ, 114 หน้า.

การสร้างหลักเกณฑ์เชื่อมโยงเป็นเทคนิคหนึ่งที่สำคัญในการค้นความรู้จากข้อมูล ซึ่งถูกใช้วิเคราะห์ข้อมูลมาเก็ตบาสเก็ต วัตถุประสงค์หลักคือการดึงหลักเกณฑ์ที่แสดงความสัมพันธ์ที่เชื่อมโยงระหว่างสินค้าภายในการซื้อหนึ่งใบเสร็จ วิธีการแบบตรงไปตรงมาที่สร้างหลักเกณฑ์ที่เป็นไปได้ทั้งหมดเป็นเรื่องที่ทำได้ยากเนื่องจากปัญหาการระเบิดออกของความเป็นไปได้ของหลักเกณฑ์ ดังนั้นนักวิจัยจึงใช้ตัววัดความน่าสนใจเพื่อลดปริมาณการพิจารณาหลักเกณฑ์ที่ไม่น่าสนใจตามค่าสนับสนุนต่ำสุด และค่าความเชื่อมั่นต่ำสุด ในที่นี้ค่าสนับสนุนของหลักเกณฑ์แสดงถึงการนำหลักเกณฑ์ไปใช้ได้กับข้อมูลเมื่อเทียบกับข้อมูลทั้งหมด โดยคำนวณสัดส่วนของข้อมูลที่มีสินค้าที่ปรากฏในหลักเกณฑ์กับปริมาณข้อมูลการซื้อทั้งหมด ในอีกมุมมองหนึ่งคือค่าความเชื่อมั่นของหลักเกณฑ์แสดงถึงความน่าเชื่อถือของหลักเกณฑ์ ในงานวิจัยนี้นำเสนอตัววัดความน่าสนใจอีกหนึ่งตัวคือค่าสนับสนุนแบบอ่อน จากพื้นฐานทางตรรกศาสตร์จริงหรือเท็จ ข้อมูลที่ขัดแย้งกับหลักเกณฑ์เชื่อมโยงจะเหมือนกับข้อมูลที่ขัดแย้งกับประพจน์ถ้า-แล้ว ดังนั้นเรานิยามค่าสนับสนุนแบบอ่อนคือค่าที่พิจารณาเฉพาะระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์เปรียบเทียบกับระเบียนทั้งหมด สภาพไวถูกใช้ในการเปรียบเทียบหลักเกณฑ์ที่ได้จากค่าสนับสนุนแบบอ่อนบวกค่าความเชื่อมั่นกับค่าสนับสนุนดังเดิมบวกค่าความเชื่อมั่น การเลือกใช้สภาพไวก็เนื่องมาจากกลุ่มของหลักเกณฑ์ที่ถูกสร้างขึ้นจากทั้งสองรูปแบบมีความแตกต่างกันมาก ผลลัพธ์ที่ได้จากการเปรียบเทียบสภาพไว พบว่ากลุ่มของหลักเกณฑ์ที่เกิดขึ้นจากค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่นให้ค่าที่ดีกว่าค่าที่ได้จากกลุ่มของหลักเกณฑ์ที่เกิดจากค่าสนับสนุนและค่าความเชื่อมั่น

ภาควิชา คณิตศาสตร์ ลายมือชื่อนิสิต.....
สาขาวิชา วิทยาการคณนา ลายมือชื่ออาจารย์ปรึกษา.....
ปีการศึกษา 2550

4772262723 : MAJOR COMPUTATIONAL SCIENCE
KEY WORD: ASSOCIATION RULE / MINIMUM SUPPORT / MINIMUM CONFIDENCE / MINIMUM WEAK SUPPORT / SENSITIVITY
CHAREEWAN SIRISRISUMRIT : ASSOCIATION RULE SELECTION ALGORITHM BASED ON WEAK SUPPORT. THESIS ADVISOR : ASST. PROF. KRUNG SINAPIROMSARAN, Ph.D. [114] pp.

Association rule mining is one of a knowledge discovery methodology that has been used to analyze the market basket data. The aim of this methodology is to extract rules that address a strong association among purchased items appearing within transactional data. The naïve method of searching for all possibilities are prohibitive because of the combinatorial explosions of rules. Therefore, researchers use two interesting measures to narrow the search based on the minimum support and minimum confidence. The support of the rule exhibits the applicable of the rule with respect to the whole transaction by computing the ratio of transactions containing related items versus the total number of transactions. The confidence, on the other hand, demonstrates the reliability of the rule. This thesis suggests another numerical interesting measure called the weak support. Based on the true-false logic, the contradiction of association rule is the same as the contradiction of if-then rules and hence we define the weak support measure as the non-contradictory of the rule from the whole transactions. The sensitivity is used to compare rules from the weak support together with confidence against rules from the original support with confidence. The sensitivity is selected due to distinct rules and unequal number of generated rules. The results from the sensitivity comparison between the group of rules from weak support and confidence show better value than the sensitivity from the group of rules from support and confidence.

Department **Mathematics** Student's signature.....

Field of study **Computational Science** Advisor's signature.....

Academic year **2007**

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. กรุง สินอภิรมย์สราญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ท่านได้กรุณาให้ความรู้ คำแนะนำ และคำปรึกษาต่างๆ ที่ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดี

ขอกราบขอบพระคุณ อาจารย์ ดร. จารุโลจน์ จงสถิตย์วัฒนา ประธานกรรมการ อาจารย์ ดร. สิริพันธุ์ สงวนสินธุกุล กรรมการ ที่ได้ให้คำปรึกษาและการช่วยเหลือตลอดระยะเวลาในการทำงานวิจัยนี้ ซึ่งทำให้วิทยานิพนธ์ฉบับนี้มีความสมบูรณ์มากยิ่งขึ้น

ขอกราบขอบพระคุณบิดา มารดา ตลอดจนพี่น้องในครอบครัวที่คอยเป็นกำลังใจและช่วยเหลือผู้วิจัยมาโดยตลอด และขอบคุณเพื่อนๆ ทุกคนที่เป็นกำลังใจให้

ชรีย์วรรณ สิริศรีสัมฤทธิ์

ผู้เขียนวิทยานิพนธ์

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1

บทนำ

ข้อมูลเป็นหัวใจสำคัญของการทำธุรกิจ ทำให้องค์กรต่างๆ ไม่ว่าจะเป็นองค์กรขนาดใหญ่ ขนาดกลาง หรือขนาดเล็ก ต่างก็ต้องจัดเก็บข้อมูลที่สำคัญขององค์กร โดยนำมาใช้สรุป รวบรวม ประมวลผล และวิเคราะห์ เพื่อให้เกิดประโยชน์แก่องค์กรสูงสุด ในอดีตที่ผ่านมาการวิเคราะห์ข้อมูลหรือการจัดการเกี่ยวกับข้อมูลต้องอาศัยผู้ที่มีความเชี่ยวชาญ เพื่อให้สามารถจัดเก็บและวิเคราะห์ข้อมูลผ่านฐานข้อมูลได้สะดวกและรวดเร็ว แต่เนื่องจากข้อมูลที่มีมากจนเกินไป ทำให้มนุษย์คิดค้นวิธีการทำคลังข้อมูล (data warehousing) [1,2] และการทำเหมืองข้อมูล (data mining) [1,3,4,5,6] เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์และจัดการข้อมูลขนาดใหญ่

การทำคลังข้อมูลเป็นเทคโนโลยีที่เน้นการรวบรวมข้อมูลจากหลายฐานข้อมูล และประมวลผลข้อมูลอย่างมีประสิทธิภาพ โดยเรียกข้อมูลที่ถูกรวบรวมว่า คลังข้อมูล (data warehouse) แต่อย่างไรก็ตามการประมวลข้อมูลซึ่งได้แก่การทำโอแลป (online analytic processing : OLAP) ยังคงต้องใช้มนุษย์ที่มีความเชี่ยวชาญเพื่อกำหนดทิศทางและพิจารณาการประมวลผลสรุปที่นำไปสู่ข้อมูลสารสนเทศที่สำคัญ

การทำเหมืองข้อมูลเป็นการวิเคราะห์และประมวลผลข้อมูลขนาดใหญ่โดยอัตโนมัติ เพื่อค้นหาความรู้ หรือรูปแบบความสัมพันธ์ที่แอบซ่อนอยู่ในข้อมูล โดยการทำเหมืองข้อมูลอาศัยความรู้และความเข้าใจจากศาสตร์หลายแขนง เช่น ระบบฐานข้อมูล การเรียนรู้ด้วยเครื่อง สถิติ คณิตศาสตร์ คอมพิวเตอร์ การตลาดและอื่นๆ ส่งผลให้การทำเหมืองข้อมูลสามารถปรับ ประยุกต์และตอบคำถามในเชิงธุรกิจในปัจจุบัน ทำให้ได้รับความนิยม และเกิดการพัฒนาขั้นตอนวิธี เทคนิค และสายงานที่เกี่ยวข้องกับการทำเหมืองข้อมูลขึ้นมากมาย ตัวอย่างเช่น การวิเคราะห์หลักเกณฑ์เชื่อมโยง (association rule analysis) การเกาะกลุ่ม (clustering) และ การจำแนกประเภท (classification) ซึ่งในงานวิจัยนี้จะพิจารณาเฉพาะการวิเคราะห์หลักเกณฑ์เชื่อมโยง

การวิเคราะห์หลักเกณฑ์เชื่อมโยงถูกกล่าวถึงเป็นครั้งแรกในปี ค.ศ. 1993 โดย Agrawal, Imielinski และ Swami [7] ซึ่งในงานวิจัยนี้พบว่า การพิจารณาความน่าสนใจของหลักเกณฑ์เชื่อมโยงใช้ตัววัด 2 ตัว คือ ค่าสนับสนุน (support) และค่าความเชื่อมั่น (confidence) หลัง

จากนั้นได้มีงานวิจัยเกี่ยวกับการพัฒนาขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยง และพัฒนาตัววัดความน่าเชื่อถือของหลักเกณฑ์เชื่อมโยงอีกหลายงานวิจัย [8,9,10,11,12,13,14] แต่ขั้นตอนวิธีพื้นฐานที่เป็นที่นิยมในการสร้างหลักเกณฑ์เชื่อมโยงคือ ขั้นตอนวิธีเอไพออรี (apriori algorithm) [8] ซึ่งถูกพัฒนาขึ้นในปี ค.ศ. 1994 โดย Agrawal และ Srikant

วัตถุประสงค์ในการวิเคราะห์หลักเกณฑ์เชื่อมโยง คือ การค้นหาหลักเกณฑ์ที่แสดงถึงความสัมพันธ์ระหว่างลักษณะประจำของกลุ่มของข้อมูล ซึ่งเรียกว่า หลักเกณฑ์เชื่อมโยง (association rule) หลักเกณฑ์เชื่อมโยงจะเขียนอยู่ในรูปของหลักเกณฑ์ $A \rightarrow C$ โดยที่ A และ C เป็นเซตของกลุ่มข้อมูลที่เรียกว่า ไอเทมเซต (itemset) [7] การวิเคราะห์หลักเกณฑ์เชื่อมโยงเกิดขึ้นจากความต้องการทางด้านการวิเคราะห์ข้อมูลการขายสินค้าปลีก หรือ การวิเคราะห์มาเก็ตบาสเก็ต (market basket analysis)

การวิเคราะห์มาเก็ตบาสเก็ตเป็นการวิเคราะห์เพื่อสร้างหลักเกณฑ์เชื่อมโยงที่แสดงความสัมพันธ์ของสินค้าแต่ละชนิด ซึ่งอธิบายลักษณะหรือพฤติกรรมการซื้อสินค้าของลูกค้า โดยผลลัพธ์ที่ได้จะนำมาใช้ในการวางแผนทางด้านการตลาดเพื่อเพิ่มยอดขายของสินค้าหรือเพิ่มกำไรให้กับองค์กร เช่น การออกแบบรูปแบบของการวางสินค้าบนชั้นวางของ และบนสื่อสิ่งพิมพ์ โดยวางสินค้าที่มีความสัมพันธ์กันไว้ในบริเวณใกล้เคียงกัน เพื่อให้ลูกค้าสามารถเลือกซื้อสินค้าได้ง่ายขึ้น การวางแผนจัดทำรายการส่งเสริมการขายสินค้า โดยอาจจัดชุดของสินค้าที่มีความสัมพันธ์กันขายร่วมกัน เพื่อกระตุ้นยอดขายของสินค้า หรือในธุรกิจขายตรงจะใช้หลักเกณฑ์เชื่อมโยงเพื่อวางแผนการซื้อสินค้าครั้งต่อไปของลูกค้าจากการซื้อสินค้าปัจจุบัน

วัตถุประสงค์ในการวิเคราะห์มาเก็ตบาสเก็ตก็เพื่ออธิบายลักษณะและพฤติกรรมการซื้อสินค้าของลูกค้าโดยพิจารณาว่าสินค้าชนิดใดถูกซื้อไปพร้อมกันในการซื้อสินค้าแต่ละครั้งของลูกค้า ซึ่งข้อมูลที่ใช้เป็นบันทึกการซื้อสินค้าของลูกค้าจากหนึ่งใบเสร็จ โดยเรียกข้อมูลการซื้อสินค้าว่า ข้อมูลมาเก็ตบาสเก็ต (market basket data) ซึ่งแบ่งออกเป็น 2 ส่วน ส่วนแรกคือลักษณะประจำหมายเลขระเบียน ซึ่งระบุถึงการระเบียนการซื้อสินค้าที่บันทึกค่าไม่ซ้ำ และส่วนที่สองคือลักษณะประจำสินค้าของแต่ละสินค้า ซึ่งค่าในลักษณะประจำเหล่านี้จะเป็น 1 เมื่อลูกค้าซื้อสินค้า และ 0 ในกรณีที่ไม่พบการซื้อสินค้า ตัวอย่างเช่น ร้านขายสินค้ามีสินค้าทั้งหมด 4 ชนิด คือ ขนมปัง, นม, ผ้าอ้อม และกาแฟ ตัวอย่างข้อมูลการซื้อสินค้าของลูกค้าของร้านขายสินค้าแสดงดังตารางที่ 1.1

หมายเลขระเบียน	ขนมปัง	นม	ผ้าอ้อม	กาแฟ
1	1	0	1	1
2	0	0	1	1
3	1	1	1	0
4	0	1	0	0

ตารางที่ 1.1: ตัวอย่างข้อมูลการขายสินค้าของร้านขายสินค้าแห่งหนึ่ง

จากตารางที่ 1.1 สามารถสร้างหลักเกณฑ์เชื่อมโยงได้หลายหลักเกณฑ์ เช่น {ผ้าอ้อม}→{กาแฟ}, {นม}→{กาแฟ} ซึ่งแต่ละหลักเกณฑ์จะอธิบายความหมายได้แตกต่างกัน ยกตัวอย่างเช่น {ผ้าอ้อม}→{กาแฟ} อธิบายความสัมพันธ์ที่เกิดขึ้นของผ้าอ้อมและกาแฟคือในกลุ่มลูกค้าที่ซื้อผ้าอ้อมจะพบการซื้อกาแฟปรากฏขึ้นเสมอ ซึ่งมีความหมายแตกต่างกับหลักเกณฑ์ {กาแฟ}→{ผ้าอ้อม} ที่มีความหมายว่า ในกลุ่มลูกค้าที่ซื้อกาแฟจะพบการซื้อผ้าอ้อมเกิดขึ้นพร้อมกันเสมอ ดังนั้นตำแหน่งในการเขียนหลักเกณฑ์จะมีความสำคัญต่อการแปลความหมายของหลักเกณฑ์

การพิจารณาว่าหลักเกณฑ์เชื่อมโยงที่สร้างขึ้นมีความน่าสนใจ น่าเชื่อถือ ยอมรับได้โดยผู้ใช้ หรือ ก่อให้เกิดประโยชน์ในการนำไปใช้ ต้องอาศัยตัววัดที่แสดงถึงความสัมพันธ์ระหว่างพจน์หน้าและพจน์หลังของหลักเกณฑ์ ซึ่งตัววัดในปัจจุบันมีมากมาย แต่ตัววัดที่เป็นที่นิยมและใช้กันอย่างแพร่หลายในปัจจุบันได้แก่ ค่าสนับสนุน และค่าความเชื่อมั่น ซึ่งเขียนได้ดังสมการต่อไปนี้คือ

$$\text{ค่าสนับสนุนของหลักเกณฑ์} = \frac{\text{จำนวนระเบียนที่สอดคล้องกับหลักเกณฑ์}}{\text{จำนวนระเบียนทั้งหมด}}$$

ค่าสนับสนุนแสดงความบ่อยครั้งของหลักเกณฑ์ที่เกิดขึ้นเมื่อเทียบกับข้อมูลทั้งหมด จากตารางที่ 1.1 สมมติให้หลักเกณฑ์ที่สร้างคือ {ผ้าอ้อม}→{กาแฟ} จะมีค่าสนับสนุนเป็น $\frac{2}{4}=5$ เพราะจำนวนระเบียนที่สอดคล้องกับหลักเกณฑ์คือ จำนวนระเบียนที่ผ้าอ้อมและกาแฟถูกซื้อไปพร้อมกัน ซึ่งมีอยู่ 2 ระเบียน และในฐานข้อมูลมีเพียง 4 ระเบียน

ส่วนค่าความเชื่อมั่นแสดงความบ่อยครั้งของหลักเกณฑ์ที่เกิดขึ้นเทียบกับความบ่อยครั้งของการเกิดเฉพาะพจน์หน้าของหลักเกณฑ์ ซึ่งเขียนได้ดังนี้

$$\text{ค่าความเชื่อมั่นของหลักเกณฑ์} = \frac{\text{จำนวนระเบียนที่สอดคล้องกับหลักเกณฑ์}}{\text{จำนวนระเบียนที่สอดคล้องกับพจน์หน้า}}$$

ดังนั้นจากตารางที่ 1.1 {ผ้าอ้อม}→{กาแฟ} จะมีค่าความเชื่อมั่นเป็น $\frac{2}{3}=0.67$ เนื่องจากจำนวนระเบียนที่สอดคล้องกับพจน์หน้าคือ จำนวนระเบียนที่ผ้าอ้อมถูกซื้อมีจำนวนเท่ากับ 3 ในทำนองเดียวกัน {กาแฟ}→{ผ้าอ้อม} จะมีค่าสนับสนุนเป็น $\frac{2}{4}=0.5$ และมีค่าความเชื่อมั่นเป็น $\frac{2}{2}=1$ ซึ่งหลักเกณฑ์ทั้งสองจะมีค่าสนับสนุนเท่ากัน แต่มีค่าความเชื่อมั่นแตกต่างกัน เพราะค่าสนับสนุนเป็นตัววัดที่สมมาตร การสลับที่ระหว่างพจน์หน้าและพจน์หลังของหลักเกณฑ์ไม่ส่งผลต่อค่าสนับสนุนที่คำนวณได้ แต่ค่าความเชื่อมั่นเป็นตัววัดที่ไม่สมมาตรโดยจะค่าที่คำนวณได้จากหลักเกณฑ์ที่สลับพจน์หน้าและพจน์หลังอาจแตกต่างกัน

เมื่อสร้างหลักเกณฑ์แล้วผู้ใช้จะต้องเป็นผู้กำหนดค่าต่ำสุดของค่าสนับสนุนและค่าต่ำสุดของค่าความเชื่อมั่นที่ยอมรับได้ แล้วจึงนำหลักเกณฑ์ที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนต่ำสุด และมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับค่าความเชื่อมั่นต่ำสุดมาใช้ในการวางแผนการขายสินค้า

แนวคิดในการค้นหาหลักเกณฑ์เชื่อมโยงคือการใช้ขั้นตอนวิธีเอไพออรี (Apriori algorithm) ที่เริ่มจากการสร้างไอเทมเซตปรากฏบ่อย (frequent itemset) ที่เป็นไปได้ทั้งหมด ซึ่งไอเทมเซตปรากฏบ่อยคือไอเทมเซตที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนต่ำสุดที่กำหนดโดยผู้ใช้ แล้วนำไอเทมเซตปรากฏบ่อยที่ได้มาสร้างเป็นหลักเกณฑ์ ซึ่งหลักเกณฑ์ที่สร้างจะยอมรับได้ก็ต่อเมื่อมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับค่าความเชื่อมั่นต่ำสุดที่กำหนดโดยผู้ใช้ เนื่องจากในขั้นตอนของการสร้างไอเทมเซตปรากฏบ่อย จะต้องมีการตัดไอเทมเซตที่ไม่น่าสนใจออก จึงเกิดปัญหาการกำหนดค่าสนับสนุนต่ำสุดที่เหมาะสม กล่าวคือถ้ากำหนดค่าสนับสนุนต่ำสุดมีค่ามากจนเกินไป จะทำให้หลักเกณฑ์ที่ค้นได้มีปริมาณน้อยและพลาดหลักเกณฑ์ที่มีค่าความเชื่อมั่นสูง แต่ถ้ากำหนดค่าสนับสนุนต่ำสุดมีค่าน้อยจนเกินไป จะทำให้ได้ปริมาณหลักเกณฑ์เชื่อมโยงมากซึ่งใช้เวลาในการค้นหานาน ผู้วิเคราะห์จึงต้องมีความรู้และความเข้าใจข้อมูลเพื่อกำหนด

ค่าสนับสนุนต่ำสุดที่เหมาะสมโดยไม่ทำให้หลักเกณฑ์ที่มีค่าสนับสนุนต่ำและค่าความเชื่อมั่นสูงที่น่าสนใจถูกตัดออก ซึ่งเมื่อนำหลักเกณฑ์ที่ได้ไปสร้างหรือวางแผนการตลาดจะทำให้สูญเสียลูกค้าที่เป็นกลุ่มเป้าหมายบางกลุ่มซึ่งมีแนวโน้มในการซื้อสินค้าที่สำคัญขององค์กรหรือบริษัท ส่งผลให้ทางองค์กรหรือบริษัทต้องสูญเสียกำไรหรือผลประโยชน์ที่ควรจะได้รับจากกลุ่มดังกล่าว

จากเหตุผลดังกล่าว ทำให้ผู้วิจัยสนใจศึกษาการคัดเลือกหลักเกณฑ์เชื่อมโยงโดยใช้ตัววัดใหม่คือ “ค่าสนับสนุนแบบอ่อน” (weak support) [15] แทนค่าสนับสนุน โดยค่าสนับสนุนแบบอ่อนเกิดจากแนวคิดทางตรรกศาสตร์ เนื่องจากข้อมูลที่ใช้ในการสร้างหลักเกณฑ์เชื่อมโยงจะเก็บอยู่ในรูปแบบของตัวเลขฐานสองคือ 0,1 ซึ่งหมายความถึงการไม่ซื้อสินค้า และซื้อสินค้าตามลำดับ และเมื่อพิจารณาหลักเกณฑ์เชื่อมโยงพบว่าหลักเกณฑ์เชื่อมโยงเขียนอยู่ในรูปแบบเดียวกับประพจน์ถ้า-แล้วในทางตรรกศาสตร์ จึงทำให้เกิดแนวคิดการนำทฤษฎีทางตรรกศาสตร์มาประยุกต์ใช้เข้ากับการสร้างหลักเกณฑ์เชื่อมโยง โดยประพจน์ถ้าแล้ว $p \rightarrow q$ มีข้อขัดแย้งคือ $p \wedge \neg q$ ในทำนองเดียวกันหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ จะมีข้อขัดแย้งคือ $A \cup \neg C$ ในปริภูมิลักษณะประจำ ซึ่งปริภูมิดังกล่าวจะอธิบายเพิ่มเติมในบทที่ 3 ดังนั้นค่าสนับสนุนแบบอ่อนจึงเป็นตัววัดที่ไม่พิจารณาตัวอย่างที่ขัดแย้งกับหลักเกณฑ์ ซึ่งเขียนได้เป็น

$$\text{ค่าสนับสนุนแบบอ่อนของหลักเกณฑ์} = \frac{\text{จำนวนระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์}}{\text{จำนวนระเบียนทั้งหมด}}$$

จากตารางที่ 1.1 จะได้ว่า {ผ้าอ้อม}→{กาแฟ} จะมีค่าสนับสนุนแบบอ่อนเป็น $\frac{3}{4}=0.75$ เนื่องจากระเบียนที่ขัดแย้งกับหลักเกณฑ์คือ ระเบียนที่มีการซื้อผ้าอ้อม แต่ไม่ซื้อกาแฟ ซึ่งมีจำนวน 1 ระเบียน ดังนั้นจำนวนระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์จึงมี 3 ระเบียน ในทำนองเดียวกัน {กาแฟ}→{ผ้าอ้อม} จะมีค่าสนับสนุนแบบอ่อนเป็น $\frac{4}{4}=1$

ดังนั้นในงานวิจัยนี้จะสร้างหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ ตามค่าสนับสนุนแบบอ่อนโดยพิจารณากรณีที่ C มีสมาชิกเพียงหนึ่งเดียวเท่านั้น และ A จะมีจำนวนสมาชิกหนึ่งตัวหรือมากกว่าขึ้นอยู่กับการกำหนดของผู้ใช้ จากนั้นผู้วิจัยวิเคราะห์ความสัมพันธ์ระหว่างค่าสนับสนุน ค่าความเชื่อมั่น และค่าสนับสนุนแบบอ่อน แล้วจึงทดลองสร้างหลักเกณฑ์เชื่อมโยงที่เกิดจากค่าสนับสนุนกับค่าความเชื่อมั่นเปรียบเทียบกับหลักเกณฑ์เชื่อมโยงที่เกิดจากค่าสนับสนุนแบบอ่อน

กับค่าความเชื่อมั่นโดยใช้ข้อมูลที่เป็นมาตรฐานกลางที่นักวิจัยรู้จัก หลังจากนั้นนำผลลัพธ์ที่ได้มาเปรียบเทียบโดยใช้สภาพไว (sensitivity)

โดยในบทที่ 2 จะอธิบายที่มาและความสำคัญ ความหมาย ขั้นตอนวิธีการทำเหมืองข้อมูล ประโยชน์ของการทำเหมืองข้อมูลที่นำมาใช้ในด้านธุรกิจและองค์กรต่างๆ ความหมายของการวิเคราะห์มาเก็ตบาสเก็ต ประโยชน์ของการนำการวิเคราะห์มาเก็ตบาสเก็ตมาใช้ทางธุรกิจ จากนั้นอธิบายถึงความรู้เบื้องต้นเกี่ยวกับการสร้างหลักเกณฑ์เชื่อมโยง ตัววัดที่ใช้บ่งบอกถึงความน่าเชื่อถือของหลักเกณฑ์เชื่อมโยง และขั้นตอนวิธีที่ใช้ในการสร้างหลักเกณฑ์เชื่อมโยง

ในบทที่ 3 อธิบายถึงการพิจารณาข้อมูลในความหมายเชิงตรรกศาสตร์ ซึ่งสามารถอธิบายถึงความหมายและที่มาของค่าสนับสนุนแบบอ่อน การสร้างหลักเกณฑ์โดยใช้ค่าสนับสนุนแบบอ่อน นิยามและทฤษฎีบทที่สามารถอธิบายถึงความสัมพันธ์ระหว่างค่าสนับสนุนแบบอ่อน ค่าสนับสนุน และค่าความเชื่อมั่น จากนั้นผู้วิจัยจะอธิบายถึงการพัฒนาขั้นตอนวิธีใหม่ในการสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน

ในบทที่ 4 อธิบายถึงวิธีการทดลอง ข้อมูลที่ใช้ในการทดลอง การเปรียบเทียบหลักเกณฑ์เชื่อมโยงที่เกิดจากค่าสนับสนุนและค่าความเชื่อมั่นกับหลักเกณฑ์เชื่อมโยงที่เกิดจากค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่น หลังจากนั้นผู้วิจัยจะแสดงผลการทดลองที่ได้เปรียบเทียบในรูปแบบแผนภูมิแท่งแบบ 3 มิติ

บทที่ 5 อธิบายถึงผลสรุปจากงานวิจัยและแนวทางของงานวิจัยในอนาคต

บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงพื้นฐานของการทำเหมืองข้อมูล การวิเคราะห์มาเก็ตบาสเก็ต และการสร้างหลักเกณฑ์เชื่อมโยง ซึ่งประกอบด้วยที่มาและความสำคัญ ความหมาย ขั้นตอนวิธี และ ประโยชน์ที่ได้รับทั้งจากการทำเหมืองข้อมูลและจากการวิเคราะห์มาเก็ตบาสเก็ต

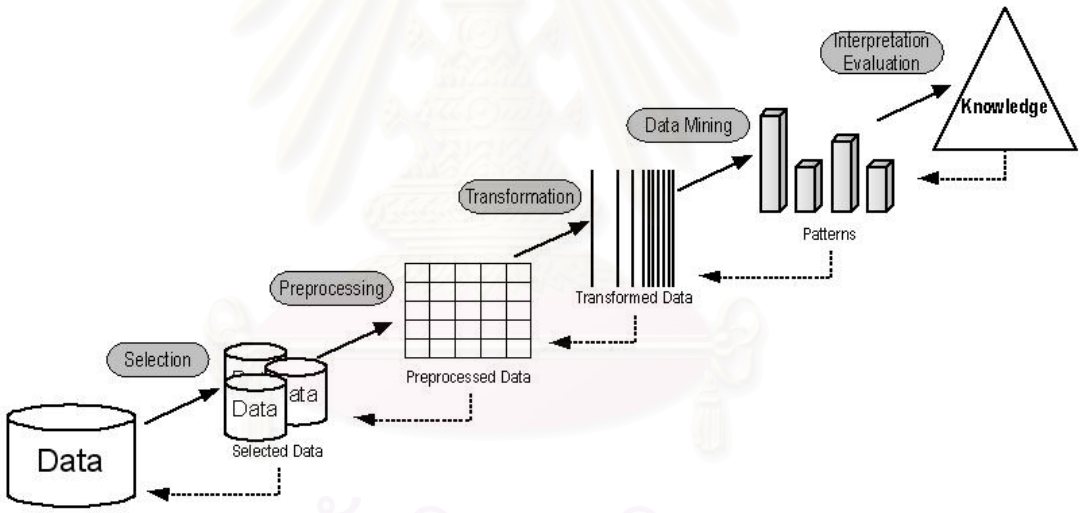
ในช่วงคริสต์ศตวรรษที่ 1960 เทคโนโลยีเกี่ยวกับระบบฐานข้อมูลได้ถูกพัฒนาขึ้นมาแทนที่ระบบแฟ้มข้อมูล ทำให้นักวิจัยต่างก็พยายามที่จะพัฒนาระบบฐานข้อมูลที่มีประสิทธิภาพ และนำมาใช้งานได้กับธุรกิจจริง ไม่ว่าจะเป็นการออกแบบระบบฐานข้อมูลแบบระดับชั้นและแบบเครือข่าย (hierarchical and network database systems) ระบบฐานข้อมูลเชิงสัมพันธ์ (relational database system) ซึ่งเมื่อฐานข้อมูลมีขนาดใหญ่มากขึ้น การพัฒนาซอฟต์แวร์บริหารจัดการทางด้านฐานข้อมูลจึงเพิ่มมากขึ้น ทำให้มีการพัฒนาโปรแกรมที่ใช้จัดการฐานข้อมูล (database management system: DBMS) ขึ้น และให้ผู้ใช้เป็นผู้วิเคราะห์ฐานข้อมูลและสรุปผล

ในการจัดเก็บข้อมูลลงฐานข้อมูลหลายฐานข้อมูล อาจทำให้เกิดความล่าช้าในการแลกเปลี่ยนข้อมูลระหว่างฐานข้อมูล ความผิดพลาดในการจัดเก็บข้อมูล เกิดปัญหาในการปรับข้อมูลให้เป็นปัจจุบัน และการเปลี่ยนแปลงข้อมูล ดังนั้นนักวิจัยจึงริเริ่มคิดค้นการทำคลังข้อมูล (data warehousing) [1,2] ขึ้นในปี ค.ศ. 1990 เพื่อให้สามารถจัดเก็บข้อมูลจากหลายฐานข้อมูลมารวมไว้ในคลังข้อมูลภายใต้เค้าร่าง (schema) เดียวกัน แต่ในการวิเคราะห์ฐานข้อมูลและสรุปผลออกมาเป็นสารสนเทศ (information) ยังคงต้องใช้มนุษย์ที่มีความรู้และความเข้าใจในข้อมูลอย่างลึกซึ้งเป็นผู้วิเคราะห์ข้อมูล

เทคโนโลยีในการทำคลังข้อมูล และการจัดการระบบฐานข้อมูลที่ดีขึ้น ส่งผลให้ฐานข้อมูลมีประสิทธิภาพในการทำงาน และสามารถจัดเก็บข้อมูลขนาดใหญ่ได้มากยิ่งขึ้น แต่เมื่อข้อมูลที่เก็บอยู่ในฐานข้อมูลมีขนาดใหญ่มาก จนเกินกำลังที่มนุษย์จะสามารถวิเคราะห์ข้อมูลที่มีอยู่ได้ทั้งหมด ทำให้เกิดเทคโนโลยีในการวิเคราะห์ฐานข้อมูลขนาดใหญ่โดยอัตโนมัติ ซึ่งเรียกว่าการทำเหมืองข้อมูล (data mining) [1,3,4,5,6]

2.1. การทำเหมืองข้อมูล

การทำเหมืองข้อมูล หรือ การสืบค้นความรู้ในฐานข้อมูล (knowledge discovery in databases :KDD) เป็นกระบวนการ วิธี หรือ รูปแบบในการสกัดหรือสืบค้นความรู้จากฐานข้อมูลขนาดใหญ่โดยอัตโนมัติ โดยความรู้ที่ได้อาจอยู่ในรูปแบบของความสัมพันธ์ ข้อสรุป แบบจำลอง หรือลักษณะเฉพาะของข้อมูลในฐานข้อมูลนั้นๆ การทำเหมืองข้อมูลทุกครั้งต้องกำหนดวัตถุประสงค์หรือเป้าหมายในการทำเหมืองข้อมูล เพื่อให้สามารถกำหนดขั้นตอนการทำงานและผลลัพธ์ที่ได้อย่างชัดเจน โดยการทำเหมืองข้อมูลจะสามารถเลือกใช้กระบวนการในการสืบค้นความรู้ในฐานข้อมูล (knowledge discovery in database process : KDD process) ซึ่งประกอบด้วยกระบวนการทำงานดัง รูปที่ 2.1 [16,17]



รูปที่ 2.1: กระบวนการในการสืบค้นความรู้ในฐานข้อมูล

กระบวนการในการสืบค้นความรู้ในฐานข้อมูลอธิบายได้ดังนี้คือ

1. ขั้นตอนการคัดเลือกข้อมูล (data selection) หลังจากที่กำหนดวัตถุประสงค์ในการทำเหมืองข้อมูลได้อย่างชัดเจนแล้ว การคัดเลือกข้อมูลที่ใช้ในการทำเหมืองข้อมูลก็เป็นส่วนสำคัญ เพราะข้อมูลที่เลือกขึ้นอยู่กับวัตถุประสงค์ที่ตั้งไว้ในข้างต้น เพื่อให้ผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลอยู่ในขอบเขตของวัตถุประสงค์ที่ตั้งไว้

2. ขั้นตอนการเตรียมข้อมูล (prepocessing) ขั้นตอนนี้จะรวมการทำความสะอาดข้อมูล (data cleansing) ซึ่งเทคนิคการเตรียมข้อมูลและการทำความสะอาดข้อมูลได้แก่ การเติมค่าข้อมูลที่หายไป (missing value) การบ่งชี้ข้อมูลที่ผิดปรกติ (outlier) และการแก้ไขข้อมูลซึ่งขัดแย้งกัน (inconsistency) เพื่อให้ข้อมูลที่จะนำมาใช้งานมีความถูกต้องและน่าเชื่อถือ

3. ขั้นตอนการแปลงข้อมูล(transformation) ขั้นตอนนี้จะนำข้อมูลที่ได้จากขั้นตอนที่ 2 มาแปลงให้อยู่ในรูปแบบที่ต้องการ เพื่อให้สามารถนำไปวิเคราะห์โดยการทำเหมืองข้อมูลได้ ขั้นตอนนี้จะรวมถึงเทคนิคในการลดจำนวนข้อมูล (data reduction)

4. ขั้นตอนการทำเหมืองข้อมูล (data mining) เป็นการคัดเลือกขั้นตอนวิธีที่เหมาะสมกับวัตถุประสงค์ที่ตั้งไว้ โดยจะนำขั้นตอนวิธีที่เลือกไว้มาวิเคราะห์ เพื่อค้นหาผลลัพธ์ที่เป็นรูปแบบหรือความรู้ที่ซ่อนในข้อมูล

5. ขั้นตอนการตีความหมายและประเมินผล(interpretation evaluation) ขั้นตอนนี้จะนำผลลัพธ์ที่เป็นรูปแบบหรือความรู้ที่ได้จากขั้นตอนที่ 4 มาแปลความหมาย เพื่อพิจารณาว่าผลลัพธ์ที่ได้ตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ โดยประเมินผลจากการนำข้อมูลที่ไม่ได้ใช้ในการวิเคราะห์ไปตรวจสอบความถูกต้อง

การทำเหมืองข้อมูลแบ่งออกได้เป็น 2 ประเภทคือ การทำเหมืองข้อมูลแบบพรรณนา (descriptive mining) และการทำเหมืองข้อมูลแบบทำนาย (predictive mining) [1] โดยการทำเหมืองข้อมูลแบบพรรณนาจะเป็นการอธิบายและสรุปข้อมูลในรูปแบบที่สั้นและกระชับ เช่น การเกาะกลุ่ม (clustering) การสร้างหลักเกณฑ์เชื่อมโยง (association rule) ส่วนการทำเหมืองข้อมูลแบบทำนายเป็นการวิเคราะห์ข้อมูล โดยสร้างตัวแบบสำหรับข้อมูลที่สนใจ เพื่อไปทำนายแนวโน้มหรือสมบัติของตัวอย่างใหม่ ซึ่งการทำเหมืองข้อมูลแบบทำนายได้แก่ การจำแนกประเภท (classification) เป็นต้น เทคนิคในการทำเหมืองข้อมูลแบบต่างๆ สรุปได้ดังนี้

1. หลักเกณฑ์เชื่อมโยง เป็นการค้นหาความสัมพันธ์ระหว่างกลุ่มของข้อมูล ซึ่งเขียนอยู่ในรูปของหลักเกณฑ์ถ้า-แล้ว คือ $A \rightarrow C$ เมื่อ A และ C เป็นเซตของกลุ่มของข้อมูลที่ต้องการ การวิเคราะห์หลักเกณฑ์เชื่อมโยงนิยมใช้กับข้อมูลการขายสินค้าของร้านขายปลีก ซึ่งมีวัตถุประสงค์เพื่อค้นหาสินค้าที่ลูกค้าซื้อพร้อมๆ กันในแต่ละครั้ง โดยผลลัพธ์ที่ได้จะใช้ในการวางแผนการทางด้านการตลาด และจัดทำรายการส่งเสริมการขาย เพื่อเพิ่มยอดขายของสินค้าแต่ละชนิด

2. การจำแนกประเภท เป็นเทคนิคที่ใช้ในการแยกแยะข้อมูลออกเป็นกลุ่มโดยใช้สมบัติของข้อมูล ซึ่งการจำแนกประเภทแบ่งการทำงานออก เป็น 2 ขั้นตอน คือ ขั้นตอนแรกเป็นการสร้างตัวแบบโดยใช้ข้อมูลที่มีอยู่ ขั้นตอนที่สองเป็นการนำตัวแบบที่ได้จากขั้นตอนแรกมาจำแนกตัวอย่างที่ไม่ทราบชนิด ซึ่งเทคนิคในการจำแนกประเภทที่ใช้บ่อยได้แก่ การสร้างต้นไม้ตัดสินใจ (decision tree) เป็นเทคนิคที่ให้ผลลัพธ์ในลักษณะของโครงสร้างต้นไม้ การจำแนกประเภทแบบเบย์ (Bayes classification) เป็นต้น การจำแนกประเภทนิยมใช้ในการวิเคราะห์งานทางด้านธุรกิจและวิทยาศาสตร์ เช่น การวิเคราะห์ความเสี่ยงของลูกหนี้ การวินิจฉัยทางการแพทย์ และการวิเคราะห์กลุ่มดาวบนท้องฟ้า

3. การเกาะกลุ่ม เป็นการวิเคราะห์การเกาะกลุ่มของข้อมูลที่มีลักษณะเหมือนกัน และสามารถแยกประเภทของกลุ่มข้อมูลออกจากกันได้ โดยไม่ต้องการการกำหนดโดยผู้ใช้ ซึ่งการเกาะกลุ่มจะนำไปใช้เพื่อวิเคราะห์การกระจายข้อมูล หรือเป็นขั้นตอนหนึ่งในการเตรียมข้อมูลเพื่อนำไปสู่การวิเคราะห์ข้อมูลแบบอื่น นอกจากนี้ยังสามารถประยุกต์ใช้กับงานได้หลากหลาย เช่น ด้านการตลาดใช้การเกาะกลุ่มเพื่อหากลุ่มของลูกค้าที่มีลักษณะเหมือนกัน ทำให้ผู้ผลิตสามารถวางแผนเพื่อเพิ่มความพึงพอใจหรือเพิ่มยอดขายสำหรับลูกค้าแต่ละกลุ่มให้แตกต่างกันได้

เทคนิคในการทำเหมืองข้อมูลมีมากมาย และมีแนวโน้มที่จะเพิ่มมากขึ้น เพราะข้อมูลที่ต้องการวิเคราะห์มีมากและหลากหลาย จึงก่อให้เกิดแนวคิด และวิธีการใหม่ เพื่อตอบสนองต่อความต้องการทางการวิเคราะห์ในแต่ละสายงานและในเชิงธุรกิจ

2.2. ประโยชน์ของการทำเหมืองข้อมูล

เหมืองข้อมูลเป็นศาสตร์ที่ได้รับความสนใจ และได้รับการยอมรับอย่างกว้างขวาง โดยในการวิเคราะห์เชิงธุรกิจจะทำเหมืองข้อมูลเพื่อค้นหาแนวโน้มการซื้อสินค้าของลูกค้า การวางแผนการลงทุน และการตรวจหาการลงทุนที่ผิดปรกติ เป็นต้น นอกจากนี้ยังสามารถนำไปประยุกต์ให้เข้ากับธุรกิจต่างๆ เช่น การปรับปรุงแผนการโฆษณาสินค้า การวางแผนการตลาดเฉพาะกลุ่ม เพื่อเพิ่มความพึงพอใจให้แก่ลูกค้าในกลุ่ม ประโยชน์ของการทำเหมืองข้อมูลสำหรับการนำไปใช้ในธุรกิจแต่ละด้าน [18,19,20,21] ได้แก่

1. ธุรกิจทางด้านการขายสินค้า เป็นธุรกิจที่ต้องทำให้ลูกค้าประทับใจในสินค้า และบริการ เพื่อให้ลูกค้าซื้อสินค้าชนิดนั้นต่อไปและไม่เปลี่ยนไปใช้สินค้าของบริษัทคู่แข่ง ซึ่งการทำเหมืองข้อมูลช่วยให้ค้นพบแนวโน้มของการซื้อสินค้าของลูกค้า ปรับปรุงการวางแผนทางการตลาดให้เหมาะสมกับลูกค้าแต่ละกลุ่มหรือที่เรียกว่าการทำการตลาดแบบมีกลุ่มเป้าหมาย (targeted marketing) ปรับปรุงรายการส่งเสริมการขายสินค้าให้เหมาะสม ทำนายแนวทางการซื้อสินค้าของลูกค้าในอนาคต และทำนายแนวทางการลงทุนทางด้านธุรกิจอื่นๆ ทำให้แต่ละองค์กรมีกำไรและยอดขายเพิ่มขึ้น เพราะสามารถตอบสนองต่อความต้องการของลูกค้าแต่ละกลุ่ม และทำให้ลูกค้าเกิดความพึงพอใจ

2. การแพทย์ การวินิจฉัยโรคเป็นเรื่องที่สำคัญสำหรับการแพทย์ เพราะการวินิจฉัยว่าคนไข้ที่มารักษาเป็นโรคหรือไม่เป็นโรคต้องอาศัยประสบการณ์และความเชี่ยวชาญของแพทย์เจ้าของคนไข้ ซึ่งการทำเหมืองข้อมูลช่วยให้สามารถวิเคราะห์อาการที่ผิดปรกติของคนไข้ วิเคราะห์วิธีการรักษา

โรค และทำนายความเสี่ยงที่จะเป็นโรคของกลุ่มผู้ป่วยที่มีอาการใกล้เคียงกันได้

3. การธนาคาร และการรักษาความปลอดภัย การทำเหมืองข้อมูลช่วยให้ธนาคารสามารถวิเคราะห์การให้สินเชื่อ (credit scoring) และตรวจสอบบัตรเครดิตปลอม เพื่อเป็นแนวทางในการตรวจสอบลูกค้าของธนาคาร และเพิ่มความปลอดภัยให้กับธนาคาร นอกจากนี้การทำเหมืองข้อมูลยังมีส่วนช่วยในเรื่องการรักษาความปลอดภัยเช่น การวิเคราะห์การปลอมแปลงเอกสารและข้อมูล เครือข่ายขององค์กรยาเสพติด พฤติกรรมของอาชญากรและผู้ก่อการร้าย การหลีกเลี่ยงภาษี การยักยอกเงิน และงานทางด้านการรักษาความปลอดภัยอื่นๆ ตั้งแต่ระดับองค์กรไปจนถึงระดับประเทศ

4. การประกันภัย ทุกครั้งที่เกิดการเรียกร้องสิทธิ์ในการประกันภัย บริษัทผู้ประกันภัยจะต้องตรวจสอบว่าผู้เรียกร้องเป็นผู้ที่เกี่ยวข้องกับผู้ประกันภัยจึงจะเรียกร้องสิทธิ์ได้ แต่จากการตรวจสอบพบว่าส่วนใหญ่ผู้เรียกร้องไม่ได้มีความเกี่ยวข้องกับผู้ประกันภัย ทำให้บริษัทต้องสูญเสียเวลา เงิน และชื่อเสียง ดังนั้นการทำเหมืองข้อมูลจะช่วยในการวิเคราะห์การอ้างสิทธิ์ของผู้เรียกร้อง และช่วยวิเคราะห์การเรียกร้องสิทธิ์แบบไม่ถูกต้อง

ผู้ใช้สามารถนำการทำเหมืองข้อมูลไปประยุกต์ใช้กับสายงานที่เหมาะสมได้ แต่ต้องกำหนดเป้าหมายในการทำเหมืองข้อมูลแต่ละครั้งให้ชัดเจน เพื่อให้การทำเหมืองข้อมูลในแต่ละครั้งมีขอบเขต ทำให้ผลลัพธ์ที่ได้สามารถนำไปใช้ประโยชน์ได้อย่างเต็มที่ โดยขั้นตอนหรือกระบวนการในการทำเหมืองข้อมูลจะแตกต่างกันไปตามแต่ละองค์กร แต่ยังคงมีขั้นตอนหลักในการทำงานเหมือนกับกระบวนการในการสืบค้นความรู้ในฐานข้อมูลซึ่งอธิบายดังรูปที่ 2.1 ในข้างต้น

เนื่องจากในงานวิจัยนี้พิจารณาการวิเคราะห์หลักเกณฑ์เชื่อมโยง ดังนั้นผู้วิจัยจึงอธิบายเรื่องของการวิเคราะห์หลักเกณฑ์เชื่อมโยง ตัววัดที่น่าสนใจ การสร้างหลักเกณฑ์เชื่อมโยง และขั้นตอนวิธีที่ใช้ในการสร้างหลักเกณฑ์เชื่อมโยง

2.3. การวิเคราะห์หลักเกณฑ์เชื่อมโยง

การวิเคราะห์หลักเกณฑ์เชื่อมโยง เป็นการวิเคราะห์เพื่อหาความสัมพันธ์ระหว่างกลุ่มของลักษณะประจำของข้อมูล โดยหลักเกณฑ์เชื่อมโยงจะเขียนอยู่ในรูปแบบของหลักเกณฑ์ถ้า-แล้ว $A \rightarrow C$ [1,7,8,9] เมื่อ A และ C เป็นเซตของกลุ่มของลักษณะประจำของข้อมูลที่เรียกว่าไอเทมเซต (itemset) การวิเคราะห์หลักเกณฑ์เชื่อมโยงมีแนวคิดจากการวิเคราะห์ข้อมูลการขายสินค้าปลีก โดยมีวัตถุประสงค์เพื่อพิจารณาว่าสินค้าชนิดใดที่ถูกซื้อพร้อมๆกันในการซื้อสินค้าแต่ละครั้งของลูกค้า ซึ่งเรียกว่า การวิเคราะห์มาเก็ตบาสเก็ต (market basket analysis)

2.3.1. การวิเคราะห์มาเก็ตบาสเก็ต

การวิเคราะห์มาเก็ตบาสเก็ตเป็นการวิเคราะห์เพื่อทำนายพฤติกรรมการซื้อสินค้าของลูกค้า โดยพิจารณาจากสินค้าที่ลูกค้าซื้อผ่านการหยิบใส่ในตระกร้าหรือรถเข็น (shopping basket) ในรูปที่ 2.2 [22] เป็นตัวอย่างของการแสดงสินค้าที่ลูกค้าซื้อในหนึ่งครั้ง โดยพบว่าลูกค้ารายนี้ซื้อกล้วย น้ำส้ม น้ำยาล้างจาน และน้ำยาทำความสะอาด และในรูปนี้ได้อธิบายถึงการวิเคราะห์มาเก็ตบาสเก็ตสำหรับสินค้าแต่ละชนิดด้วย

กลุ่มของสินค้าที่ลูกค้าแต่ละคนซื้อจะมีความแตกต่างกัน ซึ่งเมื่อนำข้อมูลของการซื้อสินค้าของลูกค้ามารวมกัน ก็จะได้ข้อมูลการซื้อสินค้าของลูกค้าทั้งหมดในช่วงระยะเวลาหนึ่ง การวิเคราะห์มาเก็ตบาสเก็ตจะนำข้อมูลการซื้อสินค้าของลูกค้าทั้งหมดมาสร้างหลักเกณฑ์เชื่อมโยง ทำให้ทราบถึงความสัมพันธ์ระหว่างกลุ่มของข้อมูล และสามารถวางแผนกลยุทธ์ทางด้านการตลาด ซึ่งเป็นประโยชน์ในเชิงธุรกิจ เช่น การปรับโครงสร้างการจัดวางสินค้าที่มีความสัมพันธ์กันไว้บริเวณใกล้เคียงกันในร้านค้า และสื่อสิ่งพิมพ์ การวางแผนจัดทำรายการส่งเสริมการขาย โดยการนำสินค้าที่มีความสัมพันธ์กันมาทำรายการส่งเสริมการขาย ได้แก่ การจับคู่ขายสินค้าร่วมกัน หรือเมื่อซื้อสินค้าชนิดแรก จะลดราคาสินค้าชิ้นที่สอง เป็นต้น และในธุรกิจขายตรง เมื่อมีสินค้าตัวใหม่ ผู้ขายสามารถติดต่อกับลูกค้าที่มีแนวโน้มว่าจะซื้อสินค้าตัวใหม่ได้ โดยวิเคราะห์จากข้อมูลการซื้อสินค้าของลูกค้า ทำให้ผู้ขายไม่ต้องไปติดต่อกับลูกค้าที่มีอยู่ทั้งหมด

รูปที่ 2.2: สินค้าที่ลูกค้าซื้อในการซื้อหนึ่งครั้ง

ผลลัพธ์ที่ได้จากการวิเคราะห์มาเก็ตบาสเก็ตจะเขียนอยู่ในรูปแบบของหลักเกณฑ์เชื่อมโยง เช่น {น้ำส้ม}→{โซดา} ซึ่งมีความหมายคือ ในฐานข้อมูลการซื้อสินค้าลูกค้าที่พบการซื้อน้ำส้มมักจะพบการซื้อโซดาด้วย แต่หลักเกณฑ์ {โซดา}→{น้ำส้ม} สื่อความหมายว่าบรรดาลูกค้าที่พบการซื้อโซดาจะพบการซื้อน้ำส้ม ดังนั้นตำแหน่งของการเขียนหลักเกณฑ์เชื่อมโยงจะมีความสำคัญในการแปลความหมายของหลักเกณฑ์ที่ได้เป็นผลลัพธ์จากการวิเคราะห์มาเก็ตบาสเก็ต

นอกจากนี้การวิเคราะห์มาเก็ตบาสเก็ตยังสามารถนำมาประยุกต์เข้ากับธุรกิจอื่นๆ ที่ไม่ใช้ธุรกิจการขายปลีก แต่ลูกค้าต้องซื้อสินค้าในช่วงเวลาเดียวกัน หรือทำบางสิ่งบางอย่างที่มีลักษณะใกล้เคียงกัน [22] เช่น

1. การซื้อสินค้าโดยผ่านบัตรเครดิต เช่น การเช่ารถ การจองห้องโรงแรม ทำให้สามารถเข้าใจพฤติกรรมการซื้อสินค้าของลูกค้าโดยผ่านบัตรเครดิตได้

2. การเปิดบริการเสริมสำหรับการติดต่อสื่อสาร เช่น บริการถือสายซ้อน บริการรับฝากข้อความ บริการเตือนเมื่อลูกค้าไม่ได้รับสาย ซึ่งเมื่อทางบริษัททราบพฤติกรรมส่วนใหญ่ของลูกค้า ก็สามารถรวมบริการเสริมเปิดเป็นแบบแพคเกจ เพื่อเพิ่มรายได้ให้กับบริษัท

3. การทำธุรกรรมกับธนาคาร เช่น การเปิดบัญชีธนาคารที่เกี่ยวกับการลงทุน การบริการทางด้านการลงทุนต่างๆ การเปิดสินเชื่อเพื่อซื้อรถ ซึ่งใช้การวิเคราะห์มาเก็ตบาสเก็ตเพื่อพิจารณาความต้องการทางด้านการทำธุรกรรมต่างๆ ของลูกค้า นอกเหนือไปจากบริการที่มีอยู่เดิม

4. รูปแบบของการกล่าวอ้างสิทธิ์ที่ผิดปรกติต่างๆ ในบริษัทประกันภัยจะถูกบันทึกให้เป็นสัญญาณที่บ่งบอกว่าเป็นการโกง และเป็นต้นแบบในการคาดเดาเหตุการณ์ที่มีลักษณะเดียวกันได้

5. การนำประวัติของคนไข้ในการรักษาโรคมาวิเคราะห์ด้วยการวิเคราะห์มาเก็ตบาสเก็ต จะช่วยให้ทราบถึงแนวโน้มของโรคแทรกซ้อนโดยอาศัยรูปแบบของการรักษาต่าง ๆ ที่มีอยู่ในปัจจุบัน

ในการวิเคราะห์มาเก็ตบาสเก็ตผู้วิเคราะห์ไม่จำเป็นต้องทราบรูปแบบเฉพาะหรือผลลัพธ์ของข้อมูลที่ต้องการ แต่การวิเคราะห์มาเก็ตบาสเก็ตจะพิจารณาจากข้อมูลที่มีอยู่ทั้งหมดและสกัด ดึง หรือสร้างรูปแบบหรือผลลัพธ์ที่น่าสนใจออกมาจากข้อมูลที่วิเคราะห์ โดยรูปแบบหรือผลลัพธ์ที่ได้จากการวิเคราะห์มาเก็ตบาสเก็ตจะเขียนอยู่ในรูปของหลักเกณฑ์ ที่เรียกว่า หลักเกณฑ์เชื่อมโยง

2.3.2. ข้อมูลมาเก็ตบาสเก็ต

โดยทั่วไปการวิเคราะห์หลักเกณฑ์เชื่อมโยงจะอธิบายในความหมายของการวิเคราะห์มาเก็ตบาสเก็ต และเก็บข้อมูลให้อยู่ในรูปแบบของข้อมูลมาเก็ตบาสเก็ต (market basket data) หรือฐานข้อมูลทรานแซคชัน(transaction database) [9] ดังตารางที่ 2.2 ซึ่งเป็นการเก็บ

ข้อมูลการซื้อสินค้าแต่ละครั้งของลูกค้า โดยข้อมูลในแต่ละแถวหรือระเบียนคือการซื้อสินค้าในหนึ่งครั้ง ซึ่งนิยมเรียกว่า ทรานแซคชัน (transaction) ในแต่ละระเบียนจะมีหมายเลขที่ระบุถึงการซื้อสินค้าแต่ละครั้งที่เรียกว่า ตัวระบุทรานแซคชัน (transaction identification: TID) เมื่อลูกค้าซื้อสินค้า ข้อมูลจะบันทึกชื่อหรือรหัสรายการของสินค้าที่ลูกค้าซื้อไป

TID	สินค้าที่ลูกค้าซื้อ
100	นม, น้ำส้ม, น้ำยาทำความสะอาด
200	น้ำส้ม, น้ำยาล้างจาน
300	น้ำส้ม, โซดา
400	นม, น้ำส้ม, น้ำยาล้างจาน
500	นม, โซดา
600	น้ำส้ม, โซดา
700	นม, โซดา
800	นม, น้ำส้ม, โซดา, น้ำยาทำความสะอาด
900	นม, น้ำส้ม, โซดา
1000	นม, น้ำยาทำความสะอาด

ตารางที่ 2.2: ตัวอย่างข้อมูลการซื้อสินค้าของลูกค้าในช่วงระยะเวลาหนึ่ง

เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูล การจัดเก็บชื่อหรือรหัสรายการของสินค้าจึงเปลี่ยนให้อยู่ในรูปของตัวเลข และทำการแปลงข้อมูลให้อยู่ในรูปแบบที่สนใจ โดยนำชื่อหรือรหัสรายการที่เปลี่ยนให้อยู่ในรูปตัวเลขสร้างเป็นลักษณะประจำแบบทวิภาค นั่นคือค่าภายในแต่ละลักษณะประจำจะมีค่าเป็น 1 หรือ 0 โดยมีค่าเป็น 1 เมื่อสินค้าถูกซื้อ และมีค่าเป็น 0 ในกรณีที่ไม่ปรากฏชื่อสินค้าจากข้อมูลเดิม ซึ่งเรียกว่าการเข้ารหัสแบบดัมมี (dummy coding) [23] ดังนั้นจากตารางที่ 2.2 เมื่อกำหนดให้นมเป็นสินค้าชนิดที่ 1 น้ำส้มเป็นสินค้าชนิดที่ 2 โซดาเป็นสินค้าชนิดที่ 3 น้ำยาล้างจานเป็นสินค้าชนิดที่ 4 และน้ำยาทำความสะอาดเป็นสินค้าชนิดที่ 5 จะสามารถแปลงข้อมูลให้อยู่ในรูปแบบที่สนใจได้ดังตารางที่ 2.3

TID	1	2	3	4	5
100	1	1	0	0	1
200	0	1	0	1	0
300	0	1	1	0	0
400	1	1	0	1	0
500	1	0	1	0	0
600	0	1	1	0	0
700	1	0	1	0	0
800	1	1	1	0	1
900	1	1	1	0	0
1000	1	0	0	0	1

ตารางที่ 2.3: ตัวอย่างข้อมูลการซื้อสินค้าของลูกค้าที่ถูกแปลงให้อยู่ในรูปแบบที่สนใจ

หลังจากที่สร้างหลักเกณฑ์เชื่อมโยงที่ได้จากการวิเคราะห์ข้อมูลที่อยู่ในรูปแบบของข้อมูลมาเก็ตบาสเก็ตดังตารางที่ 2.3 จึงนำหลักเกณฑ์เชื่อมโยงที่ได้มาแปลความหมายเป็นชื่อหรือรหัสรายการของสินค้าและความสัมพันธ์ที่เกิดขึ้น เช่น 1→2 แปลความหมายได้ว่า เมื่อพบการซื้อสินค้าชนิดที่ 1 (น้ำส้ม) แล้วจะพบการซื้อสินค้าชนิดที่ 2 (โซดา) ซึ่งมีความหมายแตกต่างกับ 2→1 คือ เมื่อพบการซื้อสินค้าชนิดที่ 2 แล้วจะพบการซื้อสินค้าชนิดที่ 1

2.3.3. นิยามและหลักการทั่วไปในการสร้างหลักเกณฑ์เชื่อมโยง

กำหนดให้ $I = \{I_1, \ldots, I_m\}$ เป็นเซตของไอเทม (item) หรือลักษณะประจำทั้งหมด ให้ $A \subseteq I$ เรียก A ว่า ไอเทมเซต (itemset) และเรียกว่า k–ไอเทมเซต (k–itemset) เมื่อ A มีสมาชิกทั้งหมด k ตัว และนิยามปริภูมิลักษณะประจำเป็น $\mathscr{S}_I = \{A \mid A \subseteq I\}$ และกำหนดให้ $D = \{t_1, \ldots, t_n\}$ เป็นเซตของระเบียนทั้งหมดดังตารางที่ 2.4 ซึ่งแต่ละระเบียนจะมีตัวระบุทรานแซคชัน (TID) เป็นหมายเลขที่ใช้ระบุถึงระเบียนนั้น ๆ โดยแต่ละระเบียน t_i เมื่อ $i \in \{1, \ldots, n\}$ เป็นไอเทมเซต และนิยามปริภูมิระเบียนเป็น $\mathscr{S}_D = \{T \mid T \subseteq D\}$ โดยกำหนดให้ T_A เป็นเซตของกลุ่มระเบียนที่สอดคล้องกับไอเทมเซต A และ $T_{\neg A}$ เป็นเซตของกลุ่มระเบียนที่ไม่สอดคล้องกับไอเทมเซต A ซึ่ง $T_{\neg A} = D \setminus T_A$ และ $T_{A \cup C} = T_A \cap T_C$ เมื่อ $C \subseteq I$ และ $A \cap C = \emptyset$

TID	I_1	I_2	...	I_m
t_1	1	0	...	1
t_2	0	0	...	1
⋮	⋮	⋮		⋮
t_n	1	1		1

ตารางที่ 2.4: รูปแบบของข้อมูลที่สนใจ

หลักเกณฑ์เชื่อมโยงเขียนอยู่ในรูป $A \rightarrow C$ [M] เมื่อ $A, C \subseteq I$ และ $A \cap C = \varnothing$ โดยที่ A เป็นไอเทมเซตที่เรียกว่า พจน์หน้า (antecedent) หรือ ด้านซ้ายมือ (LHS) และ C เป็นไอเทมเซตที่เรียกว่า พจน์หลัง (consequent) หรือ ด้านขวามือ (RHS) และ M คือตัววัด (measure) ที่ใช้บ่งบอกถึงความสัมพันธ์ระหว่างพจน์หน้าและพจน์หลัง

การเลือกพจน์หน้าและพจน์หลังเพื่อสร้างหลักเกณฑ์เชื่อมโยง จะไม่สนใจในกรณีที่เป็น 0-ไอเทมเซต หรือ เซตว่าง และกรณีที่เป็นเซตของไอเทมทั้งหมด หรือ I เพราะไอเทมเซตรูปแบบนี้ไม่สามารถนำมาใช้สร้างหลักเกณฑ์ที่เป็นประโยชน์ได้ และเนื่องจากในการสร้างหลักเกณฑ์เชื่อมโยงเป็นการหาความสัมพันธ์ระหว่างกลุ่มของข้อมูล ดังนั้นพจน์หน้าและพจน์หลังจะต้องมีสมาชิกแตกต่างกันเสมอ

ตัวอย่างที่ 2.1 จากตารางที่ 2.3 ซึ่งเป็นข้อมูลตัวอย่างจะสามารถหารูปแบบทั่วไปของข้อมูลได้ดังนี้

ให้ $I = \{1, 2, 3, 4, 5\}$ จะได้ว่ามีปริภูมิลักษณะประจำคือ $\mathcal{S}_I = \{ \varnothing , \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{3, 4\}, \{3, 5\}, \{4, 5\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 3, 4\}, \{1, 3, 5\}, \{1, 4, 5\}, \{2, 3, 4\}, \{2, 3, 5\}, \{2, 4, 5\}, \{3, 4, 5\}, \{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 4, 5\}, \{1, 3, 4, 5\}, \{2, 3, 4, 5\}, \{1, 2, 3, 4, 5\}\}$ ซึ่งเท่ากับเซตกำลังของเซต I ($P(I)$) จึงมีจำนวนสมาชิกเท่ากับ 2^5 และสมมติให้ $A = \{2, 5\} \subseteq I$ ซึ่งเรียกว่า 2-ไอเทมเซต และ $C = \{1\} \subseteq I$ ซึ่งเรียกว่า 1-ไอเทมเซต

ในทำนองเดียวกันเซตของข้อมูลคือ $D = \{100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$ และปริภูมิระเบียนคือ $\mathcal{S}_D = P(D)$ เนื่องจาก T_A คือเซตของระเบียนที่สอดคล้องกับไอเทมเซต A คือ เซตของระเบียนที่ลักษณะประจำ 2 และลักษณะประจำ 5 มีค่าเป็น 1 ในทำนองเดียวกัน $T_{\neg A}$ คือเซตของระเบียนที่ไม่สอดคล้องกับไอเทมเซต A ซึ่งคือเซตของระเบียนที่ลักษณะประจำ 2 หรือลักษณะประจำ 5 มีค่าเป็น 0 ซึ่ง $T_{\neg A} = D \setminus T_A$ และ $T_{A \cup C}$ คือ

เซตของระเบียนที่สอดคล้องไอเทมเซต $A \cup C$ ในปริภูมิลักษณะประจำ ซึ่ง $T_{A\cup C} = T_A \cap T_C$ ดังนั้นจึงมีค่าดังต่อไปนี้

$$T_A = \{100, 800\}$$

$$T_{\neg A} = D \setminus T_A = \{200, 300, 400, 500, 600, 700, 900, 1000\}$$

$$T_C = \{100, 400, 500, 700, 800, 900, 1000\}$$

$$T_{\neg C} = D \setminus T_C = \{200, 300, 600\}$$

$$T_{A\cup C} = T_A \cap T_C = \{100, 800\}$$

$$T_{A\cup \neg C} = T_A \cap T_{\neg C} = \{\ \}$$

$$T_{\neg A\cup C} = T_{\neg A} \cap T_C = \{400, 500, 700, 900, 1000\}$$

$$T_{\neg A\cup \neg C} = T_{\neg A} \cap T_{\neg C} = \{200, 300, 600\}$$

2.3.4. นิยามตัววัดที่น่าสนใจ

การสร้างหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ จะต้องกำหนดทั้งพจน์หน้าและพจน์หลังของหลักเกณฑ์ ดังนั้นการพิจารณาความสัมพันธ์ของพจน์หน้าและพจน์หลังจึงต้องมีตัววัดที่ใช้อธิบายความน่าเชื่อถือ และความน่าสนใจ ซึ่งตัววัดที่ถูกสร้างใช้พร้อมกับขั้นตอนวิธีการสร้างหลักเกณฑ์เชื่อมโยงคือ ค่าสนับสนุน (support) และค่าความเชื่อมั่น (confidence) โดย Argawal, Imielinski และ Swami [7] ในปี ค.ศ. 1993 ทำให้หลักเกณฑ์เชื่อมโยงเป็นที่รู้จักและก่อให้เกิดการพัฒนาตัววัดและขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงเป็นจำนวนมาก ต่อมาในปี ค.ศ. 1997 Brin และคณะ[13] ได้ทดลองหาความสัมพันธ์ระหว่างไอเทมเซตโดยอาศัยการทดสอบด้วยไคกำลังสอง (chi-square test) ซึ่งเกิดปัญหาในการคำนวณตารางการจร (contingency table) เพราะค่าภายในตารางมีปริมาณน้อย จึงสร้างตัววัดที่มีชื่อว่า ค่าคอนวิกชัน (conviction) และค่ายกระดับ (lift) เพื่ออธิบายความสัมพันธ์ระหว่างไอเทมเซต ส่วนค่าครอบคลุม (coverage) เป็นตัววัดที่ถูกสร้างเพื่อใช้อธิบายความสัมพันธ์ของตัววัดจึงไม่ปรากฏว่าเกิดขึ้นครั้งแรกเมื่อใด และ ค่าเลเวอร์เรจ (leverage) เป็นตัววัดที่กล่าวถึงเป็นครั้งแรกในปี ค.ศ. 1991 ซึ่งนิยามตัววัดที่กล่าวถึงคือ

- ค่าสนับสนุน (support) คือ ค่าที่แสดงถึงอัตราของความถี่ของหลักเกณฑ์โดยเทียบกับข้อมูลทั้งหมด ซึ่งสามารถคำนวณได้จากจำนวนระเบียนที่สอดคล้องตามหลักเกณฑ์หารด้วยจำนวนระเบียนทั้งหมด โดยค่าสนับสนุนของหลักเกณฑ์ก็คือค่าสนับสนุนของยูเนียนเซตระหว่างพจน์หน้ากับพจน์หลัง ค่าสนับสนุนมีค่าตั้งแต่ 0 ไปจนถึง 1 เมื่อค่าสนับสนุนของหลักเกณฑ์มีค่าเป็น 0 หมายความว่าไม่มีระเบียนใดที่สอดคล้องกับหลักเกณฑ์นั้น และเมื่อค่าสนับสนุนมีค่าเป็น 1 จะหมายความว่าทุกๆ ระเบียนเป็นระเบียนที่สอดคล้องกับหลักเกณฑ์นั้น

$$sup(A \rightarrow C) = sup(A \cup C) = \frac{|T_{A \cup C}|}{|D|}$$

- ค่าความเชื่อมั่น (confidence) คือ ค่าที่แสดงถึงอัตราของพบข้อมูลของทั้งหลักเกณฑ์เทียบกับการพบเฉพาะพจน์หน้าของหลักเกณฑ์ ซึ่งสามารถคำนวณได้จากจำนวนระเบียนที่สอดคล้องตามหลักเกณฑ์หารด้วยจำนวนระเบียนที่สอดคล้องกับพจน์หน้าของหลักเกณฑ์ เมื่อจำนวนระเบียนที่สอดคล้องกับพจน์หน้าของหลักเกณฑ์ไม่เป็น 0 ค่าที่เป็นไปได้ของค่าความเชื่อมั่นจะมีค่าตั้งแต่ 0 ถึง 1 ซึ่งถ้าค่าความเชื่อมั่นของหลักเกณฑ์มีค่าเป็น 0 จะหมายความว่าไม่มีระเบียนที่สอดคล้องกับหลักเกณฑ์นี้ และในกรณีที่ค่าความเชื่อมั่นของหลักเกณฑ์เป็น 1 จะหมายความว่าทุกระเบียนที่สอดคล้องกับพจน์หน้าจะสอดคล้องกับพจน์หลังด้วย โดยค่าความเชื่อมั่นสามารถเขียนให้อยู่ในรูปของค่าสนับสนุนดังสมการด้านล่าง

$$conf(A \rightarrow C) = \frac{sup(A \cup C)}{sup(A)} = \frac{|T_{A \cup C}|}{|T_A|}$$

- ค่าครอบคลุม (coverage) ในบางครั้งเรียกว่าค่าสนับสนุนของพจน์หน้าของหลักเกณฑ์ ซึ่งคำนวณได้จากจำนวนระเบียนที่สอดคล้องกับพจน์หน้าหารด้วยจำนวนระเบียนทั้งหมด ค่าครอบคลุมจะมีค่าตั้งแต่ 0 ถึง 1

โดยเมื่อค่าครอบคลุมมีค่าเป็น 0 แสดงว่าไม่มีระเบียนใดที่สอดคล้องกับพจน์หน้า และเมื่อค่าครอบคลุมมีค่าเป็น 1 แสดงว่าทุกระเบียนของข้อมูลจะสอดคล้องกับพจน์หน้า ซึ่งสามารถเขียนค่าครอบคลุมได้ดังนี้

$$cov(A \rightarrow C) = sup(A) = \frac{|T_A|}{|D|}$$

- ค่ายกระดับ (lift) [9,10,11,12] เดิมเรียกว่า ค่าความสนใจ (interest) เกิดจากแนวคิดทางสถิติของความน่าจะเป็นของสองเหตุการณ์ที่เป็นอิสระต่อกัน ซึ่งพบว่าหลักเกณฑ์จะไม่เป็นหลักเกณฑ์ที่น่าสนใจ เมื่อค่าสนับสนุนของหลักเกณฑ์มีค่าใกล้เคียงกับค่าสนับสนุนของพจน์หน้าคูณกับค่าสนับสนุนของพจน์หลัง หรือกล่าวในเชิงสถิติว่า พจน์หน้ากับพจน์หลังของหลักเกณฑ์เป็นอิสระต่อกัน ดังนั้นจึงนิยามค่ายกระดับของหลักเกณฑ์เป็นจำนวนระเบียนที่สอดคล้องกับพจน์หน้าและพจน์หลังหารด้วยผลคูณของจำนวนระเบียนที่สอดคล้องกับพจน์หน้าและจำนวนระเบียนที่สอดคล้องกับพจน์หลัง โดยค่ายกระดับของหลักเกณฑ์ $A \rightarrow C$ จะมีค่าเท่ากับค่า ยกระดับของหลักเกณฑ์ $C \rightarrow A$ เพราะค่ายกระดับเป็นตัววัดที่สมมาตร ซึ่งเมื่อพิจารณาค่าที่เป็นไปได้ของค่ายกระดับของหลักเกณฑ์ $A \rightarrow C$ จะแบ่งออกเป็น 3 กรณี คือ กรณีที่ค่ายกระดับของหลักเกณฑ์มีค่าเท่ากับ 1 แสดงว่า A และ C เป็นอิสระต่อกัน กรณีที่สองคือค่ายกระดับมีค่ามากกว่า 1 แสดงว่า A และ C ไม่เป็นอิสระต่อกันแบบบวก (positively dependent) และกรณีสุดท้ายเมื่อค่ายกระดับของหลักเกณฑ์มีค่าน้อยกว่า 1 แสดงว่าหลักเกณฑ์ที่สร้างไม่เป็นอิสระต่อกันแบบลบ (negatively dependent) นั่นคือ A และ $\neg C$ หรือ $\neg A$ และ C ไม่เป็นอิสระต่อกัน

$$lift(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A) \times sup(C)} = \frac{|T_{A \cup C}| \times |D|}{|T_A| \times |T_C|}$$

- ค่าเลเวอร์เรจ (leverage) [10,11] คือผลต่างของการเกิดทั้งหลักเกณฑ์กับค่าคาดหวังที่เกิดจากความเป็นอิสระต่อกันของพจน์หน้าและพจน์หลังของหลักเกณฑ์ ซึ่งค่าเลเวอร์เรจจะมีความหมายคล้ายกับค่ายกระดับ คือ เมื่อค่าเลเวอร์เรจของหลักเกณฑ์มีค่ามากกว่า 0 แสดงว่าหลักเกณฑ์ที่พิจารณาไม่เป็นอิสระต่อกันแบบบวก กรณีที่ค่าเลเวอร์เรจเป็น 0 แสดงว่าพจน์หน้าและพจน์หลังของหลักเกณฑ์เป็นอิสระต่อกัน และกรณีที่ค่าเลเวอร์เรจน้อยกว่า 0 แสดงว่าพจน์หน้าและพจน์หลังของหลักเกณฑ์ไม่เป็นอิสระต่อกันแบบลบ โดยค่าเลเวอร์เรจจะนิยามเป็นจำนวนระเบียนที่สอดคล้องกับหลักเกณฑ์ลบด้วยผลคูณของจำนวนระเบียนที่สอดคล้องกับพจน์หน้าและจำนวนระเบียนที่สอดคล้องกับพจน์หลัง

$$lev(A \rightarrow C) = sup(A \rightarrow C) - sup(A) \times sup(C)$$

$$= \frac{|T_{A \cup C}|}{|D|} - \frac{|T_A| \times |T_C|}{|D^2|}$$

ซึ่งสามารถเขียนให้อยู่ในอีกรูปแบบหนึ่งที่เรียกว่าความแม่นสัมพัทธ์ถ่วงน้ำหนัก (weighted relative accuracy) ได้ดังสมการด้านล่าง

$$lev(A \rightarrow C) = sup(A) \times (conf(A \rightarrow C) - sup(C))$$

$$= \frac{|T_A|}{|D|} \times \left(\frac{|T_{A \cup C}|}{|T_A|} - \frac{|T_C|}{|D|} \right)$$

- ค่าคอนวิกชัน (conviction) [10,13] เกิดจากแนวความคิดเดียวกับค่ายกระดับ แต่เป็นการพิจารณาความเป็นอิสระต่อกันของหลักเกณฑ์ในกรณีที่เป็นพจน์หลังเป็นนิเสธ โดยถ้าพจน์หน้าและพจน์หลังแบบนิเสธเป็นอิสระต่อกันค่าคอนวิกชันจะมีค่าเป็น 1 และค่าคอนวิกชันจะมีค่าเป็นอนันต์เมื่อ $sup(A \rightarrow C)=0$ ซึ่งหมายความว่าพจน์หน้าและพจน์หลังจะปรากฏพร้อมกันเสมอ ซึ่งเหตุการณ์นี้จะมีค่าความเชื่อมั่นค่าเป็น 1

$$conv(A \rightarrow C) = \frac{sup(A) \times sup(\neg C)}{sup(A \rightarrow \neg C)} = \frac{|T_A| \times |T_{\neg C}|}{|T_{A \cup \neg C}| \times |D|}$$

การพิจารณาว่าหลักเกณฑ์ที่สร้างมีความน่าเชื่อถือ หรือน่าสนใจมากหรือน้อยจะขึ้นอยู่กับผู้ใช้หรือผู้วิเคราะห์ โดยผู้ใช้หรือผู้วิเคราะห์สามารถเลือกตัววัดได้ตามความต้องการและตามความเหมาะสมของข้อมูล แต่ตัววัดที่เป็นที่นิยมและใช้กันอย่างแพร่หลายในปัจจุบันคือ ค่าสนับสนุน และค่าความเชื่อมั่น เนื่องจากเป็นตัววัดที่เข้าใจง่าย และขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงโดยทั่วไปนิยมใช้ 2 ตัววัดนี้

เนื่องจากหลักเกณฑ์เชื่อมโยงเป็นการบ่งบอกถึงความสัมพันธ์ระหว่างกลุ่มของลักษณะประจำของข้อมูล ดังนั้นในการสร้างหลักเกณฑ์เชื่อมโยงจะต้องระบุพจน์หน้าและพจน์หลังของหลักเกณฑ์ก่อน แล้วจึงหาค่าของตัววัดที่ต้องการ หลักเกณฑ์เชื่อมโยงที่ได้จะน่าสนใจหรือยอมรับได้โดยผู้ใช้จะขึ้นกับการกำหนดค่าต่ำสุดของผู้ใช้ ดังตัวอย่างต่อไปนี้

ตัวอย่างที่ 2.2 จากตารางที่ 2.3 ซึ่งเป็นข้อมูลตัวอย่าง และตัวอย่างที่ 2.1 จะสามารถสร้างหลักเกณฑ์ได้ดังต่อไปนี้

สมมติให้หลักเกณฑ์ที่ต้องการสร้างคือ $\{2, 5\}\rightarrow\{1\}$ และ $\{1\}\rightarrow\{2, 5\}$

หลักเกณฑ์ $\{2, 5\}\rightarrow\{1\}$ จะมีค่าของตัววัดต่าง ๆ ดังนี้

$$sup\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}\cup\{1\}}|}{|D|} = \frac{2}{10} = 0.2$$

$$conf\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}\cup\{1\}}|}{|T_{\{2,5\}}|} = \frac{2}{2} = 1$$

$$cov\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}}|}{|D|} = \frac{2}{10} = 0.2$$

$$lift\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}\cup\{1\}}|\times|D|}{|T_{\{2,5\}}|\times|T_{\{1\}}|} = \frac{2\times10}{2\times7} = 1.43$$

$$lev\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}\cup\{1\}}|}{|D|} - \frac{|T_{\{2,5\}}|\times|T_{\{1\}}|}{|D^2|} = \frac{2}{10} - \frac{2\times7}{100} = 0.06$$

$$conv\ (\{2, 5\}\rightarrow\{1\}) = \frac{|T_{\{2,5\}}|\times|T_{\neg\{1\}}|}{|T_{\{2,5\}\cup\neg\{1\}}\times|D||} = \frac{2\times3}{0\times10} = \infty$$

ดังนั้นจึงสามารถเขียนหลักเกณฑ์ที่สร้างขึ้นได้ดังนี้คือ {2, 5}→{1}[*sup*= 0.2, *conf* = 1, *cov* = 0.2, *lift* = 1.43, *lev* = 0.06, *conv* = ∞] ซึ่งมีความหมายดังต่อไปนี้คือ

- ค่าสนับสนุนของหลักเกณฑ์มีค่าเป็น 0.2 อธิบายว่าหลักเกณฑ์นี้เกิดขึ้น 20% เมื่อเทียบกับข้อมูลทั้งหมด
- ค่าความเชื่อมั่นของหลักเกณฑ์มีค่าเป็น 1 อธิบายว่าทุกครั้งที่เกิดพจน์หน้าจะเกิดพจน์หลังเสมอ
- ค่าครอบคลุมมีค่าเป็น 0.2 อธิบายว่ามีระเบียนที่สอดคล้องกับพจน์หน้าคิดเป็น 20% เมื่อเทียบกับข้อมูลทั้งหมด
- ค่ายกระดับมีค่ามากกว่า 1 และค่าเลเวอร์เรจมีค่าเป็นบวก อธิบายได้ว่าหลักเกณฑ์นี้มีความสัมพันธ์กัน เพราะพจน์หน้าและพจน์หลังไม่เป็นอิสระต่อกันแบบบวก
- ค่าคอนวิกชันมีค่าเป็นอนันต์ ซึ่งอธิบายว่าพจน์หน้าและพจน์หลังของหลักเกณฑ์จะเกิดขึ้นพร้อมกันเสมอ

ในทำนองเดียวกันหลักเกณฑ์ {1}→{2, 5} จะมีค่าของตัววัดต่าง ๆ ดังนี้

$$sup\ (\{1\}\rightarrow\{2,5\}) = \frac{|T_{\{1\}\cup\{2,5\}}|}{|D|} = \frac{2}{10} = 0.2$$

$$conf\ (\{1\}\rightarrow\{2,5\}) = \frac{|T_{\{1\}\cup\{2,5\}}|}{|T_{\{1\}}|} = \frac{2}{7} = 0.29$$

$$cov\ (\{1\}\rightarrow\{2,5\}) = \frac{|T_{\{1\}}|}{|D|} = \frac{7}{10} = 0.7$$

$$lift\ (\{1\}\rightarrow\{2,5\}) = \frac{|T_{\{1\}\cup\{2,5\}}|\times|D|}{|T_{\{1\}}|\times|T_{\{2,5\}}|} = \frac{2\times10}{7\times2} = 1.43$$

$$lev\ (\{1\}\rightarrow\{2,5\}) = \frac{|T_{\{1\}\cup\{2,5\}}|}{|D|} - \frac{|T_{\{1\}}|\times|T_{\{2,5\}}|}{|D^2|} = \frac{2}{10} - \frac{7\times2}{100} = 0.06$$

$$conv\,(\{1\}\rightarrow\{2,5\}) = \frac{\left|T_{\{1\}}\right|\times\left|T_{\neg\{2,5\}}\right|}{\left|T_{\{1\}\cup\neg\{2,5\}}\times|D|\right|} = \frac{7\times 8}{5\times 10} = 1.12$$

จึงสรุปเป็นหลักเกณฑ์ได้ว่า {1}→{2, 5}[*sup*= 0.2, *conf* = 0.29, *cov* = 0.7, *lev* = 0.06, *lift* = 1.43, *conv* = 1.12] จะเห็นได้ชัดว่าค่าสนับสนุน ค่ายกระดับ และค่าเลเวอร์เรจของหลักเกณฑ์ {1}→{2, 5} และ {2, 5}→{1} มีค่าเท่ากัน เนื่องจากตัววัดดังกล่าวเป็นตัววัดที่สมมาตร การสลับที่ระหว่างพจน์หน้าและพจน์หลังของหลักเกณฑ์จะไม่ส่งผลต่อค่าที่คำนวณได้ แต่ค่าความเชื่อมั่น ค่าครอบคลุม และค่าคอนวิกชันของทั้งสองหลักเกณฑ์จะมีค่าแตกต่างกัน เพราะตัววัดกลุ่มนี้เป็นตัววัดที่ไม่สมมาตร ซึ่งหลักเกณฑ์ {1}→{2, 5} จะมีความหมายแตกต่างกับหลักเกณฑ์ {2, 5}→{1} เฉพาะตัววัดที่ไม่สมมาตรดังต่อไปนี้

- ค่าความเชื่อมั่นของหลักเกณฑ์มีค่าเป็น 0.29 อธิบายว่าทุกครั้งที่เกิดพจน์หน้าจะเกิดพจน์หลังเป็นจำนวน 29%
- ค่าครอบคลุมมีค่าเป็น 0.7 อธิบายว่าพจน์หน้าของหลักเกณฑ์ปรากฏในข้อมูลเป็น 70% เมื่อเทียบกับข้อมูลทั้งหมด
- ค่าคอนวิกชันของหลักเกณฑ์มีค่าเป็น 1.12 อธิบายว่าหลักเกณฑ์นี้พจน์หน้าและพจน์หลังแบบนิเสธไม่เป็นอิสระต่อกันแบบบวก

โดยทั่วไปในการพิจารณาว่าหลักเกณฑ์เชื่อมโยงที่สร้างมีความน่าเชื่อถือ จะขึ้นอยู่กับค่าต่ำสุดที่กำหนดโดยผู้ใช้ สมมติให้ค่าสนับสนุนต่ำสุดมีค่าเป็น 0.1 และค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 0.8 จะพบว่าหลักเกณฑ์ {2, 5}→{1} เป็นหลักเกณฑ์ที่น่าเชื่อถือ เพราะมีค่าสนับสนุนและค่าความเชื่อมั่นมากกว่าค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุดที่กำหนดไว้ แต่หลักเกณฑ์ {1}→{2, 5} ไม่เป็นหลักเกณฑ์ที่น่าเชื่อถือเพราะมีค่าความเชื่อมั่นเพียง 0.29 ซึ่งน้อยกว่าค่าความเชื่อมั่นต่ำสุดที่กำหนด

ในการสร้างหลักเกณฑ์เชื่อมโยงจะต้องกำหนดพจน์หน้าและพจน์หลังของหลักเกณฑ์ แล้วจึงหาค่าความสัมพันธ์ของหลักเกณฑ์โดยคำนวณจากตัววัด ดังนั้นจำนวนหลักเกณฑ์ที่สร้างขึ้นได้ทั้งหมด จะขึ้นอยู่กับปริมาณของไอเทมทั้งหมด จากปริภูมิลักษณะประจำพบว่าจำนวน

ไอเทมเซตที่สามารถนำมาสร้างเป็นหลักเกณฑ์ได้มีอยู่ 2^{m-1} ตัว เพราะไม่พิจารณาการเลือกพจน์หน้าหรือพจน์หลังที่เป็นเซตว่างและเซตของไอเทมเซตทั้งหมด เมื่อนำมาสร้างหลักเกณฑ์จะมีจำนวนหลักเกณฑ์ที่สร้างได้เป็นจำนวนมาก ทำให้เกิดขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงดังหัวข้อถัดไป

2.3.5. ขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยง

ขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงในปัจจุบันมีอยู่หลากหลาย ซึ่งผู้วิเคราะห์หรือผู้ใช้งานสามารถเลือกตามความเหมาะสมของข้อมูลและผลลัพธ์ที่ต้องการ โดยทั่วไปขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงแบ่งออกเป็น 2 ขั้นตอนหลักคือ

1. ขั้นตอนสร้างไอเทมเซตปรากฏบ่อย (frequent itemset generation) มีจุดมุ่งหมายคือสร้างไอเทมเซตปรากฏบ่อย (frequent itemset) ทั้งหมดจากข้อมูล เมื่อไอเทมเซตปรากฏบ่อยคือเซตของไอเทมที่มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนต่ำสุดที่กำหนดโดยผู้ใช้
2. ขั้นตอนสร้างหลักเกณฑ์ (rule generation) มีจุดมุ่งหมายเพื่อสร้างหลักเกณฑ์เชื่อมโยงทั้งหมด โดยต้องผ่านการสร้างไอเทมเซตปรากฏบ่อยจากขั้นตอนที่ 1 แล้วนำไอเทมเซตปรากฏบ่อยที่ได้มาสร้างเป็นหลักเกณฑ์เชื่อมโยง โดยอาศัยตัววัดอื่น เช่น ความเชื่อมั่น เป็นตัวตัดสินความน่าเชื่อถือของหลักเกณฑ์

ขั้นตอนวิธีที่สอดคล้องกับขั้นตอนในการสร้างหลักเกณฑ์เชื่อมโยงดังข้างต้น จะเป็นขั้นตอนวิธีที่ใช้ค่าสนับสนุนและค่าความเชื่อมั่นในการตัดสินความน่าเชื่อถือของหลักเกณฑ์ ซึ่งเรียกว่า โครงข่ายค่าสนับสนุนค่าความเชื่อมั่น (support-confidence framework) [9] ได้แก่ ขั้นตอนวิธีเอไพออรี (Apriori algorithm) และขั้นตอนวิธีต้นไม้ Fp (frequent pattern tree algorithm: Fp-tree algorithm) [14] เป็นต้น

แต่ในปัจจุบันการสร้างหลักเกณฑ์เชื่อมโยงมีการศึกษาและวิจัยออกไปหลายแนวทาง [10] ได้แก่ การทำเหมืองเงื่อนไขบังคับพื้นฐาน (constraint-based mining) ซึ่งเป็นการสร้างหลักเกณฑ์เชื่อมโยงโดยกำหนดเงื่อนไขเพิ่มเติมเพื่อลดปริมาณหลักเกณฑ์ที่พิจารณา ในขณะที่เพิ่มความเร็วในการทำงาน เช่น ขั้นตอนวิธี KORD (k-optimal rule discovery algorithm: KORD algorithm) [11] ขั้นตอนวิธีนี้จะสร้างหลักเกณฑ์เชื่อมโยงโดยพิจารณาการค้นหาพจน์หน้าของหลักเกณฑ์แบบ OPUS (optimised pruning for unordered search-spaces: OPUS) [24] ผลลัพธ์ที่ได้คือหลักเกณฑ์จำนวน *k* หลักเกณฑ์ที่เรียงลำดับค่าเลเวอร์เรจจากมากไปน้อย เพราะหลักเกณฑ์ที่มีค่าเลเวอร์เรจมากแสดงว่าพจน์หน้ากับพจน์หลังของหลักเกณฑ์มีความสัมพันธ์กันมาก การทำเหมืองเพื่อเลือกตัวแทนแบบกระชับของไอเทมเซตปรากฏบ่อย (mining concise representations of frequent itemset) ซึ่งเป็นการค้นหาไอเทมเซตแบบปิด และมากที่สุด เช่น ขั้นตอนวิธีแบบปิด (closed algorithm) [25] โดยจะค้นหาไอเทมเซตแบบปิด (closed itemset) ซึ่งเป็นไอเทมเซตปรากฏบ่อยที่ใหญ่ที่สุดที่เกิดขึ้นในฐานข้อมูล เป็นต้น

นอกจากนี้หลักเกณฑ์เชื่อมโยงได้นำไปประยุกต์ใช้กับปัญหาการวัดจำแนกประเภทเพื่อสร้างตัวแบบจำแนกประเภท เช่น ขั้นตอนวิธี CBA (classification based on associations: CBA) ซึ่งกำหนดให้ลักษณะประจำที่ใช้ในการจำแนกอยู่ที่พจน์หลัง และกำหนดให้ค่าความเชื่อมั่นมีค่าสูง ผลลัพธ์ที่ได้พบว่ามีค่าความแม่นยำมากกว่า C4.5 แต่ขั้นตอนวิธีที่ได้รับความนิยม และเป็นมาตรฐาน คือ ขั้นตอนวิธีเอไพออรี ซึ่งจะกล่าวถึงในหัวข้อถัดไป

2.3.6. ขั้นตอนวิธีเอไพออรี

ขั้นตอนวิธีในการสร้างหลักเกณฑ์เชื่อมโยงที่เป็นที่รู้จักและนิยมใช้กันอย่างแพร่หลายในปัจจุบันคือ ขั้นตอนวิธีเอไพออรี (Apriori algorithm) [8] ซึ่งแบ่งการทำงานออกเป็น 2 ขั้นตอนหลัก คือ ขั้นตอนแรกเป็นการหาไอเทมเซตปรากฏบ่อย (frequent itemsets) $\{A \mid A \subseteq I, sup(A) \geq minSup\}$ เมื่อ *minSup* คือค่าสนับสนุนต่ำสุดที่กำหนดโดยผู้ใช้ และในขั้นตอนที่สองคือการนำไอเทมเซตปรากฏบ่อยที่ได้จากขั้นตอนแรกมาสร้างเป็นหลักเกณฑ์ ซึ่งหลักเกณฑ์ที่สร้างจะยอมรับได้ก็ต่อเมื่อมีค่าความเชื่อมั่นของหลักเกณฑ์มากกว่าหรือเท่ากับค่าความเชื่อมั่นต่ำสุดที่กำหนดโดยผู้ใช้ โดยในการทำงานของขั้นตอนที่สองไม่จำเป็นต้องเข้าถึงฐานข้อมูล เพราะสามารถ

คำนวณค่าความเชื่อมั่นได้จากไอเทมเซตปรากฏบ่อยที่ได้จากขั้นตอนแรก ทำให้การทำงานของขั้นตอนวิธีนี้รวดเร็วมากยิ่งขึ้น

ขั้นตอนที่ 1. ขั้นตอนในการสร้างไอเทมเซตปรากฏบ่อย

เริ่มจากการสร้างไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิก $k = 1$ แล้วสร้างไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิก $k+1$ จากไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิกเท่ากับ k จนไม่สามารถสร้างไอเทมเซตปรากฏบ่อยได้อีก อธิบายดังรูปที่ 2.3 [1] โดยกำหนดให้ เซตของไอเทมเซตตัวเลือก (candidate itemset) ที่มีจำนวนสมาชิกเท่ากับ k คือ C_k เซตของไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิกเท่ากับ k คือ L_k และค่าสนับสนุนต่ำสุดที่นับได้ (minimum support count) คือ *minSupCount* ซึ่งคำนวณได้จากการนำค่าสนับสนุนต่ำสุดที่กำหนดโดยผู้ใช้คูณกับจำนวนของข้อมูล ซึ่งสามารถเขียนได้เป็น $minSupCount = minSup \times |D|$

1. L_1 = find frequent 1–itemset from D
2. for ($k = 2$; $L_{k-1} \neq \emptyset$; k++)
3. C_k= apriori_gen(L_{k-1}, *minSupCount*)
4. for each transaction $t \in D$ { //scan D for counts
5. C_t=subset(C_k, t) //get the subset of t that are candidates
6. for each candidate $c \in C_t$
7. c.count++
8. $L_k = \{c \in C_k \mid c.count \geq minSupCount\}$
9. **return** $L = \cup_k L_k$

รูปที่ 2.3: ขั้นตอนการสร้างไอเทมเซตปรากฏบ่อยของขั้นตอนวิธีเอไพออรี

จากรูปที่ 2.3 จะอธิบายขั้นตอนในการสร้างไอเทมเซตปรากฏบ่อยได้ดังนี้

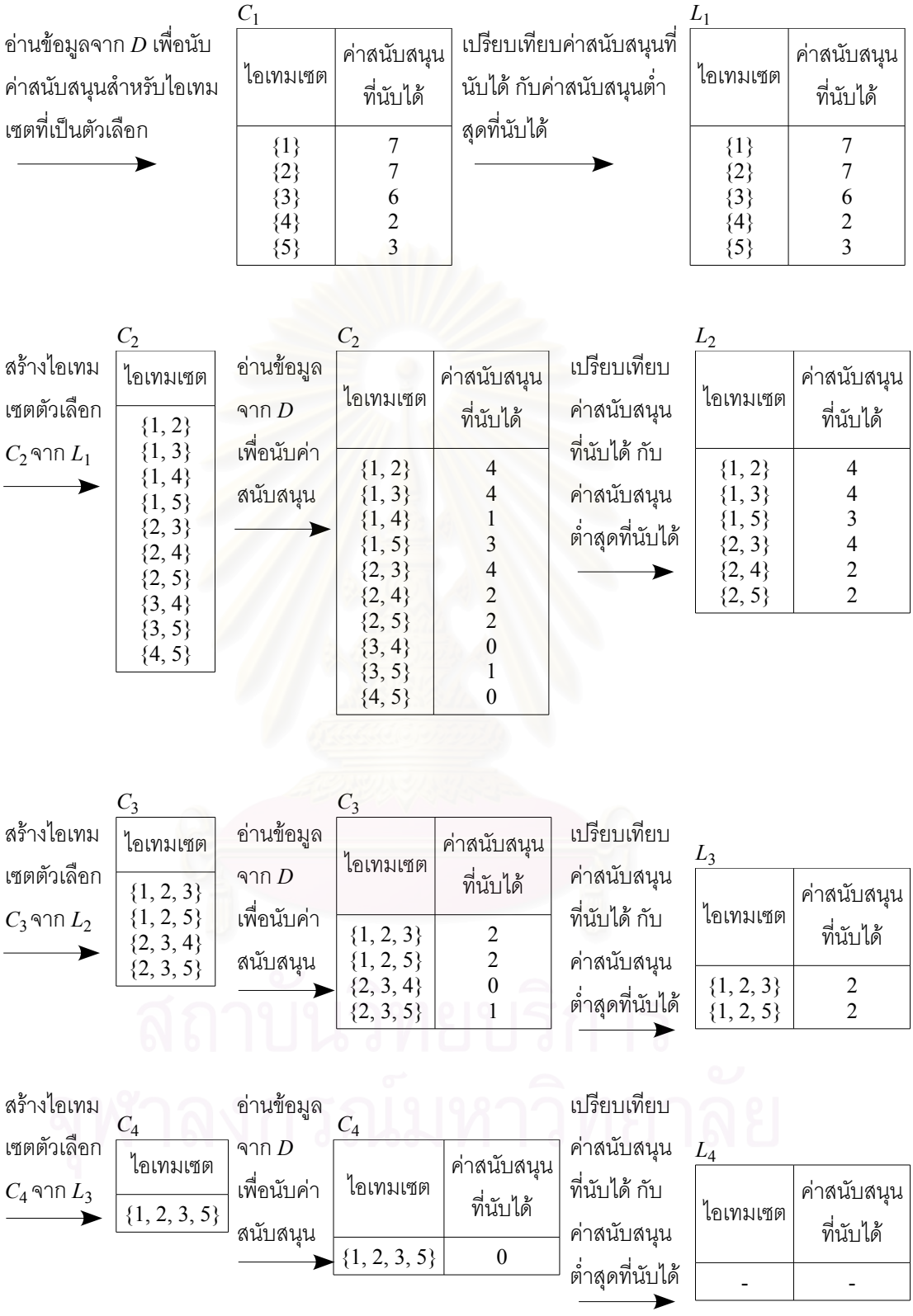
1. สร้างเซตของไอเทมเซตที่มีจำนวนสมาชิกเท่ากับ 1 (1–ไอเทมเซต) โดยเรียกว่าเซตของไอเทมเซตตัวเลือกที่มีจำนวนสมาชิกเท่ากับ 1 หรือ C_1 จากข้อมูลที่ต้องการวิเคราะห์ แล้วพิจารณาสร้างเซตของไอเทมเซตปรากฏบ่อยที่มีขนาดเท่ากับ 1 หรือ L_1 จากไอเทมเซตใน C_1 ที่มีค่าสนับสนุนที่นับได้มากกว่าหรือเท่ากับค่าสนับสนุนต่ำสุดที่นับได้

2. ในแต่ละรอบของการทำงานที่ k จะสร้างเซตของไอเทมเซตตัวเลือก k หรือ C_k จากเซตของไอเทมเซตปรากฏบ่อย L_{k-1} โดยกำหนดให้ค่า k เริ่มต้นมีค่าเป็น 2

3. นำ C_k ที่ได้จากขั้นตอนที่ 2 มาหาค่าสนับสนุน แล้วพิจารณาสร้างไอเทมเซตปรากฏบ่อย L_k จากไอเทมเซตที่เป็นสมาชิกใน C_k ที่มีค่าสนับสนุนที่นับได้มากกว่าหรือเท่ากับค่าสนับสนุนต่ำสุดที่นับได้ และกำหนดให้ $k = k+1$

4. ทำซ้ำในขั้นตอนที่ 2 และ 3 ไปจนกระทั่งไม่สามารถสร้างเซตของ ไอเทมเซตปรากฏบ่อยได้อีก

ดังนั้นเมื่อนำข้อมูลตัวอย่างจากตารางที่ 2.3 มาผ่านขั้นตอนแรกของขั้นตอนวิธีเอไพออรี จะได้ผลลัพธ์เป็นไอเทมเซตปรากฏบ่อยดังรูปที่ 2.4 เมื่อกำหนดให้ค่าสนับสนุนต่ำสุดคือ 0.2 ดังนั้นค่าสนับสนุนต่ำสุดที่นับได้ จะมีค่าเป็น $0.2 \times 10 = 2$ ดังนั้นไอเทมเซตปรากฏบ่อยคือไอเทมเซตที่มีค่าสนับสนุนที่นับได้มากกว่าหรือเท่ากับ 2



รูปที่ 2.4: ตัวอย่างการสร้างไอเทมเซตปรากฏบ่อยของขั้นตอนวิธีเอไพออรี

จากรูปที่ 2.4 จะได้เซตของไอเทมเซตตัวเลือก และเซตของไอเทมเซตปรากฏบ่อย ในแต่ละรอบของการทำงานดังต่อไปนี้

- รอบการทำงานที่ 1

  C_1 = {{1}, {2}, {3}, {4}, {5}}

  L_1 = {{1}, {2}, {3}, {4}, {5}}

- รอบการทำงานที่ 2

  C_2 = {{1, 2}, {1, 3}, {1, 4}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {3, 4}, {3, 5}, {4, 5}}

  L_2 = {{1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 4}, {2, 5}}

- รอบการทำงานที่ 3

  C_3 = {{1, 2, 3}, {1, 2, 5}, {2, 3, 4}, {2, 3, 5}}

  L_3 = {{1, 2, 3}, {1, 2, 5}}

- รอบการทำงานที่ 4

  C_4 = {{1, 2, 3, 5}}

  L_4 = {}

ดังนั้นเซตของไอเทมเซตปรากฏบ่อยทั้งหมดคือ {{1}, {2}, {3}, {4}, {5}, {1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {1, 2, 3}, {1, 2, 5}}

ในรอบการทำงานที่ 3 และ 4 พบว่าเซตของไอเทมเซตตัวเลือกที่สร้างขึ้นจากไอเทมเซตปรากฏบ่อยที่ได้จากรอบการทำงานที่ 2 และ 3 มีจำนวน 6 และ 2 ไอเทมเซตตามลำดับ แต่เมื่อหาค่าสนับสนุนที่นับได้แล้วพบว่า บางไอเทมเซตตัวเลือกที่สร้างมีค่าสนับสนุนที่นับได้น้อยกว่าค่าสนับสนุนต่ำสุดที่นับได้ ซึ่งในกรณีที่ไอเทมเซตตัวเลือกมีจำนวนมากจะทำให้ใช้เวลาในการทำงานมากขึ้น ดังนั้นการลดจำนวนไอเทมเซตตัวเลือกจะต้องใช้สมบัติของค่าสนับสนุน ที่เรียกว่าสมบัติเอไพออรี (Apriori property) [1] นั่นคือ ทุกเซตย่อยที่เป็นไอเทมเซตปรากฏบ่อยที่ไม่ใช่เซตว่างจะเป็นไอเทมเซตปรากฏบ่อย ดังนั้นเมื่ออาศัยสมบัติเอไพออรีเพื่อลดจำนวนไอเทมเซตตัวเลือกจะทำให้ลดจำนวนครั้งในการเข้าถึงข้อมูล ทำให้การทำงานของขั้นตอนวิธีนี้รวดเร็วมากขึ้น

ดังนั้นจากรูปที่ 2.4 เมื่อพิจารณารอบการทำงานที่ 3 พบว่าไอเทมเซตตัวเลือก {2, 3, 4} ไม่สอดคล้องกับสมบัติเอไพออรี เพราะ {3, 4} ไม่เป็นไอเทมเซตปรากฏบ่อย ในทำนองเดียวกันไอเทมเซตตัวเลือก {2, 3, 5} ไม่สอดคล้องกับสมบัติเอไพออรี เพราะ {3, 5} ไม่เป็นไอเทมเซตปรากฏบ่อย และในรอบการทำงานที่ 4 จะพบว่าไอเทมเซตตัวเลือก {1, 2, 3, 5} ไม่สอดคล้องกับสมบัติเอไพออรี

ขั้นตอนที่ 2. ขั้นตอนในการสร้างหลักเกณฑ์เชื่อมโยงจากไอเทมเซตปรากฏบ่อย

สมมติให้เซตของไอเทมเซตปรากฏบ่อยที่ได้จากขั้นตอนที่ 1 คือ L และกำหนดให้ค่าความเชื่อมั่นต่ำสุดที่กำหนดโดยผู้ใช้คือ *minConf* การสร้างหลักเกณฑ์เชื่อมโยงจากไอเทมเซตปรากฏบ่อยมีขั้นตอนดังต่อไปนี้ [1]

1. กำหนดให้ไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิกมากกว่าหรือเท่ากับ 2 เป็น l โดยที่ $l \in L$ หลักเกณฑ์เชื่อมโยงที่สร้างได้จะอยู่ในรูปแบบของ $s \rightarrow (l-s)$ สำหรับทุก $s \subset l$
2. คำนวณหาค่าความเชื่อมั่นของหลักเกณฑ์ $s \rightarrow (l-s)$ เมื่อค่าความเชื่อมั่นของหลักเกณฑ์มีค่ามากกว่าหรือเท่ากับ *minConf* ก็จะสร้างหลักเกณฑ์ดังกล่าวออกมาเป็นผลลัพธ์

ดังนั้นจากรูปที่ 2.4 จะได้ว่าไอเทมเซตปรากฏบ่อยที่มีจำนวนสมาชิกมากกว่าหรือเท่ากับ 2 ได้แก่ {1, 2}, {1, 3}, {1, 5}, {2, 3}, {2, 4}, {2, 5}, {1, 2, 3}, {1, 2, 5} การสร้างหลักเกณฑ์เชื่อมโยงทำได้ดังนี้

สมมติให้ l = {1, 2, 5} เซตย่อยของ l ที่ไม่เท่ากับเซตว่าง และไม่เท่ากับ l คือ {1}, {2}, {5}, {1, 2}, {1, 5}, {2, 5} ดังนั้นหลักเกณฑ์เชื่อมโยงที่สร้างได้คือ

$\{1\} \rightarrow \{2, 5\}$ มีค่าความเชื่อมั่น = $\frac{2}{7} = 0.29$

$\{2\} \rightarrow \{1, 5\}$ มีค่าความเชื่อมั่น = $\frac{2}{7} = 0.29$

$\{5\}\rightarrow\{1, 2\}$ มีค่าความเชื่อมั่น = $\frac{2}{3} = 0.67$

$\{1, 2\}\rightarrow\{5\}$ มีค่าความเชื่อมั่น = $\frac{2}{4} = 0.5$

$\{1, 5\}\rightarrow\{2\}$ มีค่าความเชื่อมั่น = $\frac{2}{3} = 0.67$

$\{2, 5\}\rightarrow\{1\}$ มีค่าความเชื่อมั่น = $\frac{2}{2} = 1$

เมื่อกำหนด *minConf* = 0.6 จะได้ว่าหลักเกณฑ์ที่ได้เป็นผลลัพธ์จากไอเทมเซตปรากฏบ่อย {1, 2, 5} ของขั้นตอนวิธีเอไพออรีคือหลักเกณฑ์ $\{5\}\rightarrow\{1, 2\}$, $\{1, 5\}\rightarrow\{2\}$ และ $\{2, 5\}\rightarrow\{1\}$

ค่าสนับสนุนเป็นตัววัดที่สมมาตร เพราะเมื่อสลับตำแหน่งระหว่างพจน์หน้ากับพจน์หลังค่าสนับสนุนที่คำนวณได้จะมีค่าเหมือนเดิม ดังนั้นหลักเกณฑ์ที่สร้างขึ้นโดยอาศัยสมบัติของค่าสนับสนุนจะต้องอาศัยค่าความเชื่อมั่นหรือตัววัดอื่นๆ ที่มีลักษณะของตัววัดแบบไม่สมมาตร คือสลับพจน์หน้าและพจน์หลังของหลักเกณฑ์จะทำให้ค่าที่คำนวณได้เปลี่ยนแปลง เพื่อใช้ในการคัดเลือกหลักเกณฑ์ที่น่าเชื่อถือ

เนื่องจากในการหาไอเทมเซตปรากฏบ่อยในขั้นตอนแรกของขั้นตอนวิธีเอไพออรีจะต้องใช้ค่าสนับสนุนต่ำสุดเป็นตัวตัดไอเทมเซตที่ไม่สำคัญออก ทำให้หลักเกณฑ์ที่สร้างในขั้นตอนที่สองนั้นเป็นหลักเกณฑ์ที่ต้องสอดคล้องกับค่าสนับสนุนต่ำสุดที่ผู้ใช้เป็นคนกำหนดเท่านั้น ดังนั้นผู้ใช้จะต้องเป็นคนที่มีความเข้าใจในข้อมูลอย่างลึกซึ้งเพื่อให้สามารถกำหนดค่าสนับสนุนต่ำสุดได้อย่างเหมาะสม ในกรณีที่ผู้ใช้กำหนดค่าสนับสนุนต่ำสุดมากเกินไปจะทำให้ตัดไอเทมเซตที่จะสร้างหลักเกณฑ์ที่มีค่าความเชื่อมั่นสูงออกไป และในกรณีที่ผู้ใช้กำหนดค่าสนับสนุนต่ำสุดน้อยเกินไป จะทำให้สร้างไอเทมเซตปรากฏบ่อยออกมามากและได้หลักเกณฑ์ออกมามากเป็นจำนวนมาก

ดังนั้นในงานวิจัยนี้จึงสนใจที่จะสร้างตัววัดตัวใหม่ที่มีชื่อว่า ค่าสนับสนุนแบบอ่อน (weak support) ซึ่งเกิดจากแนวคิดทางตรรกศาสตร์ เนื่องจากข้อมูลที่ใช้ในการสร้างหลักเกณฑ์เชื่อมโยงจะเก็บอยู่ในรูปแบบของตัวเลขฐานสองคือ 0,1 ซึ่งหมายความถึงการไม่ซื้อสินค้า และซื้อสินค้าตามลำดับ และเมื่อพิจารณาหลักเกณฑ์เชื่อมโยงก็พบว่าหลักเกณฑ์เชื่อมโยงเขียนอยู่ในรูปแบบเดียวกับประพจน์ถ้า-แล้วในทางตรรกศาสตร์ จึงทำให้เกิดแนวคิดว่าสามารถนำทฤษฎีทางตรรกศาสตร์มาประยุกต์ใช้เข้ากับการสร้างหลักเกณฑ์เชื่อมโยงได้ ซึ่งประพจน์ถ้า-แล้ว $p\rightarrow q$ มีข้อ

ขัดแย้งคือ $p \wedge \neg q$ ในทำนองเดียวกับหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ จะมีข้อขัดแย้งคือ $A \cup \neg C$ ในปริภูมิลักษณะประจำ ซึ่งจะอธิบายเพิ่มเติมในบทที่ 3 ค่าสนับสนุนแบบอ่อนจึงเป็นตัววัดที่ไม่พิจารณาตัวอย่างที่ขัดแย้งกับหลักเกณฑ์ หลังจากนั้นจึงสร้างขั้นตอนวิธีที่ใช้ค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่นเป็นตัวกำหนดความน่าเชื่อถือของหลักเกณฑ์แล้วเปรียบเทียบหลักเกณฑ์ที่ได้จากสองขั้นตอนวิธี

บทที่ 3

การคัดเลือกหลักเกณฑ์เชื่อมโยงโดยค่าสนับสนุนแบบอ่อน

ในบทนี้จะอธิบายที่มาและแนวคิดของการสร้างค่าสนับสนุนแบบอ่อน เนื่องจากค่าสนับสนุนแบบอ่อนเกิดจากแนวคิดทางด้านตรรกศาสตร์ ดังนั้นจะกล่าวถึงการนำความรู้ทางด้านตรรกศาสตร์มาประยุกต์เข้ากับการสร้างหลักเกณฑ์เชื่อมโยง อธิบายทฤษฎีบทที่เกี่ยวข้องกับการสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน อธิบายและแสดงความสัมพันธ์ระหว่างค่าสนับสนุนแบบอ่อน ค่าสนับสนุน และค่าความเชื่อมั่น และในหัวข้อสุดท้ายจะอธิบายขั้นตอนวิธีที่ใช้ในการคัดเลือกหลักเกณฑ์เชื่อมโยงโดยค่าสนับสนุนแบบอ่อน

3.1. ค่าสนับสนุนแบบอ่อน

เนื่องจากหลักเกณฑ์เชื่อมโยงเขียนอยู่ในรูปแบบของประพจน์ถ้า-แล้ว จึงทำให้เกิดการนำแนวคิดทางตรรกศาสตร์มาประยุกต์เข้ากับการสร้างหลักเกณฑ์เชื่อมโยง กลายเป็นตัววัดใหม่ที่น่าสนใจ อธิบายได้ดังนี้

พิจารณาประพจน์ถ้า-แล้วทางตรรกศาสตร์ $p \rightarrow q$ พบว่าข้อขัดแย้งของประพจน์คือ $p \wedge \neg q$ และพิจารณาหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ บนฐานข้อมูลทั้งหมดพบว่าสามารถแบ่งข้อมูลออกเป็น 4 กลุ่มดังตารางที่ 3.1 คือ

TID	A	C
$t_i;\ \exists i \in \{1,\ldots,n\}, t_i \subseteq \neg A \cup \neg C$	0	0
$t_i;\ \exists i \in \{1,\ldots,n\}, t_i \subseteq \neg A \cup C$	0	1
$t_i;\ \exists i \in \{1,\ldots,n\}, t_i \subseteq A \cup \neg C$	1	0
$t_i;\ \exists i \in \{1,\ldots,n\}, t_i \subseteq A \cup C$	1	1

ตารางที่ 3.1: ข้อมูลที่แบ่งออกเป็นสี่กลุ่มตามตรรกะของ A, C

1. กลุ่มของกรณีพจน์หน้าไม่ปรากฏ และพจน์หลังไม่ปรากฏในระเบียนเดียวกัน ซึ่งหมายความถึงระเบียนที่ไอเทม A มีค่าเป็น 0 และ C มีค่าเป็น 0 โดยมีความหมายในการวิเคราะห์มาเก็ตบาสเก็ตว่าในการซื้อครั้งนั้นจะต้องไม่ซื้อทั้งสินค้า A และ C และเมื่ออธิบายในรูปแบบของนิยามในการสร้างหลักเกณฑ์เชื่อมโยงคือ $\neg A \cup \neg C$ ในปริภูมิลักษณะประจำ และเป็น $T_{\neg A \cup \neg C} = T_{\neg A} \cap T_{\neg C}$ ในปริภูมิระเบียน

2. กลุ่มของกรณีพจน์หน้าไม่ปรากฏ แต่พจน์หลังปรากฏในระเบียนเดียวกัน คือระเบียนที่ไอเทม A มีค่าเป็น 0 และ C มีค่าเป็น 1 นั่นคือการซื้อครั้งนั้นจะต้องไม่ซื้อสินค้า A แต่ซื้อสินค้า C ซึ่งหมายถึงไอเทมเซต $\neg A \cup C$ ในปริภูมิลักษณะประจำ หรือ $T_{\neg A \cup C} = T_{\neg A} \cap T_C$ ในปริภูมิระเบียน

3. กลุ่มของกรณีพจน์หน้าปรากฏ แต่พจน์หลังไม่ปรากฏในระเบียนเดียวกัน คือระเบียนที่ไอเทม A มีค่าเป็น 1 และ C มีค่าเป็น 0 นั่นคือการซื้อครั้งนั้นจะต้องซื้อสินค้า A แต่ไม่ซื้อสินค้า C ซึ่งหมายถึงไอเทมเซต $A \cup \neg C$ ในปริภูมิลักษณะประจำ หรือ $T_{A \cup \neg C} = T_A \cap T_{\neg C}$ ในปริภูมิระเบียน

4. กลุ่มของกรณีพจน์หน้าปรากฏ และพจน์หลังปรากฏในระเบียนเดียวกัน คือระเบียนที่ไอเทม A มีค่าเป็น 1 และ C มีค่าเป็น 1 นั่นคือการซื้อครั้งนั้นจะต้องซื้อสินค้าทั้งสองชนิด ซึ่งหมายถึงไอเทมเซต $A \cup C$ ในปริภูมิลักษณะประจำ หรือ $T_{A \cup C} = T_A \cap T_C$ ในปริภูมิระเบียน

เมื่อพิจารณาตารางที่ 3.1 พบว่าข้อมูลสามารถมองเป็นตารางค่าความจริงได้ โดยค่า 1 ของลักษณะประจำในแต่ละกลุ่มของข้อมูล จะหมายถึงค่าความจริงเป็นจริงในความหมายทางตรรกศาสร์ ในทำนองเดียวกันกรณีที่มีค่าเป็น 0 จะมีค่าความจริงเป็นเท็จ ดังนั้นกลุ่มของข้อมูลที่ขัดแย้งกับหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ คือกรณีที่ A เป็น 1 และ C เป็น 0 ซึ่งก็คือกลุ่มของระเบียนที่สอดคล้องกับไอเทมเซต $A \cup \neg C$ ในปริภูมิลักษณะประจำ หรือ $T_{A \cup \neg C} = T_A \cap T_{\neg C}$ ในปริภูมิระเบียน ดังนั้นจึงนิยามค่าสนับสนุนแบบอ่อนให้เป็นตัววัดที่ไม่พิจารณาระเบียนหรือไอเทมเซตที่ขัดแย้งกับหลักเกณฑ์ โดยสามารถคำนวณได้จากสัดส่วน

ระหว่างจำนวนระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์เชื่อมโยงต่อระเบียนทั้งหมด ดังนั้นค่าสนับสนุนแบบอ่อนจะเป็นตัววัดที่ไม่สมมาตร ซึ่งเขียนได้ดังนี้

$$weak\ sup(A \rightarrow C) = \frac{|T_{A \cup C}| + |T_{\neg A \cup C}| + |T_{\neg A \cup \neg C}|}{|D|}$$

หรือเขียนให้อยู่ในอีกรูปแบบหนึ่ง คือ

$$weak\ sup(A \rightarrow C) = 1 - \frac{|T_{A \cup \neg C}|}{|D|}$$

ดังนั้นเมื่อค่าสนับสนุนแบบอ่อนของหลักเกณฑ์มีค่าเป็น 1 จะหมายความว่าไม่มีระเบียนใดที่ขัดแย้งกับหลักเกณฑ์ และเมื่อค่าสนับสนุนแบบอ่อนมีค่าเป็น 0 แสดงว่าทุกระเบียนเป็นระเบียนที่ขัดแย้งกับหลักเกณฑ์ที่ต้องการพิจารณา โดยตัวอย่างในการหาค่าสนับสนุนแบบอ่อนของหลักเกณฑ์แสดงดังตัวอย่างที่ 3.1

ตัวอย่างที่ 3.1 จากข้อมูลในตารางที่ 2.3 และตัวอย่างที่ 2.1 สมมติให้หลักเกณฑ์ที่ต้องการหาค่าสนับสนุนแบบอ่อนคือ $\{2, 5\} \rightarrow \{1\}$ และ $\{1\} \rightarrow \{2, 5\}$ จะมีค่าสนับสนุนแบบอ่อนดังนี้ คือ

$$weak\ sup(\{2, 5\} \rightarrow \{1\}) = 1 - \frac{|T_{\{2,5\} \cup \neg\{1\}}|}{|D|} = 1 - \frac{0}{10} = 1$$

$$weak\ sup(\{1\} \rightarrow \{2, 5\}) = 1 - \frac{|T_{\{1\} \cup \neg\{2,5\}}|}{|D|} = 1 - \frac{5}{10} = 0.5$$

หลักเกณฑ์ $\{2, 5\} \rightarrow \{1\}$ มีค่าสนับสนุนแบบอ่อนเป็น 1 หมายความว่าไม่มีระเบียนใดที่ขัดแย้งกับหลักเกณฑ์ ส่วนหลักเกณฑ์ $\{1\} \rightarrow \{2, 5\}$ มีค่าสนับสนุนแบบอ่อนเป็น 0.5 จะหมายความว่าข้อมูลที่พิจารณามีระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์คิดเป็น 50% ของข้อมูลทั้งหมด และมีระเบียนที่ขัดแย้งกับหลักเกณฑ์เป็น 50% เช่นเดียวกัน

จะเห็นได้ชัดว่า ค่าสนับสนุนแบบอ่อนของ $\{2, 5\} \rightarrow \{1\}$ และ $\{1\} \rightarrow \{2, 5\}$ จะมีค่าแตกต่างกัน ซึ่งแสดงให้เห็นว่าค่าสนับสนุนแบบอ่อนเป็นตัววัดที่ไม่สมมาตร

3.2. ทฤษฎีบทของค่าสนับสนุนแบบอ่อน

เนื่องจากค่าสนับสนุนแบบอ่อนเป็นตัววัดที่ไม่พิจารณาระเบียนหรือไอเทมเซตที่ขัดแย้งกับหลักเกณฑ์ และเป็นตัววัดที่ไม่สมมาตร ทำให้สามารถสร้างทฤษฎีบทที่เกี่ยวข้องกับค่าสนับสนุนแบบอ่อน และตัววัดอื่นได้ดังต่อไปนี้ คือ

ทฤษฎีบทที่ 1. ถ้า $weak\ sup(A \rightarrow C) = a$ แล้ว $weak\ sup(\ A \cup B \rightarrow C) \geq a$ เมื่อ $A \cap B \cap C = \emptyset$

<u>พิสูจน์</u> ให้ $weak\ sup(A \rightarrow C\ ) = a$

เนื่องจาก $\left|T_{A \cup B \cup \neg C}\right| \leq \left|T_{A \cup \neg C}\right|$

จะได้ว่า $1 - \frac{\left|T_{A \cup B \cup \neg C}\right|}{|D|} \geq 1 - \frac{\left|T_{A \cup \neg C}\right|}{|D|}$

สรุปได้ว่า $weak\ sup(\ A \cup B \rightarrow C) \geq a$

ทฤษฎีบท 1 แสดงสมบัติของค่าสนับสนุนแบบอ่อน คือ หลักเกณฑ์เชื่อมโยงที่สร้างจากเซตขยายของพจน์หน้าของหลักเกณฑ์เชื่อมโยงเดิม จะมีค่าสนับสนุนแบบอ่อนเพิ่มมากขึ้น แสดงว่าเมื่อหลักเกณฑ์เชื่อมโยงผ่านค่าสนับสนุนแบบอ่อนต่ำสุดที่กำหนดโดยผู้ใช้ จะทำให้หลักเกณฑ์ที่สร้างจากเซตขยายของพจน์หน้าของหลักเกณฑ์ผ่านค่าสนับสนุนแบบอ่อนต่ำสุดด้วย ซึ่งในการสร้างหลักเกณฑ์ควรพิจารณาเฉพาะหลักเกณฑ์ที่มีค่าสนับสนุนไม่เท่ากับ 0 เพราะเมื่อค่าสนับสนุนของหลักเกณฑ์มีค่าเป็น 0 จะหมายถึงไม่มีระเบียนใดที่สอดคล้องกับหลักเกณฑ์ ซึ่งทำให้หลักเกณฑ์นั้นไม่เป็นประโยชน์ต่อการนำไปใช้

ทฤษฎีบทที่ 2. สำหรับทุกหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ ได้ว่า

$weak\ sup(A \rightarrow C) = 1$ ก็ต่อเมื่อ $conf(A \rightarrow C) = 1$

<u>พิสูจน์</u> ให้ $A \rightarrow C$ เป็นหลักเกณฑ์ที่สนใจ

สมมติให้ $weak\ sup(A \rightarrow C)=1$

จะได้ว่า $|T_{A \cup \neg C}|=0$

ซึ่ง $\frac{|T_{A \cup C}|}{|T_A|}=\frac{|T_{A \cup C}|}{|T_{A \cup C}|+|T_{A \cup \neg C}|}=\frac{|T_{A \cup C}|}{|T_{A \cup C}|}=1$

ดังนั้น $conf(A \rightarrow C) = 1$

ในทำนองเดียวกัน ให้ $conf(A \rightarrow C) = 1$

จะได้ว่า $|T_{A \cup \neg C}|=0$

ดังนั้น $weak\ sup(A \rightarrow C)=1$

จากทฤษฎีบทที่ 2 จะได้ว่าค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่นจะมีค่าเป็น 1 พร้อมกัน ซึ่งจากทฤษฎีบทนี้จะเห็นได้ชัดว่าค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่นมีลักษณะสอดคล้องกัน ดังนั้นในกรณีของเหตุการณ์หายากนั่นคือมีค่าสนับสนุนต่ำ แต่มีค่าความเชื่อมั่นเป็น 1 จะมีค่าสนับสนุนแบบอ่อนเป็น 1 ด้วย

ทฤษฎีบทที่ 3. สำหรับทุกหลักเกณฑ์เชื่อมโยง $A \rightarrow C$ ได้ว่า

$$weak\ sup(A \rightarrow C) \geq conf\ (A \rightarrow C)$$

<u>พิสูจน์</u> ให้ $A \rightarrow C$ เป็นหลักเกณฑ์

เนื่องจาก $\frac{1}{|D|} \leq \frac{1}{|T_A|}$

$$\frac{|T_{A \cup \neg C}|}{|D|} \leq \frac{|T_{A \cup \neg C}|}{|T_A|}$$

$$1-\frac{|T_{A \cup \neg C}|}{|D|} \geq 1-\frac{|T_{A \cup \neg C}|}{|T_A|}=\frac{|T_A|-|T_{A \cup \neg C}|}{|T_A|}$$

จะได้ว่า $1-\frac{|T_{A\cup\neg C}|}{|D|}\geq\frac{|T_{A\cup C}|}{|T_A|}$

ดังนั้น $weak\ sup(A\rightarrow C)\geq conf\,(A\rightarrow C)$

จากทฤษฎีบทที่ 3 สรุปได้ว่าหลักเกณฑ์ที่สร้างโดยใช้ค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่น จะมีค่าสนับสนุนแบบอ่อนมากกว่าค่าความเชื่อมั่น ดังนั้นผู้ใช้ต้องกำหนดค่าสนับสนุนแบบอ่อนต่ำสุดให้มีค่ามากกว่าหรือเท่ากับค่าความเชื่อมั่นต่ำสุดเสมอ

3.3. ความสัมพันธ์ระหว่างค่าสนับสนุนแบบอ่อน ค่าสนับสนุน และค่าความเชื่อมั่น

สมมติให้หลักเกณฑ์เชื่อมโยงที่จะพิจารณาคือ $A\rightarrow C$ ค่าความเชื่อมั่นของหลักเกณฑ์ คือ

$$conf\,(A\rightarrow C)=\frac{|T_{A\cup C}|}{|T_A|}=\frac{|T_{A\cup C}|}{|D|}\times\frac{|D|}{|T_A|}$$

ดังนั้น

$$\frac{|T_A|}{|D|}=\frac{|T_{A\cup C}|/|D|}{conf\,(A\rightarrow C)}=\frac{sup\,(A\rightarrow C)}{conf\,(A\rightarrow C)}$$

ค่าสนับสนุนแบบอ่อนของหลักเกณฑ์คือ

$$weak\ sup\,(A\rightarrow C)=\frac{|T_{A\cup C}|+|T_{\neg A\cup C}|+|T_{\neg A\cup\neg C}|}{|D|}$$

$$=\frac{|T_{A\cup C}|+|T_{\neg A}|}{|D|}$$

$$=\frac{|T_{A\cup C}|}{|D|}+(1-\frac{|T_A|}{|D|})$$

เพราะฉะนั้นจึงสามารถเขียนค่าสนับสนุนอ่อนให้อยู่ในรูปของความสัมพันธ์ระหว่างค่าสนับสนุนและค่าความเชื่อมั่นได้ดังสมการด้านล่าง

$$weak\ sup(A \rightarrow C) = sup(A \rightarrow C) + 1 - \frac{sup(A \rightarrow C)}{conf(A \rightarrow C)}$$

ดังนั้นถ้าผู้ใช้ทราบค่าสนับสนุน และค่าความเชื่อมั่น ผู้ใช้สามารถคำนวณหาค่าสนับสนุนแบบอ่อนดังตารางที่ 3.2 และแสดงความสัมพันธ์ที่เกิดขึ้นให้อยู่ในรูปกราฟสามมิติได้ดังรูปที่ 3.5 ซึ่งค่าความเชื่อมั่นที่สนใจจะต้องมีค่ามากกว่าหรือเท่ากับ 0.5 เพราะในกรณีที่มีค่าความเชื่อมั่นน้อยกว่านี้ จะทำให้หลักเกณฑ์ที่พิจารณาไม่น่าสนใจ และไม่เกิดประโยชน์ต่อการนำไปใช้

เมื่อค่าสนับสนุนของหลักเกณฑ์มีค่าเป็น 1 แสดงว่าทุกระเบียนของข้อมูลทั้งหมดจะต้องมีไอเทมเซตที่ประกอบด้วยพจน์หน้า และพจน์หลัง จึงทำให้ค่าความเชื่อมั่นมีค่าเป็น 1 เพราะจำนวนระเบียนที่สอดคล้องกับหลักเกณฑ์กับจำนวนระเบียนที่สอดคล้องกับพจน์หน้าคือจำนวนระเบียนทั้งหมด และจากทฤษฎีบทที่ 2 จะได้ว่าค่าสนับสนุนแบบอ่อนมีค่าเป็น 1 เช่นเดียวกัน ดังนั้นตารางที่ 3.2 และในรูปที่ 3.5 จึงไม่แสดงค่าในกรณีดังกล่าวเพราะค่าของตัววัดทั้งสามมีค่าเป็น 1 และในกรณีที่ค่าความเชื่อมั่นมีค่าเท่ากับค่าสนับสนุน จะได้ว่าค่าสนับสนุนแบบอ่อนที่คำนวณได้จากสูตรเท่ากับค่าความเชื่อมั่น และค่าสนับสนุนด้วย

จากตารางที่ 3.2 การคำนวณค่าสนับสนุนแบบอ่อนจะคำนวณได้จากค่าสนับสนุนและค่าความเชื่อมั่นที่กำหนด แต่เมื่อพิจารณาความสัมพันธ์ระหว่างค่าสนับสนุนกับค่าความเชื่อมั่นของหลักเกณฑ์ $A \rightarrow C$ พบว่า

$$\frac{1}{|D|} \leq \frac{1}{|T_A|}$$

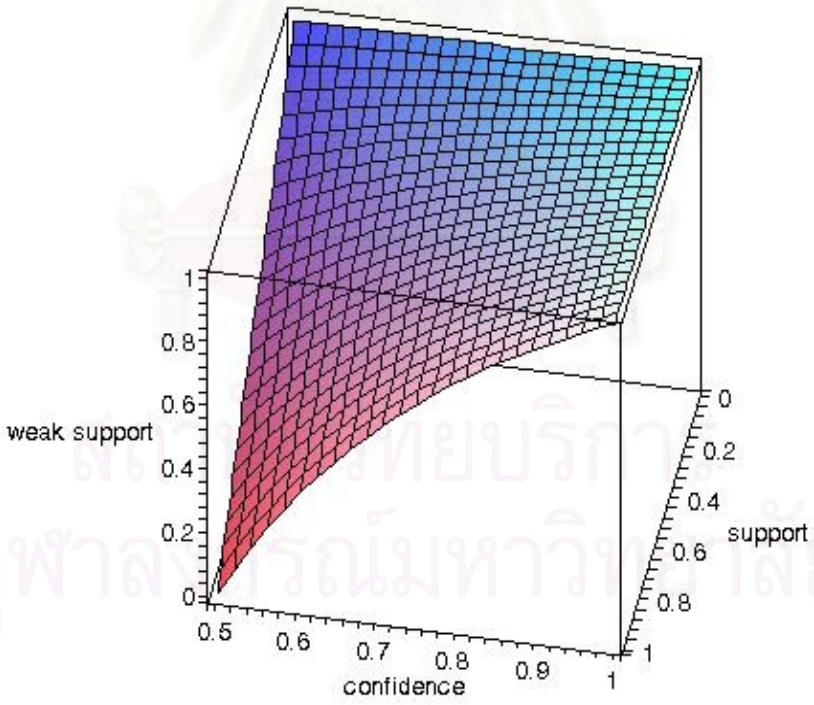
$$\frac{|T_{A \cup C}|}{|D|} \leq \frac{|T_{A \cup C}|}{|T_A|}$$

$$sup(A \rightarrow C) \leq conf(A \rightarrow C)$$

ดังนั้นจึงไม่ควรกำหนดค่าสนับสนุนให้มีค่าน้อยกว่าค่าความเชื่อมั่น ซึ่งส่งผลให้ค่าสนับสนุนแบบอ่อนที่คำนวณได้ไม่สมเหตุสมผล และไม่มีประโยชน์ต่อการนำไปใช้

conf \ *sup*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.9	0.99	0.98	0.97	0.96	0.94	0.93	0.92	0.91	0.90
0.8	0.98	0.95	0.93	0.90	0.88	0.85	0.83	0.80	0.78
0.7	0.96	0.91	0.87	0.83	0.79	0.74	0.70	0.66	0.61
0.6	0.93	0.87	0.80	0.73	0.67	0.60	0.53	0.47	0.40
0.5	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10

ตารางที่ 3.2: ค่าสนับสนุนจากสูตร เมื่อกำหนดค่าสนับสนุน และค่าความเชื่อมั่น



รูปที่ 3.5: กราฟ 3 มิติแสดงความสัมพันธ์ระหว่างค่าสนับสนุนแบบอ่อน ค่าสนับสนุน และค่าความเชื่อมั่น

3.4. ขั้นตอนวิธีค่าสนับสนุนแบบอ่อน

เนื่องจากการสร้างหลักเกณฑ์โดยใช้ค่าสนับสนุนแบบอ่อน จะคำนวณได้จากข้อมูลที่ขัดแย้งกับหลักเกณฑ์ ดังนั้นเพื่อให้สอดคล้องกับค่าสนับสนุนแบบอ่อน ขั้นตอนวิธีที่สร้างขึ้นจึงกำหนดให้พิจารณาเฉพาะหลักเกณฑ์ที่พจน์หลังเป็น 1–ไอเทมเซตเท่านั้น และเรียกขั้นตอนวิธีนี้ว่า ขั้นตอนวิธี WS (weak support algorithm)

กำหนดให้ *minWS* คือ ค่าสนับสนุนแบบอ่อนต่ำสุดที่กำหนดโดยผู้ใช้ *minC* คือ ค่าความเชื่อมั่นต่ำสุดที่กำหนดโดยผู้ใช้ *LHS* คือ เซตของพจน์หน้าของหลักเกณฑ์ที่มีค่าสนับสนุนแบบอ่อนที่คำนวณได้มากกว่าหรือเท่ากับ *minWS* และมีค่าความเชื่อมั่นที่คำนวณได้มากกว่าหรือเท่ากับ *minC* และ *CandidateLHS* คือ เซตของพจน์หน้าตัวเลือก

ขั้นตอนวิธี WS แสดงดังรูปที่ 3.1 และอธิบายได้ดังต่อไปนี้

1. พิจารณาที่แต่ละพจน์หลังของหลักเกณฑ์ $y \in I$ โดยกำหนดให้พจน์หลังที่สนใจต้องเป็น 1–ไอเทมเซต เนื่องจาก *LHS* เป็นเซตของพจน์หน้าของหลักเกณฑ์สำหรับแต่ละพจน์หลังของ *y* ดังนั้นจึงต้องกำหนดให้มีค่าเริ่มต้นเป็นเซตว่าง

2. สำหรับแต่ละพจน์หน้าของหลักเกณฑ์ $x \in I, \{x\} \cap \{y\} = \varnothing$ นำมาสร้างเป็นหลักเกณฑ์เชื่อมโยง $\{x\} \rightarrow \{y\}$ ถ้าหลักเกณฑ์ที่สร้างมีค่าสนับสนุนแบบอ่อนที่คำนวณได้มากกว่าหรือเท่ากับ *minWS* และมีค่าความเชื่อมั่นที่คำนวณได้มากกว่าหรือเท่ากับ *minC* จะกำหนดให้ $LHS = LHS \cup \{x\}$ ในขั้นตอนนี้จะได้ *LHS* เป็นเซตของพจน์หน้าที่เล็กที่สุดที่ทำให้เกิดหลักเกณฑ์ที่ผ่านค่าสนับสนุนแบบอ่อนต่ำสุดและผ่านค่าความเชื่อมั่นต่ำสุด ซึ่งเมื่อนำไปสร้างหลักเกณฑ์ที่มีเซตของพจน์หน้าให้มีขนาดใหญ่ขึ้นจากเซตเดิมที่เก็บใน *LHS* จะผ่านค่าสนับสนุนแบบอ่อนต่ำสุด (จากทฤษฎีบทที่ 1)

3. ในแต่ละรอบของการทำงานที่ *k* จะสร้างเซตของพจน์หน้าตัวเลือก *CandidateLHS* ที่มีขนาด *k* จาก *LHS* แล้วจึงกำหนดให้ *LHS* = {} เพื่อให้ *LHS* เป็นเซตของพจน์หน้าที่มีขนาด *k* ที่สร้างหลักเกณฑ์ที่มีค่า

สนับสนุนแบบอ่อนมากกว่าหรือเท่ากับ *minWS* และมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับ *minC* โดยเก็บกำหนดให้ค่า *k* เริ่มต้นมีค่าเป็น 2

4. นำ *CandidateLHS* ที่สร้างได้จากขั้นตอนที่ 3 มาทดสอบเพื่อสร้าง *LHS* โดยที่ $LHS = LHS \cup \{x\}$ เมื่อ $x \in CandidateLHS$ และหลักเกณฑ์ $\{x\} \rightarrow \{y\}$ เป็นหลักเกณฑ์ที่มีค่าสนับสนุนแบบอ่อนมากกว่าหรือเท่ากับ *minWS* และมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับ *minC* แล้วกำหนดให้ขนาดของพจน์หน้ามีค่าเพิ่มขึ้นเป็น $k = k+1$

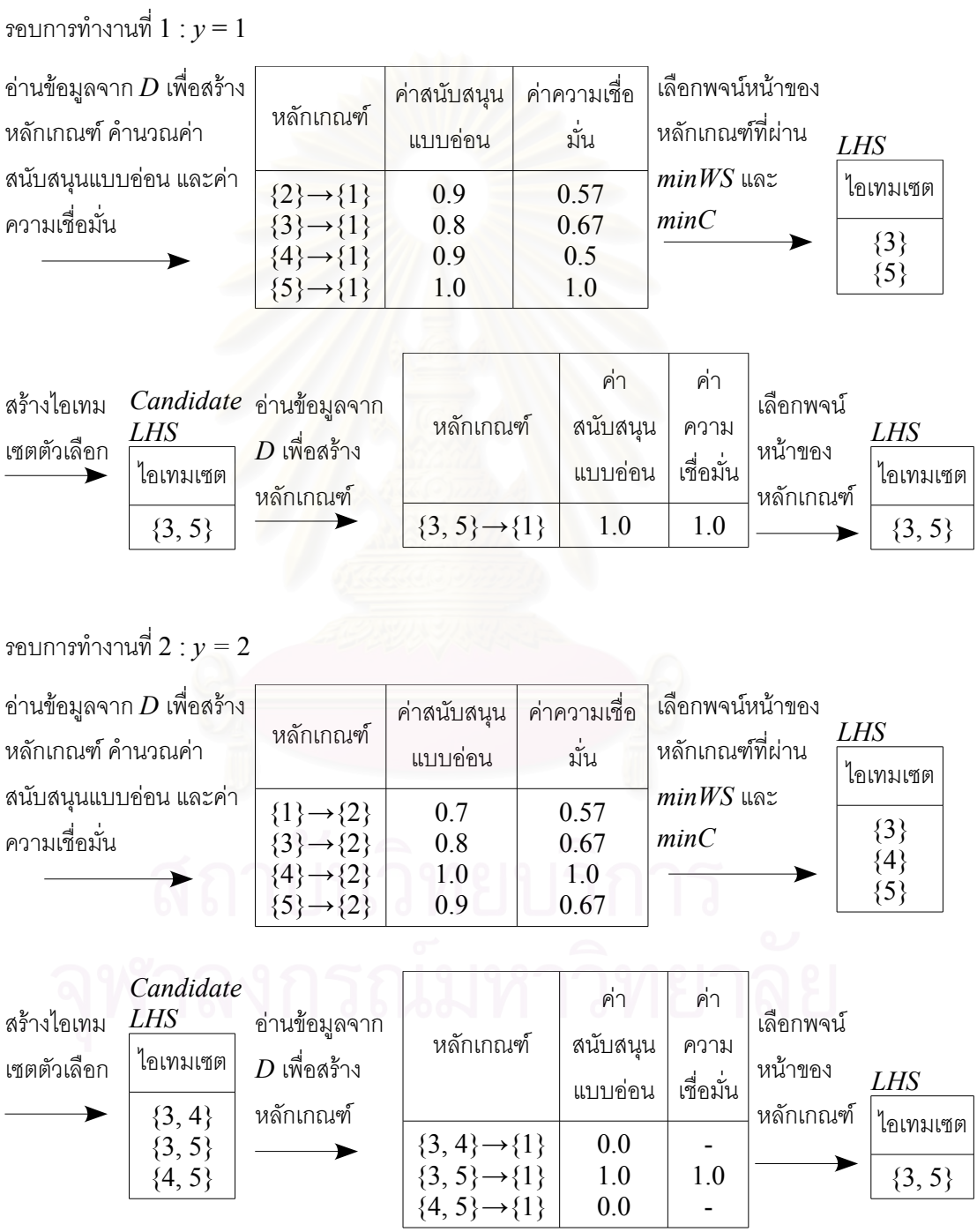
5. ทำซ้ำในขั้นตอนที่ 3 และ 4 ไปจนกระทั่งไม่สามารถสร้างเซตของพจน์หน้าที่สร้างหลักเกณฑ์ที่มีค่าสนับสนุนแบบอ่อนมากกว่าหรือเท่ากับ *minWS* และมีค่าความเชื่อมั่นมากกว่าหรือเท่ากับ *minC*

6. ทำซ้ำในขั้นตอนที่ 1 จนกระทั่งครบทุกไอเทมใน *I*

```
1. for each item y∈ I
2.     LHS = { }
3.     for each item x ∈ I and {x}∩{y}=Ø
4.         if {x}→{y} pass minWS and minC
5.             LHS = LHS ∪ {x}
6.     for(k = 2; LHS ≠ {}; k++)
7.         CandidateLHS = Generate(LHS, k)
8.         LHS = {}
9.         for each x ∈ CandidateLHS
10.            if {x}→{y} pass minWS and minC
11.                LHS = LHS ∪ {x}
```

รูปที่ 3.1: ขั้นตอนวิธี WS

ดังนั้นเมื่อนำข้อมูลตัวอย่างจากตารางที่ 2.3 มาผ่านขั้นตอนวิธี WS จะได้ผลลัพธ์เป็นหลักเกณฑ์เชื่อมโยงดังรูปที่ 3.2 เมื่อกำหนดให้ค่าสนับสนุนแบบอ่อนต่ำสุด (*minWS*) เป็น 0.6 และค่าความเชื่อมั่นต่ำสุด (*minC*) เป็น 0.6



รูปที่ 3.2: ตัวอย่างของการสร้างหลักเกณฑ์ของขั้นตอนวิธี WS

รูปที่ 3.2 เป็นตัวอย่างในการสร้างหลักเกณฑ์จากขั้นตอนวิธี WS เมื่อกำหนดให้ $y = 1$ และ 2 หลักเกณฑ์ที่ได้คือ {3}→{1}, {5}→{1}, {3, 5}→{1}, {3}→{2}, {4}→{2}, {5}→{2} และ {3, 5}→{2} ซึ่งขั้นตอนวิธี WS จะหยุดการทำงานในแต่ละพจน์หลัง เมื่อเซตของพจน์หน้าเป็นเซตว่าง และจบการทำงานเมื่อ y ทำงานครบทุกไอเทมในไอเทมเซต I

ขั้นตอนวิธี WS จะมีระยะเวลาการทำงานดังต่อไปนี้ โดยกำหนดให้

- $|I| = m$ คือ จำนวนไอเทมของเซต I
- $|D|$ คือ จำนวนระเบียน
- $|C_k|$ คือ จำนวนไอเทมเซตใน *CandidateLHS* ที่มีขนาด k
- $|L_k|$ คือ จำนวนไอเทมเซตใน *LHS* ที่มีขนาด k
- t คือ เวลาในการเข้าถึงและดึงข้อมูลออกจากฐานข้อมูล
- *maxLHS* คือ ขนาดของพจน์หน้าที่ใหญ่ที่สุดที่สร้างได้

การเข้าถึงข้อมูลในฐานข้อมูลเพื่อสร้างหลักเกณฑ์จะต้องใช้เวลา $t|D|$ ดังนั้นในแต่ละรอบของการทำงานรอบที่ k การสร้างหลักเกณฑ์จากพจน์หน้าจะใช้เวลาในการทำงานเป็น $O(|C_k| \times t|D|)$

ในบรรทัดที่ 3 – 5 จะใช้เวลาในการทำงานเป็น $O((m-1) \times t|D|)$ เนื่องจากจำนวนของพจน์หน้าของหลักเกณฑ์ที่พิจารณามีจำนวนเท่ากับ $m - 1$

ในบรรทัดที่ 7 – 11 เมื่อพิจารณาที่แต่ละรอบการทำงานที่ k เมื่อ $k \geq 2$ เนื่องจาก *CandidateLHS* ที่มีขนาด k จะสร้างจาก *LHS* ที่มีขนาด $k - 1$ โดยเกิดจากการรวม 2 ไอเทมเซตเข้าด้วยกันแบบมีเงื่อนไข ดังนั้น $|C_k| \approx |L_{k-1}|^2$ ซึ่งจะใช้เวลาในการทำงานแต่ละรอบเป็น $O(|L_{k-1}|^2 \times t|D|)$ ดังนั้นในกรณีที่ $k = 2$ จะใช้เวลาในการทำงานเป็น $O(|L_1|^2 \times t|D|)$ และในกรณีที่ $k = maxLHS$ จะใช้เวลาในการทำงานเป็น $O(|L_{maxLHS-1}|^2 \times t|D|)$

เนื่องจากขั้นตอนวิธี WS จะต้องกำหนดพจน์หลังแล้วจึงสร้างพจน์หน้าของหลักเกณฑ์ และพจน์หลังของหลักเกณฑ์มีจำนวนเท่ากับ m ทำให้เวลาในการทำงานของแต่ละขั้นตอนต้องวนซ้ำ m รอบ จะได้ว่าเวลาในการทำงานทั้งหมด คือ

$$O(\ m(m-1)\times t\ |D\ |\ ) + O(\ m\ |L_{k-1}\ |^2 \times t\ |D\ |) + \ldots + O(\ m\ |L_{maxLHS-1}\ |^2 \times t\ |D\ |))$$

และในกรณีที่พจน์หน้ามีขนาดใหญ่ขึ้นจะมีจำนวนพจน์หน้ามากขึ้น เพราะจากทฤษฎีบท 1 เมื่อเพิ่มขนาดของพจน์หน้าค่าสนับสนุนแบบอ่อนที่คำนวณได้จะมีค่าเพิ่มมากขึ้น ดังนั้นสรุปได้ว่า

$$\begin{aligned} &O(\ m(m-1)\times t\ |D\ |\ ) + \ \ldots + O(\ m\ |L_{maxLHS-1}\ |^2 \times t\ |D\ |\ ) \\ &\leq\ maxLHS \times O(\ m\ |L_{maxLHS-1}\ |^2 \times t\ |D\ |) \end{aligned}$$

เพราะฉะนั้นเวลาในการทำงานของขั้นตอนวิธี WS จะไม่เกิน

$$O(\ m\ |L_{maxLHS-1}\ |^2 \times t\ |D\ |)$$

และเนื่องจาก $t\ |D\ |$ เป็นเวลาที่ใช้ในการเข้าถึงข้อมูลในฐานข้อมูล จึงสามารถสรุปได้ว่าขั้นตอนวิธี WS จะใช้จำนวนครั้งในการเข้าถึงข้อมูลไม่เกิน

$$maxLHS\times\ m\ |L_{maxLHS-1}\ |^2$$

จะเห็นได้ชัดว่าขั้นตอนวิธี WS มีวิธีในการสร้างหลักเกณฑ์ที่แตกต่างกับขั้นตอนวิธีเอไพออรี ทำให้หลักเกณฑ์ที่ได้จากทั้งสองขั้นตอนวิธีแตกต่างกัน ดังนั้นในงานวิจัยนี้จึงต้องการเปรียบเทียบการทำงานระหว่างขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS และเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ที่สร้างขึ้นจากทั้งสองขั้นตอนวิธี ซึ่งการทดลองเพื่อเปรียบเทียบขั้นตอนวิธีทั้งสอง จะอธิบายอย่างละเอียดในบทต่อไป

บทที่ 4

วิธีการทดลอง วิธีการเปรียบเทียบผลการทดลอง และผลการทดลอง

ในบทนี้จะกล่าวถึงวิธีการทดลองและการเปรียบเทียบผลการทดลองระหว่างขั้นตอนวิธีเอไพออรีที่ใช้ค่าสนับสนุนและค่าความเชื่อมั่นในการคัดเลือกหลักเกณฑ์ กับขั้นตอนวิธี WS ที่ใช้ค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่นในการคัดเลือกหลักเกณฑ์ ซึ่งข้อมูลที่ใช้ในการทดลองเป็นข้อมูลมาตรฐานกลางจากฐานข้อมูล UCI โดยจะแสดงผลลัพธ์ที่ได้ในรูปแบบแผนภูมิแท่งแบบ 3 มิติ และตาราง เพื่อเปรียบเทียบประสิทธิภาพของทั้งสองขั้นตอนวิธี

4.1. วิธีการทดลอง

การทดลองเพื่อเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS จะใช้ข้อมูลมาตรฐานกลางที่รู้จักดังตารางที่ 4.1 ได้แก่ ข้อมูล chess และข้อมูล mushroom โดยข้อมูลทั้งสองจะอยู่ในฐานข้อมูลของ UCI [26] และข้อมูลดังกล่าวสามารถดาวน์โหลดได้จาก FIMI [27] ซึ่งข้อมูลทั้งสองจะต้องผ่านขั้นตอนการเตรียมข้อมูลก่อนนำมาทดลอง

ชื่อ	จำนวนระเบียน	จำนวนลักษณะประจำ	จำนวนค่าที่ต่างกัน	ค่าสนับสนุนเฉลี่ย
chess	3196	37	75	0.4933
mushroom	8124	23	119	0.1933

ตารางที่ 4.1: ข้อมูลสรุปบางส่วนของ chess และ mushroom สำหรับการทดลอง

ข้อมูล chess เป็นข้อมูลของการเดินหมากรุกเมื่อฝ่ายหนึ่งมีหมากรุกตัวพระราชาและเรือซึ่งเป็นหมากขาว กับอีกฝ่ายหนึ่งมีหมากรุกตัวพระราชากับเบี้ยซึ่งเป็นหมากดำ โดยตาเดินตาแรกของข้อมูลเป็นของหมากขาว เนื่องจากข้อมูลนี้เป็นข้อมูลของการเดินหมากรุก ดังนั้นข้อมูล chess จึงเป็นข้อมูลที่ค่อนข้างหนาแน่น เพราะมีจำนวนตาเดินที่แน่นอน เมื่อนำมาแปลงให้อยู่ในรูปข้อมูลมาเก็ตบาสเก็ตทำให้มีค่าสนับสนุนเฉลี่ยสูง

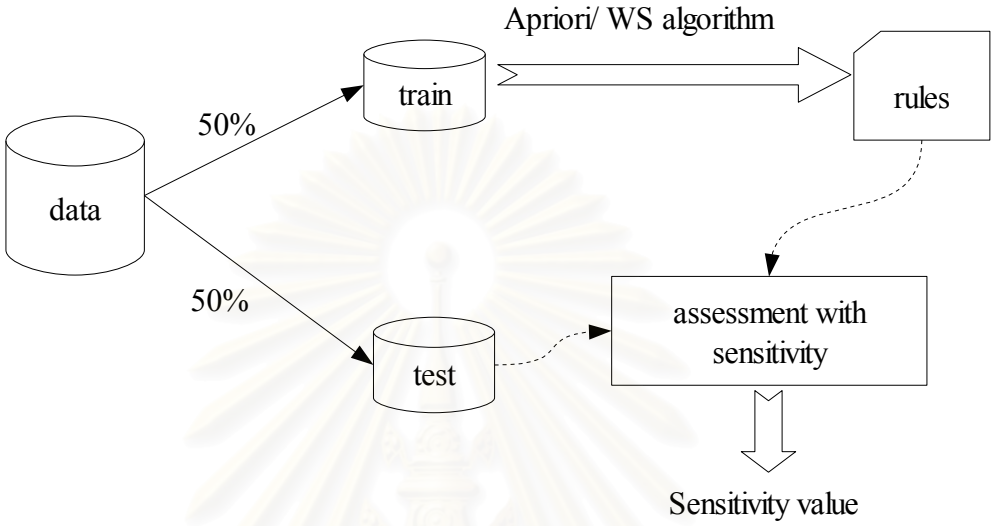
ข้อมูล mushroom เป็นข้อมูลรายละเอียดของเห็ด โดยเก็บข้อมูลเป็นลักษณะรูปร่าง และข้อมูลทั่วไปของเห็ดที่ศึกษา ดังนั้นข้อมูลนี้จึงมีปริมาณความหนาแน่นน้อยกว่าข้อมูล chess ซึ่งเมื่อแปลงข้อมูลให้อยู่ในรูปมาเก็ตบาสเก็ต จะทำให้มีค่าสนับสนุนเฉลี่ยไม่สูงมากนักเมื่อเทียบกับข้อมูล chess

เนื่องจากหลักเกณฑ์เชื่อมโยงที่สร้างจากขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS มีความแตกต่างกัน ทั้งจำนวนหลักเกณฑ์ และรูปแบบของหลักเกณฑ์ที่ได้ ดังนั้นในงานวิจัยนี้จึงต้องการตัววัดที่สามารถวัดประสิทธิภาพของหลักเกณฑ์ที่ไม่ขึ้นกับปริมาณหลักเกณฑ์และรูปแบบ ซึ่งตัววัดที่งานวิจัยนี้ใช้คือ สภาพไว (sensitivity) [28, 29] ซึ่งจะอธิบายเพิ่มเติมในหัวข้อถัดไป

ในการทดลองเพื่อเปรียบเทียบประสิทธิภาพของทั้งสองขั้นตอนวิธีแบ่งเป็นการทดลอง 2 ลักษณะ คือ

1. การทดลองเพื่อพิจารณาว่าขั้นตอนวิธีใดสร้างหลักเกณฑ์ได้ดีกว่า ซึ่งแสดงวิธีการทดลองดังรูปที่ 4.1 โดยแบ่งข้อมูลออกเป็น 2 ส่วน ส่วนละ 50% ซึ่งข้อมูลส่วนแรกจะเรียกว่า ข้อมูลทดลอง (train data) และข้อมูลส่วนที่สองจะเรียกว่าข้อมูลทดสอบ (test data) นำข้อมูลทดลองมาสร้างหลักเกณฑ์เชื่อมโยงจากทั้ง 2 ขั้นตอนวิธี เมื่อได้หลักเกณฑ์จากทั้ง 2 ขั้นตอนวิธีแล้วนำหลักเกณฑ์ของแต่ละขั้นตอนวิธีมาหาค่าสภาพไวจากข้อมูลทดสอบ เพื่อเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ของแต่ละขั้นตอนวิธี ซึ่งค่าสนับสนุนต่ำสุดที่กำหนดจะมีค่าตั้งแต่ 0.1 ไปจนถึง 0.5 โดยเพิ่มค่าครั้งละ 0.1 และค่าความเชื่อมั่นต่ำสุดจะมีค่าตั้งแต่ 0.8 ไปจนถึง 1.0 โดยเพิ่มค่าครั้งละ 0.1 แล้วคำนวณค่าสนับสนุนแบบอ่อนต่ำสุดดังตารางที่ 3.2 โดยนำค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุดไปสร้างหลักเกณฑ์โดยขั้นตอนวิธีเอไพออรี และนำค่าสนับสนุนแบบอ่อน

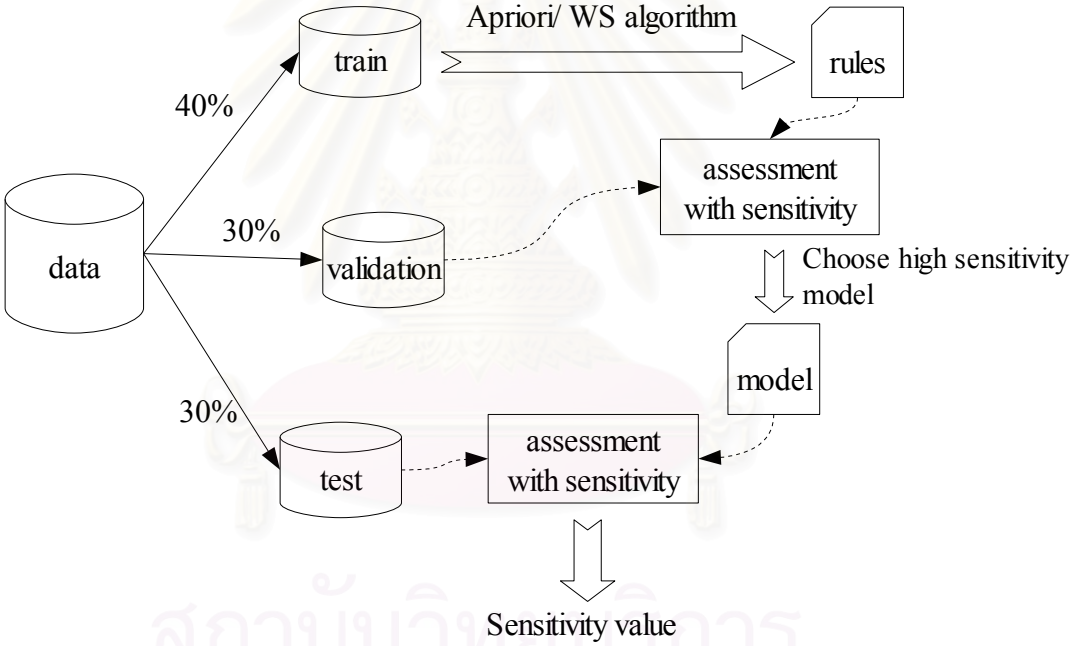
ที่คำนวณได้จากสูตรและค่าความเชื่อมั่นต่ำสุดไปสร้างหลักเกณฑ์จากขั้นตอนวิธี WS



รูปที่ 4.1: การทดลองลักษณะที่ 1

2. การทดลองเพื่อทดสอบประสิทธิภาพของกลุ่มของหลักเกณฑ์ที่มีสภาพไวสูงสุดของแต่ละขั้นตอนวิธี โดยพิจารณาว่ากลุ่มของหลักเกณฑ์ที่มีสภาพไวสูงสุดจากขั้นตอนวิธีใดมีประสิทธิภาพมากกว่า ซึ่งแสดงวิธีการทดลองดังรูปที่ 4.2 การทดลองนี้ต้องสร้างหลักเกณฑ์ให้สอดคล้องกับแต่ละขั้นตอนวิธี และคัดเลือกกลุ่มของหลักเกณฑ์ที่สร้างจากแต่ละขั้นตอนวิธีที่มีค่าสภาพไวสูงมาเปรียบเทียบกัน ดังนั้นการกำหนดค่าต่ำสุดของตัววัดต่างๆ จึงขึ้นกับขั้นตอนวิธีที่เลือกใช้ ด้วยเหตุผลดังกล่าวทำให้การทดลองนี้ต้องแบ่งข้อมูลออกเป็น 3 ส่วนคือ 40% เป็นข้อมูลทดลอง, 30% เป็นข้อมูลตรวจสอบความสมเหตุสมผล (validation data) และ 30% เป็นข้อมูลทดสอบ ขั้นตอนแรกจะนำข้อมูลทดลองมาสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS โดยขั้นตอนวิธีเอไพออรีจะกำหนดค่าสนับสนุนต่ำสุดให้มีค่าตั้งแต่ 0.1 ไปจนถึง 0.5 และกำหนดค่าความเชื่อมั่นต่ำสุดให้มีค่าตั้งแต่ 0.8 ไปจนถึง 1.0 ส่วนขั้นตอนวิธี WS จะกำหนดค่าสนับสนุนแบบอ่อนต่ำสุดให้มีค่าตั้งแต่ 0.6 ไปจนถึง 1.0 และกำหนดค่าความเชื่อมั่นต่ำสุดให้มีค่าเหมือนกับขั้นตอนวิธี

เอไพออรี ซึ่งทุกๆ ตัววัดจะเพิ่มค่าครั้งละ 0.1 ขั้นตอนที่สองจะนำกลุ่มของหลักเกณฑ์ที่สร้างจากขั้นตอนแรกมาหาค่าสภาพไวจากข้อมูลตรวจสอบความสมเหตุสมผล แล้วนำกลุ่มของหลักเกณฑ์ที่ให้ค่าสภาพไวสูงสุดของแต่ละขั้นตอนวิธี (model) มาเปรียบเทียบกับข้อมูลทดสอบในขั้นตอนที่สาม เพื่อเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ที่สร้างจากทั้งสองขั้นตอนวิธีโดยใช้สภาพไวอีกครั้งหนึ่ง ซึ่งถ้ากลุ่มของหลักเกณฑ์ที่ได้เลือกไว้มีค่าสภาพไวมาก แสดงว่าเมื่อนำหลักเกณฑ์กลุ่มนี้ไปใช้งานจริงจะมีประสิทธิภาพในการทำงานมาก



รูปที่ 4.2: การทดลองลักษณะที่ 2

4.2. การเปรียบเทียบผลการทดลอง

พิจารณาการจำแนกประเภทแบบทวิภาค (binary classification) [30] ซึ่งเป็นการจำแนกหรือแบ่งกลุ่มของข้อมูลเป็น 2 กลุ่มโดยพิจารณาการมีสมบัติในการจำแนก เช่น

การจำแนกผู้ป่วยว่าป่วยเป็นโรคหรือไม่ป่วยเป็นโรค ซึ่งคุณสมบัติในการจำแนกคือการป่วยเป็นโรคหรือในการผลิตสินค้าของโรงงานที่ต้องจำแนกสินค้าที่มีคุณภาพกับสินค้าที่ไม่ได้คุณภาพ ซึ่งสมบัติในการจำแนกประเภทคือคุณภาพที่ดีพอที่จะออกสู่ตลาด

ตัวแบบจำแนกประเภทที่ใช้อยู่ในปัจจุบันมีอยู่หลายตัวแบบ ซึ่งตัวแบบจำแนกประเภทแบบทวิภาคที่เป็นที่นิยมได้แก่ ต้นไม้ตัดสินใจ (decision tree) โครงข่ายแบบเบส์ (bayesian network) เครื่องเวกเตอร์สนับสนุน (support vector machine) โครงข่ายประสาท (neural network) เป็นต้น ดังนั้นในการวัดประสิทธิภาพของตัวแบบจำแนกประเภทแบบทวิภาคจะต้องอาศัยตัววัดซึ่งอธิบายได้ดังต่อไปนี้

ตารางที่ 4.2 แสดงเทอมที่ใช้เพื่อวัดประสิทธิภาพของตัวแบบจำแนกประเภทแบบทวิภาคกับเหตุการณ์จริงของข้อมูล

		Fact	
		True	**False**
Model	**Positive**	TP	FP
	Negative	FN	TN

ตารางที่ 4.2: การจำแนกประเภทแบบทวิภาค

- TP (True Positive) คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นบวก และเหตุการณ์จริงเป็นบวก
- FP (False Positive) คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นบวก แต่เหตุการณ์จริงเป็นลบ
- FN (False Negative) คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นลบ แต่เหตุการณ์จริงเป็นบวก
- TN (True Negative) คือ จำนวนของผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นลบ และเหตุการณ์จริงเป็นลบ

สภาพไว หรือ อัตราบวกถูก (true positive rate) เป็นอัตราส่วนระหว่างผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นบวก และเหตุการณ์จริงเป็นบวก หารด้วยผลลัพธ์ที่ได้จากเหตุการณ์จริงที่เป็นบวกทั้งหมด ซึ่งถ้าค่าสภาพไวมีค่าสูงแสดงว่าตัวแบบที่ใช้มีประสิทธิภาพสูงในการทำนายเหตุการณ์ที่เกิดขึ้นแบบบวก ซึ่งเขียนได้เป็น

$$sensitivity = \frac{TP}{TP + FN}$$

สภาพจำเพาะ(specificity) เป็นอัตราส่วนระหว่างผลลัพธ์ที่ได้จากตัวแบบที่ทายว่าเป็นลบ และเหตุการณ์จริงเป็นลบ หารด้วยผลลัพธ์ที่ได้จากเหตุการณ์จริงที่เป็นลบทั้งหมด ซึ่งถ้าค่าสภาพจำเพาะมีค่าสูงแสดงว่าตัวแบบที่ใช้มีประสิทธิภาพสูงในการทำนายเหตุการณ์ที่เกิดขึ้นแบบลบ แสดงได้ดังสมการด้านล่าง

$$specificity = \frac{TN}{FP + TN}$$

แต่เนื่องจากข้อมูลที่ใช้วิเคราะห์หลักเกณฑ์เชื่อมโยงเป็นข้อมูลที่เกิดขึ้นจริงเท่านั้น เมื่อพิจารณาเปรียบเทียบกับการทดสอบตัวแบบจำแนกประเภทแบบทวิภาคแล้วพบว่า เหตุการณ์จริงเป็นบวกคือการปรากฏการซื้อ แต่เหตุการณ์จริงเป็นลบไม่มี (การไม่ซื้อปกติจะไม่มีการบันทึก) จึงทำให้การทดสอบประสิทธิภาพของหลักเกณฑ์ใช้ตัววัดได้เพียงตัวเดียว คือสภาพไวเท่านั้น ซึ่งเทอมที่ใช้แสดงการทดสอบประสิทธิภาพของหลักเกณฑ์เมื่อเปรียบเทียบกับข้อมูลจริงแสดงดังตารางที่ 4.3

		Database
		True
Model	**Satisfying rule (P)**	TP
	Contradiction rule (N)	FN

ตารางที่ 4.3: การจำแนกประเภทแบบทวิภาคของหลักเกณฑ์เชื่อมโยง

- TP (True Positive) คือ จำนวนของหลักเกณฑ์สอดคล้อง (satisfying rule)
- FN (False Negative) คือ จำนวนของหลักเกณฑ์ขัดแย้ง (contradiction rule)

หลักเกณฑ์จะเป็นหลักเกณฑ์สอดคล้อง เมื่อนำหลักเกณฑ์ไปเปรียบเทียบกับข้อมูลแล้วพบว่าหลักเกณฑ์ต้องสอดคล้องกับข้อมูลทั้งพจน์หน้าและพจน์หลัง หรือข้อมูลนั้นเกิดการซื้อทั้งพจน์หน้าและพจน์หลัง ตัวอย่างเช่น หลักเกณฑ์ {5}→{2} จะเป็นหลักเกณฑ์ที่สอดคล้อง เมื่อลักษณะประจำหรือไอเทมที่ 5 และ 2 มีค่าเป็น 1 หรือระเบียนนั้นเกิดการซื้อ 5 และซื้อ 2

หลักเกณฑ์ขัดแย้งจะเป็นหลักเกณฑ์ที่สอดคล้องกับข้อมูลเฉพาะพจน์หน้าแต่ไม่สอดคล้องกับพจน์หลัง หรือ ข้อมูลนั้นเกิดการซื้อพจน์หน้าแต่ไม่ซื้อพจน์หลัง ตัวอย่างเช่น หลักเกณฑ์ {5}→{2} จะเป็นหลักเกณฑ์ขัดแย้ง เมื่อไอเทมที่ 5 มีค่าเป็น 1 แต่ไอเทมที่ 2 มีค่าเป็น 0 หรือระเบียนนั้นเกิดการซื้อ 5 แต่ไม่ซื้อ 2

ดังนั้นการทดสอบประสิทธิภาพของหลักเกณฑ์โดยใช้สภาพไว จะต้องอาศัยข้อมูลจริง โดยพิจารณาที่ข้อมูลแต่ละระเบียนหรือแต่ละทรานแซคชัน แล้วนำหลักเกณฑ์ที่สร้างจากขั้นตอนวิธีที่เลือกมาหาค่าสภาพไว เมื่อคำนวณหาค่าสภาพไวของข้อมูลแต่ละระเบียนแล้วนำค่าสภาพไวที่คำนวณได้ทั้งหมดมาหาค่าเฉลี่ย จะได้เป็นสภาพไวเฉลี่ยของกลุ่มของหลักเกณฑ์ที่สร้างขึ้นจากขั้นตอนวิธีที่เลือกไว้ ซึ่งในงานวิจัยนี้จะใช้ค่าสภาพไวเฉลี่ยในการเปรียบเทียบประสิทธิภาพของแต่ละขั้นตอนวิธี

ถ้ากลุ่มของหลักเกณฑ์ที่ได้จากขั้นตอนวิธีที่เลือกไว้ที่มีค่าสภาพไวเฉลี่ยสูง แสดงว่ากลุ่มของหลักเกณฑ์ดังกล่าวมีแนวโน้มในการทำนายเหตุการณ์ที่เกิดขึ้นในอนาคตได้มากกว่าและมีประสิทธิภาพในการนำหลักเกณฑ์ที่ได้ไปใช้งานได้จริง

ตัวอย่างที่ 4.1 การหาสภาพไวของกลุ่มของหลักเกณฑ์ โดยสมมติให้หลักเกณฑ์ที่ได้จากข้อมูลทดลอง คือ {2}→{1}, {3}→{1}, {2}→{3} และ {2, 3}→{1} และกำหนดให้ข้อมูลทดสอบเป็นดังนี้คือ

TID	1	2	3
105	0	1	1
255	1	1	0
400	1	1	1

ตารางที่ 4.4: ตัวอย่างของข้อมูลทดสอบ

การหาค่าสภาพไว จะต้องพิจารณาข้อมูลแต่ละระเบียน สมมติให้พิจารณาระเบียนที่ 105 จะพบว่า

- หลักเกณฑ์สอดคล้องคือ {2}→{3} เพราะเมื่อนำหลักเกณฑ์นี้ไปเปรียบเทียบกับระเบียน พบว่าในลักษณะประจำที่ 2 และ 3 มีค่าเป็น 1 หรือระเบียนนี้ซื้อ 2 และ 3
- หลักเกณฑ์ขัดแย้งคือ {2}→{1}, {3}→{1} และ {2, 3}→{1} เพราะเมื่อนำหลักเกณฑ์ไปเปรียบเทียบกับระเบียนพบว่า เกิดการซื้อพจน์หน้าของระเบียน แต่ไม่เกิดการซื้อพจน์หลังของระเบียน

ดังนั้นค่าสภาพไวของระเบียนที่ 105 คือ $\frac{1}{1+3}=\frac{1}{4}=0.25$ ในทำนองเดียวกันระเบียนที่ 255 จะมีหลักเกณฑ์สอดคล้อง คือ {2}→{1} และมีหลักเกณฑ์ขัดแย้งคือ {2}→{3} จึงมีค่าสภาพไวเป็น $\frac{1}{1+1}=\frac{1}{2}=0.5$ และที่ระเบียน 400 จะมีหลักเกณฑ์สอดคล้อง คือ {2}→{1}, {3}→{1}, {2}→{3} และ {2, 3}→{1} และไม่มีหลักเกณฑ์ขัดแย้ง ดังนั้นค่าสภาพไวของระเบียนที่ 400 คือ 1 เพราะฉะนั้นค่าสภาพไวเฉลี่ยของข้อมูลทดสอบดังตารางที่ 4.4 คือ $\frac{0.25+0.5+1}{3}=0.583$

4.3. ผลการทดลอง

ในหัวข้อนี้จะแสดงผลการทดลองที่ได้จากการทดลองทั้ง 2 ลักษณะ และทดสอบประสิทธิภาพของหลักเกณฑ์ที่ได้จากขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS โดยใช้สภาพไว เพื่อให้การเปรียบเทียบหลักเกณฑ์ระหว่างสองขั้นตอนวิธีมีลักษณะของหลักเกณฑ์เป็นแบบเดียวกัน จึงกำหนดให้หลักเกณฑ์ที่เป็นผลลัพธ์จากทั้งสองขั้นตอนวิธีมีจำนวนสมาชิกของพจน์หลังเพียงหนึ่งเดียว และกำหนดให้พจน์หน้าของหลักเกณฑ์มีขนาดที่ใหญ่ที่สุดคือ 1, 2 และ 3 ตามลำดับ โดยเรียกขนาดของพจน์หน้าของหลักเกณฑ์ที่ใหญ่ที่สุดที่สร้างได้เป็น *maxLHS* ดังนั้นขั้นตอนวิธีเอไพออรีขนาดของไอเทมเซตปรากฏบ่อยที่ใหญ่ที่สุดจะกำหนดให้มีขนาดเป็น *maxLHS*+1 ซึ่งมีค่าเท่ากับ 2, 3 และ 4 ตามลำดับ สำหรับขั้นตอนวิธี WS จะกำหนดให้ขนาดของพจน์หน้าเป็น *maxLHS* โดยแสดงผลการทดลองในรูปแบบของแผนภูมิแท่งแบบ 3 มิติ ซึ่งผลการทดลองในรูปแบบตารางจะแสดงในภาคผนวก ค

ผลการทดลองที่ได้เป็นผลการทดลองจากโปรแกรมที่เขียนขึ้นจากภาษาจาวา และใช้ซอฟต์แวร์จัดการฐานข้อมูล MySQL 5.0 ในการจัดเก็บข้อมูลทดลอง ซึ่งทดลองบนเครื่องคอมพิวเตอร์ Pentium 4 2.0 GHz และมีหน่วยความจำหลัก 1 GB

4.3.1. ผลการทดลองลักษณะที่ 1

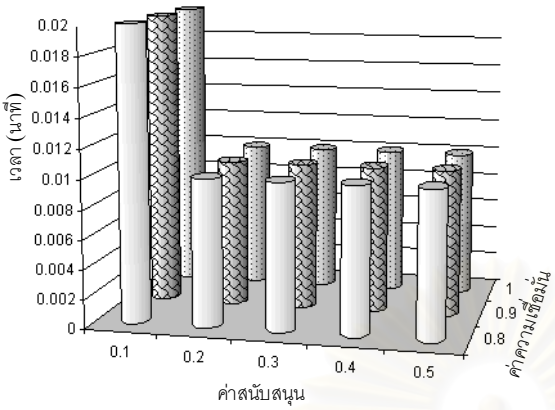
การทดลองนี้จะทดลองเพื่อเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ที่สร้างขึ้นจากข้อมูลทดลองของขั้นตอนวิธี WS และขั้นตอนวิธีเอไพออรี โดยกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด แล้วคำนวณค่าสนับสนุนแบบอ่อนต่ำสุดดังตารางที่ 2.4 กำหนดให้ค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 0.8, 0.9 และ 1.0 โดยแต่ละค่าจะกำหนดให้สร้างหลักเกณฑ์จากค่าสนับสนุนต่ำสุดเป็น 0.1 ไปจนถึง 0.5 โดยเพิ่มค่าครั้งละ 0.1 ซึ่งขั้นตอนวิธี WS จะใช้ค่าสนับสนุนแบบอ่อนต่ำสุดและค่าความเชื่อมั่นต่ำสุดในการสร้างหลักเกณฑ์ ส่วนขั้นตอนวิธีเอไพออรีจะใช้ค่าสนับสนุนต่ำสุด และค่าความเชื่อมั่นต่ำสุด

เมื่อนำข้อมูลส่วนแรก ซึ่งเป็นข้อมูลทดลองของข้อมูล chess และข้อมูล mushroom มาสร้างหลักเกณฑ์โดยใช้ขั้นตอนวิธีเอไพออรี และขั้นตอนวิธี WS จะได้ผลการทดลองดังรูปที่ 4.3 ถึงรูปที่ 4.8 โดยจะแบ่งกลุ่มตามเวลา จำนวนหลักเกณฑ์ และสภาพไว เนื่องจากในการทดลองลักษณะที่ 1 จะกำหนดค่าสนับสนุนต่ำสุด และค่าความเชื่อมั่นต่ำสุด แล้ว

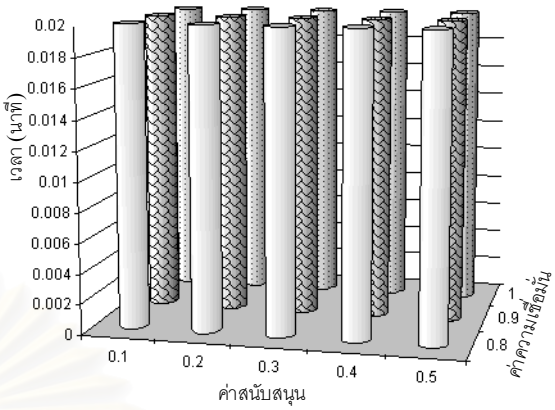
คำนวณค่าสนับสนุนแบบอ่อนจากสูตร ดังนั้นแผนภูมิแท่งแบบ 3 มิติจึงแสดงแกนของค่าสนับสนุนต่ำสุด ค่าความเชื่อมั่นต่ำสุด และผลการทดลองที่ต้องการเปรียบเทียบ ซึ่งเมื่อพิจารณาจากแผนภูมิเมื่อแกนของค่าสนับสนุนต่ำสุดมีค่าเป็น 0.3 และแกนค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 0.8 แสดงว่าจะมีค่าสนับสนุนแบบอ่อนที่คำนวณได้เป็น 0.93 ซึ่งขั้นตอนวิธีเอไพออรีจะสร้างหลักเกณฑ์โดยใช้ค่าสนับสนุนต่ำสุดเป็น 0.3 และค่าความเชื่อมั่นต่ำสุดเป็น 0.8 ส่วนขั้นตอนวิธี WS จะสร้างหลักเกณฑ์โดยใช้ค่าสนับสนุนแบบอ่อนต่ำสุดเป็น 0.93 และค่าความเชื่อมั่นต่ำสุดเป็น 0.8

จากรูปที่ 4.3 และรูปที่ 4.4 พบว่าเวลาในการทำงานของขั้นตอนวิธีเอไพออรีจะขึ้นกับการกำหนดค่าสนับสนุนต่ำสุด เนื่องจากขั้นตอนวิธีเอไพออรีจะสร้างไอเทมเซตปรากฏบ่อยจากค่าสนับสนุนต่ำสุดที่กำหนด ดังนั้นเมื่อกำหนดให้ค่าสนับสนุนต่ำสุดมีค่ามากจะทำให้ไอเทมเซตปรากฏบ่อยมีจำนวนน้อย ซึ่งส่งผลให้การทำงานรวดเร็วมากยิ่งขึ้น ส่วนขั้นตอนวิธี WS จากรูปพบว่าเมื่อกำหนดให้ค่าสนับสนุนต่ำสุดมีค่าน้อย จะทำให้ขั้นตอนวิธี WS ทำงานเร็วกว่าการกำหนดให้ค่าสนับสนุนต่ำสุดมีค่ามาก เนื่องจากเมื่อกำหนดให้ค่าสนับสนุนต่ำสุดมีค่าน้อย ค่าสนับสนุนแบบอ่อนต่ำสุดที่คำนวณได้จะมีค่ามาก นอกจากนี้ยังพบว่าหลักเกณฑ์ที่สร้างขึ้นจากขั้นตอนวิธี WS มีจำนวนมากกว่าขั้นตอนวิธีเอไพออรีดังรูปที่ 4.5 และรูปที่ 4.6 ทำให้ใช้เวลาในการทำงานนานกว่า

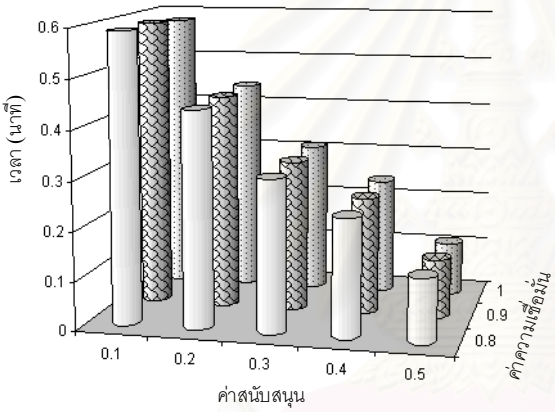
เมื่อนำค่าสภาพไวมาสร้างเป็นแผนภูมิแท่งแบบ 3 มิติ ดังรูปที่ 4.7 และรูปที่ 4.8 พบว่าเมื่อกำหนดขนาดของพจน์หน้าสูงสุดที่สร้างได้ ขั้นตอนวิธี WS จะมีค่าสภาพไวค่อนข้างคงที่และจะมีค่าเพิ่มขึ้นเมื่อค่าสนับสนุนแบบอ่อนต่ำสุด และค่าความเชื่อมั่นต่ำสุดมีค่าสูง ซึ่งค่าสภาพไวของขั้นตอนวิธี WS จะมีค่าสูงสุดเมื่อกำหนดให้ค่าความเชื่อมั่นต่ำสุดเป็น 1 ซึ่งทำให้ได้ว่าค่าสนับสนุนแบบอ่อนต่ำสุดจะมีค่าเป็น 1 ตามทฤษฎีบทที่ 2 แต่ขั้นตอนวิธีเอไพออรีจะมีค่าสภาพไวเปลี่ยนแปลงไปโดยขึ้นอยู่กับค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุด แต่เมื่อค่าสนับสนุนต่ำสุดมีค่าเพิ่มมากขึ้น จะไม่สามารถสรุปได้ว่ามีค่าสภาพไวเพิ่มมากขึ้น โดยจะเห็นความแตกต่างของสภาพไวของขั้นตอนวิธีเอไพออรีได้ชัดในกรณีของข้อมูล mushroom ดังรูปที่ 4.8 ซึ่งมีปริมาณข้อมูลมาก และมีความแตกต่างของข้อมูลสูง เมื่อกำหนดให้ค่าสนับสนุนต่ำสุดมีค่าเป็น 0.4 จะมีค่าสภาพไวต่ำกว่ากรณีที่กำหนดค่าสนับสนุนต่ำสุดมีค่าเป็น 0.3 แต่จากการทดลองเมื่อกำหนดให้ค่าสนับสนุนต่ำสุดมีค่าเป็น 0.5 และกำหนดให้ค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 1 ขั้นตอนวิธีเอไพออรีจะมีค่าสภาพไวสูงสุด และเมื่อกำหนดขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์ให้ค่าเพิ่มมากขึ้น ขั้นตอนวิธี WS จะมีค่าสภาพไวเพิ่มมากขึ้น ในขณะที่ขั้นตอนวิธีเอไพออรีจะมีค่าสภาพไวลดลง ถึงแม้ว่าจำนวนหลักเกณฑ์ที่ได้จากขั้นตอนวิธี WS จะมีจำนวนหลักเกณฑ์มากกว่า



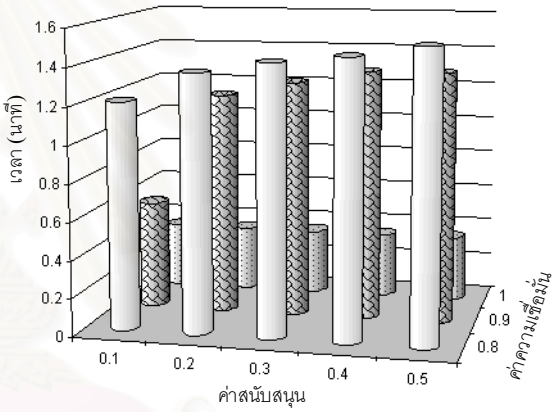
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



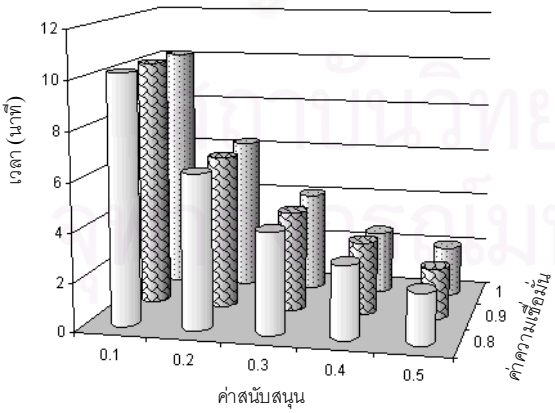
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



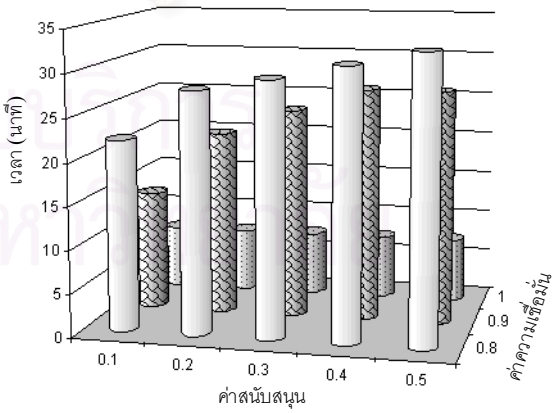
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2

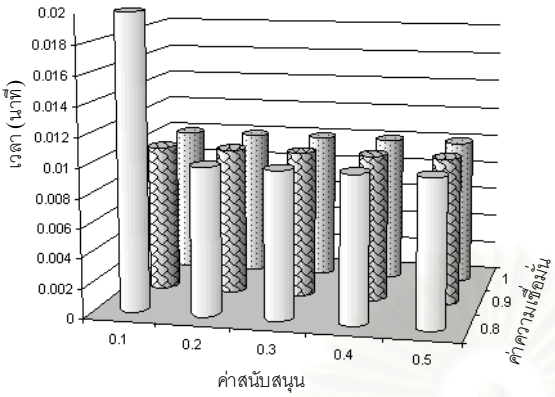


(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3

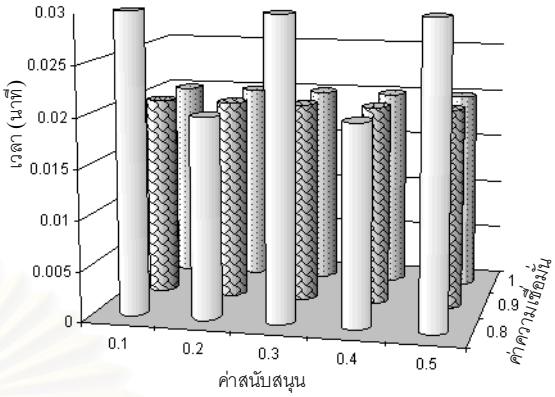


(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

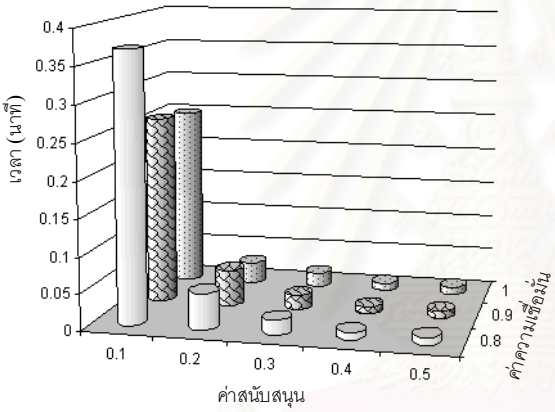
รูปที่ 4.3: (ก) – (ฉ) แสดงเวลาที่ใช้ในการทดลองลักษณะที่ 1 บนข้อมูล chess



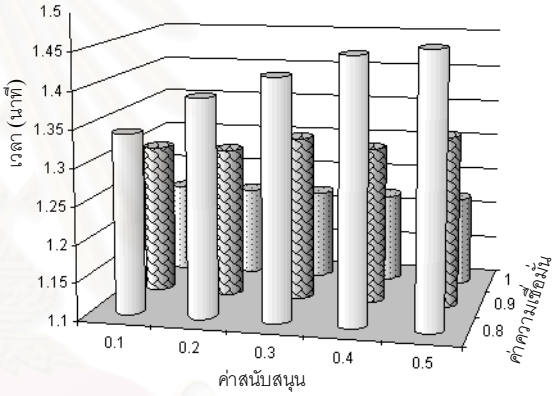
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



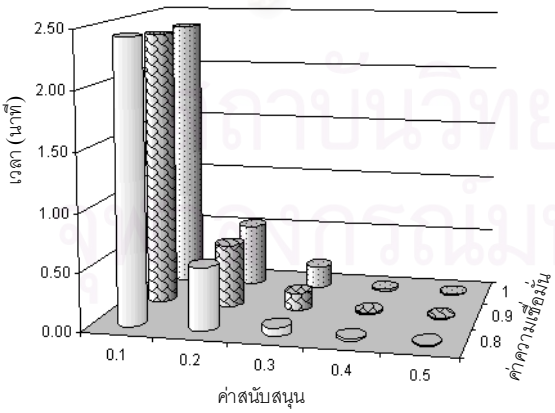
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



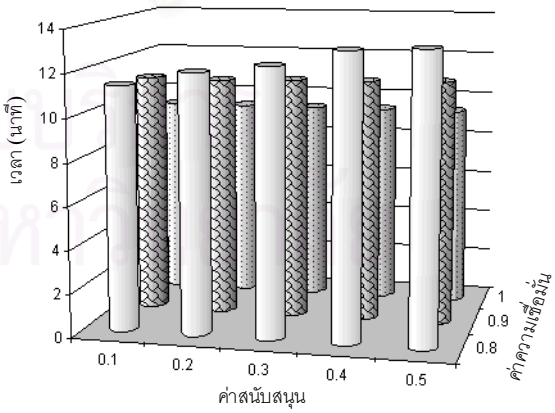
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2

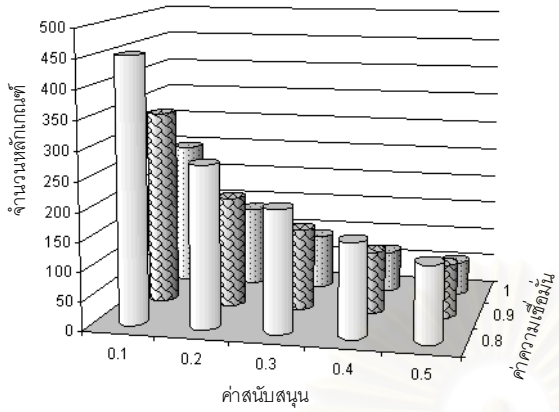


(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3

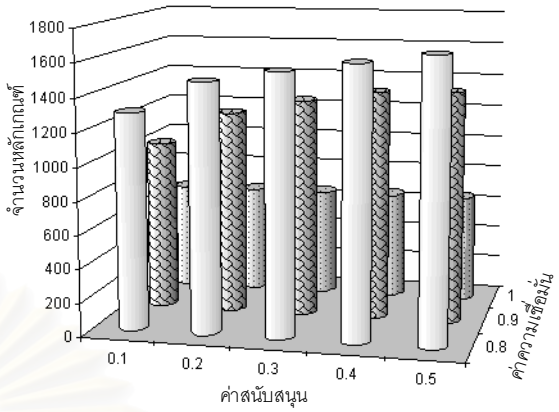


(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

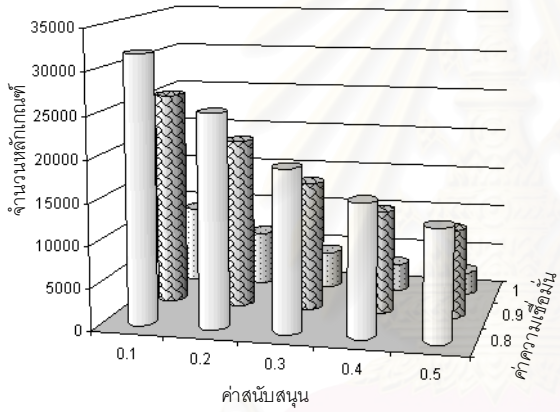
รูปที่ 4.4: (ก) – (ฉ) แสดงเวลาที่ใช้ในการทดลองลักษณะที่ 1 บนข้อมูล mushroom



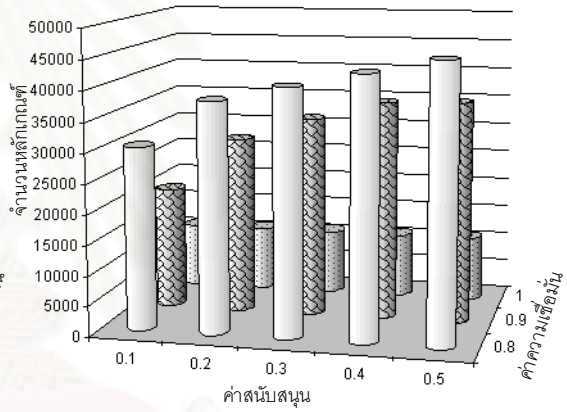
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



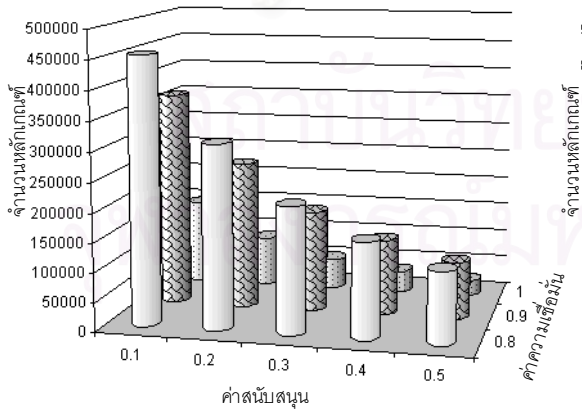
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



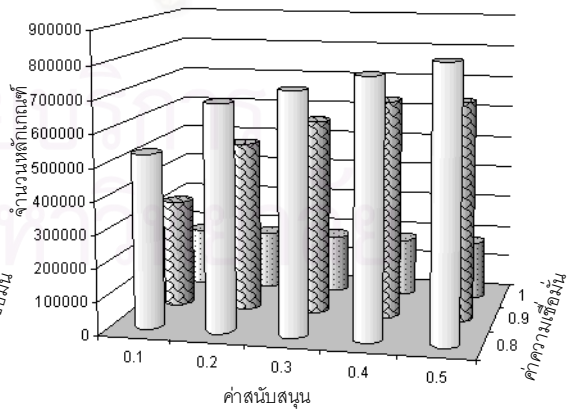
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2

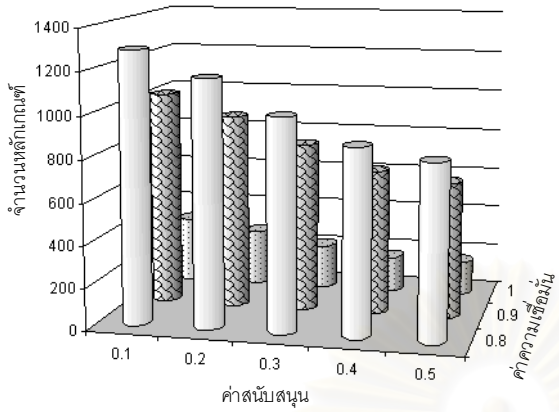


(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3

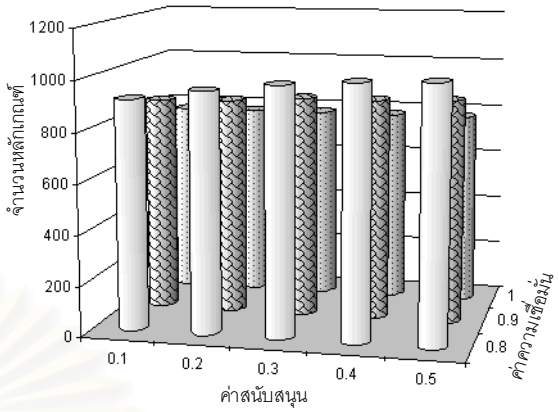


(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

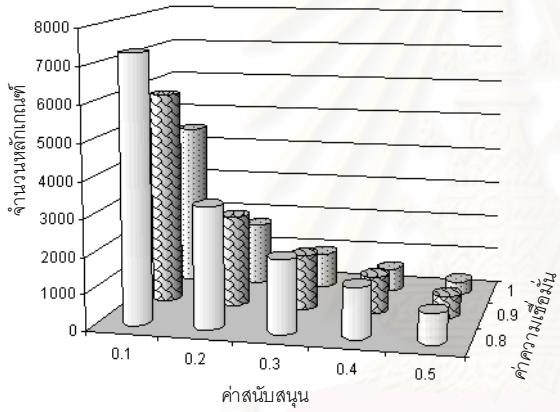
รูปที่ 4.5: (ก) – (ฉ) แสดงจำนวนหลักเกณฑ์ที่ได้จากการทดลองลักษณะที่ 1 บนข้อมูล chess



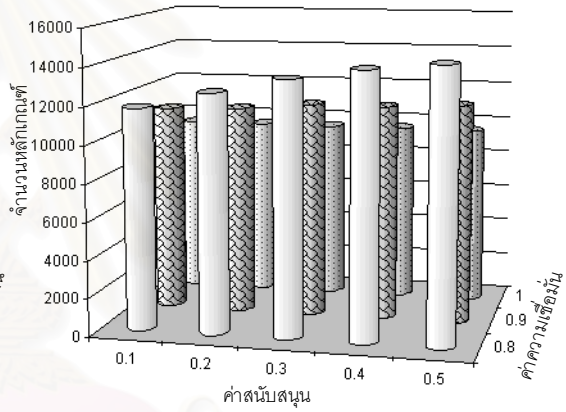
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



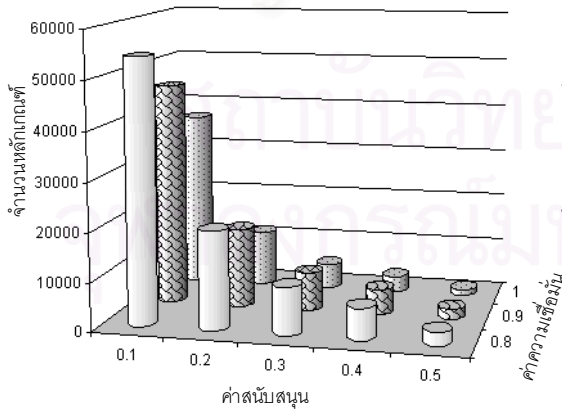
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



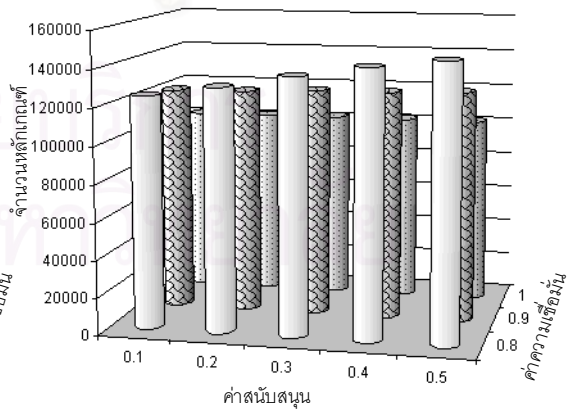
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2

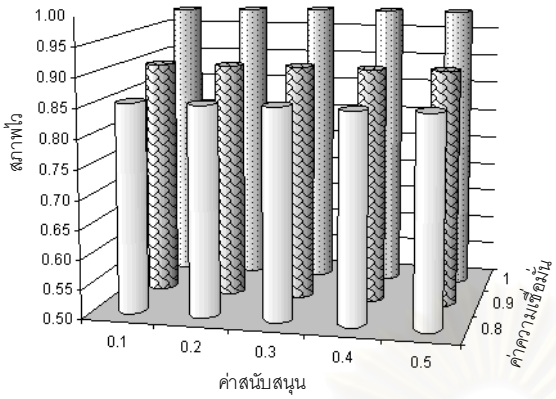


(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3

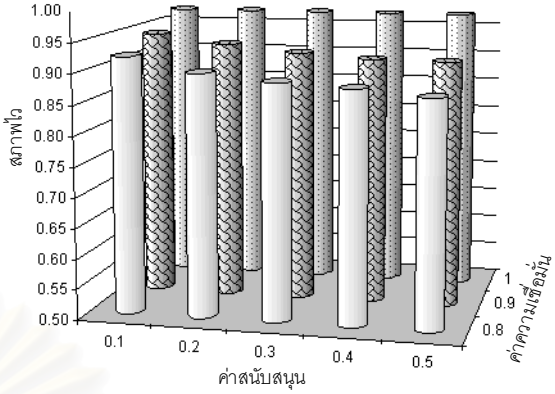


(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

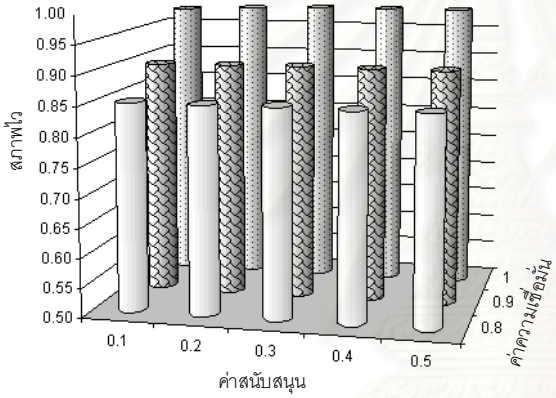
รูปที่ 4.6: (ก) – (ฉ) แสดงจำนวนหลักเกณฑ์ที่ได้จากการทดลองลักษณะที่ 1 บนข้อมูล mushroom



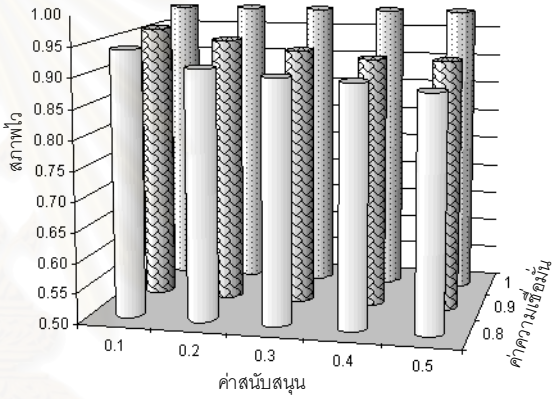
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



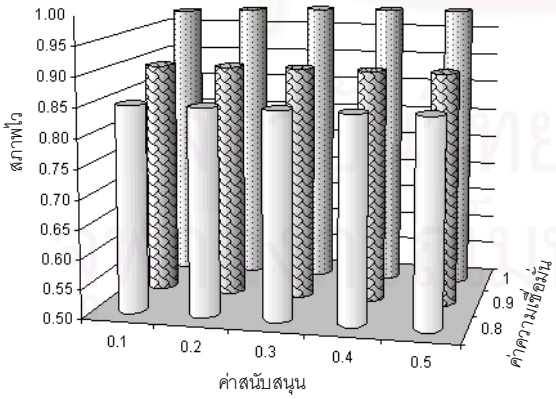
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



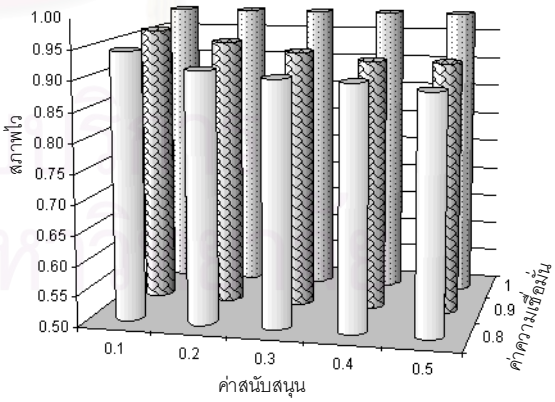
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2

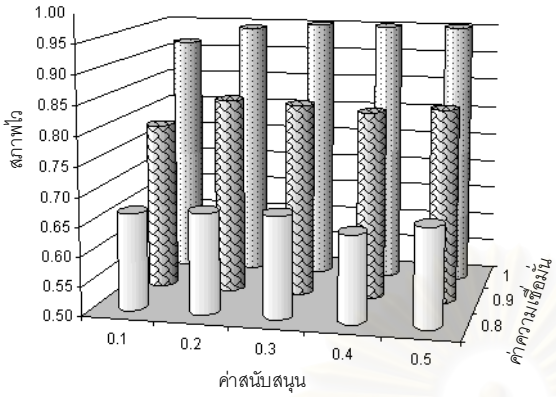


(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3

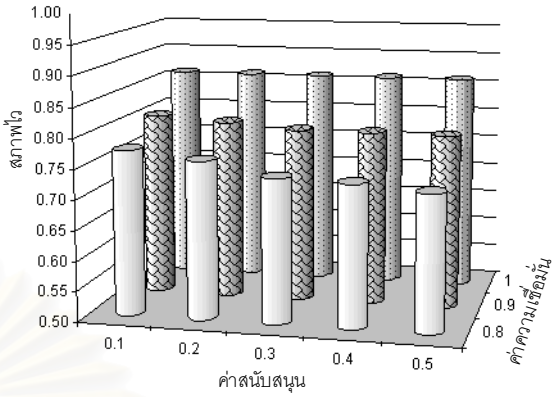


(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

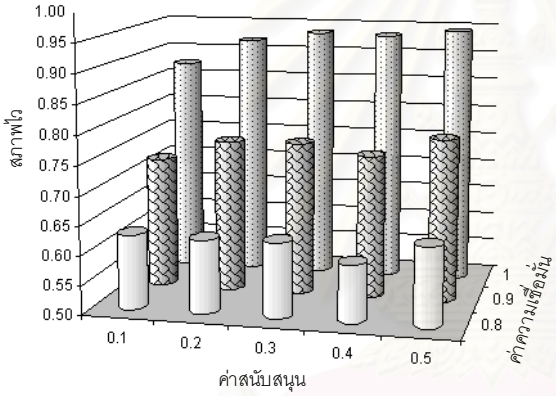
รูปที่ 4.7: (ก) – (ฉ) แสดงสภาพไวเฉลี่ยจากการทดลองลักษณะที่ 1 บนข้อมูล chess



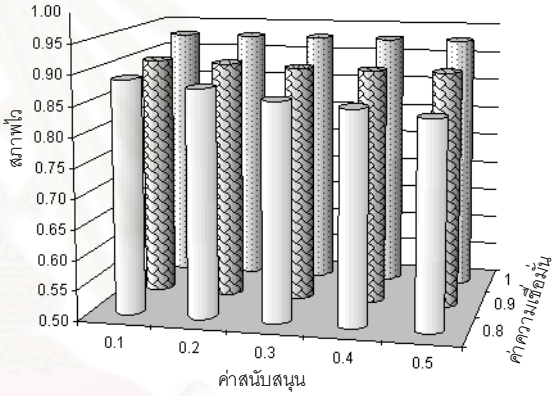
(ก) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 1



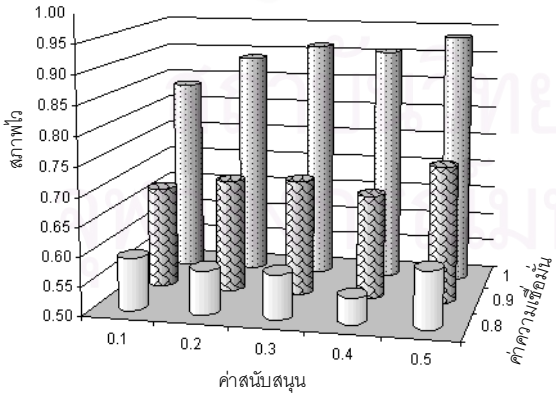
(ข) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 1



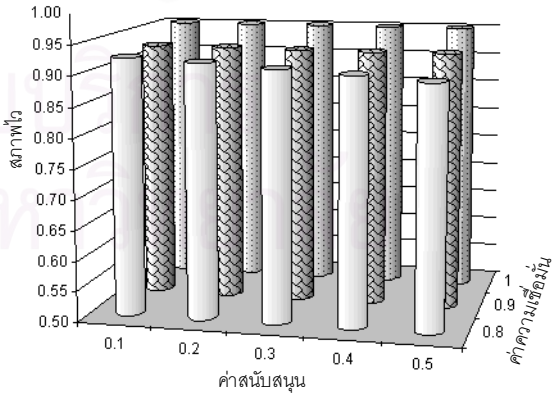
(ค) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 2



(ง) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 2



(จ) ขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* = 3



(ฉ) ขั้นตอนวิธี WS เมื่อ *maxLHS* = 3

รูปที่ 4.8: (ก) – (ฉ) แสดงสภาพไวเฉลี่ยจากการทดลองลักษณะที่ 1 บนข้อมูล mushroom

ดังนั้นการวิเคราะห์หลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธี WS จะสร้างหลักเกณฑ์ที่มีประสิทธิภาพในการทำงานใกล้เคียงกันเมื่อกำหนดขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์ และเมื่อขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์เพิ่มขึ้น สภาพไวจะมีค่าเพิ่มมากขึ้น โดยจากการทดลองพบว่าเมื่อค่าสนับสนุนแบบอ่อนต่ำสุดเป็น 1 หรือค่าความเชื่อมั่นต่ำสุดเป็น 1 หลักเกณฑ์ที่สร้างจะมีค่าสภาพไวสูงสุดสำหรับแต่ละขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์ที่กำหนดไว้ แต่ประสิทธิภาพของหลักเกณฑ์ที่สร้างจากขั้นตอนวิธีเอไพออรีจะขึ้นกับค่าสนับสนุนต่ำสุด ค่าความเชื่อมั่นต่ำสุด และเมื่อขนาดของพจน์หน้าเพิ่มมากขึ้น ค่าสภาพไวจะมีค่าลดลง ดังนั้นขั้นตอนวิธีเอไพออรีจะต้องกำหนดให้ค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุดให้สอดคล้องกับข้อมูล และขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์น้อย ถึงจะสร้างหลักเกณฑ์ที่มีประสิทธิภาพในการทำงานมาก ซึ่งจากการทดลองพบว่าขั้นตอนวิธีเอไพออรีเมื่อกำหนดให้ค่าสนับสนุนต่ำสุดเป็น 0.5 และค่าความเชื่อมั่นต่ำสุดเป็น 1 หลักเกณฑ์ที่สร้างจากขั้นตอนวิธีเอไพออรีจะมีค่าสภาพไวสูงสุดสำหรับแต่ละขนาดของพจน์หน้าที่กำหนดไว้

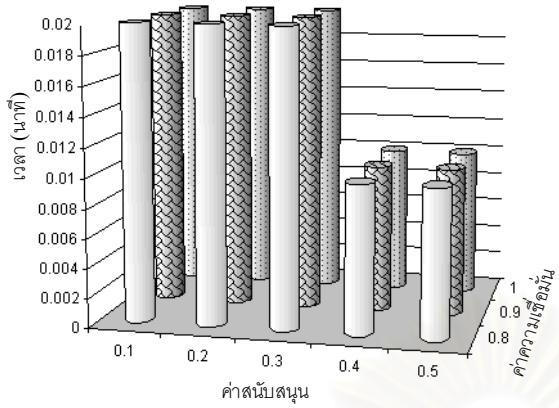
4.3.2. ผลการทดลองลักษณะที่ 2

การทดลองนี้ต้องการเปรียบเทียบประสิทธิภาพของกลุ่มหลักเกณฑ์ที่มีค่าสภาพไวสูงสุดของแต่ละขั้นตอนวิธี โดยต้องสร้างหลักเกณฑ์ที่สอดคล้องสำหรับแต่ละขั้นตอนวิธี ดังนั้นสำหรับขั้นตอนวิธีเอไพออรีจะกำหนดค่าสนับสนุนต่ำสุดให้มีค่าตั้งแต่ 0.1 ไปจนถึง 0.5 ส่วนขั้นตอนวิธี WS กำหนดให้ค่าสนับสนุนแบบอ่อนต่ำสุดมีค่าตั้งแต่ 0.6 ถึง 1.0 และกำหนดให้ค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 0.8, 0.9 และ 1.0 สำหรับทั้งสองขั้นตอนวิธี

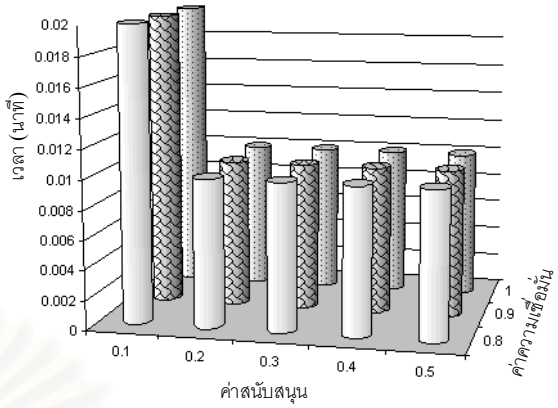
เมื่อนำข้อมูลส่วนแรกซึ่งเป็นข้อมูลทดลองของ chess และ mushroom จะได้ผลการทดลองดังรูปที่ 4.9 ถึงรูปที่ 4.14 โดยจะแบ่งกลุ่มตามเวลา จำนวนหลักเกณฑ์ และสภาพไวตามลำดับ โดยรูปที่ 4.9 ถึงรูปที่ 4.11 เป็นผลการทดลองของขั้นตอนวิธีเอไพออรี ซึ่งประกอบด้วยแกนของค่าสนับสนุนต่ำสุด ค่าความเชื่อมั่นต่ำสุด และผลการทดลองที่ต้องการเปรียบเทียบ และรูปที่ 4.12 ถึงรูปที่ 4.14 เป็นผลการทดลองของขั้นตอนวิธี WS ซึ่งประกอบด้วยแกนของค่าสนับสนุนแบบอ่อนต่ำสุด ค่าความเชื่อมั่นต่ำสุด และผลการทดลองที่ต้องการเปรียบเทียบ

จากผลการทดลองของขั้นตอนวิธี WS ซึ่งแสดงดังรูปที่ 4.12 ถึงรูปที่ 4.14 พบว่าเมื่อกำหนดค่าสนับสนุนแบบอ่อนต่ำสุดให้มีค่าน้อยกว่าค่าความเชื่อมั่นต่ำสุด จะทำให้ได้กลุ่มของหลักเกณฑ์แบบเดียวกัน ซึ่งตรงกับทฤษฎีบทที่ 3 ดังนั้นในการวิเคราะห์หลักเกณฑ์เชื่อมโยงโดยใช้ค่าสนับสนุนแบบอ่อน จึงควรกำหนดค่าสนับสนุนแบบอ่อนต่ำสุดให้มีค่ามากกว่าค่าความเชื่อมั่นต่ำสุดเสมอ และผลการทดลองที่ได้ทั้งหมดมีความสอดคล้องกับการทดลองในลักษณะที่ 1 นั่นคือขั้นตอนวิธี WS สร้างหลักเกณฑ์ออกมาเป็นจำนวนมากกว่าขั้นตอนวิธีเอไพออรี จึงทำให้ใช้เวลานานกว่า ดังรูปที่ 4.9 และรูปที่ 4.10 ที่แสดงเวลาและปริมาณของหลักเกณฑ์ที่ได้จากขั้นตอนวิธีเอไพออรี และรูปที่ 4.12 และรูปที่ 4.13 ที่แสดงเวลาและปริมาณของหลักเกณฑ์ของขั้นตอนวิธี WS

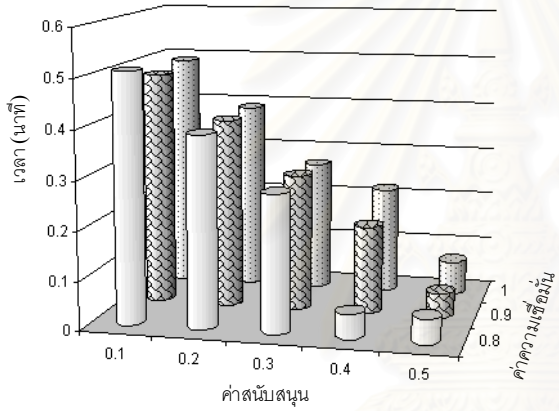
จากรูปที่ 4.11 พบว่ากลุ่มของหลักเกณฑ์ของขั้นตอนวิธีเอไพออรี ซึ่งสร้างจากค่าสนับสนุนต่ำสุดเท่ากับ 0.5 และค่าความเชื่อมั่นต่ำสุดที่มีค่าเท่ากับ 1 เป็นหลักเกณฑ์ที่ให้ค่าสภาพไวสูงสุดทั้งข้อมูล chess และ mushroom สำหรับทุกค่า *maxLHS* ที่นำมาทดสอบ และจากรูปที่ 4.14 พบว่ากลุ่มของหลักเกณฑ์ของขั้นตอนวิธี WS ที่สร้างจากค่าสนับสนุนแบบอ่อนต่ำสุดเท่ากับ 1 หรือค่าความเชื่อมั่นต่ำสุดเท่ากับ 1 มีค่าสภาพไวสูงสุดทั้งข้อมูล chess และ mushroom สำหรับทุกค่า *maxLHS* ที่นำมาทดสอบ เพราะหลักเกณฑ์ที่สร้างขึ้นเป็นหลักเกณฑ์กลุ่มเดียวกัน จากทฤษฎีบทที่ 2 ซึ่งเมื่อเปรียบเทียบสภาพไวของแต่ละขนาดของพจน์หน้าของทั้งสองขั้นตอนวิธีพบว่า เมื่อกำหนดให้ขนาดของพจน์หน้าสูงสุดเป็น 1 สภาพไวของกลุ่มหลักเกณฑ์ที่เลือกมาจากข้อมูล chess และ mushroom ของขั้นตอนวิธี WS จะมีค่าน้อยกว่าขั้นตอนวิธีเอไพออรี แต่เมื่อกำหนดให้ขนาดของพจน์หน้าเป็น 2 พบว่าสภาพไวของกลุ่มของหลักเกณฑ์ที่เลือกมาจากข้อมูล mushroom ของขั้นตอนวิธี WS จะมีค่ามากกว่าขั้นตอนวิธีเอไพออรี แต่สภาพไวของกลุ่มของหลักเกณฑ์ที่เลือกมาจากข้อมูล chess ของขั้นตอนวิธี WS จะน้อยกว่าขั้นตอนวิธีเอไพออรี และเมื่อขนาดของพจน์หน้าเป็น 3 พบว่าสภาพไวของกลุ่มของหลักเกณฑ์ที่เลือกของขั้นตอนวิธี WS จะมีค่ามากกว่า



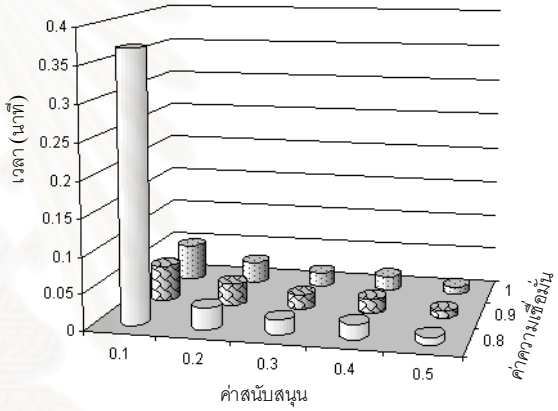
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



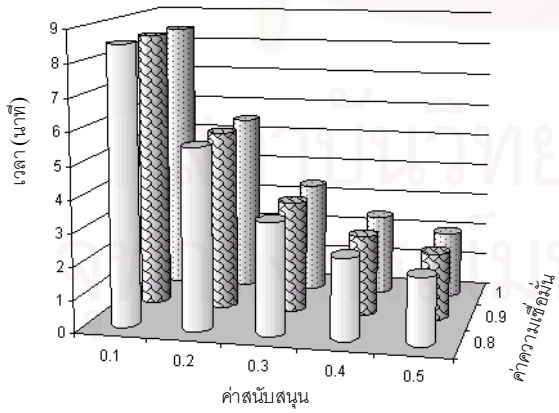
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



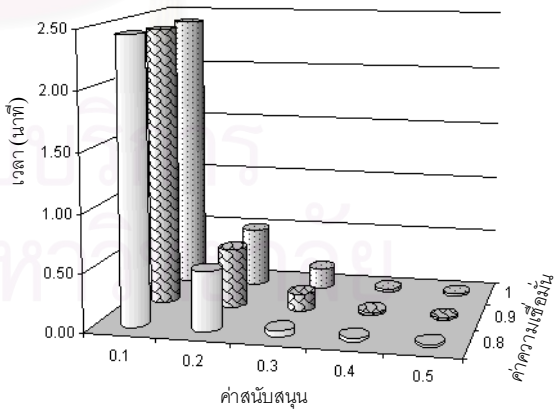
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2

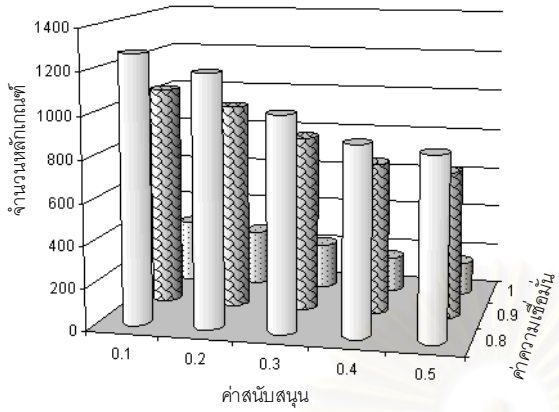


(จ) ข้อมูล chess เมื่อ *maxLHS* = 3

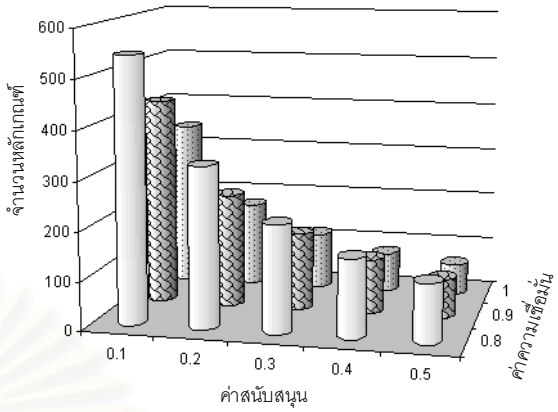


(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

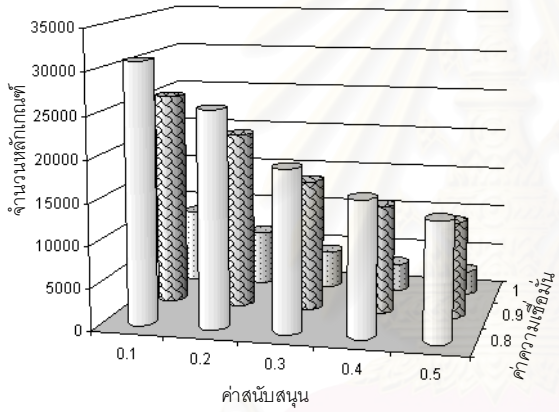
รูปที่ 4.9: (ก) – (ฉ) แสดงเวลาที่ใช้ในการทดลองลักษณะที่ 2 ของขั้นตอนวิธีเอไพออรี



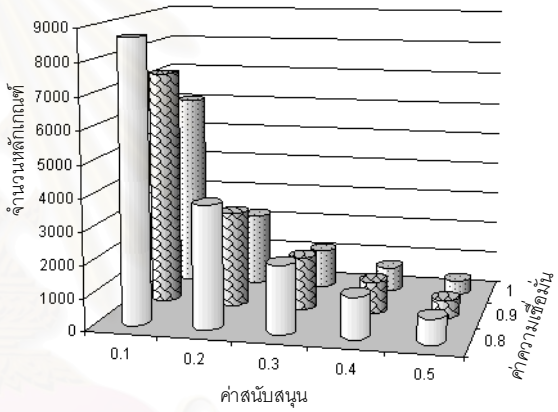
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



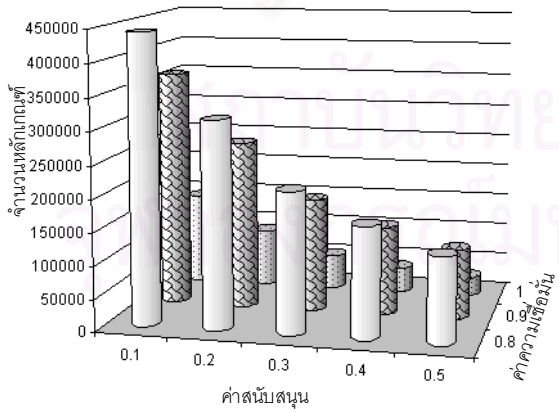
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



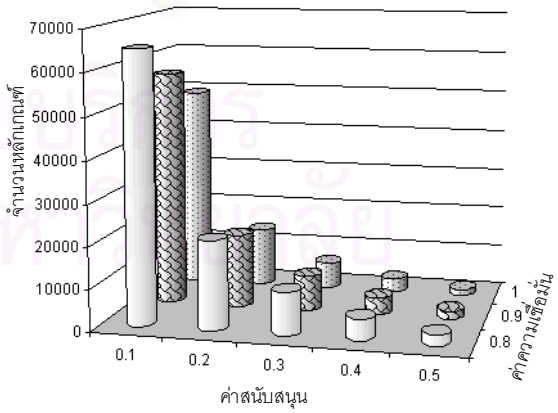
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2

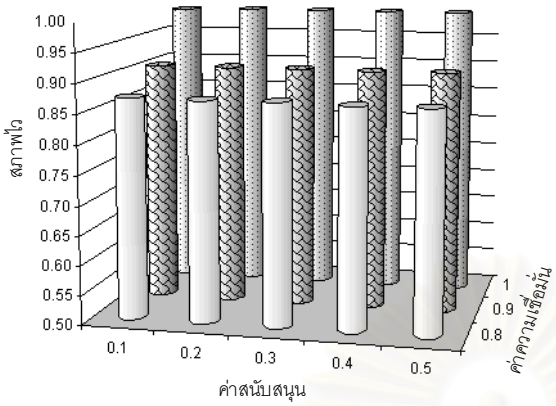


(จ) ข้อมูล chess เมื่อ *maxLHS* = 3

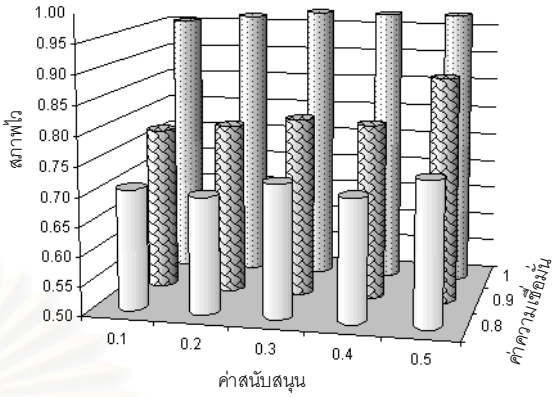


(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

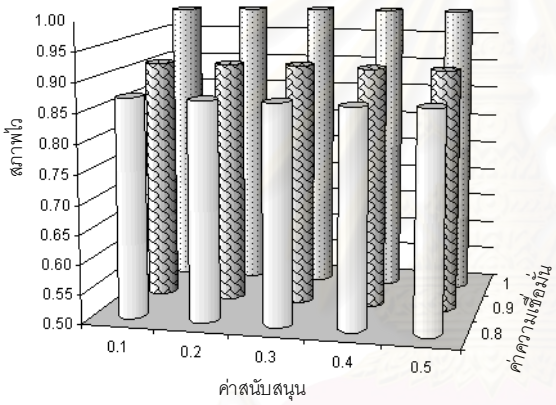
รูปที่ 4.10: (ก) – (ฉ) แสดงจำนวนหลักเกณฑ์จากการทดลองลักษณะที่ 2 ของขั้นตอนวิธีเอไพออริ



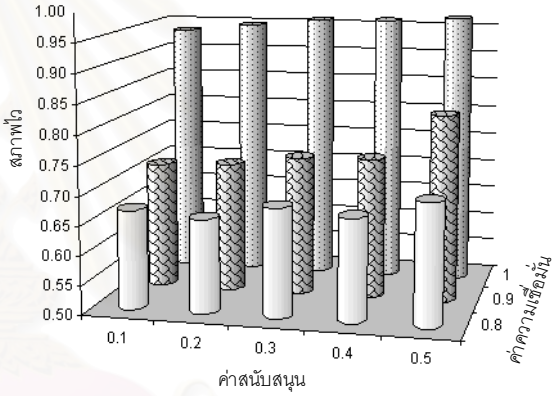
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



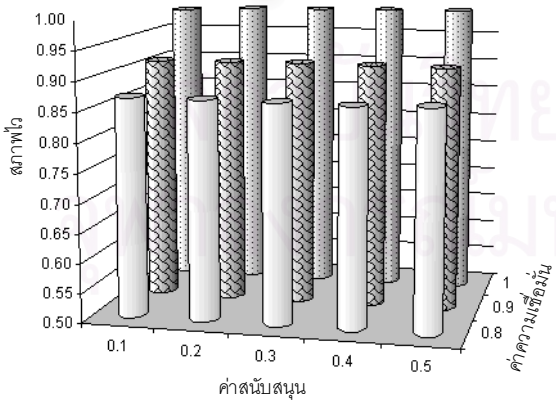
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



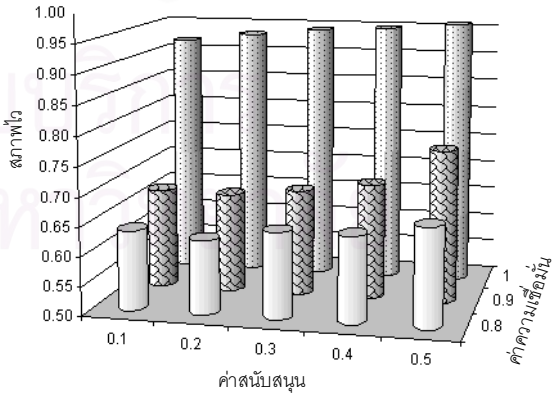
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2

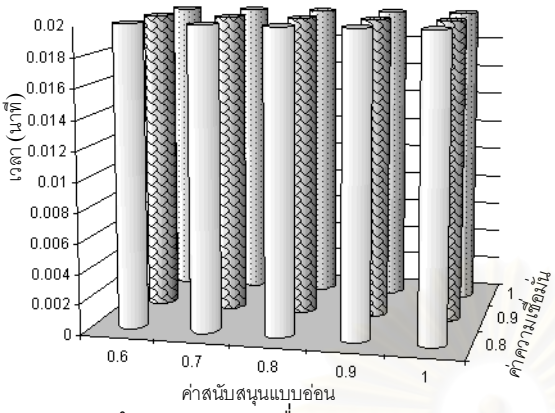


(จ) ข้อมูล chess เมื่อ *maxLHS* = 3

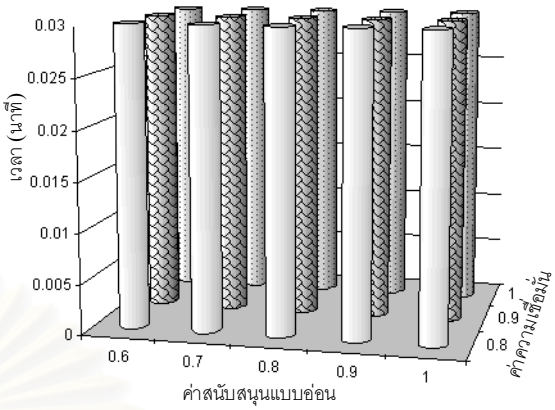


(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

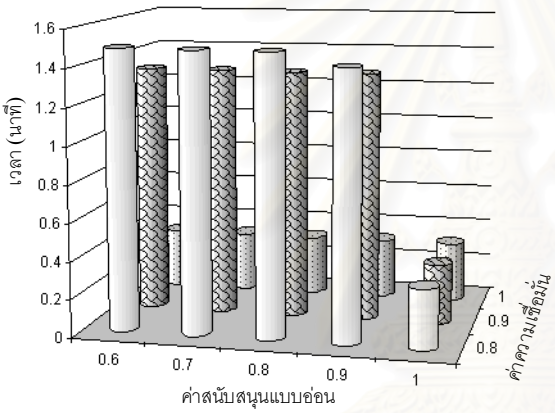
รูปที่ 4.11: (ก) – (ฉ) แสดงสภาพไวเฉลี่ยจากการทดลองลักษณะที่ 2 ของขั้นตอนวิธีเอไพออริ



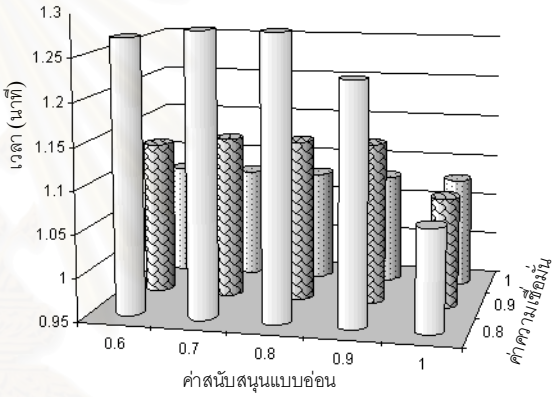
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



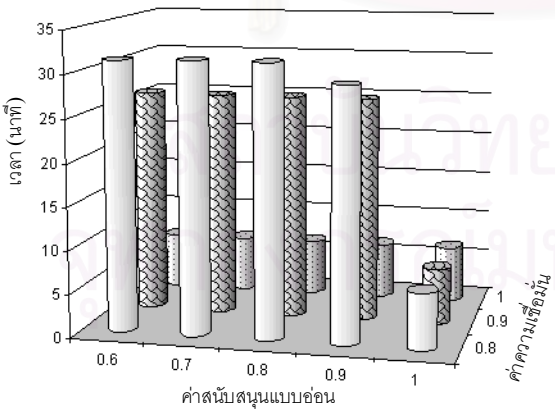
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



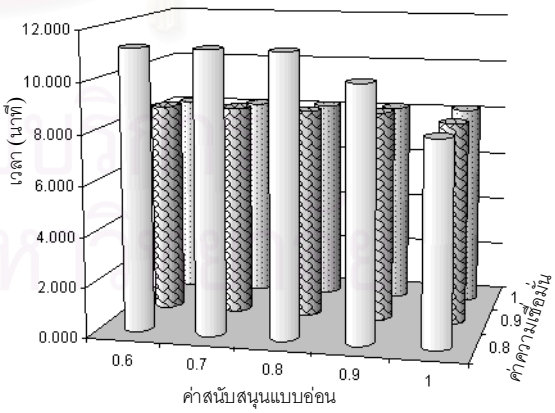
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2

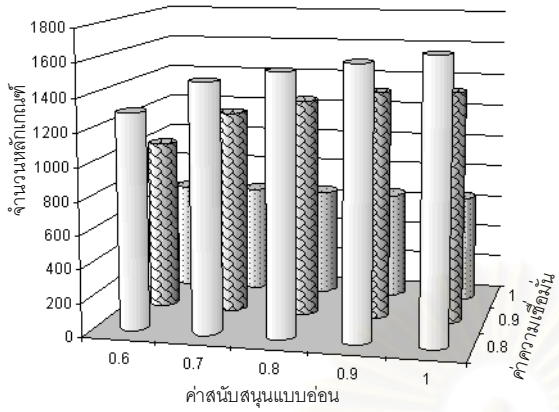


(จ) ข้อมูล chess เมื่อ *maxLHS* = 3

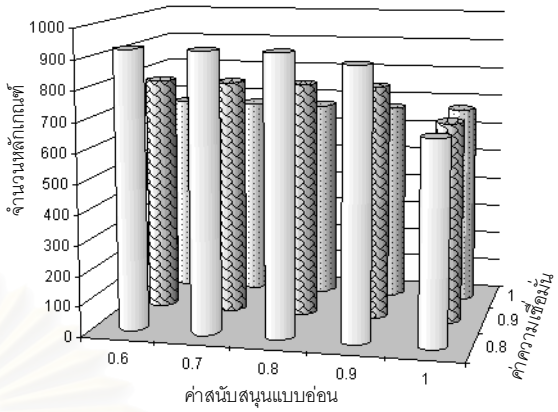


(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

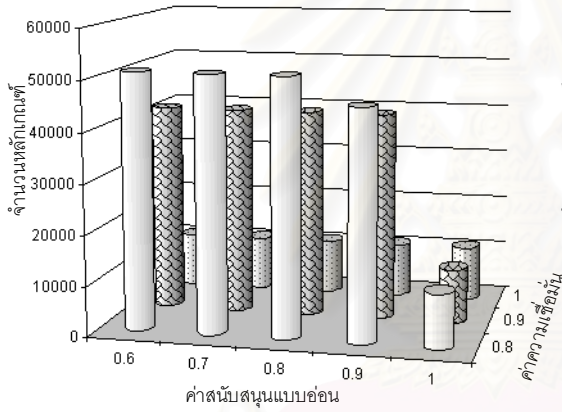
รูปที่ 4.12: (ก) – (ฉ) แสดงเวลาที่ใช้ในการทดลองลักษณะที่ 2 ของขั้นตอน WS



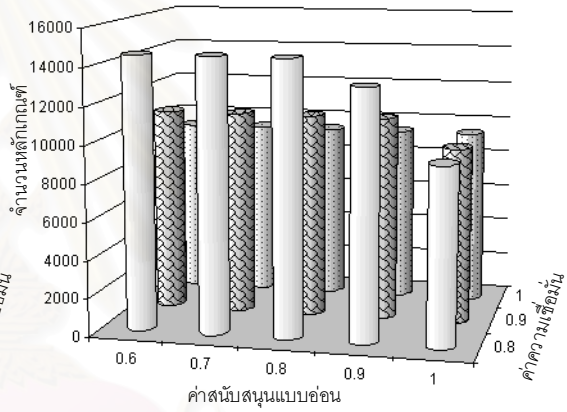
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



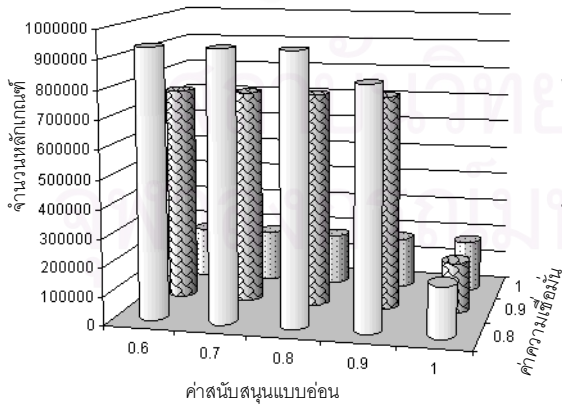
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



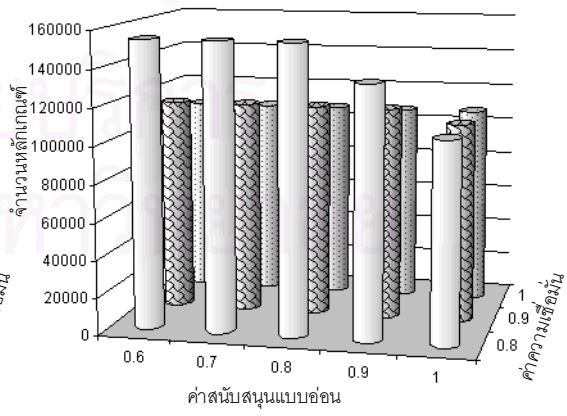
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2

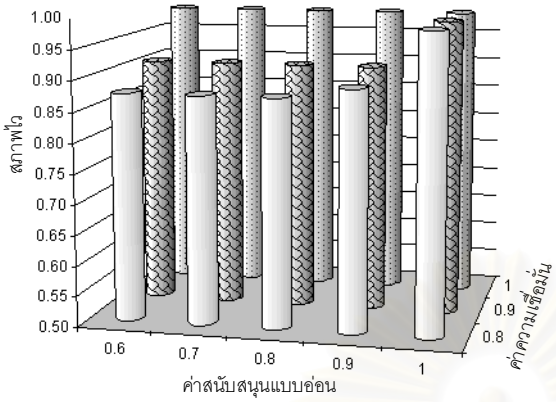


(จ) ข้อมูล chess เมื่อ *maxLHS* = 3

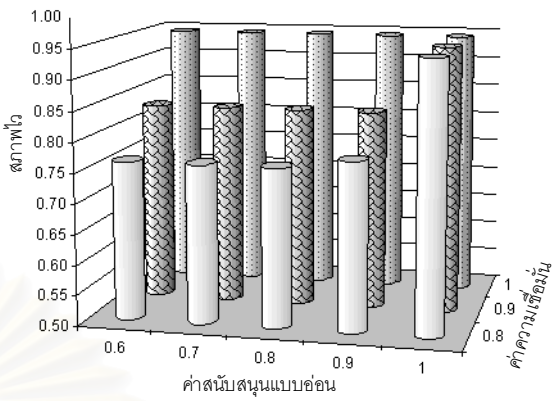


(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

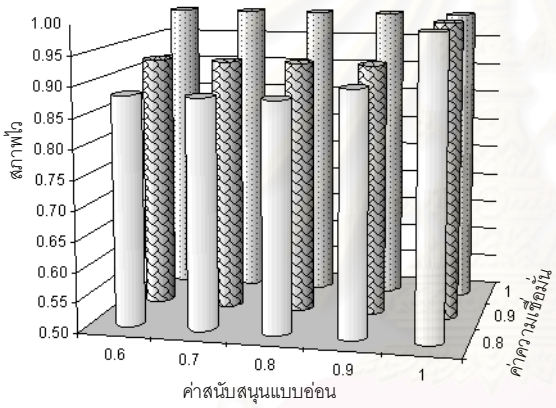
รูปที่ 4.13: (ก) – (ฉ) แสดงจำนวนหลักเกณฑ์จากการทดลองลักษณะที่ 2 ของขั้นตอนวิธี WS



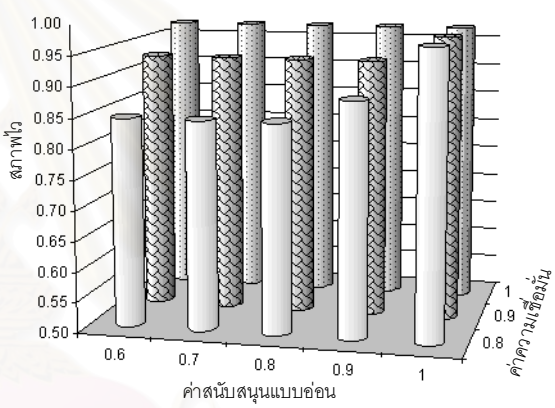
(ก) ข้อมูล chess เมื่อ *maxLHS* = 1



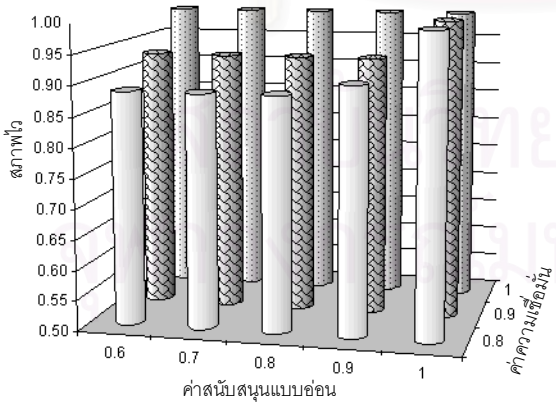
(ข) ข้อมูล mushroom เมื่อ *maxLHS* = 1



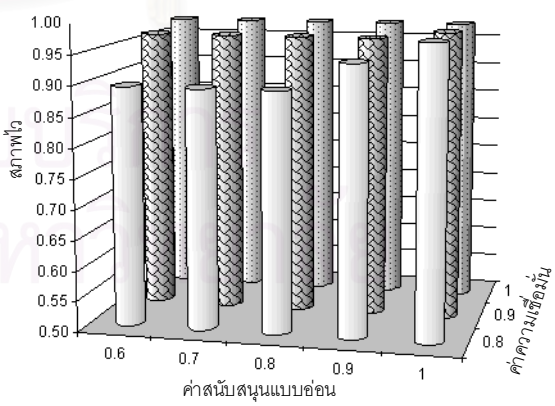
(ค) ข้อมูล chess เมื่อ *maxLHS* = 2



(ง) ข้อมูล mushroom เมื่อ *maxLHS* = 2



(จ) ข้อมูล chess เมื่อ *maxLHS* = 3



(ฉ) ข้อมูล mushroom เมื่อ *maxLHS* = 3

รูปที่ 4.14: (ก) – (ฉ) แสดงสภาพไวเฉลี่ยจากการทดลองลักษณะที่ 2 ของขั้นตอนวิธี WS

ดังนั้นเมื่อนำหลักเกณฑ์ที่มีค่าสภาพไวสูงสุดจากทั้ง 2 ขั้นตอนวิธีมาเปรียบเทียบโดยใช้ข้อมูลทดสอบจะได้ผลดังตารางที่ 4.5 ตารางที่ 4.6 และตารางที่ 4.7 เมื่อกำหนด *maxLHS* เป็น 1, 2 และ 3 ตามลำดับ

	สภาพไว	
	ข้อมูล chess	ข้อมูล mushroom
ขั้นตอนวิธีเอไพออรี	0.9810	0.8836
ขั้นตอนวิธี WS	0.9650	0.7333

ตารางที่ 4.5: ผลลัพธ์ที่ได้จากการเปรียบเทียบจากขั้นตอนวิธีเอไพออรีและขั้นตอนวิธี WS โดยใช้สภาพไว เมื่อ *maxLHS* =1

	สภาพไว	
	ข้อมูล chess	ข้อมูล mushroom
ขั้นตอนวิธีเอไพออรี	0.9805	0.8561
ขั้นตอนวิธี WS	0.9787	0.8589

ตารางที่ 4.6: ผลลัพธ์ที่ได้จากการเปรียบเทียบจากขั้นตอนวิธีเอไพออรีและขั้นตอนวิธี WS โดยใช้สภาพไว เมื่อ *maxLHS* =2

	สภาพไว	
	ข้อมูล chess	ข้อมูล mushroom
ขั้นตอนวิธีเอไพออรี	0.9805	0.8239
ขั้นตอนวิธี WS	0.9801	0.8877

ตารางที่ 4.7: ผลลัพธ์ที่ได้จากการเปรียบเทียบจากขั้นตอนวิธีเอไพออรีและขั้นตอนวิธี WS โดยใช้สภาพไว เมื่อ maxLHS =3

จากตารางที่ 4.5 ถึงตารางที่ 4.7 พบว่าค่าของสภาพไวของกลุ่มของหลักเกณฑ์ที่สร้างจากขั้นตอนวิธี WS และขั้นตอนวิธีเอไพออรี เมื่อเปรียบเทียบกับข้อมูลทดสอบมีความสอดคล้องกับค่าของสภาพไวของกลุ่มของหลักเกณฑ์เมื่อเปรียบเทียบกับข้อมูลตรวจสอบความสมเหตุสมผล โดยสภาพไวของขั้นตอนวิธี WS จะมีค่าเพิ่มขึ้นเมื่อกำหนดให้ขนาดของพจน์หน้าเพิ่มขึ้น ซึ่งตรงข้ามกับขั้นตอนวิธีเอไพออรี

จากผลการทดลองแสดงให้เห็นว่าข้อมูล chess ซึ่งมีปริมาณข้อมูล และความแตกต่างของข้อมูลที่น้อย จะมีค่าสภาพไวของขั้นตอนวิธีเอไพออรีมากกว่าขั้นตอนวิธี WS เมื่อขนาดของพจน์หน้าเป็น 1 และ 2 และเมื่อขนาดของพจน์หน้ามีค่าเป็น 3 สภาพไวของหลักเกณฑ์ที่ได้จากขั้นตอนวิธีเอไพออรีจะมีค่าใกล้เคียงกับขั้นตอนวิธี WS แม้ว่าจำนวนหลักเกณฑ์ที่ออกมาจากขั้นตอนวิธี WS จะมีมากกว่า ในขณะที่ข้อมูล mushroom ที่มีปริมาณข้อมูลและความแตกต่างของข้อมูลสูง จะมีค่าสภาพไวของขั้นตอนวิธีเอไพออรีสูงกว่าขั้นตอนวิธี WS ในกรณีที่กำหนดขนาดของพจน์หน้าเป็น 1 และในกรณีที่ขนาดของพจน์หน้าเป็น 2 และ 3 ขั้นตอนวิธี WS จะมีค่าสภาพไวมากกว่าขั้นตอนวิธีเอไพออรี

ดังนั้นในการวิเคราะห์หลักเกณฑ์เชื่อมโยงที่ใช้กับข้อมูลที่มีความหลากหลายของสินค้าที่มาก จึงควรเลือกใช้ขั้นตอนวิธี WS พร้อมกับกำหนดค่าสนับสนุนแบบอ่อนต่ำสุดเป็น 1 ซึ่งทำให้ค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 1 โดยทฤษฎีบทที่ 2 และควรกำหนดให้พจน์หน้าของหลักเกณฑ์ที่สร้างมีขนาดใหญ่ เพราะแนวโน้มของสภาพไวของหลักเกณฑ์ที่สร้างจากขั้นตอนวิธี WS จะมีค่าเพิ่มมากขึ้นเมื่อพจน์หน้ามีขนาดใหญ่ขึ้น แต่ในการวิเคราะห์หลักเกณฑ์เชื่อมโยงที่ใช้กับข้อมูลที่มีความแตกต่างของข้อมูลน้อย ควรเลือกใช้ขั้นตอนวิธีเอไพออรี พร้อมกับกำหนดให้มีค่าสนับสนุนต่ำสุดเป็น 0.5 ความเชื่อมั่นต่ำสุดเป็น 1 และกำหนดพจน์หน้าของหลักเกณฑ์ที่สร้างได้ให้มีขนาดเล็ก เพราะเมื่อขนาดของพจน์หน้ามีขนาดใหญ่ขึ้น จะทำให้ค่าสภาพไวที่ได้จากขั้นตอนวิธีเอไพออรีมีค่าลดลง

บทที่ 5

สรุปผลการทดลอง

งานวิจัยนี้นำเสนอตัววัดตัวใหม่ที่มีชื่อว่า ค่าสนับสนุนแบบอ่อน และสร้างขั้นตอนวิธีที่สอดคล้องกับตัววัดนี้ที่มีชื่อว่าขั้นตอนวิธี WS เพื่อสร้างหลักเกณฑ์เชื่อมโยงที่สมเหตุสมผลและน่าสนใจจากฐานข้อมูล โดยค่าสนับสนุนแบบอ่อนเป็นตัววัดที่เกิดจากแนวคิดทางตรรกศาสตร์ เนื่องจากข้อมูลที่ขัดแย้งกับหลักเกณฑ์เชื่อมโยงจะเหมือนกับข้อมูลที่ขัดแย้งกับประพจน์ถ้า-แล้วในทางตรรกศาสตร์ ซึ่งนำมาใช้นิยามค่าสนับสนุนแบบอ่อนให้พิจารณาเฉพาะระเบียนที่ไม่ขัดแย้งกับหลักเกณฑ์เทียบกับระเบียนทั้งหมด

ถึงแม้ว่าผู้ใช้จะกำหนดให้ค่าสนับสนุนแบบอ่อนต่ำสุดและค่าความเชื่อมั่นต่ำสุดแตกต่างกัน แต่หลักเกณฑ์เชื่อมโยงที่สร้างจากค่าสนับสนุนแบบอ่อนจะมีประสิทธิภาพในการทำงานใกล้เคียงกัน และเพื่อให้ผลลัพธ์เป็นหลักเกณฑ์ที่มีประสิทธิภาพ ผู้ใช้ควรกำหนดให้ค่าสนับสนุนแบบอ่อนต่ำสุดมีค่าเป็น 1 ทำให้ค่าความเชื่อมั่นต่ำสุดมีค่าเป็น 1 โดยทฤษฎีบทที่ 2 ในขณะที่ขั้นตอนวิธีเอไพออรี ผู้ใช้จะต้องพิจารณาและเลือกค่าสนับสนุนต่ำสุด และค่าความเชื่อมั่นต่ำสุดที่สอดคล้องกับข้อมูล ดังนั้นในกรณีที่สร้างหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธีเอไพออรี ผู้ใช้ต้องมีความรู้และความเข้าใจข้อมูลที่ต้องการวิเคราะห์อย่างลึกซึ้ง

ข้อมูลที่เหมาะสมในการสร้างหลักเกณฑ์โดยใช้ค่าสนับสนุนแบบอ่อน คือข้อมูลที่มีปริมาณมากและมีความแตกต่างของข้อมูลสูง หรือมีจำนวนระเบียนและจำนวนลักษณะประจำมาก ข้อมูลในลักษณะนี้จะเป็นข้อมูลที่มีขนาดใหญ่ โดยผลการทดลองแสดงหลักเกณฑ์ที่สร้างจากขั้นตอนวิธี WS ของข้อมูล mushroom จะมีประสิทธิภาพในการทำงานมากกว่าหลักเกณฑ์ที่สร้างจากขั้นตอนวิธีเอไพออรี

เนื่องจากการเลือกใช้ตัววัดค่าสนับสนุนแบบอ่อนโดยผ่านขั้นตอนวิธี WS ช่วยให้ผู้ใช้สร้างหลักเกณฑ์ที่น่าสนใจออกมาเป็นจำนวนมาก จึงทำให้ใช้เวลาในการทำงานนานกว่าขั้นตอนวิธีเอไพออรี ดังนั้นงานในอนาคตคือ การลดจำนวนหลักเกณฑ์ที่สร้างจากค่าสนับสนุนแบบอ่อน และปรับปรุงความเร็วในการคัดเลือกหลักเกณฑ์จากขั้นตอนวิธี WS โดยยังคงสภาพไวที่สูงอยู่ นอกจากนี้จากผลการทดลองจะเห็นได้ชัดว่าขั้นตอนวิธีเอไพออรีและขั้นตอนวิธี WS มีลักษณะ

ค่าของสภาพไวที่ตรงข้ามกันคือ เมื่อกำหนดพจน์หน้าของหลักเกณฑ์ให้มีขนาดเล็ก ค่าสภาพไวของขั้นตอนวิธีเอไพออรีจะมีค่าสูง แต่ขั้นตอนวิธี WS จะมีค่าน้อย และในกรณีที่กำหนดพจน์หน้าของหลักเกณฑ์ให้มีขนาดเพิ่มขึ้น ขั้นตอนวิธีเอไพออรีจะมีค่าสภาพไวลดลง ในขณะที่ขั้นตอนวิธี WS จะมีค่าสภาพไวเพิ่มขึ้น ดังนั้นงานที่น่าสนใจคือ การสร้างหลักเกณฑ์เชื่อมโยงโดยอาศัยคุณสมบัติหรือข้อดีของขั้นตอนวิธีทั้งสอง เพื่อให้หลักเกณฑ์ที่สร้างขึ้นใหม่มีค่าสภาพไวสูงมากยิ่งขึ้น

รายการอ้างอิง

[1] Han, J., and Kamber, M. *Data mining : Concepts and techniques*. San Fransisco: Morgan Kaufmann Publishers, 2001.

[2] กิติพงศ์ กลมกล่อม. *การออกแบบและพัฒนาคลังข้อมูล (Data warehouse)*. กรุงเทพฯ: เค ทีพี คอมพ์ แอนด์ คอนซัลท์, 2546.

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* (1996): 37-54.

[4] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. The kdd process for extract in useful knowledge from volumes data. *Communications of the ACM* 11 (November 1996): 27-31.

[5] Two Crows corporations. *Introduction to data mining and knowledge discovery* [online] . (1999). Available from : http://www.twocrows.com [cited 18 January 2007].

[6] Kuonen, D. A statistical perspective of data mining. *CRM Zine* (December 2004): 1-6.

[7] Agrawal, R., Imielinski, T., and Swami, A. Mining association rules between sets of items in large databases. *Proceeding of the ACM SIGMOD International Conference on Data* (1993) : 207-216.

[8] Agrawal, R., and Srikant, R. Fast algorithms for mining association rules. *Proceeding of the 20th International Conference on Very Large Data* (1994) : 487-499.

[9] Zhang, C., and Zhang, S. *Associaiton rule mining: Models and algorithms*. Berlin: Springer, 2002.

[10] Hahsler, M. *A comparison of commonly used interest measures for association rules* [online] . (n.d.). Available from : http://wwwai.wu-wien.ac.at/~hahsler/ research/association_rules/ [cited 19 July 2007].

[11] Webb, G., I., and Zhang, S. K-optimal rule discovery. *Data Mining and Knowledge Discovery* (2005): 39-79.

[12] Brin, S., Motwani, R., and Silverstein, C. Beyond market baskets: Generalizing association rules to correlations. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1997) : 265-276.

[13] Brin, S., MotWani, R., Ullman, J.D., and Tsur, S. Dynamic itemset counting and implication rules for market basket data. *Proceedings of the ACM SIGMOD International Conference on Management of Data* (1997) : 255-264.

[14] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. *In 2000 ACM SIGMOD International Conference on Management of Data* (2000) : 1-12.

[15] Sinapiromsaran, K. and Sirisrisumrit, C. Association rule extraction based on the weak support measure. *The Second Graduate Congress of Mathematics and Physical Science* (2006) : .

[16] Fayyad, U., Shapiro, G., P., and Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine* (1996): 37-54.

[17] Fayyad, U., Shapiro, G., P., and Smyth, P. The kdd process for extractin useful knowledge from volumes data. *Communications of the ACM* 11 (November 1996): 27-31.

[18] Teradata. *Data mining for enterprise solutions- A business perspective on mining data for corperate intelligence* [online] . (n.d.). Available from : http://www.teradata.com [cited 25 April 2007].

[19] Westphal, C., and Blaxton, T. *Data mining solutions: Methods and tools for solving real-world problems* [book] . (1998). Available from : http://as.wiley.com [cited 17 April 2007].

[20] Kleissner, C. Data mining for the enterprise. *Proceedings of the Annual International Conference on System Science* (1998) : 295-304.

[21] Apte, C., Liu, B., and Pednault, E., P.D. Business applications of data mining. *Communications of the ACM* (May 22, 2002): 49-53.

[22] Berry, M., J.A., and Linoff, G. *Data mining techniques for marketing*. United States: Wiley Computer, 1997.

[23] Kerlinger, F.N., and Pedhazur, E.J. *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, 1973.

[24] Webb, G., I. OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research* (1995): 431-465.

[25] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. Efficient mining of associaiton rules using closed itemset lattices. *Information Systems* (1999): 25-46.

[26] Blake, C., L. and Merz, C., L. *UCI repository of machine learning databases* [data] . (1998). Available from : www.ics.uci.edu/~mlearn/MLRepository.html [cited 10 October 2006].

[27] FIMI. *Frequent Itemset Mining Dataset Repository* [data] . (2003-2004). Available from : http://fimi.cs.helsinki.fi/data/ [cited 10 November 2006].

[28] Beck, J.R. and Schultz, E.K. The use of ROC curves in test performance evaluation. (1986): 13-20.

[29] Swet, J. Measuring the accuracy of diagnostic systems. *Science* (1988): 1285-1293.

[30] Wikipeida. *Binary classification* [online] . (n.d.). Available from : http://en.wikipedia.org/ wiki/Binary_classification [cited 7 February 2007].

ภาคผนวก

ภาคผนวก ก

โปรแกรมสร้างตารางและนำเข้าฐานข้อมูล

โปรแกรมการสร้างตารางข้อมูลบนฐานข้อมูลเขียนจากภาษาจาวา โดยซอฟต์แวร์ฐานข้อมูลที่เลือกใช้คือ MySQL 5.0 มีชื่อว่า CreateTable.java ซึ่งข้อมูลที่ใช้ในการทดลองต้องมีรูปแบบเดียวกับตารางที่ 2.4 และเมื่อนำไปสร้างตารางในฐานข้อมูลจะมีโครงสร้างดังตารางที่ ก.1

ชื่อลักษณะประจำ	ชนิด	ขนาด	ความหมาย
tid	int	4	หมายเลขระเบียน
i_1	boolean	1	สินค้าชนิดที่ 1
⋮	⋮	⋮	⋮
i_m	boolean	1	สินค้าชนิดที่ m

ตารางที่ ก.1: โครงสร้างข้อมูลของตารางที่สร้างในฐานข้อมูล

การทำงานของโปรแกรมนี้จะสร้างตารางในฐานข้อมูล และนำข้อมูลที่อยู่ในรูปแบบแฟ้มข้อความไปสร้างเป็นฐานข้อมูล โดยผู้ใช้จะต้องตั้งชื่อและสร้างฐานข้อมูล ข้อมูลที่ต้องการนำเข้า และแฟ้มข้อความอธิบายข้อมูล

การสร้างฐานข้อมูลใน MySQL ผู้ใช้จะใช้คำสั่งผ่านภาษา SQL สมมติให้ฐานข้อมูลที่ต้องการสร้างชื่อ MYDB รูปที่ ก.1 แสดงคำสั่งที่ใช้สร้างฐานข้อมูลที่มีชื่อว่า MYDB ผ่านคำสั่ง SQL

```
CREATE DATABASE MYDB;
```

รูปที่ ก.1: คำสั่งที่ใช้ในการสร้างฐานข้อมูล

ข้อมูลที่ใช้ในการทดลองจะต้องเก็บอยู่ในรูปแบบแฟ้มข้อความและผ่านการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกับตารางที่ 2.4 การนำข้อมูลเข้าตารางในฐานข้อมูลจะต้องให้ข้อมูลอยู่ในลักษณะที่สอดคล้องกับการเก็บข้อมูลใน MySQL ซึ่งแต่ละหลักของข้อมูลต้องเว้นระยะด้วยเครื่องหมายแท็บเท่านั้น สมมติให้ข้อมูลที่ต้องการนำเข้ามีชื่อว่า myTable.txt ซึ่งเป็นข้อมูลเดียวกับตารางที่ 2.3 โดยมีจำนวนไอเทมเท่ากับ 5 และมี 10 ระเบียน ข้อมูลจะต้องจัดอยู่ในลักษณะดังรูปที่ ก.2

100	1	1	0	0	1
200	0	1	0	1	0
300	0	1	1	0	0
400	1	1	0	1	0
500	1	0	1	0	0
600	0	1	1	0	0
700	1	0	1	0	0
800	1	1	1	0	1
900	1	1	1	0	0
1000	1	0	0	0	1

รูปที่ ก.2: ลักษณะของข้อมูลที่นำเข้าตารางในฐานข้อมูล

แฟ้มข้อความอธิบายข้อมูลเป็นแฟ้มที่ใช้อธิบายฐานข้อมูล ชื่อตารางที่สร้าง ชื่อแฟ้มข้อมูลและตำแหน่งของข้อมูลที่ต้องการนำเข้าตารางในฐานข้อมูล ค่าต่ำสุดของไอเทม และค่าสูงสุดของไอเทม สมมติให้ชื่อตารางที่ต้องการสร้างในฐานข้อมูลคือ myTable โดยจะสร้างแฟ้มข้อความอธิบายข้อมูลชื่อ fileDescription.txt ได้ดังรูปที่ ก.3 ซึ่งสมมติให้ myTable.txt จัดเก็บไว้บน D:/ และในกรณีที่ผู้ใช้ต้องการใส่ความข้อความอธิบายเพิ่มเติม ให้ใส่เครื่องหมาย # เป็นตัวอักษรแรกของบรรทัดที่ต้องการ ซึ่งโปรแกรมจะอ่านข้ามบรรทัดดังกล่าวไป

```
#This is my first table
    MYDB myTable        D:/myTable.txt        1        5
...
```

รูปที่ ก.3: ตัวอย่างของแฟ้มอธิบายข้อมูล

จากรูปที่ ก.3 โปรแกรมจะอ่านข้ามบรรทัดแรก ไปทำงานที่บรรทัดที่สอง ซึ่งในการเรียกใช้โปรแกรมจะต้องใส่ชื่อแฟ้มอธิบายข้อมูลเป็นอาร์กิวเมนต์เสมอ ดังรูปที่ ก.4 ซึ่งบรรทัดแรกเป็นการแปลโปรแกรมและบรรทัดที่สองเป็นตัวอย่างในการดำเนินงานของโปรแกรม โดยความหมายของตัวแปรที่สำคัญของโปรแกรม แสดงดังตารางที่ ก.2

```
javac CreateTable.java
java CreateTable fileDescription.txt
```

รูปที่ ก.4: ตัวอย่างการแปลโปรแกรม และการสั่งประมวลผลโปรแกรมสร้างตารางและนำเข้าข้อมูล

ชื่อ	ชนิด	ความหมาย
username	string	ชื่อผู้ใช้ฐานข้อมูล
password	string	รหัสผ่านของผู้ใช้ฐานข้อมูล
dbName	string	ชื่อของฐานข้อมูลที่ได้สร้างไว้แล้ว
tableName	string	ชื่อของตารางที่ต้องการสร้างในฐานข้อมูล
readFile	string	ชื่อของแฟ้มอธิบายข้อมูล
fileName	string	ชื่อและตำแหน่งของแฟ้มข้อมูลที่ต้องการนำเข้า
minItem	int	หมายเลขไอเทมที่เล็กที่สุดของข้อมูล
maxItem	int	หมายเลขไอเทมที่มากที่สุดของข้อมูล
sql	string	เก็บคำสั่ง SQL ที่ใช้สร้างตารางและนำเข้าข้อมูล

ตารางที่ ก.2: ชื่อ ชนิด และความหมายของตัวแปรที่สำคัญในโปรแกรม

```
import java.sql.*;
import java.io.*;
import java.util.*;
import java.lang.*;

public class CreateTable{
    //JDBC driver name and database URL
    static String JDBC_DRIVER = "com.mysql.jdbc.Driver";
    static String DATABASE_URL = "jdbc:mysql://localhost/";
```

```
//launch the application
public static void main(String args[]){
    //set username and password for using mysql
    String username = "root";
    String password = "";
    String dbName,tableName,fileName,readFile,sql;
    int minItem, maxItem;
    Connection connection = null;
    Statement statement = null;
    Scanner fText;

    //check JDBC driver
    try{
        Class.forName(JDBC_DRIVER); //load jdbc driver class
    }
    catch(ClassNotFoundException classNotFound){
        classNotFound.printStackTrace();
        System.exit(1);
    }

    try{
        //open the file to read descriptions
        readFile = args[0];
        fText = new Scanner(new File(readFile));
        System.out.println("Open the file : "+readFile);

        while(fText.hasNext()){
            dbName = fText.next();
            tableName = fText.next();
            fileName = fText.next();
            minItem = Integer.parseInt(fText.next());
            maxItem = Integer.parseInt(fText.next());
            //if the input file has comment with by #, we pass it to the next read
            if(dbName.startsWith("#"))
                continue;
            System.out.println("\tDatabase Name is : "+dbName);
            System.out.println("\tTable Name is : "+tableName);
            System.out.println("\tFile name for load data to mysql \t:
             "+fileName+"\n");
            //establish connection to database
            connection = DriverManager.getConnection(DATABASE_URL
            +dbName,username,password);

            //create statement for querying database
            statement = connection.createStatement();
            statement.executeUpdate("drop table if exists "+tableName);
            sql = "create table "+tableName+" (tid int  not null auto_increment" ;
            for(int i = minItem ; i<=maxItem; i++)
```

```
                {       sql = sql.concat(", i").concat(Integer.toString(i))
                        .concat(" boolean");
                }
                sql = sql.concat(", primary key(tid))");
                statement.executeUpdate(sql);
                statement.executeUpdate("load data local infile \'"+fileName+
                "\' into table "+tableName);

            }//end of while
            if(fText != null){
                fText.close();
                System.out.println("Close the read file : "+readFile);
            }
        }
        catch(SQLException sqlException){
            sqlException.printStackTrace();
            System.exit(1);
        }
        catch(SecurityException securityException){
            System.err.println("You do not have write access to this file");
            System.exit(1);
        }
        catch(FileNotFoundException fileNotfound){
            System.err.println("Error opening file.");
            System.exit(1);
        }
        try{
            statement.close();
            connection.close();
        }
        catch(Exception exception){
            exception.printStackTrace();
            System.exit(1);
        }
    }
}
```

ภาคผนวก ข

โปรแกรมสร้างหลักเกณฑ์เชื่อมโยงโดยขั้นตอนวิธี WS

โปรแกรมการสร้างหลักเกณฑ์เชื่อมโยงโดยขั้นตอนวิธี WS เขียนจากภาษาจาวา มีชื่อว่า WS.java โดยการทำงานของโปรแกรมนี้จะอ่านตารางจากฐานข้อมูล MySQL แล้วนำไปสร้างไปหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธี WS ซึ่งข้อมูลที่ใช้ในการทดลองต้องมีโครงสร้างดังตารางที่ ก.1 โดยตารางที่ ข.1 แสดงชื่อ ชนิด และความหมายของตัวแปรที่สำคัญของโปรแกรม และตารางที่ ข.2 แสดงชื่อ และการทำงานของฟังก์ชันที่สำคัญในโปรแกรม

ชื่อ	ชนิด	ความหมาย
username	string	ชื่อผู้ใช้ฐานข้อมูล
password	string	รหัสผ่านของผู้ใช้ฐานข้อมูล
dbName	string	ชื่อของฐานข้อมูลที่ได้สร้างไว้แล้ว
tableName	string	ชื่อของตารางข้อมูลที่ต้องการสร้างหลักเกณฑ์เชื่อมโยง
minWS	double	ค่าสนับสนุนแบบอ่อนต่ำสุด
minC	double	ค่าความเชื่อมั่นต่ำสุด
maxSizeLHS	int	ขนาดของพจน์หน้าสูงสุดของหลักเกณฑ์ที่สร้าง
directory	string	ตำแหน่งของแฟ้มข้อมูลที่ต้องการให้บันทึก
minItem	int	หมายเลขไอเทมที่เล็กที่สุดของข้อมูลที่เลือก
maxItem	int	หมายเลขไอเทมที่ใหญ่ที่สุดของข้อมูลที่เลือก
RHS	int	พจน์หลังของหลักเกณฑ์ที่เป็น 1-ไอเทมเซต
LHS	Array list of string	เก็บไอเทมเซตของพจน์หน้าของหลักเกณฑ์ สำหรับแต่ละพจน์หลังที่กำหนดไว้

ตารางที่ ข.1: ชื่อ ชนิด และความหมายของตัวแปรที่สำคัญในโปรแกรมสร้างหลักเกณฑ์เชื่อมโยง

ชื่อ	ความหมาย
findMinItem	ค้นหาหมายเลขไอเทมที่เล็กที่สุดของข้อมูลในฐานข้อมูล
findMaxItem	ค้นหาหมายเลขไอเทมที่ใหญ่ที่สุดของข้อมูลในฐานข้อมูล
findNumTransaction	คำนวณจำนวนระเบียนของข้อมูลในฐานข้อมูล
findWeakS	คำนวณค่าสนับสนุนแบบอ่อน
findConf	คำนวณค่าความเชื่อมั่น
generateSQL	สร้างคำสั่งภาษา SQL
createRule	สร้างหลักเกณฑ์เชื่อมโยงจากค่าสนับสนุนแบบอ่อนและค่าความเชื่อมั่น

ตารางที่ ข.2: ชื่อ และการทำงานของฟังก์ชันที่สำคัญในโปรแกรมสร้างหลักเกณฑ์เชื่อมโยง

ในการเรียกใช้โปรแกรมผู้ใช้จะต้องใส่ค่าอาร์กิวเมนต์ดังต่อไปนี้คือ

1. ชื่อฐานข้อมูลที่จะทำการทดลอง (dbName)
2. ชื่อตารางที่จะใช้ในการสร้างหลักเกณฑ์เชื่อมโยง (tableName)
3. ค่าสนับสนุนแบบอ่อนต่ำสุด (minWS)
4. ค่าความเชื่อมั่นต่ำสุด (minC)
5. ขนาดของพจน์หน้าที่ใหญ่ที่สุดของหลักเกณฑ์ที่สร้างได้ (maxSizeLHS)
6. ตำแหน่งของแฟ้มข้อมูลที่ต้องการบันทึก (directory)

ซึ่งผลลัพธ์ที่ได้จะเป็นแฟ้มข้อมูลที่บันทึกหลักเกณฑ์เชื่อมโยงที่สร้างจากขั้นตอนวิธี WS โดยใช้ชื่อว่า ชื่อตาราง-ค่าสนับสนุนแบบอ่อนต่ำสุด-ค่าความเชื่อมั่นต่ำสุด.txt และแฟ้มที่มีชื่อว่า time.txt ซึ่งเป็นแฟ้มข้อมูลที่เก็บชื่อแฟ้มข้อมูลที่บันทึก เวลาในการทำงาน และจำนวนหลักเกณฑ์ที่สร้าง

สมมติให้ชื่อฐานข้อมูลที่ต้องการใช้งานคือ MYDB ชื่อตารางคือ myTable กำหนดให้ค่าสนับสนุนแบบอ่อนต่ำสุดคือ 0.9 ค่าความเชื่อมั่นต่ำสุดคือ 0.8 ขนาดของพจน์หน้าที่ใหญ่ที่สุดคือ 2 และตำแหน่งของข้อมูลที่ต้องการบันทึกคือ D:/ การเรียกใช้โปรแกรมแสดงดังรูปที่ ข.1 โดยบรรทัดแรกเป็นการแปลโปรแกรมและบรรทัดที่สองเป็นตัวอย่างในการดำเนินงานของโปรแกรม

```
javac WS.java
java WS MYDB myTable 0.9 0.8 2 D:/
```

รูปที่ ข.1: ตัวอย่างการแปลโปรแกรม และการดำเนินงานของโปรแกรมการสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธี WS

จากรูปที่ ข.1 ผลลัพธ์ที่ได้จะเป็นแฟ้มข้อมูลที่มีชื่อว่า myTable-0.9-0.8.txt และ time.txt ซึ่งจะมีรูปแบบผลลัพธ์ที่ได้เป็นดังรูปที่ ข.2 และ ตามลำดับ

```
Table name : chess
The number of item : 5 (start with 1)
The number of transaction : 10
Minimum weak support: 0.9
Minimum confidence : 0.8
LHS    RHS    WeakS Conf
5      1      1.0   1.0
4      2      1.0   1.0

time is 47 millisecond.      numRule : 2
```

รูปที่ ข.2: ตัวอย่างผลลัพธ์ที่ได้จากการดำเนินงานของโปรแกรมการสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธี WS

```
...
ex4-0.8-0.9.txt          47 millisecond          numRule : 2
...
```

รูปที่ ข.3: ตัวอย่างแฟ้มเก็บเวลาที่ได้จากการดำเนินงานของโปรแกรมการสร้างหลักเกณฑ์เชื่อมโยงโดยใช้ขั้นตอนวิธี WS

```
import java.sql.*;
import java.io.*;
import java.util.*;
import java.lang.*;
public class WS{
    static Connection connection;
    static PreparedStatement prepared;
    static ResultSet resultSet;
    static FileWriter gText;
    static String directory;
    //launch the application
    public static void main(String args[]){
        long start=0L,elapsed=0L;//variables for compute time
        long numRule=0L; //the number of extracted rules
        int maxSizeLHS=0;//the maximum size of the antecedent set
        String dbName=null,tableName=null,writeFile,temp1=null, temp2=null;

        //declare real variable : min weak support , min confidence, number of data
        double minC=0.0, minWS=0.0, numT=0.0;

        //declare integer variable : min number of item , max number of item
        int minItem=0, maxItem=0;

        //the variable for use the algorithm
        int i,j,RHS,k,numLHS;

        ArrayList<String> LHS = new ArrayList<String>();

        //read the command line argument of the program
        try{
            dbName = args[0];
            tableName = args[1];
            minWS = Double.parseDouble(args[2]);
            minC = Double.parseDouble(args[3]);
            maxSizeLHS = Integer.parseInt(args[4]);
            directory = args[5];
        }catch(ArrayIndexOutOfBoundsException arrayOutOfBound){
            arrayOutOfBound.printStackTrace();
            System.err.println("\nThis program needs the 6 - command line arguments
            : ");
            System.err.println("\t1. (name of database)\n”);
            System.err.println("\t2. (name of table)\n”);
            System.err.println("\t3. (minimum weak support)");
            System.err.println("\t4. (minimum confidence)\n”);
            System.err.println("\t5. (maximum number of LHS)");
            System.err.println("\t6. (the name of directory to save file)\n");
            System.exit(1);
        }

        //start program
```

```
try{
    Class.forName( "com.mysql.jdbc.Driver"); //load jdbc driver class

    //establish connection to database
    connection =DriverManager.getConnection("jdbc:mysql://localhost/"
    +dbName,"root","");

    minItem = findMinItem(tableName);
    maxItem = findMaxItem(tableName);
    numT = findNumTransaction(tableName);

    //create the text file for writing a data.
    writeFile = tableName+"-"+minWS+"-"+minC+".txt";
    gText = new FileWriter(directory+writeFile);
    writeHeadInFile(tableName, maxItem,  minItem, numT, minWS,  minC);
    writeHeadInMonitor(tableName, maxItem,  minItem, numT, minWS,  minC);


    //launch WS algorithm
    numRule = 0;
    System.out.print("RHS = ");
    start = System.currentTimeMillis(); //start the time
    for(RHS = minItem; RHS<= maxItem; RHS++){
        System.out.print("  "+RHS);
        LHS.clear();

        //the first step of the algorithm
        prepared = connection.prepareStatement(generateSQL(tableName,1,
        RHS));
        for(i=minItem; i<=maxItem; i++){
            if(i!=RHS){
                prepared.setInt(1,i);
                numRule += createRule(RHS,LHS,Integer.toString(i),numT
                ,minWS,minC);
            }
        }

        //the second step of the algorithm
        numLHS = LHS.size();
        k=2;
        if(k>maxSizeLHS) continue;
        prepared = connection.prepareStatement(generateSQL(tableName,
        2,RHS));
        for(i=0; i<numLHS-1; i++)
        {   temp1 = LHS.get(i);
            //System.out.println(temp1);
            prepared.setInt(1, Integer.parseInt(temp1));
            for(j=i+1; j<numLHS; j++)
            {   temp2 = LHS.get(j);
                prepared.setInt(2, Integer.parseInt(temp2));
```

```
                temp2 = temp1.concat(",").concat(temp2);
                numRule += createRule(RHS,LHS,temp2,numT,minWS,minC);
            }
        }
        LHS.subList(0,numLHS-1).clear();

        //the third step of the algorithm
        k=3;
        if(k>maxSizeLHS) continue;
        while(LHS.size()>1 && k<=maxSizeLHS)
        {   numLHS = LHS.size();
            prepared = connection.prepareStatement(generateSQL(tableName,
            k,RHS));
            for(i=0; i<numLHS-1; i++)
            {   temp1 = LHS.get(0);
                //System.out.println("Temp 1 = "+temp1);
                for(j=i+1; j<numLHS; j++)
                {       temp2 = LHS.get(j);
                 if(chkSimilarOfTwoItem(temp1,temp2))
                 {      temp2 = setPrepared_getItem(temp1,temp2);
                        numRule += createRule(RHS,LHS,temp2,numT,
                        minWS,minC);
                 }
                 else
                        break;
                }
            }
            LHS.subList(0,numLHS-1).clear();
            k++;
        }
    }
    System.out.println("\nNumrule = "+numRule+"\n");

    if(gText != null)
    {   elapsed = System.currentTimeMillis() - start;
        writeTimeInTimeFile(writeFile,elapsed,numRule);
        gText.write("\n\ntime is "+elapsed+" millisecond.\tnumRule :
        "+numRule+"\n");
        gText.close();
        System.out.println("\nThe file : "+writeFile+" is created.\n\n");
    }
}
catch(SQLException sqlException){
    sqlException.printStackTrace();
    System.exit(1);
}
catch(SecurityException securityException){
    System.err.println("You do not have write access to this file");
    System.exit(1);
}
```

```
    catch(FileNotFoundException fileNotfound){
        System.err.println("Error opening file.");
        System.exit(1);
    }
    catch(ArrayIndexOutOfBoundsException arrayOutOfBound){
        arrayOutOfBound.printStackTrace();
        System.exit(1);
    }
    catch(IOException ioException){
        ioException.printStackTrace();
        System.err.println("Cannot append the file\n");
        System.exit(1);
    }
    catch(ClassNotFoundException classNotFound){
        classNotFound.printStackTrace();
        System.exit(1);
    }
    catch(Exception exception){
        exception.printStackTrace();
        System.exit(1);
    }
    finally{
        try{
            prepared.close();
            connection.close();
        }
        catch(Exception exception){
            exception.printStackTrace();
            System.exit(1);
        }
    }
    System.exit(0);
}//end of main program

public static void writeHeadInFile(String tableName, int maxItem, int minItem,
double numT, double minWS, double minC) throws IOException{
    gText.write("Table name : "+tableName+"\n");
    gText.write("The number of item : "+(maxItem-minItem+1)+" (start with
     "+minItem+")\n");
    gText.write("The number of transaction : "+(int)numT+"\n");
    gText.write("Minimum weak support: "+minWS+"\n");
    gText.write("Minimum confidence : "+minC+"\n");
    gText.write("LHS \t RHS \t WeakS \t Conf  \n");
}

public static void writeHeadInMonitor(String tableName, int maxItem, int minItem,
double numT, double minWS, double minC){
    System.out.println("\nTable name : "+tableName);
    System.out.println("The number of item : "+(maxItem-minItem+1)+" (start with
    "+minItem+")");
```

```
    System.out.println("The number of transaction : "+(int)numT);
    System.out.println("Minimum weak support: "+minWS);
    System.out.println("Minimum confidence : "+minC);
}

public static int findMinItem(String tableName)
throws SQLException, Exception{
    String temp;
    prepared = connection.prepareStatement("desc "+tableName);
    resultSet = prepared.executeQuery();
    resultSet.absolute(2);
    temp = resultSet.getString("Field");
    temp = temp.substring(1);
    return Integer.parseInt(temp);
}

public static int findMaxItem(String tableName)
throws SQLException, Exception{
    String temp;
    prepared = connection.prepareStatement("desc "+tableName);
    resultSet = prepared.executeQuery();
    resultSet.afterLast();
    resultSet.previous();
    temp = resultSet.getString("Field");
    temp = temp.substring(1);
    return Integer.parseInt(temp);
}
public static double findNumTransaction(String tableName)
throws SQLException, Exception{
    prepared = connection.prepareStatement("select count(tid) from "+tableName);
    //create statement for querying database
    resultSet = prepared.executeQuery();//query database
    resultSet.next();
    return resultSet.getDouble(1);
}

public static void writeTimeInTimeFile(String fileName, long timeMilliSecond,
long numRule) throws IOException, Exception{
    FileWriter timeText;
            //open the file "time.txt"  for append
    timeText =new FileWriter(directory+"time.txt",true);
    timeText.append(fileName+"\t");
    timeText.append(timeMilliSecond+" millisecond \tnumRule : "+numRule+" \n");
    timeText.close();
}
```

```
public static void writeRuleInFile(String LHS, int RHS, double weakS, double conf)
throws IOException{
    gText.write(LHS+"\t"+RHS+"\t"+weakS+"\t"+conf+"\n");
    //gText.write(LHS+"\t"+RHS+"\n");
}

public static double findWeakS(double numWS, double numT){
    //return : the weak support value
    return 1.0-(numWS/numT);
}

public static double findConf(double numWS, double numS){
//return : the confidence value
    double numA = numWS+numS;
    if(numA!=0)
        return numS/numA;
    else return 0;
}

public static String generateSQL(String tableName, int k, int RHS){
    String temp = "select i"+RHS+", count(tid) from ";
    temp = temp.concat(tableName).concat(" where i? = 1 ");
    for(int i=2; i<=k; i++)
    {   temp = temp.concat(" and i? = 1 ");
    }
    temp = temp.concat(" group by i"+RHS);
    return temp;
}

public static long createRule(int RHS,ArrayList<String> LHS, String temp,
double numT, double minWS, double minC) throws IOException, SQLException{
    double weakS, conf,numWS=0,numS=0;
    int chk;
    resultSet = prepared.executeQuery();
    while(resultSet.next())
    {   chk = resultSet.getInt(1);
        if(chk==0)
            numWS = resultSet.getDouble(2);
        else if(chk==1)
            numS = resultSet.getDouble(2);
    }
    weakS = findWeakS(numWS, numT);
    conf = findConf(numWS, numS);
    if(weakS>=minWS && conf>=minC){
            LHS.add(temp);
            writeRuleInFile(temp,RHS,weakS,conf);
            return 1L;
    }
    else return 0L;
}
```

```
    public static boolean chkSimilarOfTwoItem(String temp1, String temp2)
    {   boolean chk=true;
        String item1[], item2[];
        item1 = temp1.split(",");
        item2 = temp2.split(",");
        for(int i=0; i<item1.length-1; i++)
        {    if(!item1[i].equals(item2[i]))
             {    chk = false;
                  break;
             }
        }
        return chk;
    }

    public static String setPrepared_getItem(String temp1, String temp2)
    throws SQLException{
        String item1[], item2[];
        int itemSize;
        //System.out.println("First String = "+temp1);
        item1 = temp1.split(",");
        item2 = temp2.split(",");
        itemSize = item1.length;
        //System.out.println("Item size = "+itemSize);
        for(int i=0; i<itemSize; i++)
             prepared.setInt(i+1, Integer.parseInt(item1[i]));
        prepared.setInt(itemSize+1, Integer.parseInt(item2[itemSize-1]));
        temp2 = temp1.concat(",").concat(item2[itemSize-1]);
        return temp2;
    }

}
```

ภาคผนวก ค

ผลการทดลอง

การทดลองเพื่อเปรียบเทียบขั้นตอนวิธี WS กับขั้นตอนวิธีเอไพออรี จะมีการทดลอง 2 ลักษณะ คือ การทดลองลักษณะที่ 1 ซึ่งเป็นการเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ที่สร้างจากทั้งสองขั้นตอนวิธี โดยแบ่งข้อมูลที่ใช้ทดลองออกเป็น 2 ส่วน คือ ข้อมูลทดลองและข้อมูลทดสอบ และการทดลองลักษณะที่ 2 ซึ่งเป็นการเปรียบเทียบประสิทธิภาพของหลักเกณฑ์ที่มีค่าสภาพไวสูงสุดของทั้งสองขั้นตอนวิธี ในการทดลองนี้จะแบ่งข้อมูลออกเป็น 3 ส่วน คือ ข้อมูลทดลอง ข้อมูลตรวจสอบความสมเหตุสมผล และข้อมูลทดสอบ โดยผลการทดลองจากทั้งสองลักษณะจะแสดงในรูปแบบตารางดังต่อไปนี้

ผลการทดลองลักษณะที่ 1

เมื่อนำข้อมูลส่วนแรกซึ่งเป็นข้อมูลทดลองของ chess และ mushroom จะได้ผลการทดลองดังตารางที่ ค.1 ถึงตารางที่ ค.6 ซึ่งผลการทดลองของข้อมูล chess แสดงดังตารางที่ ค.1 ถึงตารางที่ ค.3 และผลการทดลองของข้อมูล mushroom แสดงดังตารางที่ ค.4 ถึงตารางที่ ค.6 เมื่อ *minC* คือค่าความเชื่อมั่นต่ำสุด *minS* คือค่าสนับสนุนต่ำสุด *minWS* คือค่าสนับสนุนแบบอ่อนต่ำสุด เวลาคือเวลาที่ใช้ในการทดลองมีหน่วยเป็นนาที #หลักเกณฑ์ คือจำนวนหลักเกณฑ์ที่ได้จากการทดลอง และสภาพไวคือค่าสภาพไวเฉลี่ยของกลุ่มหลักเกณฑ์ที่สร้างขึ้นโดยเปรียบเทียบกับข้อมูลทดสอบ

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	0.02	1,291	0.8542	0.02	1,298	0.9239
	0.2	0.95	0.01	1,170	0.8543	0.02	1,489	0.8998
	0.3	0.93	0.01	1,006	0.8560	0.02	1,559	0.8885
	0.4	0.90	0.01	882	0.8535	0.02	1,618	0.8826
	0.5	0.88	0.01	827	0.8546	0.02	1,674	0.8723
0.9	0.1	0.99	0.02	1,030	0.8985	0.02	1,024	0.9435
	0.2	0.98	0.01	932	0.8994	0.02	1,228	0.9297
	0.3	0.97	0.01	807	0.9004	0.02	1,317	0.9175
	0.4	0.96	0.01	692	0.8995	0.02	1,382	0.9099
	0.5	0.94	0.01	651	0.9003	0.02	1,394	0.9085
1.0	0.1	1.00	0.02	320	0.98	0.02	650	0.97
	0.2		0.01	270	0.98			
	0.3		0.01	213	0.99			
	0.4		0.01	170	0.99			
	0.5		0.01	165	0.99			

ตารางที่ ค.1: การทดสอบลักษณะที่ 1 สำหรับข้อมูล chess เมื่อ *maxLHS* =1

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	0.59	31,830	0.8516	1.21	30,457	0.9407
	0.2	0.95	0.44	25,310	0.8516	1.37	38,280	0.9135
	0.3	0.93	0.31	19,303	0.8526	1.43	40,854	0.9028
	0.4	0.90	0.24	15,784	0.8504	1.47	43,355	0.8985
	0.5	0.88	0.13	13,333	0.8512	1.53	45,759	0.8872
0.9	0.1	0.99	0.59	25,626	0.8965	0.58	20,496	0.9595
	0.2	0.98	0.44	20,317	0.8977	1.19	29,631	0.9425
	0.3	0.97	0.31	15,483	0.8986	1.27	33,526	0.9274
	0.4	0.96	0.24	12,475	0.8973	1.34	36,668	0.9170
	0.5	0.94	0.12	10,594	0.8976	1.34	36,982	0.9165
1	0.1	1.00	0.58	9,330	0.9782	0.35	10,833	0.9837
	0.2		0.44	6,566	0.9820			
	0.3		0.31	4,455	0.9847			
	0.4		0.24	3,297	0.9850			
	0.5		0.11	2,757	0.9854			

ตารางที่ ค.2: การทดสอบลักษณะที่ 1 สำหรับข้อมูล chess เมื่อ maxLHS =2

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	10.21	454,708	0.8493	22.21	530,819	0.9437
	0.2	0.95	6.34	312,884	0.8488	28.11	688,292	0.9154
	0.3	0.93	4.15	215,918	0.8488	29.56	732,143	0.9058
	0.4	0.90	3.00	162,901	0.8476	31.30	779,852	0.9030
	0.5	0.88	2.08	121,838	0.8487	33.04	825,111	0.8931
0.9	0.1	0.99	10.17	368,030	0.8949	14.03	329,205	0.9619
	0.2	0.98	6.32	251,694	0.8961	21.57	523,371	0.9446
	0.3	0.97	4.14	172,753	0.8966	24.59	601,816	0.9297
	0.4	0.96	3.00	129,753	0.8952	27.28	670,323	0.9187
	0.5	0.94	2.08	98,013	0.8950	27.30	673,550	0.9186
1	0.1	1.00	10.15	147,483	0.9763	7.46	178,833	0.9850
	0.2		6.30	86,550	0.9812			
	0.3		4.12	52,686	0.9840			
	0.4		2.58	36,310	0.9845			
	0.5		2.07	26,422	0.9853			

ตารางที่ ค.3: การทดสอบลักษณะที่ 1 สำหรับข้อมูล chess เมื่อ *maxLHS* =3

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	0.02	452	0.6647	0.03	917	0.7745
	0.2	0.95	0.01	277	0.6717	0.02	958	0.7619
	0.3	0.93	0.01	209	0.6735	0.03	988	0.7399
	0.4	0.90	0.01	161	0.6493	0.02	1,005	0.7356
	0.5	0.88	0.01	131	0.6683	0.03	1,014	0.7263
0.9	0.1	0.99	0.01	334	0.7864	0.02	863	0.8068
	0.2	0.98	0.01	190	0.8350	0.02	871	0.7995
	0.3	0.97	0.01	140	0.8300	0.02	888	0.7906
	0.4	0.96	0.01	107	0.8215	0.02	888	0.7906
	0.5	0.94	0.01	91	0.8310	0.02	896	0.7899
1	0.1	1.00	0.01	249	0.9159	0.02	778	0.8624
	0.2		0.01	138	0.9441			
	0.3		0.01	93	0.9521			
	0.4		0.01	70	0.9517			
	0.5		0.01	58	0.9522			

ตารางที่ ค.4: การทดสอบลักษณะที่ 1 สำหรับข้อมูล mushroom เมื่อ *maxLHS* =1

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	0.59	7,298	0.6274	1.34	11,781	0.8880
	0.2	0.95	0.44	3,328	0.6248	1.39	12,659	0.8774
	0.3	0.93	0.31	2,020	0.6270	1.42	13,474	0.8614
	0.4	0.90	0.24	1,385	0.5975	1.45	14,056	0.8531
	0.5	0.88	0.13	824	0.6345	1.46	14,417	0.8423
0.9	0.1	0.99	0.59	5,870	0.7251	1.3	11,061	0.9003
	0.2	0.98	0.44	2,517	0.7607	1.3	11,181	0.8980
	0.3	0.97	0.31	1,520	0.7623	1.32	11,506	0.8930
	0.4	0.96	0.24	1,027	0.7457	1.31	11,506	0.8930
	0.5	0.94	0.12	608	0.7788	1.33	11,689	0.8927
1	0.1	1.00	0.58	4,529	0.8754	1.22	9,597	0.9284
	0.2		0.44	1,740	0.9199			
	0.3		0.31	977	0.9360			
	0.4		0.24	626	0.9334			
	0.5		0.11	366	0.9440			

ตารางที่ ค.5: การทดสอบลักษณะที่ 1 สำหรับข้อมูล mushroom เมื่อ *maxLHS* =2

minC	*minS*	*minWS*	ขั้นตอนวิธีเอไพออรี			ขั้นตอนวิธี WS		
			เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.98	2.42	54,243	0.5895	11.36	125,003	0.9241
	0.2	0.95	0.54	20,245	0.5759	12.02	130,480	0.9196
	0.3	0.93	0.07	9,734	0.5762	12.4	137,171	0.9124
	0.4	0.90	0.02	6,493	0.5447	13.18	143,069	0.9065
	0.5	0.88	0.01	2,686	0.5977	13.32	147,281	0.8994
0.9	0.1	0.99	2.37	46,077	0.6747	11.2	121,801	0.9272
	0.2	0.98	0.54	16,435	0.6930	11.17	122,269	0.9266
	0.3	0.97	0.15	7,934	0.6985	11.27	123,878	0.9252
	0.4	0.96	0.02	5,280	0.6791	11.3	123,878	0.9252
	0.5	0.94	0.01	2,074	0.7339	11.36	124,900	0.9253
1	0.1	1.00	2.37	36,858	0.8372	9.38	101,978	0.9517
	0.2		0.54	11,562	0.8892			
	0.3		0.19	5,144	0.9136			
	0.4		0.02	3,259	0.9067			
	0.5		0.01	1,195	0.9353			

ตารางที่ ค.6: การทดสอบลักษณะที่ 1 สำหรับข้อมูล mushroom เมื่อ *maxLHS* =3

ผลการทดลองลักษณะที่ 2

เมื่อนำข้อมูลส่วนแรกซึ่งเป็นข้อมูลทดลองของ chess และ mushroom มาผ่านขั้นตอนวิธีเอไพออรีจะได้ผลดังตารางที่ ค.7 ตารางที่ ค.8 และ ตารางที่ ค.9 และเมื่อใช้ขั้นตอนวิธี WS จะได้ผลดังตารางที่ ค.10 ตารางที่ ค.11 และตารางที่ ค.12 เมื่อ *minC* คือค่าความเชื่อมั่นต่ำสุด *minS* คือค่าสนับสนุนต่ำสุด *minWS* คือค่าสนับสนุนแบบอ่อนต่ำสุด เวลาคือเวลาที่ใช้ในการทดลองมีหน่วยเป็นนาที #หลักเกณฑ์ คือจำนวนหลักเกณฑ์ที่ได้จากการทดลอง ซึ่งค่าสภาพไวที่ได้ในตารางเป็นค่าสภาพไวเฉลี่ยที่เกิดจากการนำหลักเกณฑ์ที่สร้างขึ้นมาเปรียบเทียบกับข้อมูลตรวจสอบความสมเหตุสมผล

minC	*minS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.02	1,275	0.8722	0.02	543	0.7054
	0.2	0.02	1,193	0.8713	0.01	330	0.6985
	0.3	0.02	1,016	0.8723	0.01	221	0.7262
	0.4	0.01	894	0.8698	0.01	161	0.7092
	0.5	0.01	862	0.8707	0.01	120	0.7431
0.9	0.1	0.02	1,053	0.9066	0.02	427	0.7783
	0.2	0.02	981	0.9071	0.01	234	0.7900
	0.3	0.02	841	0.9076	0.01	160	0.8057
	0.4	0.01	728	0.9059	0.01	113	0.8001
	0.5	0.01	703	0.9067	0.01	80	0.8830
1	0.1	0.02	302	0.9922	0.02	345	0.9544
	0.2	0.02	267	0.9940	0.01	176	0.9654
	0.3	0.02	217	0.9946	0.01	117	0.9746
	0.4	0.01	167	0.9947	0.01	82	0.9733
	0.5	0.01	159	0.9947	0.01	65	0.9740

ตารางที่ ค.7: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* =1

minC	*minS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	0.51	31,013	0.8712	0.37	8,664	0.6684
	0.2	0.39	25,686	0.8701	0.03	3,791	0.6595
	0.3	0.28	19,252	0.8704	0.02	2,106	0.6851
	0.4	0.05	16,137	0.8680	0.02	1,283	0.6745
	0.5	0.05	14,148	0.8697	0.01	770	0.7061
0.9	0.1	0.48	25,532	0.9098	0.05	7,261	0.7159
	0.2	0.39	21,106	0.9107	0.03	2,948	0.7210
	0.3	0.28	15,767	0.9112	0.02	1,644	0.7373
	0.4	0.18	13,042	0.9095	0.02	987	0.7399
	0.5	0.05	11,537	0.9099	0.01	551	0.8197
1	0.1	0.49	8,883	0.9899	0.05	6,054	0.9359
	0.2	0.39	6,640	0.9930	0.03	2,276	0.9482
	0.3	0.27	4,496	0.9938	0.02	1,199	0.9591
	0.4	0.22	3,340	0.9940	0.02	741	0.9607
	0.5	0.07	2,878	0.9941	0.01	435	0.9664

ตารางที่ ค.8: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* =2

minC	*minS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.1	8.50	442,358	0.6364	2.44	65,042	0.8699
	0.2	5.54	316,540	0.6260	0.51	21,408	0.8689
	0.3	3.43	215,594	0.6461	0.05	10,260	0.8682
	0.4	2.48	169,951	0.6463	0.03	5,117	0.8668
	0.5	2.04	131,253	0.6683	0.02	2,537	0.8690
0.9	0.1	8.49	363,744	0.6720	2.41	56,699	0.9118
	0.2	5.53	259,708	0.6693	0.51	17,654	0.9131
	0.3	3.43	174,805	0.6817	0.15	8,487	0.9139
	0.4	2.47	137,312	0.6977	0.03	4,165	0.9124
	0.5	2.04	107,765	0.7618	0.02	1,900	0.9118
1	0.1	8.45	142,828	0.9191	2.41	49,285	0.9880
	0.2	5.51	89,259	0.9326	0.51	14,241	0.9918
	0.3	3.41	53,589	0.9430	0.18	6,428	0.9930
	0.4	2.47	38,396	0.9485	0.03	3,276	0.9933
	0.5	2.03	28,842	0.9584	0.02	1,493	0.9936

ตารางที่ ค.9: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีเอไพออรี เมื่อ *maxLHS* =3

minC	*minWS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.6	0.02	1,799	0.8754	0.03	543	0.7625
	0.7	0.02	1,799	0.8754	0.03	330	0.7625
	0.8	0.02	1,799	0.8754	0.03	221	0.7625
	0.9	0.02	1,675	0.8935	0.03	161	0.7783
	1.0	0.02	665	0.9864	0.03	120	0.9436
0.9	0.6	0.02	1,529	0.9089	0.03	427	0.8316
	0.7	0.02	1,529	0.9089	0.03	234	0.8316
	0.8	0.02	1,529	0.9089	0.03	160	0.8316
	0.9	0.02	1,529	0.9089	0.03	113	0.8316
	1.0	0.02	665	0.9864	0.03	80	0.9436
1	0.6	0.02	665	0.9864	0.03	345	0.9436
	0.7	0.02	665	0.9864	0.03	176	0.9436
	0.8	0.02	665	0.9864	0.03	117	0.9436
	0.9	0.02	665	0.9864	0.03	82	0.9436
	1.0	0.02	665	0.9864	0.03	65	0.9436

ตารางที่ ค.10: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีWS เมื่อ *maxLHS* =1

minC	*minWS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.6	1.49	51,135	0.8818	1.27	14,524	0.8448
	0.7	1.49	51,135	0.8818	1.28	14,524	0.8448
	0.8	1.49	51,135	0.8818	1.28	14,524	0.8448
	0.9	1.42	45,799	0.9032	1.23	13,188	0.8843
	1.0	0.32	10,916	0.9939	1.07	9,410	0.9712
0.9	0.6	1.33	41,610	0.9215	1.13	10,893	0.9283
	0.7	1.33	41,610	0.9215	1.14	10,893	0.9283
	0.8	1.33	41,610	0.9215	1.14	10,893	0.9283
	0.9	1.33	41,610	0.9215	1.14	10,893	0.9283
	1.0	0.32	10,916	0.9939	1.08	9,410	0.9712
1	0.6	0.32	10,916	0.9939	1.08	9,410	0.9712
	0.7	0.32	10,916	0.9939	1.08	9,410	0.9712
	0.8	0.32	10,916	0.9939	1.08	9,410	0.9712
	0.9	0.32	10,916	0.9939	1.08	9,410	0.9712
	1.0	0.32	10,916	0.9939	1.08	9,410	0.9712

ตารางที่ ค.11: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีWS เมื่อ *maxLHS* =2

minC	*minWS*	ข้อมูล chess			ข้อมูล mushroom		
		เวลา	#หลักเกณฑ์	สภาพไว	เวลา	#หลักเกณฑ์	สภาพไว
0.8	0.6	31.36	934,686	0.8863	11.25	154,563	0.8945
	0.7	31.54	934,686	0.8863	11.26	154,563	0.8945
	0.8	31.53	934,686	0.8863	11.24	154,563	0.8945
	0.9	29.26	831,508	0.9066	10.14	134,678	0.9411
	1.0	6.56	177,441	0.9948	8.15	107,624	0.9764
0.9	0.6	26.29	748,259	0.9304	8.47	114,990	0.9641
	0.7	26.25	748,259	0.9304	8.49	114,990	0.9641
	0.8	26.28	748,259	0.9304	8.47	114,990	0.9641
	0.9	26.28	748,259	0.9304	8.48	114,990	0.9641
	1.0	6.54	177,441	0.9948	8.15	107,624	0.9764
1	0.6	6.54	177,441	0.9948	8.17	107,624	0.9764
	0.7	6.54	177,441	0.9948	8.17	107,624	0.9764
	0.8	6.54	177,441	0.9948	8.17	107,624	0.9764
	0.9	6.54	177,441	0.9948	8.17	107,624	0.9764
	1.0	6.54	177,441	0.9948	8.17	107,624	0.9764

ตารางที่ ค.12: การทดสอบลักษณะที่ 2 ของขั้นตอนวิธีWS เมื่อ *maxLHS* =3

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวชรีย์วรรณ สิริศรีสัมฤทธิ์ เกิดเมื่อวันที่ 7 กันยายน พุทธศักราช 2526 สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ จากภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ เมื่อปีการศึกษา 2547 และเข้าศึกษาต่อในหลักสูตรปริญญามหาบัณฑิต สาขาวิทยาการคณนา ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เมื่อปีการศึกษา 2547