การพัฒนาขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นเพื่อรักษาความเป็นส่วนตัวของการจำแนกแบบความสัมพันธ์

Incremental Data Transformation Algorithm Development for Privacy Preserving of Associative Classification

นายบวรศักดิ์ ศรีสังสิทธิสันติ

รหัส 500631109

หัวข้อโครงร่างนี้เสนอต่อบัณฑิตวิทยาลัยเพื่อขออนุมัติ
สำหรับทำวิทยานิพนธ์ตามหลักสูตรปริญญา
วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย
มหาวิทยาลัยเชียงใหม่
กุมภาพันธ์ กุมภาพันธ์2552

แบบเสนอหัวข้อและโครงร่างวิทยานิพนธ์

1. ชื่อ/สกุล/รหัส นายบวรศักดิ์ ศรีสังสิทธิสันติ รหัส 500631109

2. ชื่อเรื่องวิทยานิพนธ์

การพัฒนาขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นเพื่อรักษาความเป็นส่วนตัวของการจำแนกแบบความสัมพันธ์

Incremental Data Transformation Algorithm Development for Privacy Preserving of Associative Classification

3. อาจารย์ที่ปรึกษา ดร.จักรพงศ์ นาทวิชัย

4. หลักการและเหตุผล

4.1 ปัญหาและที่มาของการศึกษา

ปัญหา ความเป็นส่วนตัวของข้อมูล

ผลจากการสำรวจเว็บไซต์ในประเทศไทยจาก [1] พบว่าข้อมูลที่ถูกเก็บรวบรวมและเกี่ยวข้องกับเว็บไซต์นั้นมี ตัวบ่งชี้บุคคล เช่น รหัสประจำตัวประชาชน หรือ ชื่อ-สกุล (Identifier) ถึง 64% และมีตัวบ่งชี้บุคคลทางอ้อม เช่น ที่อยู่ อายุ หรือ อาชีพ (quasi-identifier) 21% ไม่เพียงเท่านั้น 1 ใน 3 ข้อมูลเหล่านั้นยังถูกองค์กรที่รวบรวมข้อมูลนำไปตรวจสอบกับแหล่งข้อมูลต้นฉบับว่าเป็นข้อมูลจริงหรือไม่ (Verified Data) และจากการสำรวจยังพบว่า 53% ของนโยบายความเป็นส่วนตัวของข้อมูล (Privacy Policy) มีนโยบายที่อาจเผยแพร่ข้อมูลส่วนบุคคลโดยไม่มีการขออนุญาตจากเข้าเจ้าของข้อมูล และ 82% ยังเป็นนโยบายที่ไม่มีนโยบายความปลอดภัยของข้อมูล (Security Policy) ผลสรุปจากการสำรวจจะเห็นได้ว่าในประเทศไทยนั้นยังไม่ได้ให้ความสนใจในเรื่องของปัญหาความเป็นส่วนตัวของข้อมูลเท่าที่ควร

จาก [1] ผู้เขียนให้ข้อคิดเห็นว่าปัญหาความเป็นส่วนตัวของข้อมูลสามารถแก้ไขได้โดยมีแนวทางหลักในการแก้ปัญหาจำนวน 3 แนวทาง คือ 1) ออกกฎหมายเพื่อแก้ปัญหาดังกล่าว 2) เพิ่มความรับผิดชอบของผู้เก็บข้อมูล 3) การนำเทคโนโลยีเข้ามาช่วยในการแก้ปัญหา ในโครงร่างวิทยานิพนธ์นี้จะกล่าวถึงแนวทางที่ 3 เท่านั้น เนื่องจากแนวทางที่ 1 และ 2 นั้นอยู่นอกเหนือขอบเขตงานวิจัยของวิศวกรรมคอมพิวเตอร์

ในการแก้ปัญหาความเป็นส่วนตัวของข้อมูลโดยแนวทางที่ 3 วิธีการที่ง่ายที่สุดที่สามารถปฏิบัติได้ก่อนที่จะทำการเผยแพร่ข้อมูลคือ การลบข้อมูลที่เป็นตัวบ่งชี้บุคคลออกก่อน เพื่อให้ข้อมูลในแต่ละระเบียนนั้นไม่มีการชี้ตัวบุคคล วิธีการนี้ขึ้นอยู่กับการเลือกคอลัมน์ ให้ถูกต้องว่าคอลัมน์ใดเป็นตัวบ่งชี้บุคคล

ในตารางที่ 1 แสดงตัวอย่างรายละเอียดโครงสร้างข้อมูลของแฟ้มที่ 2 ของข้อมูลสาธารณสุข 12 แฟ้มชื่อ PAT (ดูรายละเอียดเพิ่มเติมในภาคผนวก 1) ซึ่งเป็นชุดข้อมูลมาตรฐานของการประกันสุขภาพของประเทศไทย ตัวอย่างข้อมูลของโครงสร้างข้อมูลดังกล่าวแสดงในตารางที่ 2

ตารางที่ 1 โครงสร้างข้อมูลของแฟ้มที่ 2

FIELD NAME	TYPE	LENGTH	DECIMAL	QUALIFICATION
HCODE	C	5	0	รหัสสถานพยาบาล (Left justified)
HN	C	9	0	หมายเลขประจำตัวผู้รับบริการ ควรใช้หมายเลขเดิมให้นานกว่า 5 ปี (Left justified)
CHANGWAT	C	2	0	ตามรหัสมหาดไทย
AMPHUR	C	2	0	ตามรหัสมหาดไทย
DOB	D	8	0	บันทึกวันเดือนปีเกิด ปีมีค่าเป็น ค.ศ.
SEX	C	1	0	1 หมายถึง เพศชาย 2 หมายถึง เพศหญิง
MARRIAGE	C	1	0	รหัสสภาพสมรส
OCCUPA	C	3	0	อาชีพ
NATION	C	2	0	สัญชาติ
PERSON_ID	C	13	0	รหัสประจำตัวประชาชนตามสำนักทะเบียนราษฎร์

ตารางที่ 2 ตัวอย่างการลบข้อมูลที่เป็นตัวบ่งชี้บุคคล

HCODE	HN	CHANGWAT	AMPHUR	DOB	SEX	MARRIAGE	OCCUPA	NATION	PERSONID
12	Null	48	1	2/01/1977	1	1	111	99	null
12	Null	48	1	5/02/1970	1	1	112	99	null
12	Null	48	1	9/06/1968	2	3	113	99	null

จากตารางที่ 1 ซึ่งเป็นโครงสร้างที่ใช้จริงในโรงพยาบาล ข้อมูลที่อยู่ในตารางเป็นค่ารหัสซึ่งสามารถตรวจสอบค่าจริงได้จากภาคผนวก 1 จากโครงสร้างดังกล่าวนี้ในการระบุว่าคอลัมน์ใดเป็นตัวบ่งชี้บุคคลอาจขึ้นอยู่กับการวิเคราะห์ของแต่ละงาน ซึ่งในแฟ้มที่ 2 พิจารณาว่า PERSON_ID และ HN เป็นตัวบ่งชี้ เพราะค่าในแต่ละระเบียนไม่มีการซ้ำกัน จากที่กล่าวมาถ้าทำการลบคอลัมน์ PERSON_ID และ HN ออกจะไม่สามารถชี้ตัวบุคคลได้ ฉะนั้นข้อมูลหลังการลบจึงสามารถเผยแพร่ได้ แต่นั้นหมายถึงการพิจารณาเพียงข้อมูลจากตารางนี้เท่านั้น ซึ่งวิธีการดังกล่าวยังไม่สามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้ (Re-Identification) การระบุตัวบุคคลอีกครั้งนั้นสามารถทำได้โดยการหาข้อมูลจากแหล่งอื่นมาทำการเปรียบเทียบกับข้อมูลในคอลัมน์อื่นที่เป็นตัวบ่งชี้บุคคลทางอ้อมของตารางที่ถูกปกปิดตัวบ่งชี้บุคคลไปแล้ว โดยข้อมูลจากแหล่งอื่นนั้นอาจเป็นข้อมูลที่มีตัวบ่งชี้บุคคล และถ้าข้อมูลทั้งสองแหล่งมีการเหลื่อมซ้อนกัน (Overlap) อาจทำให้ข้อมูลที่ได้ทำการปกปิดไปแล้วนั้นสามารถกลับมาระบุตัวบุคคลอีกครั้งได้

จากการศึกษาข้อมูลสาธารณสุข 12 แฟ้มซึ่งเป็นชุดข้อมูลมาตรฐานของการประกันสุขภาพของประเทศไทย สามารถแสดงให้เห็นว่าหลังจากการตัดตัวบ่งชี้บุคคลออกแล้ว ข้อมูลยังสามารถถูกระบุตัวบุคคลอีกครั้งได้ ให้พิจารณาจากข้อมูลสมมุติต่อไปนี้ สมมุติให้หน่วยงานประกันสังคมต้องการเปิดเผยข้อมูลบางส่วนให้นักวิจัยนำข้อมูลมาวิเคราะห์เพื่อการเฝ้าระวังการทุจริตของผู้ประกันตน โดยข้อมูลอาจเป็นไปตามตารางที่ 3

ตารางที่ 3 ข้อมูลจากประกันสังคม

ชื่อ-นามสกุล	ชื่อ นามสกุล ของผู้ประกันตนแล้วมีการเบิกเงิน
อาชีพ	อาชีพของผู้ประกันตน
วันที่เบิก	วันที่ยื่นเรื่องทำการเบิก
จำนวนเงิน	จำนวนเงินที่จ่าย

จากการพิจารณาจะเห็นว่าข้อมูลเหล่านี้ ดูเหมือนไม่ได้เป็นข้อมูลที่มีความสำคัญ การเปิดเผยข้อมูลเหล่านี้ มีความเป็นไปได้ที่จะเกิดขึ้น เมื่อกลับมาพิจารณาข้อมูลสาธารณสุข 12 แฟ้ม ทุกแฟ้มสามารถเชื่อมต่อกันโดยคอลัมน์ HN (Hospital number, หมายเลขประจำตัวผู้ป่วย) หรือ คอลัมน์ AN (หมายเลขประจำตัวผู้ป่วยใน) แต่เนื่องจาก HN เป็นตัวบ่งชี้บุคคลทำให้ถูกลบค่าทิ้งไป ส่วน AN นั้นไม่ได้เป็นตัวบ่งชี้บุคคลเนื่องจากสามารถมีระเบียนหลายระเบียนที่มีค่า AN เหมือนกันได้ ดังเช่นในตัวอย่างรูปที่ 1 ฉะนั้นสามารถใช้ AN เพื่อเข้าถึงทุกแฟ้มที่เกี่ยวข้อง ตัวอย่างเช่นข้อมูลการวินิจฉัยโรค ข้อมูลการผ่าตัด ข้อมูลการรักษาเป็นต้น ฉะนั้นจากข้อมูลทั้งสองแหล่งถ้านำข้อมูลมาวิเคราะห์แล้วเชื่อมโยงจะสามารถเปรียบเทียบ ให้ AN กลับมาเป็นชื่อและนามสกุลได้ แล้วเชื่อมโยงไปถึงข้อมูลที่ต้องการได้

ต่อไปนี้เป็นตัวอย่างการวิเคราะห์และเชื่อมโยง ให้พิจารณาข้อมูลแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูลการเงินซึ่งมีโครงสร้างตามตารางที่ 5 มีคอลัมน์ PTTYPE เป็นสิทธิการรักษาซึ่งมีค่ารหัสตามตารางที่ 4

จากการพิจารณาสามารถรู้ได้ว่า AN ที่มี PTTYPE เป็น AI, AJ, AK และ AL อาจมีข้อมูลตรงกับข้อมูลจากประกันสังคม และเนื่องจากแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูลการเงินมีคอลัมน์ DATE, TOTAL (จำนวนเงินค่ารักษา) และ PAID (จำนวนเงินที่ผู้ป่วยจ่ายเอง) จะเห็นว่าถ้าข้อมูลจำนวนเงินที่มาจากประกันสังคมมีจำนวนเงินตรงกับ TOTAL ที่มาจากข้อมูลแฟ้มที่ 11 เป็นไปได้ว่าเป็นการเบิกเงินจากประกันสังคมทั้งหมดหรือถ้าข้อมูลจำนวนเงินที่มาจากประกันสังคมตรงกับค่า TOTAL ลบด้วยค่า PAID แสดงว่าค่า PAID เป็นการจ่ายส่วนต่างที่ประกันสังคมไม่ได้ครอบคลุมถึง แล้วยังสามารถพิจารณาได้อีกว่าเกิดขึ้นในเดือนเดียวกันหรือไม่ กล่าวคือวันที่เบิกในข้อมูลประกันสังคมกับค่า DATE ในข้อมูลแฟ้มที่ 11 เป็นเดือนเดียวกันหรือไม่เพราะระบบการเบิกประกันสังคมนั้นต้องเบิกภายในเดือนเดียวกัน กับการเรียกชำระเงินจากทางสถานพยาบาล

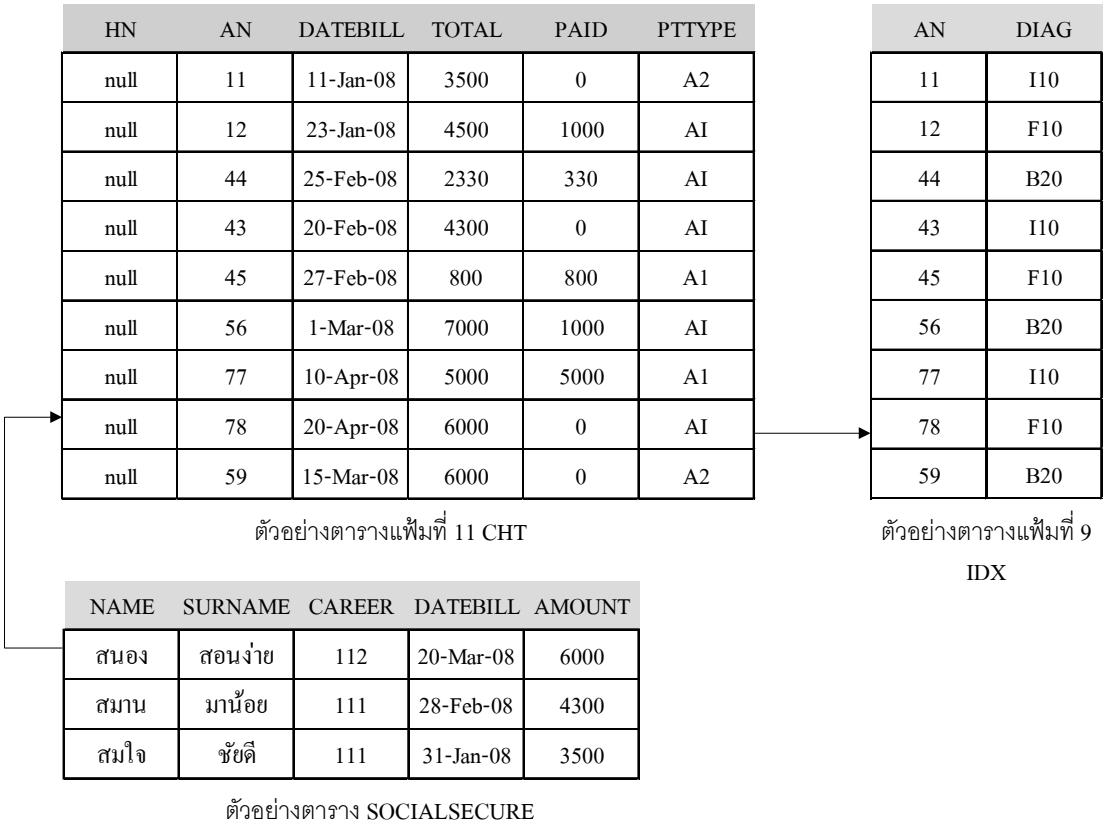
ตารางที่ 4 ตัวอย่างการให้รหัส (ยังไม่ใช้เป็นทางการ)

รหัส	สิทธิการรักษา
A1	ชำระเงินเองโดยไม่มีสิทธิเบิกคืน
A2	ใช้สิทธิเบิกหน่วยงานต้นสังกัดราชการ
A3	สิทธิลดหย่อนประเภท ก. *
A4	สิทธิลดหย่อนประเภท ข. *
A5	สิทธิลดหย่อนประเภท ค. *
A6	สิทธิลดหย่อนประเภท ง. *
A7	ผู้ประกันตนตาม พ.ร.บ.ประกันสังคม
A8	กองทุนเงินทดแทน
A9	ประกันภัยรถ ตาม พรบ.บุคคลที่ 3
AA	เด็ก 0-12 ปี
AB	ผู้มีรายได้น้อย
AC	นักเรียน
AD	ผู้พิการ
AE	ทหารผ่านศึก
AF	ภิกษุ/ผู้นำศาสนา
AG	ผู้สูงอายุ
AH	บัตรชั่วคราว
AI	บัตรประกันสุขภาพ ประชาชนทั่วไป
AJ	บัตรประกันสุขภาพ อาสาสมัครสาธารณสุข
AK	บัตรประกันสุขภาพ ผู้นำชุมชน
AL	บัตรประกันสุขภาพ คนต่างด้าว
UC	บัตรประกันสุขภาพถ้วนหน้า

ตารางที่ 5 โครงสร้างข้อมูลแฟ้มที่ 11 มาตรฐานแฟ้มข้อมูลการเงิน

FIELD NAME	TYPE	LENGTH	DECIMAL	QUALIFICATION
HN	C	9	0	หมายเลขประจำตัวผู้รับบริการ ควรใช้หมายเลขเดิมให้นานกว่า 5 ปี (Left justified)
AN	C	9	0	หมายเลขประจำตัวผู้ป่วยใน ไม่ควรใช้หมายเลขนี้ซ้ำ (Left justified)
DATE	DATE	8	0	วันที่คิดค่ารักษา วันจำหน่าย หรือวันที่ผู้ป่วยเปลี่ยนสิทธิการรักษา บันทึก ปีในค่า ค.ศ.
TOTAL	N	7	0	จำนวนเงินค่ารักษารวมเป็นบาท ที่เรียกเก็บ
PAID	N	7	0	จำนวนเงินที่ผู้ป่วยจ่ายเอง ในกรณีที่โรงพยาบาลไม่ได้รับเงินไว้ = 0
PTTYPE	C	2	0	ชนิดการชำระเงิน ถ้าชำระเงินเอง = 10

หมายเหตุ คอลัมน์ AN, PTTYPE, DATE และ PAID เมื่อรวมกันพิจารณาเป็นตัวบ่งชี้บุคคลทางอ้อม (Quasi-Identifier) ซึ่งสามารถนำมาเปรียบเทียบกับตารางข้อมูลจากประกันสังคมเพื่อทำการระบุตัวบุคคลอีกครั้ง ในกรณีนี้คือการเปรียบเทียบค่า AN ให้เป็นชื่อและนามสกุลและเชื่อมโยงไปข้อมูลอื่นที่ต้องการได้



HN	AN	DATEBILL	TOTAL	PAID	PTTYPE
null	11	11-Jan-08	3500	0	A2
null	12	23-Jan-08	4500	1000	AI
null	44	25-Feb-08	2330	330	AI
null	43	20-Feb-08	4300	0	AI
null	45	27-Feb-08	800	800	A1
null	56	1-Mar-08	7000	1000	AI
null	77	10-Apr-08	5000	5000	A1
null	78	20-Apr-08	6000	0	AI
null	59	15-Mar-08	6000	0	A2

ตัวอย่างตารางแฟ้มที่ 11 CHT

AN	DIAG
11	I10
12	F10
44	B20
43	I10
45	F10
56	B20
77	I10
78	F10
59	B20

ตัวอย่างตารางแฟ้มที่ 9 IDX

NAME	SURNAME	CAREER	DATEBILL	AMOUNT
สนอง	สอนง่าย	112	20-Mar-08	6000
สมาน	มาน้อย	111	28-Feb-08	4300
สมใจ	ชัยดี	111	31-Jan-08	3500

ตัวอย่างตาราง SOCIALSECURE

รูปที่ 1 ตัวอย่างการระบุตัวตนอีกครั้งเพื่อเชื่อมโยงข้อมูลจากการวิเคราะห์ความสัมพันธ์

จากรูปที่ 1 กำหนดให้ ตัวอย่างตาราง SOCIALSECURE เป็นตัวอย่างตารางที่มาจากประกันสังคม ตัวอย่างตารางแฟ้มที่ 11 CHT เป็นตัวอย่างตารางแฟ้มการเงินที่มาจากข้อมูลสาธารณสุข 12 แฟ้มและตัวอย่างตารางแฟ้มที่ 9 IDX และเป็นตัวอย่างตารางแฟ้มการวินิจฉัยโรคที่มาจากข้อมูลสาธารณสุข 12 แฟ้ม ซึ่งในตารางนี้ค่า DIAG เป็นรหัสที่เรียกว่า ICD-10 ซึ่งเป็นรหัสชื่อโรค สามารถเข้าไปดูชื่อโรคได้ที่ [2] ในที่นี้ได้นำตัวอย่างมา 3 โรคคือ I10 = High blood pressure ,F10 = ภาวะและพฤติกรรมแปรปรวนเนื่องจากการใช้ alcohol และ B20 = Human Immunodeficiency Virus (HIV)

สมมุติให้มีเพื่อนรวมงานของนายสนอง สอนง่าย ทราบว่านายสนอง ได้ไปโรงพยาบาลมา แล้วต้องการทราบว่านายสนองเป็นโรคอะไร ถ้าข้อมูลทั้ง 3 ตารางในรูปที่ 1 สามารถค้นหาได้จาก Internet เขาอาจใช้หลักการระบุตัวบุคคลอีกครั้งเพื่อทำให้ทราบว่านายสนองเป็นโรคอะไรได้ โดยมีหลักการวิเคราะห์คือ พิจารณาตาราง SOCIALSECURE สนอง สอนง่ายมีค่า AMOUNT เท่ากับ 6000 ซึ่งตรงกับระเบียนที่มีค่า AN เท่ากับ 56, 78 และ 59 ของตัวอย่างตารางแฟ้มที่ 11 CHT โดยระเบียนที่ค่า AN เท่ากับ 78 และ 59 มีค่า TOTAL เท่ากับ 6000 ส่วนระเบียนที่ค่า AN เท่ากับ 56 มีค่า TOTAL ลบ PAID เท่ากับ 6000 จาก 3 ระเบียนนี้ มีเพียงระเบียนที่ค่า AN เท่ากับ 78 และ 56 ที่มี PTTYPE เป็น AI ซึ่งแสดงว่าเป็นการเบิกเงินจาก

ประกันสังคมและมีเพียงระเบียนที่ค่า AN เท่ากับ 78 เท่านั้นที่ DATEBILL เป็นเดือนเดียวกันกับ DATABILL ของสนอง สอนง่าย

จากที่กล่าวมานั้นแสดงให้เห็นว่า สนองสอนง่ายนั้นมีค่า AN เท่ากับ 78 และเมื่อนำค่า AN เท่ากับ 78 นี้ไปเชื่อมโยงในตาราง IDX จะได้ว่า สนอง สอนง่ายนั้นเป็นโรค F10 ซึ่งคือโรคภาวะและพฤติกรรมแปรปรวนเนื่องจากการใช้ Alcohol

ฉะนั้นจึงสามารถสรุปได้ว่า การปกปิดข้อมูลโดยการลบตัวบ่งชี้บุคคลอย่างเดียวไม่สามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้

เทคนิค *k*-Anonymity

อย่างไรก็ตามจาก [3] และ [4] ผู้เขียนได้นำเสนอวิธีการที่เรียกว่า *k*-Anonymity ซึ่งสามารถแก้ปัญหาการระบุตัวบุคคลอีกครั้งได้ โดยการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity มีหลักการพื้นฐานอยู่ว่าให้ทำการเปลี่ยนแปลงค่าของข้อมูลในแต่ละระเบียน เมื่อเปลี่ยนแปลงค่าของข้อมูลแล้วต้องมีระเบียนที่มีตัวบ่งชี้ทางอ้อม (Quasi-Identifier) เหมือนกันไม่ต่ำกว่า k ระเบียน ในการเปลี่ยนแปลงค่าของข้อมูลในแต่ละระเบียนอาจทำการกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น (Hierarchy) ให้กับค่าของข้อมูลในแต่ละคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมไว้ก่อน แล้วจึงทำการเปลี่ยนค่าของแต่ละระเบียนไปทีละคอลัมน์จากระดับปัจจุบันไประดับที่สูงขึ้น จนมีคุณสมบัติ *k*-Anonymity จึงหยุดทำการเปลี่ยนแปลง ตัวอย่างในการกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นเช่นคอลัมน์ DATE ใน ระดับที่ 0 เป็นข้อมูลที่มีหน่วยเป็น วัน/เดือน/ปี ระดับที่ 1 เป็นข้อมูลที่มีหน่วยเป็น สัปดาห์/เดือน/ปี และระดับที่ 2 เป็นข้อมูลที่มีหน่วยเป็น เดือน/ปี เป็นต้น ในการทำข้อมูลให้มีคุณสมบัติ *k*-Anonymity สามารถเลือกทำได้อีก 2 วิธีคือ วิธี Alternative กับวิธี Full-Domain วิธี Alternative คือในแต่ละระเบียนสามารถเลือกตัวบ่งชี้บุคคลทางอ้อมตัวใดก็ได้ ที่อยู่ในระดับเดียวกันในการทำการเปลี่ยนแปลง ส่วนการทำ Full-Domain เป็นการเลือกคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมแล้วทำการเปลี่ยนทุกระเบียนในคอลัมน์นั้นตามขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่กำหนดไว้ ซึ่งในงานวิจัยนี้สนใจในกรณีของ Full-Domain

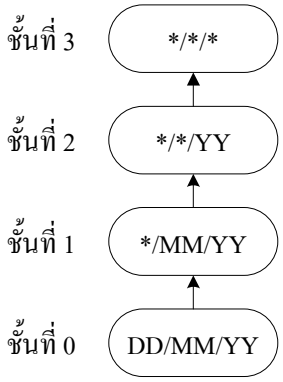
จากตัวอย่างในรูปที่ 2 ตารางข้อมูล ให้คอลัมน์ AN, DATEBILL, TOTAL และ PAID เป็นตัวบ่งชี้ทางอ้อมสามารถเปลี่ยนแปลงค่าในคอลัมน์ DATEBILL ตามการเปลี่ยนแปลงตามลำดับขั้นของ DATEBILL โดยเปลี่ยนค่าที่อยู่ในตารางข้อมูลซึ่งอยู่ในระดับที่ 0 ให้เป็นระดับที่ 1 ส่วนค่าในคอลัมน์อื่นก็สามารถทำได้โดยวิธีการเดียวกัน ในการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จากในตัวอย่างได้กำหนดให้ $k = 2$ แสดงว่าต้องทำการเปลี่ยนแปลงค่าในแต่ละคอลัมน์จนมีตัวบ่งชี้ทางอ้อมซ้ำกันเป็นจำนวน 2 ระเบียนขึ้นไป จึงจะถือว่าข้อมูลมีคุณสมบัติ *2*-Anonymity

AN	DATEBILL	TOTAL	PAID	PTTYPE
11	11 ม.ค. 2008	4500	0	A2
12	23 ม.ค. 2008	4500	1000	AI
44	25 ก.พ. 2008	2330	330	AI
43	20 ก.พ. 2008	2300	0	AI
45	27 ก.พ. 2008	2800	2800	A1
56	01 มี.ค. 2008	6000	1000	AI
77	10 เม.ย. 2008	7000	7000	A1
78	20 เม.ย. 2008	7000	0	AI
59	15 มี.ค. 2008	6000	0	A1

ตารางข้อมูล

AN	DATEBILL	TOTAL	PAID	PTTYPE
1*	* ม.ค. 2008	4***	*	A*
1*	* ม.ค. 2008	4***	*	A*
4*	* ก.พ. 2008	2***	*	A*
4*	* ก.พ. 2008	2***	*	A*
4*	* ก.พ. 2008	2***	*	A*
5*	* มี.ค. 2008	6***	*	A*
7*	* เม.ย. 2008	7***	*	A*
7*	* เม.ย. 2008	7***	*	A*
5*	* มี.ค. 2008	6***	*	A*

ตารางข้อมูลหลังมีการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่มีคุณสมบัติ *2*-Anonimity



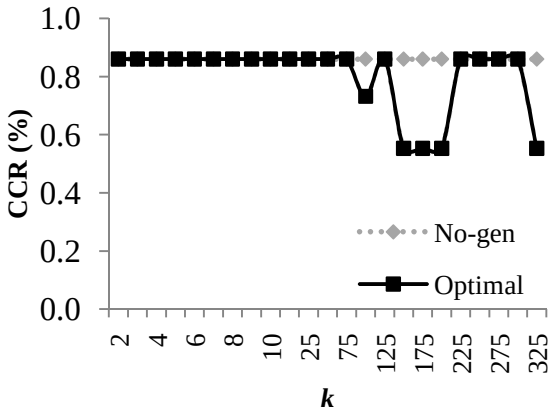
ขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของคอลัมน์ DATEBILL

รูปที่ 2 ตัวอย่างการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity

การนำข้อมูลที่มีคุณสมบัติ *k*-Anonymity ไปใช้งานต่อนั้น มีปัญหาอยู่ว่าการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity นั้นสามารถกำหนดระดับของการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของแต่ละคอลัมน์ที่เป็นตัวบ่งชี้ทางอ้อม หรือ ระดับการเจนเนอรัลไลเซชัน (Generalization Level) ได้หลายรูปแบบ ตัวอย่างระดับการเจนเนอรัลไลเซชัน เช่น (AN:2, DATEBILL:1, TOTAL:3, PAID:5, PTTYPE:1) จากรูปที่ 2 ถ้าเปลี่ยนแปลงข้อมูลของคอลัมน์ DATEBILL ให้เพิ่มขึ้นไปอีก 1 ระดับ (ระดับการเจนเนอรัลไลเซชัน เท่ากับ (AN:2, DATEBILL:2, TOTAL:3, PAID:5, PTTYPE:1))ข้อมูลก็ยังคงคุณสมบัติ *k*-Anonymity อยู่ ทำให้ยากต่อการเลือกว่าจะใช้ข้อมูลที่มีคุณสมบัติ *k*-Anonymity ที่เกิดจากระดับการเจนเนอรัลไลเซชันที่เท่าไหร่จึงจะเหมาะสมกับงานที่ต้องนำข้อมูลนี้ไปใช้

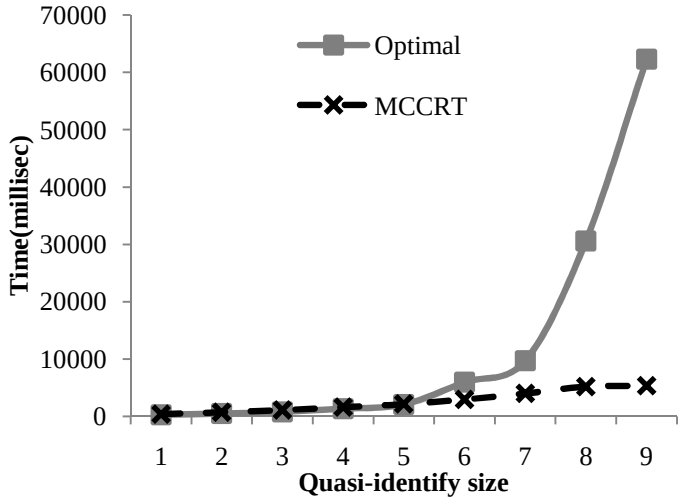
งานวิจัย [5] ได้เสนอวิธีการขั้นตอนวิธีที่เหมาะที่สุด (Optimal Algorithm) ซึ่งเป็นการหาระดับการเจนเนอรัลไลเซชันที่มีผลกระทบต่อคุณภาพข้อมูลน้อยที่สุดโดยการวัดการบิดเบือนของข้อมูล (Distortion

Ratio: C_{GM}) และตัววัดคุณภาพข้อมูลสำหรับการจำแนกแบบความสัมพันธ์ (Frequency-based Classification: C_{FCM}) โดยจะเลือกระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ k-Anonymity ที่ให้ค่า C_{GM} และค่า C_{FCM} น้อยที่สุด ข้อมูลที่ทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นด้วยระดับการเจนเนอรัลไลเซชันที่ได้จากวิธีนี้เมื่อนำไปทำเหมืองข้อมูลที่เรียกว่า Associative Classification ปรากฏว่าได้ผลการทดลองตามรูปที่ 3 ซึ่งเป็นการวัดค่าอัตราความถูกต้องในการจำแนก (Classification Correction Rate: CCR) เปรียบเทียบระหว่างการไม่เปลี่ยนแปลงค่าของข้อมูล (No-gen) กับ เมื่อเปลี่ยนแปลงค่าของข้อมูลใช้ขั้นตอนวิธีที่เหมาะที่สุด (Optimal)

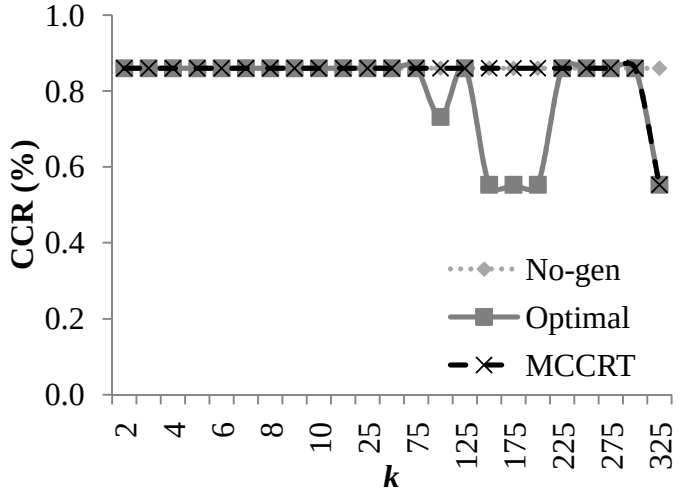


รูปที่ 3 ผลกระทบของค่า k ต่อค่า CCR [5]

แต่เนื่องจากปัญหาการแปลงข้อมูลเพื่อที่จะได้มาซึ่งฐานข้อมูลที่ดีที่สุดโดยยังคงคุณสมบัติ k-Anonymity และมีผลกระทบต่อคุณภาพข้อมูลน้อยที่สุด เป็นปัญหาเอ็นพีแบบยาก (NP-hard) โดยขั้นตอนวิธีนี้มีระดับความซับซ้อนเชิงคำนวณที่เป็นเลขชี้กำลัง (Exponential) ใน [6] จึงได้มีการคิดค้น MCCRT (Minimum Classification Correction Rate Transformation algorithm) ที่มีขั้นตอนวิธีเป็นแบบศึกษาสำนึกที่จะแก้ปัญหานี้ได้อย่างมีประสิทธิภาพ และประสิทธิผล โดยสามารถทำงานได้อย่างรวดเร็วเมื่อมีจำนวนข้อมูลมากขึ้นและยังมีการทำงานที่เร็วรวดเร็วกว่า ซึ่งความซับซ้อนเชิงคำนวณคือ $O(n\log n)$ ในงานวิจัยนี้จึงให้สนใจที่จะปรับปรุงการทำงานของ MCCRT



รูปที่ 4 ผลกระทบของจำนวนคอลัมน์ที่เหลื่อมซ้อนกันต่อเวลากระทำการ [6]



รูปที่ 5 ผลกระทบค่า *k* ต่อค่า *CCR* [6]

ลำดับการทำงานของขั้นตอนวิธี MCCRT

MCCRT มีหลักการทำงานคือการนำอัตราความถูกต้องในการจำแนก (Classification Correction Rate: *CCR*) มาใช้ในการเรียงลำดับเพื่อนำไปเปลี่ยนค่าของแต่ละคอลัมน์โดย จะเรียงจากคอลัมน์ที่มี *CCR* น้อยไปหาคอลัมน์ที่มี *CCR* มาก โดยในกรณีที่คอลัมน์มี *CCR* เท่ากันจะเรียงจากค่าความสูงของขั้นตอนการเปลี่ยนแปลงตามลำดับขั้นของคอลัมน์นั้นๆ จากมากไปหาน้อย ซึ่งเรียกการเรียงตัวกันของคอลัมน์นี้ว่า *S* (Sequent of novel heuristic algorithm MCCRT) ต่อมาทำการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์จากตัวแรกของ *S* โดยเริ่มจากระดับที่ 0 ขึ้นไปจนถึงระดับบนสุดแล้วจึงเปลี่ยนแปลงค่าในตัวถัดไปของ *S* ทำซ้ำในรูปแบบเดียวกันจนข้อมูลมีคุณสมบัติ *k*-Anonymity ผลลัพธ์ที่ได้นี้คือระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จากการทดลองปรากฏว่าเมื่อนำข้อมูลที่ได้จากการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นในระดับการเจนเนอรัลไลเซชันนี้ไปหา Associative Classification

ปรากฏว่า Associative Classification ที่ได้มีความแม่นยำที่มากกว่า (ดังรูปที่ 5) และมีความเร็วที่มากกว่า (ดังรูปที่ 4) ระดับการเจนเนอรัลไลเซชันที่ได้จากแบบขั้นตอนวิธีที่เหมาะที่สุด สาเหตุที่มีความแม่นยำที่มากกว่าเนื่องจาก MCCRT ใช้ค่า CCR มาคิดโดยตรงแต่ขั้นตอนวิธีที่เหมาะที่สุดคิดจากค่า C_{FCM} และสาเหตุที่ MCCRT มีความเร็วที่มากกว่ามาจาก การทดสอบคุณสมบัติ *k*-Anonymity ของ MCCRT เป็นการหาระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity เพียงแค่ค่าแรกที่เจอ แต่ของขั้นตอนวิธีที่เหมาะที่สุดต้องทดสอบคุณสมบัติ *k*-Anonymity ในทุกระดับการเจนเนอรัลไลเซชันที่เป็นไปได้เพื่อหาระดับการเจนเนอรัลไลเซชันที่มีคุณสมบัติ *k*-Anonymity ทุกค่า

การประมวลผลแบบ One-time fashion ของขั้นตอนวิธี MCCRT

อย่างไรก็ตาม ในการหาระดับการเจนเนอรัลไลเซชันของ MCCRT นั้น ต้องทำการอ่านข้อมูลทั้งหมดหนึ่งครั้ง จึงได้ผลลัพธ์ S ออกมาและทำการ ทดสอบคุณสมบัติ *k*-Anonymity ไปจนกว่า จะได้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ถ้าในภายหลังมีการเพิ่มข้อมูลเข้ามาใหม่จำนวนหนึ่งแล้วต้องการระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ใหม่ต้องทำการอ่านข้อมูลทั้งหมดและทำการทดสอบ คุณสมบัติ *k*-Anonymity ตั้งแต่เริ่มอีกครั้ง แสดงว่าถ้ามีการเพิ่มข้อมูลใหม่ 10 ครั้งต้อง อ่านข้อมูลทั้งหมดและทำการทดสอบ คุณสมบัติ *k*-Anonymity ตั้งแต่เริ่ม ใหม่ 10 ครั้ง ซึ่งเป็นการเสียเวลาอย่างมาก ในงานวิจัยนี้ต้องการหาวิธีที่ไม่ต้อง อ่านข้อมูลใหม่ทั้งหมดและลดการทดสอบ คุณสมบัติ *k*-Anonymity ให้น้อยลงโดยอาจ อ่านข้อมูลที่เข้ามาใหม่และข้อมูลเก่าบางส่วนและทำการทดสอบ คุณสมบัติ *k*-Anonymityที่น้อยลงกล่าวคือไม่ต้องทำการทดสอบตั้งแต่เริ่มต้น ซึ่งในการทำ MCCRT ครั้งแรกอาจเก็บข้อมูลบางอย่างไว้ ซึ่งการทำแบบนี้ย่อมเร็วกว่า และผลลัพธ์ที่ได้ จะเหมือนการอ่านและทำการทดสอบ คุณสมบัติ *k*-Anonymityใหม่ทั้งหมด

4.2 แนวทางการแก้ปัญหา

ในงานนี้จะเสนอขั้นตอนวิธีที่สามารถรักษาความเป็นส่วนตัวของข้อมูลเมื่อมีการเพิ่มขึ้นของข้อมูลในสถานการณ์ที่การวิเคราะห์ข้อมูลเหล่านั้นคือการจำแนกแบบความสัมพันธ์ โดยจะเริ่มจากการวิเคราะห์สถานการณ์ที่จะเกิดขึ้นกับข้อมูลในแง่ของความเป็นส่วนตัวในรูปแบบของ *k*-Anonymity และในแง่ของคุณภาพข้อมูล ซึ่งการวิเคราะห์สถานการณ์ในลักษณะนี้เป็นหลักการพื้นฐานที่ใช้ในการพัฒนาขั้นตอนวิธีแบบเพิ่มขึ้น [7] เมื่อเสร็จสิ้นการวิเคราะห์ศึกษาแล้วจะทำการพัฒนาขั้นตอนวิธีต่อไป โดยขั้นตอนวิธีที่จะพัฒนาขึ้นจะต้องมีความซับซ้อนเชิงคำนวณต่ำกว่าการนำขั้นตอนวิธีที่มีลักษณะเป็น One-time fashion

จากการศึกษาการทำงานของ MCCRT ใน [6] และแนวทางการแก้ปัญหาของ IncSpan ใน [7] พบว่าเมื่อมีการเพิ่มข้อมูลเข้าใหม่อาจทำให้เกิดการเปลี่ยนแปลงระดับการเจนเนอรัลไลเซชันเพื่อให้ยังคงคุณสมบัติ *k*-Anonymity ซึ่งมีสาเหตุมาจาก 1) เนื่องจากค่าของข้อมูลที่เพิ่มเข้ามาใหม่มีค่าซ้ำค่าเดิมหรือมี

ค่าที่ไม่เคยปรากฏมาก่อน 2) เกิดการเปลี่ยนแปลงลำดับใน S (Sequent of novel heuristic algorithm MCCRT) เพราะข้อมูลที่เพิ่มเข้ามาใหม่ทำให้ค่า CCR เปลี่ยนแปลงไป จากที่กล่าวมาสามารถแยกย่อยปัญหาลงไปเป็นกรณีได้สองกลุ่มดังต่อไปนี้

1) เนื่องจากค่าของข้อมูลที่เพิ่มเข้ามาใหม่
   - ทำให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลยังคงคุณสมบัติ *k*-Anonymity ลดลง
   - ทำให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลยังคงคุณสมบัติ *k*-Anonymity เท่าเดิม
   - ทำให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลยังคงคุณสมบัติ *k*-Anonymity เพิ่มขึ้น
2) มีการเปลี่ยนแปลงลำดับใน S (ให้ Point: P คือคอลัมน์ สุดท้ายที่มีการเปลี่ยนค่าของข้อมูลตามลำดับขั้น)
   - มีการสลับลำดับก่อนหน้า P
   - มีการสลับลำดับหลัง P
   - มีการสลับลำดับจากก่อนหน้า P ไปหลัง P (P ถูกเลื่อนไปข้างหน้า)
   - มีการสลับลำดับจากหลัง P ไปหน้า P (P ถูกเลื่อนไปข้างหลัง)
   - สลับลำดับแบบผสม

จากทุกกรณีที่กล่าวมาสามารถลดการ อ่านข้อมูลได้โดยแยกเป็นกรณี ทำให้โดยรวมแล้วไม่จำเป็นต้องทำการ อ่านข้อมูลทั้งหมดใหม่อีกครั้ง

การวิเคราะห์และอธิบายวิธีการแก้ปัญหาในแต่ละกรณี

4.2.1 ปัญหาที่เกิดจากค่าของข้อมูลที่เพิ่มเข้ามา

4.2.1.1 ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลยังคงคุณสมบัติ *k*-Anonymity ลดลง

ปัญหาในกรณีนี้นอกจากทำการตรวจสอบว่าเกิดกรณีนี้หรือไม่ แล้วยังต้องการทราบว่าระดับการเจนเนอรัลไลเซชันลดลงเป็นเท่าไหร่ พิจารณาสาเหตุของกรณีนี้เกิดจากข้อมูลที่เพิ่มเข้ามาบางระเบียนมีค่าซ้ำกับข้อมูลเดิมซึ่งอาจทำให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ลดลงได้ ยกตัวอย่างเช่น จากตารางที่ 6 กำหนดให้ k เท่ากับ 2 ให้ S เท่ากับ (A,B,C,D,E) และให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity เท่ากับ (A:2, B:3, C:1, D:0, E:0) ตารางนี้แสดงถึงข้อมูลของระดับการเจนเนอรัลไลเซชันก่อนมีคุณสมบัติ *k*-Anonymity 1 ระดับซึ่งเท่ากับ (A:2, B:3, C:0, D:0, E:0) จะเห็นว่ามีเพียงระเบียนที่ 05,09,10,13,14 และ 15 ที่ยังไม่มีคุณสมบัติ *k*-Anonymity ถ้าข้อมูลใหม่ที่เพิ่มเข้ามาทำให้ 6 ระเบียนนี้มีคุณสมบัติ *k*-Anonymity แล้วระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จะลดลงเป็นระดับการเจนเนอรัลไลเซชันนี้แทน จากที่กล่าวมาแสดงให้เห็นว่า

ข้อมูลที่จะต้องอ่านมีเพียง ข้อมูลที่เพิ่มเข้ามาใหม่กับ 6 ระเบียนที่กล่าวมา ฉะนั้นถ้าเก็บข้อมูลระเบียนที่ทำให้ข้อมูลไม่มีคุณสมบัติ *k*-Anonymity ในแต่ละระดับการเจนเนอรัลไลเซชันไว้ แล้วทำการตรวจสอบกับข้อมูลที่เพิ่มเข้ามาโดยย้อนกลับจากระดับการเจนเนอรัลไลเซชันเดิมที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ลงไปทีละขั้น ก็จะทราบระดับการเจนเนอรัลไลเซชันที่ลดลงได้

ตารางที่ 6 ข้อมูลก่อนมีคุณสมบัติ *k*-Anonymity 1 level

Tuple-ID	A	B	C	D	E	Class
1	*	*	C1	D1	E1	Class1
2	*	*	C1	D1	E1	Class1
3	*	*	C1	D1	E1	Class1
4	*	*	C1	D1	E1	Class1
5	*	*	C2	D1	E1	Class1
6	*	*	C2	D2	E2	Class2
7	*	*	C2	D2	E2	Class2
8	*	*	C2	D2	E2	Class2
9	*	*	C2	D3	E2	Class2
10	*	*	C1	D3	E2	Class2
11	*	*	C1	D3	E3	Class3
12	*	*	C1	D3	E3	Class3
13	*	*	C2	D3	E3	Class3
14	*	*	C3	D4	E2	Class3
15	*	*	C4	D4	E2	Class3

4.2.1.2 ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity เท่าเดิม

ปัญหาในกรณีนี้เพียงต้องการตรวจสอบว่าการที่ข้อมูลเพิ่มเข้ามาทำให้เกิดกรณีนี้หรือไม่ เนื่องจากเมื่อเป็นกรณีนี้ไม่ต้องทำการเปลี่ยนแปลงใดๆ พิจารณาสาเหตุของกรณีนี้เกิดจากข้อมูลที่เพิ่มเข้ามาไม่สามารถทำให้ 6 ระเบียน ซึ่งคือ 05, 09, 10, 13, 14 และ 15 ในตารางที่ 6 ที่กล่าวมามีคุณสมบัติ *k*-Anonymity ได้ทั้งหมดและเมื่อทำการเปลี่ยนแปลงข้อมูลตามลำดับขั้นของข้อมูลที่เพิ่มเข้ามาให้เป็นระดับการเจนเนอรัลไลเซชันเดิมแล้วข้อมูลที่ได้มีคุณสมบัติ *k*-Anonymity เหมือนเดิมแสดงว่าการเพิ่มเข้ามาของข้อมูลทำให้เกิดกรณีนี้ขึ้น

4.2.1.3 ระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity เพิ่มขึ้น

ปัญหาในกรณีนี้นอกจากตรวจสอบว่าเกิดกรณีขึ้นหรือไม่ ยังต้องการทราบว่าระดับการเจนเนอรัลไลเซชันเพิ่มขึ้นเป็นเท่าไหร่ พิจารณาสาเหตุของกรณีนี้เกิดจาก ค่าของข้อมูลที่เพิ่มเข้ามาใหม่อาจเป็นค่าไม่ที่เคยปรากฏมาก่อน ซึ่งข้อมูลที่เข้ามาใหม่อาจไม่มีคุณสมบัติ *k*-Anonymity ในระดับการเจนเนอรัลไลเซชันเดิม แสดงว่าต้องทำการเพิ่ม Generalization Level ขึ้นไป ซึ่งสามารถเพิ่มได้ตามวิธีของ MCCRT ทำให้ได้ Generalization Level ใหม่มา

4.2.1.4 การแก้ปัญหารวมทั้งสามกรณี (Solution of Generalization Level Changing)

ในการแก้ปัญหา สามารถทำการทดสอบข้อมูลที่เข้ามาใหม่ซึ่งสามารถแบ่งข้อมูลที่เข้ามาใหม่เป็นสองส่วน ส่วนแรกเป็นส่วนที่มีค่าซ้ำกับข้อมูลเดิม ส่วนที่สองเป็นค่าใหม่ที่ไม่เคยมีมาก่อน ในการทำงานเมื่อข้อมูลใหม่เข้ามาสามารถเริ่มทดสอบข้อมูลใหม่ในส่วนที่สองก่อน โดยทดสอบว่าเมื่อทำการเปลี่ยนแปลงค่าตามลำดับขึ้นถึงระดับการเจนเนอรัลไลเซชันเดิมที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity แล้วข้อมูลในส่วนนี้มีคุณสมบัติ *k*-Anonymity หรือไม่ถ้าไม่มี แสดงว่าเกิดกรณีที่ 4.2.3.3 ต้องเพิ่มระดับการเจนเนอรัลไลเซชันขึ้นไปตาม MCCRT ต่อมาถ้าข้อมูลส่วนที่สอง มีคุณสมบัติ *k*-Anonymity ในระดับการเจนเนอรัลไลเซชันเดิมแสดงว่าอาจเกิดกรณีที่ 4.2.3.1 หรือ 4.2.3.2 โดยวิธีการแก้ปัญหาให้ทำการตรวจสอบให้ทำการตรวจสอบข้อมูลใหม่ในส่วนแรก โดยทำการทดสอบย้อนหลังว่า ข้อมูลนี้สามารถทำให้ระเบียนที่ไม่มีคุณสมบัติ *k*-Anonymity ในระดับการเจนเนอรัลไลเซชันที่ต่ำลงไปมีคุณสมบัติ *k*-Anonymity ขึ้นมาหรือไม่ โดยในการทดสอบให้ลดระดับการเจนเนอรัลไลเซชันลงครั้งละ 1 ระดับเมื่อทำการทดสอบแล้วว่าระเบียนที่ขาดคุณสมบัติ *k*-Anonymity ของระดับการเจนเนอรัลไลเซชันนั้นๆมีคุณสมบัติ *k*-Anonymity และในข้อมูลที่เพิ่มเข้ามาใหม่ในส่วนที่สองยังคงมีคุณสมบัติ *k*-Anonymity อยู่ให้ทำการลดระดับลงอีก 1 ระดับแล้วทำการทดสอบเหมือนเดิม จนถึงระดับการเจนเนอรัลไลเซชันที่ไม่สามารถทำให้ระเบียนไม่มีคุณสมบัติ *k*-Anonymity เปลี่ยนเป็นมีคุณสมบัติ *k*-Anonymity ขึ้นมาได้ หรือข้อมูลส่วนที่สองไม่มีคุณสมบัติ *k*-Anonymity แสดงว่าระดับการเจนเนอรัลไลเซชันนี้เป็นระดับการเจนเนอรัลไลเซชันที่ต่ำลงมา 1 ระดับของที่ต้องการ กล่าวคือจากระดับการเจนเนอรัลไลเซชันนี้ เมื่อเพิ่มเข้าไป 1 ระดับจะเป็นระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลหลังการเพิ่มเข้ามาของข้อมูลใหม่ (D') มีคุณสมบัติ *k*-Anonymity

จากวิธีการแก้ปัญหานี้ต้องทำการคิดวิธีในการเก็บค่าเพื่อทำการทดสอบข้อมูลที่เข้ามาใหม่ว่าระเบียนใดเป็นข้อมูลที่ซ้ำกับข้อมูลเก่าส่วนใดเป็นข้อมูลที่เป็นค่าใหม่ และยังต้องสามารถรู้ข้อมูลระเบียนที่ไม่มีคุณสมบัติ *k*-Anonymity ในแต่ละระดับการเจนเนอรัลไลเซชันถ้าวิธีการเก็บข้อมูลนี้มีประสิทธิภาพจะทำให้การทดสอบการมีคุณสมบัติ *k*-Anonymity ทำได้รวดเร็วมากขึ้นกว่าการอ่านข้อมูลใหม่ทั้งหมด

4.2.2 ปัญหาที่เกิดจากการเปลี่ยนลำดับใน S

เมื่อข้อมูลที่เพิ่มเข้ามาให้ CCR เปลี่ยนจนทำให้ลำดับใน S เปลี่ยนจะทำให้ระดับการเจนเนอรัลไลเซชันที่ทำให้ D' (ข้อมูลหลังการเพิ่มเข้ามาของข้อมูลใหม่) มีคุณสมบัติ k-Anonymity เปลี่ยนแปลงตามไปด้วยเพราะเมื่อลำดับใน S เปลี่ยนแปลงจะทำให้ลำดับในระดับการเจนเนอรัลไลเซชันเปลี่ยนตาม จึงจำเป็นต้องทำการทดสอบการมีคุณสมบัติ k-Anonymity ของข้อมูลใหม่เพื่อให้ได้ระดับการเจนเนอรัลไลเซชันที่เป็นไปตาม MCCRT ซึ่งในงานวิจัยนี้ได้นำกรณีที่เกิดขึ้นได้ของการสลับลำดับใน S มาวิเคราะห์เพื่อให้ได้การทำงานที่เร็วกว่าการทำ MCCRT ใหม่

4.2.2.1 มีการสลับลำดับก่อนหน้า P (ให้ point: P คือคอลัมน์สุดท้ายที่มีการเปลี่ยนค่าของข้อมูลตามลำดับขั้น)

พิจารณาการสลับลำดับกรณีนี้ไม่มีผลกระทบต่อการระดับการเจนเนอรัลไลเซชันเนื่องจากทุกคอลัมน์ในลำดับก่อนหน้า P ได้ทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของตัวเองไปจนถึงระดับบนสุดแล้วก็ยังไม่ผ่านการทดสอบการมีคุณสมบัติ k-Anonymity ตัวอย่าง ถ้าลำดับเก่า S เป็น (A, B, C, D, E) มีระดับการเจนเนอรัลไลเซชันคือ (A:2, B:3, C:2, D:1, E:0) และ D คือ P ถ้าลำดับใหม่ S' คือ (A, C, B, D, E) จาก MCCRT จะเห็นว่า A, B และ C ใน S ได้ทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของตัวเองไปจนถึงระดับบนสุดแล้ว ซึ่งก็ยังไม่ผ่านการทดสอบการมีคุณสมบัติ k-Anonymity เมื่อมีการสลับลำดับเป็น S' จะเห็นว่า ไม่ว่าอย่างไรก็ต้องต้องทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นไปจนถึง D ระดับที่ 1 จึงจะผ่านการทดสอบการมีคุณสมบัติ k-Anonymity จากที่กล่าวมาสามารถสรุปได้ว่า ไม่จำเป็นต้องทำการทดสอบการมีคุณสมบัติ k-Anonymity ใหม่ เพียงทำการสลับลำดับในระดับการเจนเนอรัลไลเซชันตาม S' ก็เพียงพอ

4.2.2.2 กรณีทีมีการสลับลำดับหลัง P

จากการพิจารณาการสลับลำดับกรณีนี้ไม่มีผลกระทบต่อการระดับการเจนเนอรัลไลเซชันเนื่องจากทุกคอลัมน์ในลำดับหลัง P ไม่ได้ทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นอยู่แล้ว ตัวอย่างเช่น ลำดับเก่า S เป็น (A, B, C, D, E) มีระดับการเจนเนอรัลไลเซชันคือ (A:2, B:3, C:0, D:0, E:0) และ B คือ P ถ้าลำดับใหม่ S' คือ (A, B, E, C, D) จาก MCCRT จะเห็นว่าเมื่อทำการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่ B ระดับที่ 3 แล้วจะผ่านการทดสอบการมีคุณสมบัติ k-Anonymity แสดงว่า ส่วนหลังจากนี้ถึงมีการสลับลำดับก็ไม่เกิดผลกระทบกับระดับการเจนเนอรัลไลเซชันเนื่องจากผ่านการทดสอบการมีคุณสมบัติ k-Anonymity ไปก่อนแล้ว จากที่กล่าวมาสามารถสรุปได้ว่า ไม่จำเป็นต้องมีการทดสอบการมีคุณสมบัติ k-Anonymity ใหม่ เพียงทำการสลับลำดับในระดับการเจนเนอรัลไลเซชันตาม S' ก็เพียงพอ

4.2.2.3 มีการสลับลำดับจากก่อนหน้า P ไปหลัง P (กล่าวคือ P ถูกเลื่อนไปข้างหน้า)

จากการพิจารณาการสลับลำดับในกรณีนี้ เมื่อคอลัมน์ใดๆ ถูกย้ายไปด้านหลัง P เพื่อให้เป็นไปตาม MCCRT คอลัมน์เหล่านั้นจะลดระดับลงมาที่ระดับ 0 และสามารถใช้ระดับการเจนเนอรัลไลเซชันที่ได้นี้เป็น

จุดเริ่มต้นในการทดสอบการมีคุณสมบัติ *k*-Anonymity ใหม่ ตัวอย่าง ถ้าลำดับเก่า S เป็น (A, B, C, D, E) มีระดับการเจนเนอรัลไลเซชันคือ (A:2, B:3, C:1, D:0, E:0) และ C คือ P ถ้าลำดับใหม่ S' คือ (A, C, B, D, E) จาก MCCRT ระดับการเจนเนอรัลไลเซชันจะเท่ากับ (A:2, C:1, B:0, D:0, E:0) จากตัวอย่างจะเห็นว่าเมื่อ B ถูกย้ายไปด้านหลังและถูกลดระดับลงมาที่ระดับ 0 ทำให้ไม่สามารถรู้ได้ว่าระดับการเจนเนอรัลไลเซชันนี้ยังคงคุณสมบัติ *k*-Anonymity หรือไม่แต่สามารถแน่ใจได้ว่าระดับการเจนเนอรัลไลเซชันก่อนหน้าเช่น (A:2, C:0, B:0, D:0, E:0) นี้ไม่มีคุณสมบัติ *k*-Anonymity เพราะระดับการเจนเนอรัลไลเซชันของ S เช่น (A:2, B:0, C:0, D:0, E:0) ได้ทำการทดสอบแล้วว่าไม่มีคุณสมบัติ *k*-Anonymity เช่นกัน จากที่กล่าวมาแสดงว่าเมื่อเกิดกรณีนี้ให้ทำการสลับลำดับระดับการเจนเนอรัลไลเซชันตาม S' ลดระดับลำดับหลัง P ให้เป็น 0 แล้วใช้ระดับการเจนเนอรัลไลเซชันที่ได้ใหม่นี้เป็นจุดเริ่มต้นการทดสอบการมีคุณสมบัติ *k*-Anonymity ใหม่ตาม MCCRT

4.2.2.4 มีการสลับลำดับจากหลัง P ไปหน้า P (กล่าวคือ P ถูกเลื่อนไปข้างหลัง)

จากการพิจารณาการสลับลำดับในกรณีนี้เมื่อคอลัมน์ใดๆ ถูกย้ายมาด้านหน้า P เพื่อให้เป็นไปตาม MCCRT คอลัมน์หลังจากนั้นจะลดลงมาที่ระดับ 0 พิจารณาจากตัวอย่างต่อไปนี้ ให้ S เท่ากับ (A, B, C, D, E) มีระดับการเจนเนอรัลไลเซชันคือ (A:2, B:3, C:1, D:0, E:0) และ C คือ P ถ้าลำดับใหม่ S' คือ (A, B, D, C, E) จาก MCCRT ระดับการเจนเนอรัลไลเซชันจะเท่ากับ (A:2, B:3, D:0, C:0, E:0) จากตัวอย่างจะเห็นว่าเมื่อ D ถูกย้ายไปด้านหน้าและคอลัมน์หลังจากนั้นถูกลดระดับลงมาที่ระดับ 0 ทำให้สามารถรู้ได้ว่าระดับการเจนเนอรัลไลเซชันนี้ไม่มีคุณสมบัติ *k*-Anonymity เพราะระดับการเจนเนอรัลไลเซชันของ S เช่น (A:2, B:3, C:0, D:0, E:0) ได้ทำการทดสอบแล้วว่าไม่มีคุณสมบัติ *k*-Anonymity เช่นกัน จากที่กล่าวมาแสดงว่าเมื่อเกิดกรณีนี้ให้ทำการสลับลำดับระดับการเจนเนอรัลไลเซชันตาม S' ลดระดับลำดับหลังคอลัมน์ที่ถูกย้ายมาให้เป็น 0 แล้วใช้ระดับการเจนเนอรัลไลเซชันที่ได้ใหม่นี้เป็นจุดเริ่มต้นการทดสอบการมีคุณสมบัติ *k*-Anonymity ใหม่ตาม MCCRT

4.2.2.5 กรณีทีมีการสลับลำดับแบบผสม

จากกรณีทั้งหมดที่กล่าวมายังสามารถเกิดการสลับตำแหน่งในรูปแบบผสมได้ ซึ่งในการทำงานสามารถแยกกรณีเพื่อทำการทดสอบได้ตัวอย่างเช่น จาก S (A, B, C, D, E) มีระดับการเจนเนอรัลไลเซชันเป็น (A:2, B:3, C:1, D:0, E:0) และ C คือ P ถ้า S' คือ (B, A, D, C, E) จะเห็นว่า A, B เกิดการสลับในกรณีที่ 4.2.1 และ D, C เป็นการสลับในกรณีที่ 4.2.7 สามารถใช้วิธีการแก้ไขของทั้งสองกรณีรวมกัน คือที่ A สลับ B ไม่มีผลกระทบ ที่กรณี 4.2.7 ต้องทำการเริ่มทดสอบใหม่ที่ D เป็นต้น

5 ทฤษฎีที่ใช้ในการแก้ปัญหา

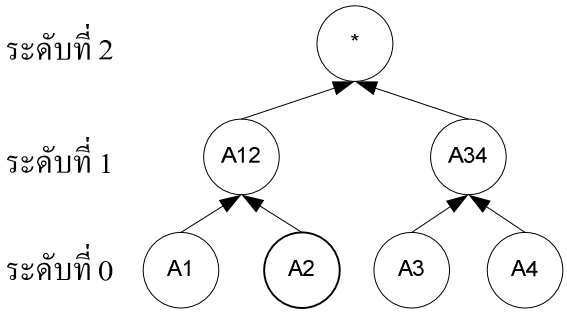
5.1 *k*-Anonymity ในปัญหาของการระบุตัวบุคคลอีกครั้งจะเกิดขึ้นได้ถ้าข้อมูลที่ถูกระบุตัวบุคคลอีกครั้งมีความเป็นเอกลักษณ์ กล่าวคือมีระเบียนที่ตัวบ่งชี้บุคคลทางอ้อมมีค่าไม่เหมือนระเบียนใดๆอยู่

ระเบียนเหล่านี้จะถูกนำไปตรวจสอบกับข้อมูลจากแหล่งอื่นได้ง่าย แต่ถ้าระเบียนเหล่านี้มีการซ้ำกันจะทำให้ไม่สามารถระบุตัวบุคคลอีกครั้งได้ หลักการทำข้อมูลให้มีคุณสมบัติ *k*-Anonymity ก็คือการเปลี่ยนข้อมูลในแต่ละระเบียนให้ทุก ทุกระเบียนมีค่าของตัวบ่งชี้บุคคลทางอ้อมที่เหมือนกันอย่างน้อย k ระเบียน ในการเปลี่ยนแปลงค่าของข้อมูลนั้น มีหลายวิธี เช่น 1) การเปลี่ยนค่าให้เป็นค่าที่ไม่รู้จัก ('*' :Unknown Value หรือเรียกว่า Suppression) 2) การเปลี่ยนให้ค่ากลายเป็นช่วงของค่า 3) การกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น ต่อไปนี้เป็นเพียงตัวอย่างข้อมูลและขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น

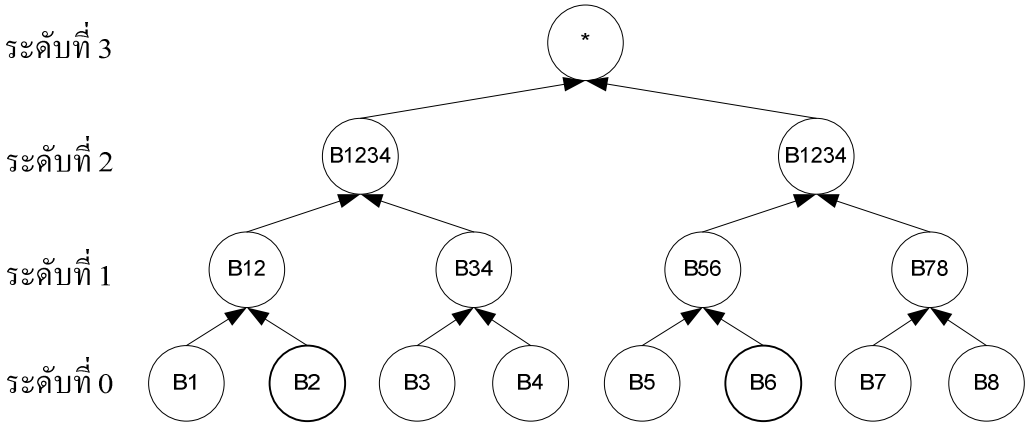
ตารางที่ 7 ตัวอย่างข้อมูล Database D

Tuple-ID	A	B	C	D	E	Class
1	A1	B1	C1	D1	E1	Class1
2	A1	B1	C1	D1	E1	Class1
3	A1	B1	C1	D1	E1	Class1
4	A1	B2	C1	D1	E1	Class1
5	A3	B3	C2	D1	E1	Class1
6	A2	B2	C2	D2	E2	Class2
7	A2	B2	C2	D2	E2	Class2
8	A3	B2	C2	D2	E2	Class2
9	A3	B3	C2	D3	E2	Class2
10	A4	B4	C1	D3	E2	Class2
11	A3	B1	C1	D3	E3	Class3
12	A3	B1	C1	D3	E3	Class3
13	A2	B3	C2	D3	E3	Class3
14	A2	B3	C3	D4	E2	Class3
15	A1	B4	C4	D4	E2	Class3

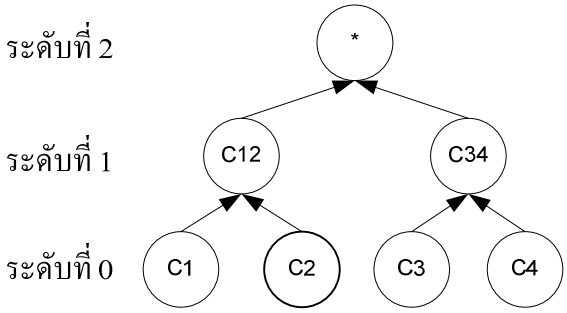
จากตารางที่ 7 สามารถกำหนดขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลในแต่ละคอลัมน์ได้ดังนี้



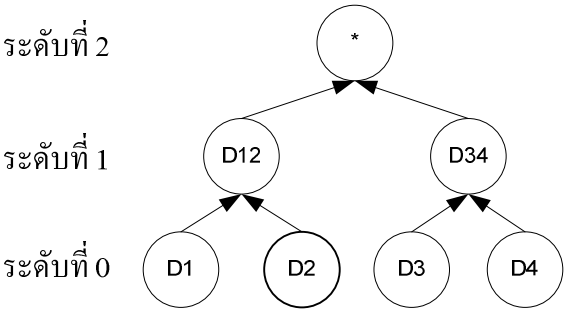
รูปที่ 6 ขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์ A



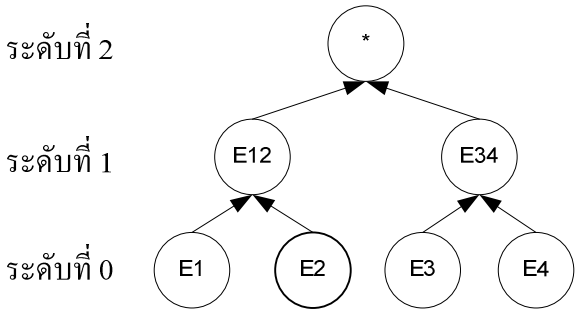
รูปที่ 7 ขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์ B



รูปที่ 8 ขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์ C



รูปที่ 9 ขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์ D



รูปที่ 10 ขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นของคอลัมน์ E

ในกรณีนี้ทำแบบ Full domain คือการเปลี่ยนค่าทั้งคอลัมน์ขึ้นไปตามระดับของขั้นตอนการเปลี่ยนแปลงค่าตามลำดับขั้นเช่นเปลี่ยน B ให้อยู่ในระดับที่ 1 กล่าวคือมีค่าระดับการเจนเนอรัลไลเซชันเท่ากับ **(A:0, B:1, C:0, D:0, E:0)** จะได้ตามตารางที่ 8

ตารางที่ 8 ตัวอย่างข้อมูลหลังการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นของคอลัมน์ B ให้อยู่ในระดับที่ 1 หรือระดับเจนเนอรัลไลเซชันเท่ากับ (A:0, B:1, C:0, D:0, E:0)

Tuple-ID	A	B	C	D	E	Class
1	A1	B12	C1	D1	E1	Class1
2	A1	B12	C1	D1	E1	Class1
3	A1	B12	C1	D1	E1	Class1
4	A1	B12	C1	D1	E1	Class1
5	A3	B34	C2	D1	E1	Class1
6	A2	B12	C2	D2	E2	Class2
7	A2	B12	C2	D2	E2	Class2
8	A3	B12	C2	D2	E2	Class2
9	A3	B34	C2	D3	E2	Class2
10	A4	B34	C1	D3	E2	Class2
11	A3	B12	C1	D3	E3	Class3
12	A3	B12	C1	D3	E3	Class3
13	A2	B34	C2	D3	E3	Class3
14	A2	B34	C3	D4	E2	Class3
15	A1	B34	C4	D4	E2	Class3

จะเห็นว่าในการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity นั้นสามารถกำหนดระดับการเจนเนอรัลไลเซชันได้หลายแบบ ยกตัวอย่างเช่น กำหนดให้ k เท่ากับ 2 A,B,C,D และ E เป็นคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม ระดับเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity จะมีหลายค่าเช่นกัน ยกตัวอย่างเช่น ที่ระดับเจนเนอรัลไลเซชันเท่ากับ (A:2, B:3, C:1, D:0, E:0) ดังตารางที่ 9 หรือ ระดับเจนเนอรัลไลเซชันเท่ากับ (A:2, B:3, C:1, D:1, E:1) ดังตารางที่ 10 ก็ได้เช่นกัน

ตารางที่ 9 และตารางที่ 10 มีคุณสมบัติ *2*-Anonymity สามารถอธิบายอีกลักษณะได้ว่า ตารางที่ 9 สามารถจัดกลุ่มที่มีค่าตัวบ่งชี้บุคคลทางอ้อมเหมือนกันได้ 5 กลุ่มคือ 1) Tuple-ID = {01, 02, 03, 04, 05} 2) Tuple-ID = {06, 07, 08} 3) Tuple-ID = {09, 10} 4) Tuple-ID = {11, 12} ทุกกลุ่มมีจำนวนสมาชิกตั้งแต่ 2 ระเบียนขึ้นไป ส่วนตารางที่ 10 สามารถจัดกลุ่มที่มีค่าตัวบ่งชี้บุคคลทางอ้อมได้ 4 กลุ่มคือ 1) Tuple-ID = {01, 02, 03, 04, 05, 06, 07, 08} 2) Tuples-ID = {09, 10} 3) Tuple-ID = {11, 12, 13} 4) Tuple-ID = {14, 15} ทุกกลุ่มมีจำนวนสมาชิกตั้งแต่ 2 ระเบียนขึ้นไป

ตารางที่ 9 ตัวอย่างข้อมูลที่มีระดับเจนเนอรัลไลเซชันเป็น (A:2, B:3, C:1, D:0, E:0)

Tuple-ID	A	B	C	D	E	Class
1	*	*	C12	D1	E1	Class1
2	*	*	C12	D1	E1	Class1
3	*	*	C12	D1	E1	Class1
4	*	*	C12	D1	E1	Class1
5	*	*	C12	D1	E1	Class1
6	*	*	C12	D2	E2	Class2
7	*	*	C12	D2	E2	Class2
8	*	*	C12	D2	E2	Class2
9	*	*	C12	D3	E2	Class2
10	*	*	C12	D3	E2	Class2
11	*	*	C12	D3	E3	Class3
12	*	*	C12	D3	E3	Class3
13	*	*	C12	D3	E3	Class3
14	*	*	C34	D4	E2	Class3
15	*	*	C34	D4	E2	Class3

ตารางที่ 10 ตัวอย่างข้อมูลที่มีระดับเจนเนอรัลไลเซชันเป็น (A:2, B:3, C:1, D:1, E:1)

Tuple-ID	A	B	C	D	E	Class
1	*	*	C12	D12	E12	Class1
2	*	*	C12	D12	E12	Class1
3	*	*	C12	D12	E12	Class1
4	*	*	C12	D12	E12	Class1
5	*	*	C12	D12	E12	Class1
6	*	*	C12	D12	E12	Class2
7	*	*	C12	D12	E12	Class2
8	*	*	C12	D12	E12	Class2
9	*	*	C12	D34	E12	Class2
10	*	*	C12	D34	E12	Class2
11	*	*	C12	D34	E34	Class3
12	*	*	C12	D34	E34	Class3
13	*	*	C12	D34	E34	Class3
14	*	*	C34	D34	E12	Class3
15	*	*	C34	D34	E12	Class3

5.2 (α, *k*)-Anonymity จากตารางที่ 10 จะเห็นว่าในกลุ่มที่ 1) นั้นมีค่าของคอลัมน์คลาสอยู่สองค่า คือ Class1 และ Class2 แต่ในกลุ่มที่ 2) 3) และ 4) มีค่าของคอลัมน์คลาสเพียงค่าเดียว ถ้ามีการเทียบค่าจากตารางข้อมูลได้เป็น 3 กลุ่มหลังจะทำให้ทราบค่าคลาสได้ ยกตัวอย่าง เมื่อมีค่า A2,B3,C3,D4,E2 เข้ามาแต่ไม่ทราบว่าเป็นคลาสใดเมื่อทำการเปลี่ยนค่าตามระดับเจนเนอรัลไลเซชันตามตารางที่ 10 จะได้เป็น *,*,C34,D34,E12 แล้วเมื่อนำไปเทียบกับตารางที่ 10 ซึ่งอาจไม่สามารถระบุได้ว่าเป็นระเบียนใดระหว่าง Tuple-ID 14 หรือ 15 แต่จะเห็นว่าค่าคอลัมน์คลาสเป็นค่าเดียวกันคือ Class3 ใน [8] จึงได้เสนอรูปแบบ (α, *k*)-Anonymity ขึ้นมาโดยนอกเหนือจากคุณสมบัติของ k แล้วยังเพิ่มคุณสมบัติของค่า α ขึ้นมาซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1 ค่า α นี้ใช้เป็นตัวกำหนดการซ้ำกันของค่าคลาสเฝ้าระวังในแต่ละกลุ่มโดย จำนวนระเบียนในกลุ่มที่คู่กับค่าคลาสเฝ้าระวังหารด้วยจำนวนระเบียนในกลุ่ม ต้องมีค่าน้อยกว่าหรือเท่ากับ α

5.3 Algorithm MCCRT เป็นการขั้นตอนวิธีที่ใช้ค้นหาระดับเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity และระดับเจนเนอรัลไลเซชันที่ได้จะเหมาะสมกับการนำข้อมูลไปใช้ในการหา Associative Classification ต่อไป โดย MCCRT ได้นำค่า CCR ของแต่ละคอลัมน์มาคิดโดยการเปลี่ยนแปลงข้อมูลตามลำดับขั้นของแต่ละคอลัมน์จะเรียงจากคอลัมน์ที่มี CCR น้อยไปหาคอลัมน์ที่มี CCR มาก โดยในกรณีที่คอลัมน์มี CCR เท่ากันจะเรียงจาก ความสูงของขั้นตอนการเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นจากมากไปหาน้อย จากตัวอย่างตารางที่ 7

การคำนวณ CCR ของแต่ละ คอลัมน์

กำหนดให้ minsup = 2 tuples (20%)

minconf = 50%

คู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่จะนำมาคิดในแต่ละคอลัมน์ ต้องมีจำนวนระเบียนมากกว่า minsup และมีค่าความเชื่อมันมากกว่า minconf

คอลัมน์ A

A1->Class1 conf = $\frac{5}{6}$ = 83%

CCR = $\frac{5}{18}$ = 28%

คอลัมน์ B

B1->Class1 conf = $\frac{4}{6}$ = 66% B2->Class2 conf = $\frac{3}{5}$ = 50%

CCR = $\frac{7}{18}$ = 39%

คอลัมน์ C

C1->Class1 conf = $\frac{5}{6}$ = 83% C2->Class2 conf = $\frac{4}{5}$ = 80%

CCR = $\frac{9}{18}$ = 50%

คอลัมน์ D

D1->Class1 conf = $\frac{6}{6}$ = 100% D2->Class2 conf = $\frac{3}{5}$ = 60% D3->Class3 conf = $\frac{5}{7}$ = 71%

CCR = $\frac{14}{18}$ = 78% (มีทั้งหมด 12 tuples และ 9 คือจำนวนของ tuples ของ D1->C1 รวมกับ D2->C2)

คอลัมน์ E

E1->Class1 conf = $\frac{6}{6}$ = 100% E2->Class2 conf = $\frac{5}{5}$ = 100% E3->Class3 conf = $\frac{5}{7}$ = 71%

CCR = $\frac{16}{18}$ = 89%

นำคอลัมน์มาเรียงตามค่า CCR จากน้อยไปหามาก จะได้ลำดับ S คือ (A,B,C,D,E) ซึ่งในกรณีนี้ได้เหมือนเดิม แล้วทำการเปลี่ยนค่าจาก คอลัมน์แรกตามการเปลี่ยนแปลงข้อมูลตามลำดับขั้นขึ้นจนถึงระดับล่างจนถึงระดับบนสุดแล้วจึงค่อยเปลี่ยนแปลงค่าในคอลัมน์ถัดมาตามลำดับ S ทำจนข้อมูลผ่านคุณสมบัติ *k*-

Anonymity จะได้ระดับการเจนเนอรัลไลเซชันที่เป็นผลลัพธ์ออกมาจากกรณีนี้ผลลัพธ์ที่ดีคือ (A:2, B:3, C:1, D:0, E:0)

5.4 IncSpan ใน [7] ได้เสนอวิธีการหารูปแบบลำดับ (Sequential Patterns) เมื่อมีการเพิ่มข้อมูลเข้ามาในภายหลัง เพื่อความเข้าใจรูปที่ 11 ได้แสดงวิธีการหารูปแบบลำดับจากตารางฐานข้อมูลซึ่งเป็นข้อมูลการขายสินค้าให้ลูกค้าประจำโดยเริ่มจากข้อมูลที่ได้จากฐานข้อมูลในรูปแบบของตารางนำมาทำการปรับเปลี่ยนให้อยู่ในลักษณะของข้อมูลอนุกรมเวลาแล้วจึงนำไปหาค่ารูปแบบลำดับต่อไป

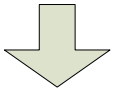
ในการตีความหมายของรูปแบบลำดับนั้น จากตัวอย่างรูปที่ 11 จะสามารถตีความหมายของรูปแบบลำดับ < (30) (40 70) > ได้ว่า มีลูกค้ามากกว่า 25% ที่ซื้อสินค้า 30 แล้วจะกลับมาซื้อสินค้า 40 พร้อมกับ 70 ซึ่ง 25% นั้นเป็นค่าสนับสนุนขั้นต่ำหรือค่ามินิมัมซัพพอท (minimum support) ที่กำหนดขึ้นมาซึ่งอาจอยู่ในลักษณะเป็นจำนวนระเบียนก็ได้ ส่วนการค้นหานั้นจะเห็นว่า < (30) (40 70) > ปรากฏในระเบียน ID ที่ 2 และ ID ที่ 4 ของตารางข้อมูลอนุกรมเวลาซึ่งมีทั้งหมด 5 ID แสดงว่า < (30) (40 70) > มีค่าสนับสนุนหรือค่าซัพพอท (Support) เท่ากับ $\frac{2}{5}$ หรือ 40% ถึงแม้ว่าในระเบียน ID ที่ 2 จะมี 60 มาขั้นแต่ก็ถือว่านับระเบียนนี้ได้ เพราะการหารูปแบบลำดับไม่สนใจการขั้นกลางของข้อมูลที่เวลาเดียวกัน และจะเห็นว่าในความเป็นจริงรูปแบบลำดับ < (30) >, < (40) >, < (70) >, < (30) (40) >, < (30) (70) > ก็มีค่าค่าสนับสนุนมากกว่า 25% แต่ในการหารูปแบบลำดับนั้นจะนับเฉพาะ รูปแบบลำดับที่ใหญ่ที่สุด ตัวอย่างเช่น < (30) (70) > เป็น ส่วนหนึ่งของ < (30) (40 70) > ฉะนั้นจึงนับแต่รูปแบบลำดับ < (30) (40 70) >

จากข้อมูลอนุกรมเวลา เมื่อมีการเพิ่มเข้ามาของข้อมูล [7] ได้แบ่งการเพิ่มเข้ามาของข้อมูลอนุกรมเวลาเป็นสองแบบคือ 1) แอพเพน (Append) หรือการเพิ่มข้อมูลการซื้อสินค้าของ Customer-ID เดิมที่มีอยู่แล้ว 2) อินเสิร์ท (Insert) หรือการเพิ่มเข้ามาของ Customer-ID กล่าวคือการมีลูกค้าประจำเพิ่มขึ้นมานั้นเอง หรือสามารถมองได้อีกลักษณะว่าเป็นการแอพเพนกับเซทว่างก็ได้ ดังตัวอย่างตารางที่ 11

ในการหารูปแบบลำดับของข้อมูลอนุกรมเวลาเมื่อมีข้อมูลเพิ่มเข้ามา [7] ได้สร้างโครงสร้างข้อมูลแบบต้นไม้ (Tree) ขึ้นมาช่วยในการเก็บข้อมูลรูปแบบลำดับและเพื่อเพิ่มความเร็วในการค้นหารูปแบบลำดับ โครงสร้างข้อมูลแบบต้นไม้นี้ สร้างจากการหารูปแบบลำดับของข้อมูลอนุกรมเวลาในครั้งแรก โดยแบ่งรูปแบบลำดับออกเป็น 3 กลุ่มคือ 1) รูปแบบลำดับที่เกิดขึ้นบ่อยครั้งหรือฟรีเควนทซีเควนท (Frequent Sequence) ซึ่งมีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ 2) รูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิหรือเซมิฟรีเควนท์ซีเควนท์ (Semi Frequent Sequence) ซึ่งมีค่าสนับสนุนเกือบถึงค่าสนับสนุนขั้นต่ำ โดยกำหนดว่าค่าสนับสนุนต้องน้อยกว่าค่าสนับสนุนขั้นต่ำ แต่มากกว่าค่าสนับสนุนขั้นต่ำคูณกับค่ามิว (μ) ซึ่งค่ามิวนี้จะมีค่าอยู่ระหว่าง 0 ถึง 1 และเป็นค่าที่กำหนดขึ้นมาเองตามความเหมาะสมของงาน 3) รูปแบบลำดับที่เกิดขึ้นไม่บ่อยครั้งหรืออินฟรีเควนทซีเควนท (Infrequent Sequence) ซึ่งมีค่าสนับสนุนน้อยกว่าค่าสนับสนุนขั้นต่ำคูณกับค่ามิว

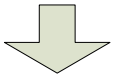
Customer ID	Transaction Time	Item Brought
1	June 25 '93	30
1	June 30 '93	90
2	June 10 '93	10, 20
2	June 15 '93	30
2	June 20 '93	40, 60, 70
3	June 25 '93	30, 50, 70
4	June 25 '93	30
4	June 30 '93	40, 70
4	July 25 '93	90
5	June 12 '93	90

ตารางข้อมูลการซื้อสินค้าของลูกค้า



Customer ID	Customer Sequence
1	< (30) (90) >
2	< (10 20) (30) (40 60 70) >
3	< (30 50 70) >
4	< (30) (40 70) (90) >
5	< (90) >

ตารางข้อมูลอนุกรมเวลา



Sequential Patterns with support > 25%
< (30) (90) >
< (30) (40 70) >

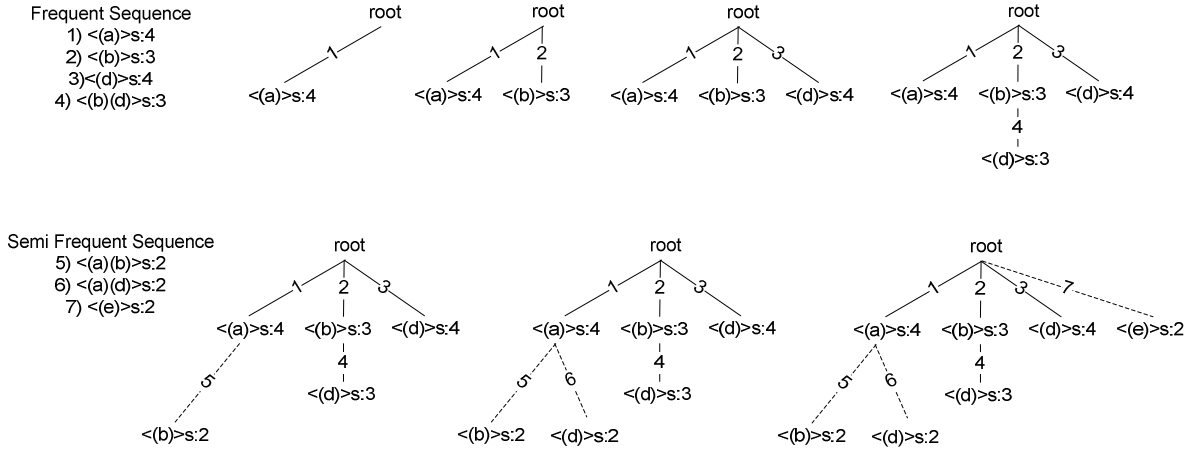
กฎความสัมพันธ์จากข้อมูลอนุกรมเวลา

รูปที่ 11 ขั้นตอนการหารูปแบบลำดับจากข้อมูลอนุกรมเวลา

ตารางที่ 11 ตัวอย่างการเติมข้อมูล

Customer-ID	Original Part	Appended Part
0	(a)(h)	(c)
1	(eg)	(a)(bce)
2	(a)(b)(d)	(ck)(l)
3	(b)(df)(a)(b)	∅
4	(a)(d)	∅
5	(be)(d)	∅
New Insert 6	∅	(cd)(e)
New Insert 7	∅	(b)(d)

ในการสร้างโครงสร้างข้อมูลรูปต้นไม้นั้นจะนำรูปแบบลำดับในกลุ่ม 1) และ 2) มาใช้ในการสร้าง พิจารณาจากตารางที่ 11 ในส่วนของ Original Part ซึ่งเป็นส่วนที่ยังไม่มีการเพิ่มเข้ามาของข้อมูล สมมติกำหนดให้ค่าสนับสนุนขั้นต่ำเท่ากับ 3 ระเบียน ค่ามิวเท่ากับ 0.5 ซึ่งแสดงว่ารูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิต้องมี 2 ระเบียน เริ่มต้นโดยทำการหารูปแบบลำดับที่เกิดขึ้นบ่อยครั้งซึ่งในที่นี้ผลลัพธ์คือ < (a) > มีค่าสนับสนุนเท่ากับ 4, < (b) > มีค่าสนับสนุนเท่ากับ 3, < (d) > มีค่าสนับสนุนเท่ากับ 4, < (b)(d) > มีค่าสนับสนุนเท่ากับ 3 และรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิคือ < (a)(b) >, < (a)(d) >, < (e) > ซึ่งทุกรูปแบบลำดับมีค่าสนับสนุนเท่ากับ 2 ต่อมานำรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งและรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิที่ได้ทำตามขั้นตอนในรูปที่ 12 จะได้โครงสร้างข้อมูลรูปต้นไม้

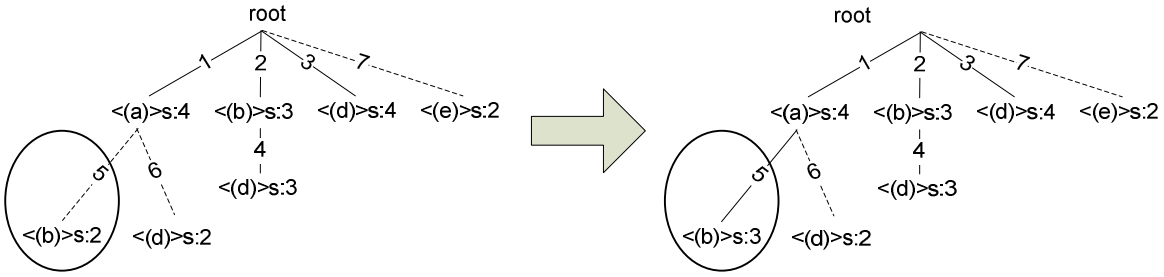


รูปที่ 12 ขั้นตอนการสร้างโครงสร้างรูปต้นไม้

เมื่อทำการสร้างโครงสร้างข้อมูลรูปต้นไม้แล้ว จะเห็นว่าเมื่อมีการเพิ่มข้อมูล ตัวอย่างเช่นตารางที่ 11 จะต้องทำการปรับโครงสร้างข้อมูลรูปต้นไม้ใหม่อีกครั้ง ใน [7] ได้นำเสนอแนวคิดในการทำโครงสร้างรูปต้นไม้ให้เป็นปัจจุบัน (Update) โดยมองว่าข้อมูลทุกข้อมูลใหม่ที่เพิ่มเข้ามาเป็นการแอพเพนและแยกกรณีการปรับโครงสร้างข้อมูลรูปต้นไม้เมื่อมีการเพิ่มเข้ามาของข้อมูลออกเป็น 6 กรณีคือ 1) ทำให้รูปแบบลำดับที่เกิดขึ้นบ่อยครั้งยังคงเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้ง 2) รูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิกลายเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้ง 3) ทำให้รูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิยังคงเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิ 4) การแอพเพนทำให้เกิดสินค้าใหม่ (New item) 5) ทำให้รูปแบบลำดับที่เกิดขึ้นไม่บ่อยครั้งกลายเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้ง 6) ทำให้รูปแบบลำดับที่เกิดขึ้นไม่บ่อยครั้งกลายเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิ

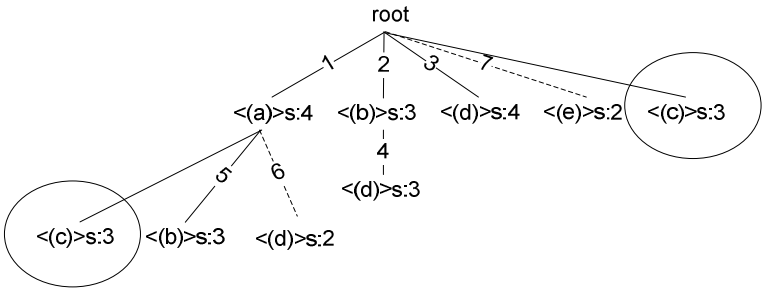
การปรับโครงสร้างข้อมูลรูปต้นไม้ให้เป็นปัจจุบันในแต่ละกรณี

<u>กรณีที่ 1) 2) และ 3)</u> สามารถใช้วิธีเดียวกันได้ โดยการอ่านเฉพาะข้อมูลที่เพิ่มเข้ามาหาค่าสนับสนุนที่เพิ่มขึ้นแล้วทำการปรับโครงสร้างข้อมูลรูปต้นไม้ได้ ยกตัวอย่างเช่น จากตารางที่ 11 ระเบียนที่ 7 การแอพเพนทำให้เกิดกรณีที่ 2) รูปแบบลำดับ < (a) (b) > มีค่าสนับสนุนจาก 2 เป็น 3 ทำให้จากเดิมที่เป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิกลายเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้ง ในการปรับโครงสร้างข้อมูลรูปต้นไม้ให้ทำการเปลี่ยนจากเส้นประเป็นเส้นทึบแล้วปรับค่าสนับสนุนเป็น 3 ดังรูปที่ 13 ส่วนในกรณีที่ 1) และ 3) เพียงปรับค่าสนับสนุนเท่านั้น



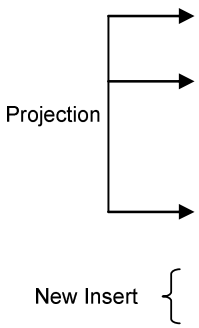
รูปที่ 13 การเปลี่ยนโครงสร้างข้อมูลรูปต้นไม้ต้นไม้ในกรณีที่ 2)

<u>กรณีที่ 4)</u> พิจารณาการแอพเพนในระเบียนที่ 0, 1 และ 2 จะเห็นว่าการแอพเพนได้เพิ่มสินค้า c ซึ่งไม่มีอยู่ในโครงสร้างข้อมูลรูปต้นไม้เมื่อพิจารณาหลังการแอพเพนจะได้รูปแบบลำดับใหม่อีกหลายลำดับที่เกิดจากเพิ่มเข้ามาของ c, k และ l แต่จะมีเพียงรูปแบบลำดับ < (c) > กับ < (a) (c) > ที่มีค่าสนับสนุนมากพอจะเพิ่มในโครงสร้างข้อมูลรูปต้นไม้ ซึ่งในกรณีนี้จะเป็นลำดับที่เกิดขึ้นบ่อยครั้งในการปรับโครงสร้างจะทำการเพิ่ม < (c) > ต่อจาก root กับ < (c) > ต่อกับ < (c) > โดยใช้เส้นทึบดังผลลัพธ์จะได้ดังรูปที่ 14



รูปที่ 14 การเปลี่ยนโครงสร้างข้อมูลรูปต้นไม้ในกรณีที่ 4)

<u>กรณีที่ 5) และ 6)</u> พิจารณาการแอพเพนในระเบียนที่ 1 จะเห็นว่า < (b e) > ไม่มีอยู่ในโครงสร้างข้อมูลรูปต้นไม้แต่ปรากฏในตารางระเบียนที่ 5 ในการปรับโครงสร้างข้อมูลจำเป็นต้องอ่านข้อมูลจากตารางอีกครั้งเพื่อค้นหาว่ารูปแบบลำดับ < (b e) > จะมีค่าสนับสนุนเป็นเท่าไร เพียงพอที่จะเป็นรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งหรือรูปแบบลำดับที่เกิดขึ้นบ่อยครั้งแบบเซมิหรือไม่ เพื่อนำไปเพิ่มในโครงสร้างต้นไม้ ในการค้นหานั้นถ้าใช้การค้นหาจากบนลงล่างจะทำให้ช้าใน [7] จึงใช้เทคนิคการโปรเจคชัน (Projection) เข้ามาช่วยซึ่งการโปรเจคชันจะเข้าถึงข้อมูลไปที่ละค่า เช่นมอง < (b e) > ก็เข้าถึงทุกระเบียนที่ มี b ก่อนจากนั้นจึงเข้าถึง < (b e) > โดยการเข้าถึงจะเป็นการเข้าถึงโดยตรง



Customer-ID	Original Part	Appended Part
0	(a)(h)	(c)
1	(eg)	(a)(bce)
2	(a)(b)(d)	(ck)(l)
3	(b)(df)(a)(b)	∅
4	(a)(d)	∅
5	(be)(d)	∅
6	∅	(cd)(e)
7	∅	(b)(d)

รูปที่ 15 การโปรเจคชันโดย b

ฉะนั้นโครงสร้างข้อมูลสุดท้ายหลังการทำงานของทั้ง 6 กรณีแล้วจะเป็นโครงสร้างข้อมูลต้นไม้ที่ให้รูปแบบลำดับครบตามที่ต้องการเหมือนการ อ่านข้อมูลใหม่ทั้งหมดเพื่อหารูปแบบลำดับอีกครั้ง

6. สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง

บวรศักดิ์ และคณะ [1] ได้ชี้ให้เห็นว่าองค์กรที่เก็บรวมรวมข้อมูลส่วนบุคคลบนระบบอินเทอร์เน็ตในประเทศไทยยังไม่ให้ความสำคัญกับปัญหาความเป็นส่วนตัวของข้อมูลเท่าไหร่นัก และในประเทศไทย

กฎหมายที่ใช้ในการปกป้องข้อมูลส่วนบุคคลยังอยู่ในขั้นตอนการร่างกฎหมายและยังคงต้องปรับปรุงอยู่ ฉะนั้นในปัจจุบันยังมีความจำเป็นที่งานทางด้านวิศวกรรมคอมพิวเตอร์ควรทำการศึกษาและพัฒนาเทคนิคในการป้องกันข้อมูลส่วนบุคคลเพื่อการเผยแพร่สู่สาธารณะ

Sweeney [3] ได้นำเสนอเทคนิค k-Anonymity ซึ่งเป็นเทคนิคการรักษาความเป็นส่วนตัวของข้อมูล งานวิจัยนี้ถูกงานวิจัยจำนวนมากนำเทคนิค k-Anonymity ไปประยุกต์ใช้ เนื่องจากเป็นเทคนิคที่เข้าใจได้โดยง่ายและสามารถนำไปใช้อย่างมีประสิทธิภาพในการรักษาความเป็นส่วนตัวของข้อมูล

Sweeney [4] ได้นำเสนอเทคนิค k-Anonymity มาใช้ในการรักษาความเป็นส่วนตัวของข้อมูล โดยวิธีการแปลงข้อมูลให้อยู่ในค่าที่ทั่วไป และวิธีการปิดบังข้อมูล และยังได้นำเสนอการกำหนดคุณภาพข้อมูลแบบทั่วไป คือ Precision metric ซึ่งจะใช้กำหนดคุณภาพข้อมูลแบบทั่วไปได้ดีเฉพาะกรณีที่โครงสร้างโดเมนในแต่ละแอตทริบิวต์ที่ถูกแปลงไม่แตกต่างกันมาก

ณัฐพลและคณะ [5] ได้เสนอขั้นตอนวิธีเพื่อใช้ในการเลือกระดับเจนเนอรัลไลเซชันโดยใช้การวัดค่าการบิดเบือนของข้อมูล (Distortion Ratio: C_{GM}) และตัววัดคุณภาพข้อมูลสำหรับการจำแนกแบบความสัมพันธ์ (Frequency-based Classification: C_{FCM}) ของข้อมูลที่ถูกเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้นที่มีค่าน้อยที่สุด วิธีการนี้ลดค่าความซับซ้อนเชิงคำนวณในการหาระดับเจนเนอรัลไลเซชันเพื่อนำข้อมูลไปหาค่า Associative Classification

ณัฐพลและคณะ [6] ได้เสนอขั้นตอนวิธีแบบศึกษาสำนึก MCCRT เพื่อใช้ในการหาระดับการเจนเนอรัลไลเซชันที่เหมาะสมกับการนำข้อมูลไปใช้ในการหา Associative Classification โดยได้ลดจำนวนการทดสอบคุณสมบัติ *k*-Anonymity ลงโดยใช้ค่า CCR ของแต่ละคอลัมน์มาใช้ในการกำหนดเส้นทางการทดสอบคุณสมบัติ *k*-Anonymity ระดับเจนเนอรัลไลเซชันที่ได้มีความแม่นยำมากกว่าขั้นตอนวิธีที่เหมาะที่สุดและมีค่าความซับซ้อนเชิงคำนวณเพียง $O(n\log n)$

Cheng และคณะ [7] ได้เสนอวิธีการในการหารูปแบบลำดับเมื่อมีการเพิ่มเข้ามาของข้อมูลโดยไม่ต้องทำการอ่านข้อมูลใหม่ทั้งหมด โดยการสร้างโครงสร้างข้อมูลรูปต้นไม้ขึ้นมาแทนการเก็บค่าแบบตาราง และยังเสนอแนวคิดในการแก้ปัญหาโดยการแยกกรณีที่เกิดขึ้นได้เป็น 6 กรณีและทำการแก้ปัญหาไปทีละกรณีจนทำให้สามารถแก้ปัญหาหลักที่ต้องการได้

Wong และคณะ [8] ได้เสนอเทคนิคในการแก้ปัญหาเพิ่มเติมที่ *k*-Anonymity ไม่สามารถแก้ไขได้ คือการมีค่าคลาสเดียวกันของกลุ่มข้อมูลที่ถูกเปลี่ยนแปลงค่าของข้อมูลตามลำดับขั้น โดยการเพิ่มค่า α ที่ใช้เป็นตัวชี้วัดค่าคลาสที่ซ้ำกันในกลุ่มข้อมูลเฝ้าระวัง

7. วัตถุประสงค์ของการศึกษา

7.1 เพื่อวิเคราะห์การเปลี่ยนแปลงของข้อมูลเมื่อมีการเพิ่มขึ้นของข้อมูลในแง่ของความเป็นส่วนตัวข้อมูลและคุณภาพข้อมูลสำหรับการจำแนกแบบสัมพันธ์

7.2 เพื่อพัฒนาโครงสร้างข้อมูลที่สามารถลดค่าความซับซ้อนเชิงคำนวณในการหาระดับเจนเนอรัลไลเซชันเมื่อมีการเพิ่มขึ้นของข้อมูล

7.2 เพื่อพัฒนาขั้นตอนวิธีในการหาระดับเจนเนอรัลไลเซชันตามขั้นตอนวิธี Minimum Classification Correction Rate Transformation (MCCRT) และขั้นตอนวิธีที่เหมาะที่สุด ซึ่งเมื่อมีการเพิ่มขึ้นของข้อมูล จะมีค่าความซับซ้อนในเชิงคำนวณน้อยกว่าและไม่จำเป็นต้องทำการประมวลผลข้อมูลใหม่ทั้งหมด

8. ประโยชน์ที่ได้รับจากการศึกษา เชิงทฤษฎี และ/หรือ เชิงประยุกต์

องค์กรใดๆ ก็ตามที่มีการรวบรวมข้อมูลส่วนบุคคลต้องการเผยแพร่ข้อมูลซึ่งจำเป็นต้องมีการรักษาความเป็นส่วนตัวของข้อมูลและข้อมูลเหล่านี้จะเพิ่มขึ้นเรื่อยๆ เช่นการเผยแพร่ข้อมูลการรักษาโรคของผู้ป่วยในโรงพยาบาลซึ่งจะเพิ่มขึ้นเรื่อยๆในทุกๆ เดือน สามารถใช้ผลลัพธ์จากวิทยานิพนธ์นี้ในการรักษาความเป็นส่วนตัวของข้อมูลได้อย่างมีประสิทธิภาพ

9. แผนดำเนินการ ขอบเขต และวิธีการทำวิจัย

9.1 แผนดำเนินการวิจัย

9.1.1 ศึกษาทฤษฎีที่เกี่ยวข้องกับการรักษาความเป็นส่วนตัวของข้อมูล

9.1.2 ศึกษาทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองคุณภาพข้อมูลที่จำเป็นต่อการจำแนกแบบความสัมพันธ์

9.1.3 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลให้มีคุณภาพข้อมูลสูง ในสถานการณ์การวิเคราะห์ข้อมูลสำหรับการจำแนกแบบความสัมพันธ์

9.1.4 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล

9.1.5 ตรวจสอบการรักษาความเป็นส่วนตัวของฐานข้อมูลที่ถูกแปลงจากคุณสมบัติ k-Anonymity

9.1.6 ตรวจสอบความถูกต้องของข้อมูลจากการแปลงข้อมูลตามทฤษฎีการเพิ่มเข้ามาของข้อมูลเปรียบเทียบกับการแปลงข้อมูลตามปรกติ

9.1.7 ตรวจสอบความเร็วในการการแปลงข้อมูลตามทฤษฎีการเพิ่มเข้ามาของข้อมูลเปรียบเทียบกับการแปลงข้อมูลตามปรกติ

9.2 ขอบเขตการทำวิจัย

9.2.1 การประเมินประสิทธิภาพของขั้นตอนวิธีที่พัฒนาขึ้นจะใช้เวลาจริง (วินาที) และความซับซ้อนเชิงคำนวณ

9.2.2 ข้อมูลที่จะใช้ทดสอบประสิทธิภาพของขั้นตอนวิธีที่พัฒนาขึ้นจะใช้ข้อมูลจาก UCI Repository [9] ซึ่งเป็นแหล่งรวบรวมข้อมูลที่มักจะถูกใช้ในการวิจัยเกี่ยวการทำเหมืองข้อมูล และข้อมูลสังเคราะห์

9.3 วิธีการทำวิจัย

9.3.1 ศึกษาทฤษฎีที่เกี่ยวข้องกับการรักษาความเป็นส่วนตัวของข้อมูล

9.3.2 ศึกษาทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองคุณภาพข้อมูลที่จำเป็นต่อการจำแนกแบบกฎความสัมพันธ์

9.3.3 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลให้มีคุณภาพข้อมูลสูงในสถานการณ์การวิเคราะห์ข้อมูลสำหรับการจำแนกแบบความสัมพันธ์

9.3.4 ศึกษาทฤษฎีที่เกี่ยวข้องกับการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล

9.3.5 ออกแบบการทดลองและพัฒนาแบบจำลองข้อมูลที่ใช้ในการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล

9.3.5 ทำการทดลองและพัฒนาขั้นตอนวิธีการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูลตามขั้นตอนวิธี MCCRT และขั้นตอนวิธีที่เหมาะที่สุด

9.3.6 วิจารณ์ สรุปผลการทำวิจัย จัดทำและเสนอรายงานวิทยานิพนธ์

10. สถานที่ที่ทำการวิจัย

ห้องวิจัยความฉลาดทางการคำนวณ (Computational Intelligence Research Laboratory) ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่

11. ระยะเวลาดำเนินงานวิจัย

ใช้ระยะเวลาในการดำเนินงานวิจัย 8 เดือน ดังรายละเอียดในตารางที่ 2

การดำเนินการ	ระยะเวลา(เดือนที่)					
	1	2	3	4	5	6
1. ศึกษาเอกสารข้อมูล ทฤษฎีและงานวิจัยที่เกี่ยวข้อง						
2. ศึกษาทฤษฎีที่เกี่ยวข้องกับงานวิจัย						
3. ออกแบบและพัฒนาแบบจำลองข้อมูลที่ใช้ในการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูล						
4. ออกแบบและสร้างขั้นตอนวิธีการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูลตามขั้นตอนวิธี MCCRT และขั้นตอนวิธีที่เหมาะ						

ที่สุด						
5. พัฒนาขั้นตอนการแปลงฐานข้อมูลเมื่อมีการเพิ่มขึ้นมาของข้อมูลตามขั้นตอนวิธี MCCRT และขั้นตอนวิธีที่เหมาะที่สุด						
6. วิเคราะห์ผลที่ได้ จากการทดลอง และสรุปผล						
7. จัดทำวิทยานิพนธ์ฉบับสมบูรณ์						

12. เอกสารอ้างอิง

[1] Seisungsittisunti, B., Natwichai, J., and Harnsamut, N. (2008). Internet Privacy Problem in Thailand. Nectec Technical Journal, 8, pp. 162-166, September 2008.

[2] รายการอาการในรหัส ICD-10. Retrieved Feb 20, 2009, from Wikipedia: http://th.wikipedia.org/wiki/รายการอาการในรหัส_ICD-10

[3] Sweeney, L. (2002). k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10, pp 557-570, 2002.

[4] Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10, pp 571-588, 2002.

[5] Harnsamut, N., Natwichai, J., Sun, X., Li, X. (2008). Data quality in privacy preserving for associative classification. In: Proceedings of The Fourth International Conference on Advanced Data Mining and Applications, pp 111-122, 2008.

[6] Harnsamut, N., and Natwichai, J. (2008). A Novel Heuristic Algorithm for Privacy Preserving of Associative Classification. In: Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence , 5351, pp. 273-283, 2008.

[7] Cheng, H., Yan, X., and Han, J. (2004). IncSpan: incremental mining of sequential patterns in large database. In: Proceedings of the Tenth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, ACM Press, pp 527-532, 2004.

[8] Wong, R.C.W., Li, J., Fu, A.W.C., Wang, K. (2006). (α, k)-anonymity: an enhanced *k*-anonymity model for privacy preserving data publishing. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, pp 754-759, 2006.

[9] UC Irvine Machine Learning Repository. (2007). Retrieved July 23, 2008, from University of California: Irvine. Web site: http://archive.ics.uci.edu/ml/

ภาคผนวก 1 ชุดข้อมูลมาตรฐานของการประกันสุขภาพ