บทที่ 3

ปัญหาและวิธีการแก้ปัญหา

การแปลงข้อมูลใดๆให้ไม่สามารถกลับมาระบุตัวบุคคลอีกครั้งได้ สามารถทำได้โดยการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity และในวิทยานิพนธ์นี้ได้ใช้ขั้นตอนวิธี MCCRT (Minimum Classification Correction Rate Transformation algorithm) ซึ่งถูกนำเสนอโดย [6] เป็นขั้นตอนวิธีในการทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity ซึ่งข้อมูลหลังการแปลงตามขั้นตอนวิธี MCCRT นี้เหมาะสมที่จะนำไปใช้กับการจำแนกแบบความสัมพันธ์ (Associative Classification) แต่ขั้นตอนวิธี MCCRT นี้ยังอยู่ในลักษณะ One-Time Fashion [7] กล่าวคือ เมื่อนำข้อมูลใดๆ มาผ่านกระบวนการตามขั้นตอนวิธี MCCRT และผลลัพธ์ที่ได้เป็นข้อมูลที่ถูกแปลงให้มีคุณสมบัติ *k*-Anonymity แล้ว ถ้าข้อมูลเริ่มต้นมีการเพิ่มเติมข้อมูลเข้ามา เมื่อต้องการผลลัพธ์ของข้อมูลทั้งหมดใหม่อีกครั้ง ต้องนำข้อมูลทั้งหมดผ่านกระบวนการตามขั้นตอน MCCRT ใหม่อีกครั้งด้วย

ในสถานการณ์จริงนั้นข้อมูลที่เก็บรวบรวมไว้จะถูกเพิ่มเข้ามาเรื่อยๆ ซึ่งทำให้การประมวลผลข้อมูลผ่านกระบวนการตามขั้นตอนวิธี MCCRT ในแต่ละครั้งที่มีการเพิ่มเข้ามาของข้อมูลนานยิ่งขึ้นเนื่องจากต้องประมวลผลใหม่ตั้งแต่ต้น จากการสังเกตกระบวนการนี้พบว่า ในการผ่านกระบวนการตามขั้นตอนวิธี MCCRT ครั้งแรกของข้อมูลเริ่มต้นที่ยังไม่มีการเพิ่มเข้ามาของข้อมูลนั้น มีผลลัพธ์ที่สามารถนำกลับมาใช้ใหม่เพื่อช่วยในการทำงานในกระบวนการรอบถัดไปเมื่อมีการเพิ่มเข้ามาของข้อมูลได้ ซึ่งอาจทำให้ไม่ต้องเริ่มต้นกระบวนการใหม่ทั้งหมดแต่ผลลัพธ์ที่ได้ยังคงเหมือนการผ่านกระบวนการใหม่ทั้งหมดอีกครั้งและอาจทำให้ประสิทธิภาพในการทำงานเพิ่มขึ้น ในวิทยานิพนธ์นี้ต้องการค้นหาขั้นตอนวิธีที่สามารถแก้ปัญหาในจุดดังกล่าว โดยได้นำหลักการการประมวลผลแบบเพิ่มขึ้น หรือ Incremental ที่นำเสนอโดย [7] มาประยุกต์ใช้ในการแก้ปัญหา ซึ่งทำให้ไม่ต้องทำการประมวลผลตาม MCCRT ใหม่ทั้งหมดเมื่อมีข้อมูลเพิ่มเข้ามาและผลลัพธ์ที่ได้เหมือนกับการประมวลผลตามขั้นตอนวิธี MCCRT ใหม่อีกครั้ง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.1 นิยามเบื้องต้น

3.1.1 ข้อมูล (Dataset)

ให้ข้อมูล $D = \{d^1, d^2, \ldots, d^n\}$ เป็นกลุ่มของระเบียนที่นิยามโดยเค้าร่าง (Schema) A สำหรับแต่ละ คอลัมน์ $A_j \in A$ แสดงถึงโดเมนของตัวเองโดยสัญลักษณ์ $dom(A_j)$ สำหรับแต่ละ $d^i \in D$ ให้ $d^i(A) = (d^i(A_1), d^i(A_2), \ldots, d^i(A_k))$ แสดงแทนโดยสัญลักษณ์ $(d^i_1, d^i_2, \ldots, d^i_k)$ และแต่ละระเบียนในข้อมูลสามารถมีค่าซ้ำกันได้

ให้ C เป็นเซตของคลาสโดย $C = \{c_1, c_2, \ldots, c_o\}$ เมื่อ $c_m \in C$ และคลาสแต่ละระเบียนแทนด้วยสัญลักษณ์ $d^i.Class$

3.1.2 การหากฎการจำแนก (Associative Classification)

ให้สัญพจน์ (Literal) p เป็นคู่ของคอลัมน์กับค่าข้อมูล ซึ่ง p ประกอบด้วยคอลัมน์ A_j กับค่าข้อมูล v ที่อยู่ใน $dom(A_j)$ โดยระเบียน d^i จะสอดคล้อง (Satisfy) กับ $p(A_j, v)$ ก็ต่อเมื่อ $d^i_j = v$

ให้ข้อมูล D เป็นกลุ่มของระเบียน ให้ C เป็นเซตของคลาส ให้ R เป็นเซตของกฎการจำแนก (Classification Rule) โดย $R = \{r_1, r_2, \ldots, r_q\}$ และให้ L เป็นตัวบ่งชี้กฎของ R โดย $L = \{1, \ldots, q\}$

สำหรับทุก $l \in L$ จะได้ว่ากฎการจำแนก r_l มีความหมายว่า $\wedge p \rightarrow c_m$ โดยด้านซ้ายของกฎการจำแนก r_l (*LHS*) คือประพจน์เชื่อม (Conjunction: $\wedge$) ของ p เขียนแทนด้วย $r_l.LHS$ และด้านขวาของกฎการจำแนก r_l (*RHS*) เป็นค่าคลาส c_m เขียนแทนด้วย $r_l.RHS$

ระเบียน d^i จะสอดคล้องกับกฎการจำแนก r_l ก็ต่อเมื่อระเบียน d^i สอดคล้องกับกฎการจำแนก $r_l.LHS$ และมีค่า $d^i.class$ เท่ากับ $r_l.RHS$

ระเบียน d^i ที่สอดคล้องกับกฎการจำแนก r_l จะถูกเรียกว่าเป็นระเบียนที่สนับสนุนกฎการจำแนก r_l ค่าสนับสนุนของกฎการจำแนก r_l เขียนแทนด้วย $Sup(r_l)$ จะหมายถึงอัตราส่วนระหว่างจำนวนระเบียนที่สนับสนุนกฎการจำแนก r_l กับจำนวนระเบียนทั้งหมด ค่าความเชื่อมั่น

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ของกฎการจำแนก r_l เขียนแทนด้วย $Conf(r_l)$ จะหมายถึงอัตราส่วนระหว่าง $Sup(r_l)$ กับจำนวนระเบียนที่สอดคล้องกับ $r_l.RHS$

3.1.3 ตัวบ่งชี้บุคคลทางอ้อม (Quasi-Identifier)

เซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมของข้อมูล D เขียนแทนด้วย Q_D ซึ่งเซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม Q_D จะเป็นซับเซตของเค้าร่างของคอลัมน์ A ที่สามารถนำมาระบุตัวบุคคลอีกครั้งได้โดยการพิจารณาร่วมกับข้อมูลอื่น

3.1.4 ข้อมูลรูปแบบ *k*-Anonymity

ข้อมูล D จากเค้าร่าง A กับเซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม Q_D จะสอดคล้องกับ *k*-Anonymity ก็ต่อเมื่อ แต่ละระเบียน $d^i \in D$ จะมีระเบียนอื่นที่มีค่าเหมือนกันอีก $k-1$ ระเบียน ให้ระเบียนเหล่านั้นเป็น $d^{i^1}, d^{i^2}, \ldots, d^{i^{k-1}} \in D$ ดังนั้น $d^{i^1}_j = d^{i^2}_j = \ldots = d^{i^{k-1}}_j$ ของ $\forall A_j \in A$

3.1.5 การแปลงค่าข้อมูล (Generalization)

ให้โดเมน $dom^l(A_j) = \{P_1, P_2, \ldots\}$ เป็นระดับการเจนเนอรัลไลเซชันระดับที่ l ของคอลัมน์ A_j ซึ่ง $\bigcup P_{jt} = dom^0(A_j)$ และ $P_{jt} \bigcap P_{jl} = \varnothing$ โดยที่ $dom^0(A_j)$ เป็นโดเมนดั้งเดิมของคอลัมน์ A_j สำหรับแต่ละคอลัมน์ A_j ให้ความสูงของขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้น (Hierarchy) เป็น h_j สำหรับค่าข้อมูล v ใน $dom^l(A_j)$ มีระดับการเจนเนอรัลไลเซชันระดับที่ l เป็นค่า P_{jt} ซึ่งต้องอยู่ใน $dom^l(A_j)$ โดยเขียนเป็น $\phi_{dom^l(A_j)}(v)$

ให้ $\prec_G$ เป็นอันดับบางส่วน (Partial order) บนโดเมน $dom^l(A_j) \prec_G dom^{l'}(A_j)$ ก็ต่อเมื่อ $dom^{l'}(A_j)$ เป็นเจนเนอรัลไลเซชันของ $dom^l(A_j)$

3.1.6 การแปลงข้อมูลให้มีคุณสมบัติ *k*-Anonymity (*k*-Anonymization)

การแปลงข้อมูลให้มีคุณสมบัติ *k*-Anonymity คือการแปลงข้อมูลเริ่มต้น D ให้เป็นข้อมูลใหม่ D' โดยที่ข้อมูลใหม่ D' เป็นการเจนเนอรัลไลเซชันของข้อมูลเริ่มต้น D ที่สอดคล้องกับคุณสมบัติของ *k*-Anonymity

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.1.7 ระดับการเจนเนอรัลไลเซชัน (Generalization Level : *GL*)

ให้ $dom(Q_D)$ เป็นเซตของโดเมนการเจนเนอรัลไลเซชันสำหรับคอลัมน์ในเซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม Q_D ให้เขียนแทนด้วย ระดับการเจนเนอรัลไลเซชัน GL ซึ่งเป็นลำดับของคู่คอลัมน์กับระดับ (A_j, l_j) โดยที่ l_k เป็นระดับในการเจนเนอรัลไลเซชันของคอลัมน์ A_k ซึ่งมีค่าระหว่าง 0 ถึง h_j

3.1.8 อัตราความถูกต้องในการจำแนก (Classification Correction Rate: *CCR*)

อัตราความถูกต้องในการจำแนกคือการวัดความแม่นยำของการจำแนกที่หาได้ โดยนำกฎการจำแนกไปทำนายค่าคลาสจากข้อมูลแต่ละระเบียนเพื่อวัดอัตราความถูกต้อง โดยในที่นี้กฎการจำแนกคือคู่ระหว่างค่าของคอลัมน์กับค่าคลาส (Frequency pair: *fp*) ที่เป็นสมาชิกของ *FP* โดย *FP* คือเซตของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่มีค่าสนับสนุนมากกว่าค่าสนับสนุนขั้นต่ำ (Minimum Support: *minsup*) และมีค่าความความเชื่อมั่นมากกว่าค่าความเชื่อมั่นขั้นต่ำ (Minimum Confidence: *minconf*) และข้อมูลที่ใช้ในการทำนายคือ *D*

เนื่องจากด้านซ้ายของกฎการจำแนกมีค่าเพียงค่าเดียวทำให้ไม่เกิดการซ้ำซ้อนในการทำนาย การหาอัตราความถูกต้องในการจำแนกจึงสามารถหาได้จากผลรวมของค่าขนาดของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่เป็นสมาชิกของ *FP* หารด้วยขนาดของข้อมูล (จำนวนระเบียนทั้งหมด) ซึ่งสามารถเขียนเป็นสมการได้ตามสมการที่ 3.1

$$CCR_{a \in Q_D} = \frac{\sum_{fp \in FP} |fp|}{|D|} \tag{3.1}$$

3.2 ขั้นตอนวิธี MCCRT (Minimum Classification Correction Rate Transformation algorithm)

จากรูปที่ 3.1 ได้แสดงรหัสเทียม (Pseudo code) ของขั้นตอนวิธี MCCRT ซึ่งเริ่มจากเรียงลำดับคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมโดยเรียงตามค่าอัตราความแม่นยำในการจำแนก *CCR* (Classification Correction Rate) ของแต่ละคอลัมน์ซึ่งคำนวณจากค่าเฉลี่ยของอัตราความแม่นยำในการบ่งบอกค่าคลาสของทุกค่าข้อมูลที่อยู่ในคอลัมน์ การเรียงลำดับจะเรียงจากคอลัมน์ที่มีค่าอัตราความแม่นยำในการจำแนกน้อยไปหาคอลัมน์ที่มีค่าอัตราความแม่นยำในการจำแนกมาก

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

โดยจะเก็บผลลัพธ์ของการจัดลำดับคู่กับระดับการเจนเนอรัลไลเซชันปัจจุบันของคอลัมน์นั้นไว้ในระดับการเจนเนอรัลไลเซชัน *GL*

หลังการเรียงลำดับจะทำการแปลงค่าข้อมูลไปที่ละระดับจนถึงระดับบนสุดของขั้นตอนการแปลงค่าข้อมูลทีละคอลัมน์ตามลำดับของคอลัมน์ที่เรียงไว้ ขั้นตอนวิธีนี้จะหยุดเมื่อข้อมูลที่ถูกแปลงแล้วมีคุณสมบัติ *k*-Anonymity

ค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT แบ่งออกเป็น 2 ส่วนคือส่วนที่ใช้ในเรียงลำดับคอลัมน์ตามค่าอัตราความแม่นยำในการจำแนก หรือ *CCR* ซึ่งคือบรรทัดเลขที่ 1 ของรูปที่ 3.1 โดยมีค่าความซับซ้อนเชิงคำนวณเท่ากับ $k \times n$ โดย n เป็นจำนวนระเบียนและ k คือจำนวนคอลัมน์ ซึ่งการคำนวณอัตราความแม่นยำในการจำแนกของแต่ละคอลัมน์ใช้วิธีการนับจำนวนระเบียนทำให้มีค่าความซับซ้อนเชิงคำนวณเท่ากับ n เมื่อต้องทำ k คอลัมน์แสดงว่าต้องนับเป็นจำนวน k รอบทำให้ค่าความซับซ้อนเชิงคำนวณของส่วนนี้เท่ากับ $k \times n$ ความซับซ้อนเชิงคำนวณส่วนต่อมาคือส่วนการหาระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลมีคุณสมบัติ *k*-Anonymity โดยส่วนนี้ความซับซ้อนเชิงคำนวณมีค่าเท่ากับ $FULL \times (n+n\log(n))$ โดย n เป็นจำนวนระเบียนและ $FULL$ เป็นจำนวนระดับทั้งหมดซึ่งก็คือ $\sum_{\forall h_j, A_j \in Q_D} h_j$ ในการคำนวณค่าความซับซ้อนเชิงคำนวณของส่วนการทำงานนี้มาจากแต่ละระดับจะทดสอบข้อมูลหลังการแปลงค่าว่ามีคุณสมบัติ *k*-Anonymity หรือไม่หนึ่งครั้ง ดังนั้นจำนวนการทดสอบที่เป็นไปได้สูงสุดจึงเท่ากับ $FULL$ โดยวิธีการทดสอบข้อมูลหลังการแปลงค่าว่ามีคุณสมบัติ *k*-Anonymity หรือไม่จะใช้การเรียงข้อมูลแล้วทำการนับ ซึ่งการเรียงข้อมูลที่เร็วที่สุดมีความซับซ้อนเชิงคำนวณเท่ากับ $n\log(n)$ และการนับระเบียนมีค่าความซับซ้อนเชิงคำนวณเท่ากับ n ดังนั้นการทดสอบการมีคุณสมบัติหนึ่งครั้งจึงมีค่าความซับซ้อนเชิงคำนวณเท่ากับ $n+n\log(n)$ ดังนั้นค่าความความซับซ้อนเชิงคำนวณของส่วนนี้เท่ากับ $FULL \times (n+n\log(n))$

จากที่กล่าวมาจะได้ว่าค่าความซับซ้อนเชิงคำนวณรวมทั้งสองส่วนเท่ากับ $FULL \times (n+n\log(n))+k \times n$ จากค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT นี้ เมื่อมีการเพิ่มเข้ามาของข้อมูลระเบียนใหม่ ΔD จะได้ว่าค่าความซับซ้อนเชิงคำนวณเมื่อการเพิ่มเข้ามาของข้อมูลระเบียนใหม่เท่ากับ $FULL \times (n+\Delta n+(n+\Delta n)\ \log\ (n+\Delta n))+k \times (n+\Delta n)$ โดยที่ Δn เป็นจำนวนระเบียนที่เพิ่มเข้ามาของ ΔD

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Input:
D : a dataset
Q_D : a quasi-identifier
minsup: a minimal support threshold
k : a condition for *k*-Anonymity
$dom(Q_D)$: a set of generalization domains for a set of attributes in Q_D
Output:
D' : the output dataset, which satisfy the *k*-Anonymity property, and the frequency-based classification impact on data quality is potentially minimal.
GL : output generalization level sequence which makes D' satisfy the *k*-Anonymity property.
Method:
Sort Q_D attribute into GL by the CCR of each attribute increasingly.
Let A_j be the first attribute in GL.
while D does not satisfy *k*-Anonymity
  increase l_j of A_j in GL by 1
  generalize D to D' using $dom_j^l(A_j)$
  if $l_j \geq h_j$ **then**
    Let A_{j+1} be the attribute to be generalized.
end while

รูปที่ 3.1 รหัสเทียมขั้นตอนวิธี MCCRT

3.3 นิยามของปัญหา (Problem Statement)

ให้ข้อมูลที่ถูกแปลงค่าโดยค่า k ตามคุณสมบัติ *k*-Anonymity เป็น D' ซึ่งจับคู่มาพร้อมกับเซตของค่าคลาส C ให้ระดับการเจนเนอรัลไลเซชันที่สอดคล้องกับ *k*-Anonymity เป็น GL ให้อัตราความแม่นยำในการจำแนกของเซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม Q_D เป็น *CCRs* และให้ข้อมูลระเบียนที่เพิ่มเข้ามาเป็น ΔD

ปัญหาในวิทยานิพนธ์นี้คือจะทำการแปลงข้อมูลที่ถูกเพิ่มเข้าซึ่งคือ $(D+\Delta D)$ ไปเป็นข้อมูลหลังการแปลง $(D+\Delta D)'$ อย่างไรให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ ที่ได้เหมือนกับการทำตามขั้นตอนวิธี MCCRT โดยการใช้ระดับการเจนเนอรัลไลเซชันเดิม GL และข้อมูลหลังการแปลง D' แทนการแปลงค่าข้อมูล $(D+\Delta D)$ ใหม่หมดตั้งแต่เริ่มต้นตามขั้นตอนวิธี MCCRT

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.4 ผลกระทบของข้อมูลจากกระบวนการการเพิ่มเข้ามาของข้อมูลระเบียนใหม่กับขั้นตอนวิธี MCCRT

หลังจากนำข้อมูล D ผ่านกระบวนการขั้นตอนวิธี MCCRT ในครั้งแรก ได้ข้อมูลที่ถูกแปลงแล้วคือ D' ซึ่งมีคุณสมบัติ k-Anonymity และได้ระดับการเจนเนอรัลไลเซชัน GL ต่อมาเมื่อมีข้อมูลระเบียนใหม่ ΔD เพิ่มเข้ามาและข้อมูลระเบียนใหม่เหล่านี้ไม่มีผลกระทบใดๆในกระบวนการ สามารถนำข้อมูลระเบียนใหม่ ΔD มาแปลงค่าข้อมูลโดยระดับการเจนเนอรัลไลเซชัน GL ได้ข้อมูลหลังการแปลง $\Delta D'$ นำ $\Delta D'$ เพิ่มเข้าไปใน D' ได้ข้อมูลหลังการแปลง $D' + \Delta D'$ ออกมาเป็นคำตอบ

แต่ข้อมูลระเบียนใหม่ ΔD เหล่านี้อาจส่งผลกระทบกับค่าอัตราความแม่นยำในการจำแนกของทุกคอลัมน์ *CCRs* หรือจำนวนระเบียนของแต่ละกลุ่มข้อมูลที่ใช้ในการทดสอบการมีคุณสมบัติ k-Anonymity ในแต่ละรอบการแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT จึงทำให้ต้องทำการตรวจสอบผลกระทบที่เกิดขึ้นและทำการปรับเปลี่ยนระดับการเจนเนอรัลไลเซชันเป็นค่าใหม่ที่ถูกต้องตามขั้นตอนวิธี MCCRT ค่าระดับการเจนเนอรัลไลเซชันใหม่นี้เรียกว่า GL' ซึ่งสามารถนำมาใช้ในการแปลงค่าข้อมูล $(D + \Delta D)$ ไปเป็นผลลัพธ์ที่ต้องการหรือ $(D + \Delta D)'$ ได้ ซึ่งจากการสังเกตพฤติกรรมของขั้นตอนวิธี MCCRT พบว่าการปรับค่าระดับการเจนเนอรัลไลเซชันเดิม GL ไปเป็นระดับเจอเนอรัลไลเซชันใหม่ GL' จะปรับตามผลกระทบซึ่งสามารถแบ่งออกเป็นกลุ่มใหญ่ได้ 2 กลุ่ม ดังต่อไปนี้

3.4.1 กลุ่มที่ 1 ระดับการเจนเนอรัลไลเซชันใหม่ GL' ไม่มีการเปลี่ยนลำดับจากระดับการเจนเนอรัลไลเซชันเดิม GL

เมื่อมีการเพิ่มเข้ามาของข้อมูลระเบียนใหม่ ΔD แล้วหลังการคำนวณค่าอัตราความแม่นยำในการจำแนก *CCR* ของแต่ละคอลัมน์ในเซตของคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อม Q_D ใหม่ปรากฏว่าลำดับการเจนเนอรัลไลเซชันตามขั้นตอนวิธี MCCRT ไม่มีการเปลี่ยนแปลงหรือกล่าวได้ว่าไม่มีการเปลี่ยนลำดับของคู่คอลัมน์กับระดับในระดับการเจนเนอรัลไลเซชันเดิม GL แต่จะพบว่าระดับเจนเนอรัลไลเซชันใหม่ GL' อาจยังต้องถูกปรับค่าระดับของคอลัมน์เนื่องจากผลกระทบจากค่าข้อมูลระเบียนใหม่ จากการสังเกตการณ์พฤติกรรมของขั้นตอนวิธี MCCRT เมื่อมีข้อมูลระเบียนใหม่เข้ามา พบว่าค่าของข้อมูลในระเบียนเหล่านั้นสามารถทำให้ระดับการเจนเนอรัลไลเซชันใหม่ GL' ลดลงจากระดับการเจนเนอรัลไลเซชันเดิม GL ได้เนื่องจากค่าของระเบียนที่เข้ามาใหม่อาจทำให้ระเบียนบางระเบียนในข้อมูลเดิมที่ยังไม่มีคุณสมบัติ k-Anonymity ในระดับเจนเนอรัลไลเซ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ชันเดิม GL ที่ลดลงสามารถมีคุณสมบัติ k-Anonymity ได้ หรืออาจทำให้ระดับการเจนเนอรัลไลเซชันใหม่ GL' เพิ่มขึ้นจากระดับการเจนเนอรัลไลเซชันเดิม GL ได้เมื่อตัวของข้อมูลระเบียนใหม่ที่เพิ่มเข้ามาเองยังไม่มีคุณสมบัติ k-Anonymity ในระดับการเจนเนอรัลไลเซชันเดิม GL งานวิจัยนี้พบว่าค่าข้อมูลระเบียนใหม่ที่เพิ่มเข้ามามี 2 ลักษณะที่อาจส่งผลกระทบทำให้ระดับการเจนเนอรัลไลเซชันใหม่ GL' เปลี่ยนไปจากระดับการเจนเนอรัลไลเซชันเดิม GL

3.4.1.1 กรณีค่าข้อมูลซ้ำเดิมหรือมีคุณสมบัติ *k*-Anonymity แล้ว (Same-value or Satisfy *k* Case)

เมื่อมีข้อมูลระเบียนใหม่ ΔD เข้ามา ถ้านำมาแปลงค่าข้อมูลโดยระดับการเจนเนอรัลไลเซชันเดิม GL ได้ข้อมูลหลังการแปลง $\Delta D'$ แล้วพบว่าถ้าทุกระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลเหมือนกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือมีบางระเบียนเป็นระเบียนที่มีค่าข้อมูลไม่เหมือนกับระเบียนใดในข้อมูลหลังการแปลง D' แต่กลุ่มระเบียนเหล่านั้นมีคุณสมบัติ k-Anonymity แสดงว่าข้อมูลหลังการแปลง $(D+\Delta D)'$ ที่ระดับการเจนเนอรัลไลเซชันเดิม GL นี้จะมีคุณสมบัติ k-Anonymity ด้วย ดังนั้นการเพิ่มเข้ามาของข้อมูลหลังการแปลง $\Delta D'$ ที่มีค่าข้อมูลเหมือนกับระเบียนในข้อมูลหลังการแปลง D' นี้จะเป็นการเพิ่มจำนวนระเบียนให้กลุ่มระเบียนที่มีในข้อมูลหลังการแปลง D' ซึ่งทำให้กลุ่มระเบียนเหล่านี้มีโอกาสที่จะมีคุณสมบัติ k-Anonymity ที่ระดับการเจนเนอรัลไลเซชันเดิม GL ลดลงด้วย ฉะนั้นระดับการเจนเนอรัลไลเซชันใหม่ GL' อาจลดลงจากระดับการเจนเนอรัลไลเซชันเดิม GL ได้

จากรูปที่ 3.2 เป็นตัวอย่างการเพิ่มเข้ามาของระเบียนที่เป็นกรณีค่าข้อมูลซ้ำเดิมหรือมีคุณสมบัติ k-Anonymity ในรูปที่ 3.2 (1) เป็นชุดข้อมูลเริ่มต้น D เมื่อนำชุดข้อมูลเริ่มต้น D ไปแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยมีขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของคอลัมน์ A เป็นดังรูปที่ 3.2 (6) และมีข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.2 (2) โดยมีระดับการเจนเนอรัลไลเซชัน GL คือ < (A, 1), (B, 0), (C, 0)> ต่อมาเมื่อมีการเพิ่มเข้ามาของระเบียนที่ 5 และ 6 ดังรูปที่ 3.2 (3) เมื่อนำระเบียนที่เพิ่มเข้ามาดังกล่าวไปแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน GL จะได้ชุดข้อมูลทั้งหมดดังรูปที่ 3.2 (4) ซึ่งจะพบว่าระเบียนที่ 5 และ 6 หลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน GL แล้วจะมีค่าเหมือนกับระเบียนที่ 3 และ 4 ซึ่งเป็นข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเจนเนอรัลไลเซชันของข้อมูลเริ่มต้น D'

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

และจากรูปที่ 3.2 (1) พบว่าระเบียนที่ 3 และ 4 เป็นระเบียนที่ไม่มีคุณสมบัติ *k*-Anonymity ที่ระดับการเจนเนอรัลไลเซชันเป็น < (A, 0), (B, 0), (C, 0)> ทำให้ต้องเพิ่มระดับการเจนเนอรัลไลเซชันเป็น < (A, 1), (B, 0), (C, 0)> เพื่อทำให้ชุดข้อมูลมีคุณสมบัติ *k*-Anonymity และเมื่อมีข้อมูลระเบียนที่ 5 และ 6 เพิ่มเข้ามาทำให้สามารถลดระดับการเจนเนอรัลไลเซชันลงมาเป็น < (A, 0), (B, 0), (C, 0)> แล้วยังคงทำให้ข้อมูลยังคงมีคุณสมบัติ *k*-Anonymity ดังรูปที่ 3.2 (5)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Id	A	B	C	Class
1	A2	B1	C2	1
2	A2	B1	C2	1
3	A2	B2	C1	0
4	A1	B2	C1	0

(1) ชุดข้อมูลเริ่มต้น *D*

กำหนดให้ k=2

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0

(2) ชุดข้อมูลหลังการแปลงค่าข้อมูล *D′* ตามขั้นตอนวิธี MCCRT โดยมีระดับการเจนเนอรัลไลเซชัน *GL* คือ <(A,1),(B,0),(C,0)>

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A1	B2	C1	1
6	A2	B2	C1	1

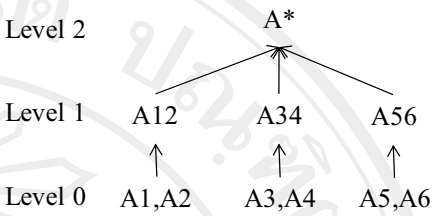
(3) ชุดข้อมูลหลังหลังการแปลงค่าเมื่อมีข้อมูลระเบียนที่ 5 และ 6 เพิ่มเข้ามา (*D′*+Δ*D*)

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A12	B2	C1	1
6	A12	B2	C1	1

(4) ชุดข้อมูลทั้งหมดหลังการแปลงค่าของข้อมูลโดยระดับการเจนเนอรัลไลเซชัน *GL* (*D′*+Δ*D′*)

Id	A	B	C	Class
1	A2	B1	C2	1
2	A2	B1	C2	1
3	A2	B2	C1	0
4	A1	B2	C1	0
5	A1	B2	C1	1
6	A2	B2	C1	1

(5) ชุดข้อมูลทั้งหมดหลังการแปลงค่าของข้อมูล (*D′*+Δ*D′*) โดยระดับการเจอนเนอรัลไลเซชัน *GL* ถูกลดมา 1 ระดับ



(6) ขั้นตอนการแปลงข้อมูลตามลำดับขั้นของคอลัมน์ A

รูปที่ 3.2 ตัวอย่างการเพิ่มเข้ามาของข้อมูลที่เป็นกรณีค่าข้อมูลซ้ำเดิมหรือมีคุณสมบัติ *k*-Anonymity

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.4.1.2 กรณีค่าข้อมูลเป็นใหม่และมีคุณสมบัติ *k*-Anonymity แล้ว (New-value and Satisfy *k* Case)

เมื่อทุกระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลไม่เหมือนกับระเบียนใดในข้อมูลหลังการแปลง D' และกลุ่มระเบียนเหล่านั้นมีคุณสมบัติ *k*-Anonymity แล้วจะแสดงว่าข้อมูลหลังการแปลง $(D+\Delta D)'$ ที่ระดับการเจนเนอรัลไลเซชันเดิม *GL* นี้จะมีคุณสมบัติ *k*-Anonymity ด้วย แต่จะไม่มีโอกาสที่ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จะลดลงจากระดับการเจนเนอรัลไลเซชันเดิม *GL* กล่าวคือระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จะเท่ากับระดับการเจนเนอรัลไลเซชันเดิม *GL* เนื่องจากไม่มีการเพิ่มจำนวนระเบียนให้กลุ่มระเบียนที่มีในข้อมูลหลังการแปลง D'

จากรูปที่ 3.3 เป็นตัวอย่างการเพิ่มเข้ามาของระเบียนที่เป็นกรณีค่าข้อมูลใหม่และมีคุณสมบัติ *k*-Anonymity ในรูปที่ 3.3 (1) เป็นชุดข้อมูลเริ่มต้น *D* เมื่อนำชุดข้อมูลเริ่มต้น *D* ไปแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยมีขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของคอลัมน์ A เป็นดังรูปที่ 3.3 (5) และมีข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.3 (2) โดยมีระดับการเจนเนอรัลไลเซชัน *GL* คือ <(A, 1), (B, 0), (C, 0)> ต่อมาเมื่อมีการเพิ่มเข้ามาของระเบียนที่ 5 และ 6 ดังรูปที่ 3.2 (3) เมื่อนำระเบียนที่เพิ่มเข้ามาดังกล่าวไปแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน *GL* จะได้ชุดข้อมูลทั้งหมดดังรูปที่ 3.3 (4) ซึ่งจะพบว่าระเบียนที่ 5 และ 6 หลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน *GL* แล้วจะมีค่าไม่เหมือนกับระเบียนใดๆตั้งแต่ระเบียนที่ 1 ถึง 4 ซึ่งเป็นข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเจนเนอรัลไลเซชันของข้อมูลเริ่มต้น D'

และจากรูปที่ 3.3 (1) พบว่าระเบียนที่ 3 และ 4 เป็นระเบียนที่ไม่มีคุณสมบัติ *k*-Anonymity ที่ระดับการเจนเนอรัลไลเซชันเป็น <(A, 0), (B, 0), (C, 0)> ทำให้ต้องเพิ่มระดับการเจนเนอรัลไลเซชันเป็น <(A, 1), (B, 0), (C, 0)> เพื่อทำให้ชุดข้อมูลมีคุณสมบัติ *k*-Anonymity ต่อมาเมื่อมีข้อมูลระเบียนที่ 5 และ 6 เพิ่มเข้ามาระเบียนเหล่านี้ไม่สามารถเพิ่มจำนวนให้กลุ่มระเบียนที่เหมือนกันของระเบียนที่ 3 และ 4 ได้ อย่างไรก็ตามระเบียนที่ 5 และ 6 นี้มีคุณสมบัติ *k*-Anonymity แล้วที่ระดับการเจนเนอรัลไลเซชันเป็น <(A, 1), (B, 0), (C, 0)> ทำให้สามารถเพิ่มข้อมูลหลังการแปลงค่าข้อมูล $\Delta D'$ เข้าไปกับข้อมูลหลังการแปลงค่าข้อมูลเดิม D' ซึ่งจะได้ข้อมูลทั้งหมดที่มีคุณสมบัติ *k*-Anonymity ดังรูปที่ 3.3 (4)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Id	A	B	C	Class
1	A2	B1	C2	1
2	A2	B1	C2	1
3	A2	B2	C1	0
4	A1	B2	C1	0

(1) ชุดข้อมูลเริ่มต้น

กำหนดให้ k=2

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0

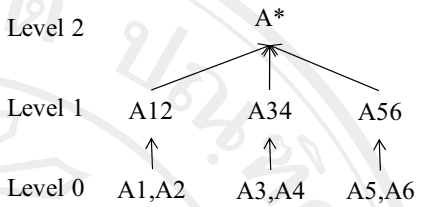
(2) ชุดข้อมูลหลังการแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยมีระดับการเจนเนอรัลไลเซชัน *GL* คือ <(A,1),(B,0),(C,0)>

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A3	B2	C1	1
6	A3	B2	C1	1

(3) ชุดข้อมูลหลังหลังการแปลงค่าเมื่อมีข้อมูลระเบียนที่ 5 และ 6 เพิ่มเข้ามา

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A34	B2	C1	1
6	A34	B2	C1	1

(4) ชุดข้อมูลทั้งหมดหลังการแปลงค่าของข้อมูล โดยระดับการเจนเนอรัลไลเซชัน *GL*



(5) ขั้นตอนการแปลงข้อมูลตามลำดับขั้นของคอลัมน์ A

รูปที่ 3.3 ตัวอย่างการเพิ่มเข้ามาของข้อมูลที่เป็นกรณีค่าข้อมูลเป็นค่าใหม่และมีคุณสมบัติ *k*-Anonymity

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.4.1.3 กรณีค่าข้อมูลเป็นค่าใหม่และไม่มีคุณสมบัติ *k*-Anonymity (New-value and not satisfy *k* Case)

เมื่อมีข้อมูลระเบียนใหม่ ΔD เข้ามา แล้วข้อมูลเหล่านั้นถูกนำมาแปลงค่าข้อมูลโดยใช้ระดับการเจนเนอรัลไลเซชันเดิม *GL* ซึ่งได้ผลลัพธ์คือข้อมูลหลังการแปลง $\Delta D'$ แล้วพบว่าบางระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลไม่เหมือนกับระเบียนใดเลยในข้อมูลหลังการแปลง D' และระเบียนเหล่านี้ไม่มีคุณสมบัติ *k*-Anonymity ระเบียนเหล่านี้จะทำให้ข้อมูลทั้งหมด

ไม่มีคุณสมบัติ *k*-Anonymity ดังนั้นจึงต้องเพิ่มระดับการเจนเนอรัลไลเซชัน *GL* เพื่อให้ข้อมูลหลังการแปลง $(D + \Delta D)'$ มีคุณสมบัติ *k*-Anonymity ในกรณีนี้ระดับการเจนเนอรัลไลเซชันใหม่ GL' จะเพิ่มขึ้นจากระดับการเจนเนอรัลไลเซชันเดิม *GL* เสมอ

จากรูปที่ 3.4 เป็นตัวอย่างการเพิ่มเข้ามาของระเบียนที่เป็นกรณีค่าข้อมูลใหม่และมีคุณสมบัติ *k*-Anonymity ในรูปที่ 3.4 (1) เป็นชุดข้อมูลเริ่มต้น *D* เมื่อนำชุดข้อมูลเริ่มต้น *D* ไปแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยมีขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของคอลัมน์ A เป็นดังรูปที่ 3.4 (5) และมีข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.4 (2) โดยมีระดับการเจนเนอรัลไลเซชัน *GL* คือ <(A, 1), (B, 0), (C, 0)> ต่อมาเมื่อมีการเพิ่มเข้ามาของระเบียนที่ 5 และ 6 ดังรูปที่ 3.4 (3) เมื่อนำระเบียนที่เพิ่มเข้ามาดังกล่าวไปแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน *GL* จะได้ชุดข้อมูลทั้งหมดดังรูปที่ 3.4 (4) ซึ่งจะพบว่าระเบียนที่ 5 และ 6 หลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน *GL* แล้วจะมีค่าไม่เหมือนกับระเบียนใดๆตั้งแต่ระเบียนที่ 1 ถึง 4 ซึ่งเป็นข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเจนเนอรัลไลเซชันของข้อมูลเริ่มต้น D'

และจากรูปที่ 3.3 (1) พบว่าระเบียนที่ 3 และ 4 เป็นระเบียนที่ไม่มีคุณสมบัติ *k*-Anonymity ที่ระดับการเจนเนอรัลไลเซชันเป็น <(A, 0), (B, 0), (C, 0)> ทำให้ต้องเพิ่มระดับการเจนเนอรัลไลเซชันเป็น <(A, 1), (B, 0), (C, 0)> เพื่อทำให้ชุดข้อมูลมีคุณสมบัติ *k*-Anonymity ต่อมาเมื่อมีข้อมูลระเบียนที่ 5 และ 6 เพิ่มเข้ามาระเบียนเหล่านี้ไม่สามารถเพิ่มจำนวนให้กลุ่มระเบียนที่เหมือนกันของระเบียนที่ 3 และ 4 ได้ และระเบียนที่ 5 และ 6 นี้ไม่มีคุณสมบัติ *k*-Anonymity ที่ระดับการเจนเนอรัลไลเซชันเป็น <(A, 1), (B, 0), (C, 0)> ทำให้ต้องเพิ่มระดับการเจนเนอรัลไลเซชันเป็น <(A, 2), (B, 0), (C, 0)> จึงทำให้ข้อมูลทั้งหมดหลังการแปลงค่าข้อมูลมีคุณสมบัติ *k*-Anonymity ดังรูปที่ 3.4 (5)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

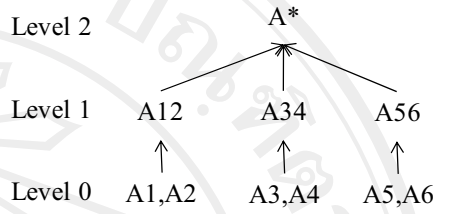
Id	A	B	C	Class
1	A2	B1	C2	1
2	A2	B1	C2	1
3	A2	B2	C1	0
4	A1	B2	C1	0

(1) ชุดข้อมูลเริ่มต้น

กำหนดให้ k=2

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0

(2) ชุดข้อมูลหลังการแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT
โดยมีระดับการเจนเนอรัลไลเซชัน *GL* คือ
<(A,1),(B,0),(C,0)>



(6) ขั้นตอนการแปลงข้อมูลตามลำดับขั้นของคอลัมน์ A

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A3	B2	C1	1
6	A5	B2	C1	1

(3) ชุดข้อมูลหลังหลังการแปลงค่าเมื่อมีข้อมูลระเบียนที่ 5
และ 6 เพิ่มเข้ามา

Id	A	B	C	Class
1	A12	B1	C2	1
2	A12	B1	C2	1
3	A12	B2	C1	0
4	A12	B2	C1	0
5	A34	B2	C1	1
6	A56	B2	C1	1

(4) ชุดข้อมูลทั้งหมดหลังการแปลงค่าของข้อมูล
โดยระดับการเจนเนอรัลไลเซชัน *GL*

Id	A	B	C	Class
1	A*	B1	C2	1
2	A*	B1	C2	1
3	A*	B2	C1	0
4	A*	B2	C1	0
5	A*	B2	C1	1
6	A*	B2	C1	1

(5) ชุดข้อมูลทั้งหมดหลังการแปลงค่าของข้อมูล
โดยระดับการเจอนเนอรัลไลเซชัน *GL* ถูกเพิ่มขึ้น 1 ระดับ

รูปที่ 3.4 ตัวอย่างการเพิ่มเข้ามาของข้อมูลที่เป็นกรณีค่าข้อมูลเป็นค่าใหม่และไม่มีคุณสมบัติ *k*-Anonymity

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

3.4.2 กลุ่มที่ 2 ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีการเปลี่ยนลำดับจากระดับการเจนเนอรัลไลเซชันเดิม *GL*

เมื่อมีข้อมูลระเบียนใหม่ ΔD เข้ามาและทำให้ค่าอัตราความแม่นยำในการจำแนก *CCR* ของแต่ละคอลัมน์เปลี่ยน ในกรณีนี้อาจทำให้ลำดับในการแปลงค่าตามขั้นตอนวิธี MCCRT เปลี่ยนแปลงตามไปด้วย กล่าวคือค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จะเกิดการเปลี่ยนแปลงลำดับจากระดับการเจนเนอรัลไลเซชันเดิม *GL* นั่นเอง โดยในที่นี้กำหนดให้คู่ของคอลัมน์กับระดับที่เป็นคู่สุดท้ายในระดับการเจนเนอรัลไลเซชันเดิม *GL* ที่มีการค่าระดับคอลัมน์มากกว่า 0 เรียกว่า *CGP* (Current Generalization Point) ตัวอย่างเช่นให้ระดับการเจนเนอรัลไลเซชัน *GL* คือ <(A_1, 2), (A_2, 3), (A_3, 2), (A_4, 0), (A_5, 0)> จะได้ *CGP* คือคู่คอลัมน์กับคลาส (A_3, 2) และจะอ้างอิงคอลัมน์ใน *CGP* ว่า Attribute(*CGP*) ซึ่งในกรณีนี้คือคอลัมน์ A_3 จากการสังเกตพฤติกรรมของขั้นตอนวิธี MCCRT พบว่าการเปลี่ยนแปลงลำดับจากระดับการเจนเนอรัลไลเซชันเดิม *GL* ไปเป็นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* สามารถเกิดขึ้นได้ 5 กรณีดังต่อไปนี้

3.4.2.1 การสลับลำดับก่อน *CGP* (Order change before *CGP*)

เมื่อลำดับของคู่คอลัมน์กับระดับของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีการเปลี่ยนตำแหน่งเกิดขึ้นเมื่อเทียบกับระดับการเจนเนอรัลไลเซชันเดิม *GL* โดยการสลับลำดับเกิดขึ้นในตำแหน่งก่อน *CGP* จะไม่มีผลกระทบต่อกระบวนการในขั้นตอนวิธี MCCRT เนื่องจากคู่คอลัมน์กับระดับที่อยู่ก่อนหน้า *CGP* เป็นคู่คอลัมน์กับระดับที่ถูกแปลงค่าข้อมูลจนถึงระดับบนสุดแล้ว อย่างไรก็ตามหลังจากการเปลี่ยนลำดับเป็น *GL′* แล้ว เนื่องจากผลกระทบจากค่าข้อมูลของข้อมูลระเบียนใหม่ ΔD ยังคงอยู่ทำให้ต้องปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'$ บางระเบียนไม่มีคุณสมบัติ *k*-Anonymity และระเบียนเหล่านั้นไม่ซ้ำกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือปรับระดับการเจนเนอรัลไลเซชันลดลงเมื่อข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนซ้ำกับระเบียนบางระเบียนในข้อมูลหลังการแปลง D' ตัวอย่างการสลับลำดับในกรณีนี้ เช่น ถ้าให้ระดับการเจนเนอรัลไลเซชันเดิม *GL* เท่ากับ <(A_1, 2), (A_2, 3), (A_3, 2), (A_4, 0), (A_5, 0)> ดังนั้นลำดับคือ <A_1, A_2, A_3, A_4, A_5> เมื่อมีข้อมูลระเบียนใหม่ ΔD เพิ่มขึ้นมา หลังการคำนวณค่าอัตราความแม่นยำในการจำแนก *CCR* แต่ละคอลัมน์ใหม่ปรากฏว่าลำดับเปลี่ยนเป็น <A_2, A_1, A_3, A_4, A_5> จะได้ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เป็น <(A_2, 3), (A_1, 2), (A_3, 2), (A_4, 0), (A_5, 0)> ซึ่งใช้จุดเริ่มต้นในการปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'$ บางระเบียนไม่มีคุณสมบัติ *k*-Anonymity และ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ระเบียนเหล่านั้นไม่ซ้ำกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ GL' ลดลงเมื่อข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนซ้ำกับระเบียนบางระเบียนในข้อมูลหลังการแปลง D'

จากรูปที่ 3.5 เป็นตัวอย่างการเพิ่มเข้ามาของระเบียนที่เป็นกรณีค่าข้อมูลใหม่และมีคุณสมบัติ k-Anonymity ในรูปที่ 3.5 (1) เป็นชุดข้อมูลเริ่มต้น D เมื่อนำชุดข้อมูลเริ่มต้น D ไปแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยกำหนดลำดับให้เป็นไปตามทิศทาง <A, B, C, D, E> โดยขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของแต่ละคอลัมน์มีลักษณะดังตัวอย่างขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของคอลัมน์ A ดังรูปที่ 3.5 (4) จะได้ข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.5 (2) โดยมีระดับการเจนเนอรัลไลเซชัน GL คือ <(A, 2), (B, 2), (C, 1), (D,0), (E,0)>

ต่อมาเมื่อมีการกำหนดลำดับให้เป็นไปตามทิศทาง <B, A, C, D, E> จะได้ข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.5 (3) โดยมีระดับการเจนเนอรัลไลเซชัน GL คือ <(B, 2), (A, 2), (C, 1), (D,0), (E,0)> ซึ่งแสดงให้เห็นว่าการสลับลำดับของคู่คอลัมน์กับระดับระหว่าง A และ B ยังคงให้ผลลัพธ์ในการแปลงคงเดิมและยังให้ค่าระดับในคู่ลำดับคอลัมน์กับระดับเป็นค่าเดิม

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Id	A	B	C	D	E	Class
1	A1	B1	C12	D1	E3	1
2	A2	B2	C12	D1	E3	1
3	A2	B1	C34	D2	E1	0
4	A1	B2	C34	D2	E1	0

กำหนดให้ $k=2$

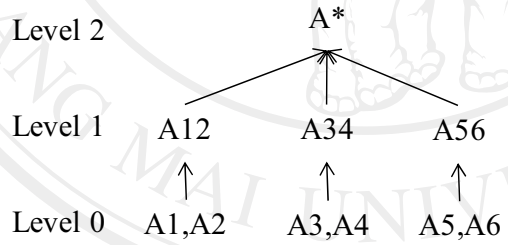
(1) ชุดข้อมูลเริ่มต้น

Id	A	B	C	D	E	Class
1	A*	B*	C12	D1	E3	1
2	A*	B*	C12	D1	E3	1
3	A*	B*	C34	D2	E1	0
4	A*	B*	C34	D2	E1	0

(2) ชุดข้อมูลที่ถูกแปลงค่าโดยระดับการเจนเนอรัลไลเซชันเท่ากับ <(A, 2), (B, 2), (C, 1), (D, 0), (E, 0)>

Id	A	B	C	D	E	Class
1	A*	B*	C12	D1	E3	1
2	A*	B*	C12	D1	E3	1
3	A*	B*	C34	D2	E1	0
4	A*	B*	C34	D2	E1	0

(3) ชุดข้อมูลที่ถูกแปลงค่าโดยระดับการเจนเนอรัลไลเซชันเท่ากับ <(B, 2), (A, 2), (C, 1), (D, 0), (E, 0)>



(4) ขั้นตอนการแปลงข้อมูลตามลำดับขั้นของคอลัมน์ A

รูปที่ 3.5 ตัวอย่างการสลับลำดับก่อน *CGP*

3.4.2.2 การสลับลำดับหลัง *CGP* (Oder change after *CGP*)

เมื่อลำดับของคู่คอลัมน์กับระดับของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีการเปลี่ยนตำแหน่งเกิดขึ้นเมื่อเทียบกับระดับการเจนเนอรัลไลเซชันเดิม *GL* โดยการสลับลำดับเกิดขึ้นหลังตำแหน่ง *CGP* พบว่าไม่มีผลกระทบต่อกระบวนการในขั้นตอนวิธี MCCRT เนื่องจากคู่คอลัมน์กับระดับที่อยู่หลัง *CGP* นั้นเป็นคู่คอลัมน์กับระดับที่ยังไม่ถูกแปลงค่าข้อมูลกล่าวคือยังมีระดับการแปลงค่าข้อมูลเป็น 0 อย่างไรก็ตามหลังจากการเปลี่ยนลำดับเป็น *GL′* แล้วเนื่องจากผลกระทบ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

จากค่าข้อมูลของข้อมูลระเบียนใหม่ ΔD ยังคงอยู่ทำให้ต้องปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ GL' เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'$ บางระเบียนไม่มีคุณสมบัติ k-Anonymity และระเบียนเหล่านั้นไม่ซ้ำกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือปรับระดับการเจนเนอรัลไลเซชันลดลงเมื่อข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนซ้ำกับระเบียนบางระเบียนในข้อมูลหลังการแปลง D' ตัวอย่างการสลับลำดับในกรณีนี้ เช่น ถ้าให้ระดับการเจนเนอรัลไลเซชันเดิม GL เท่ากับ $<(A_1, 2), (A_2, 3), (A_3, 2), (A_4, 0), (A_5, 0)>$ ดังนั้นลำดับคือ $<A_1, A_2, A_3, A_4, A_5>$ เมื่อมีข้อมูลระเบียนใหม่ ΔD เพิ่มขึ้นมาหลังการคำนวณค่าอัตราแม่นยำในการจำแนก CCR แต่ละคอลัมน์ใหม่ปรากฏว่าลำดับเปลี่ยนเป็น $<A_1, A_2, A_3, A_5, A_4>$ จะได้ระดับการเจนเนอรัลไลเซชันใหม่ GL' เป็น $<(A_1, 2), (A_2, 3), (A_3, 2), (A_5, 0), (A_4, 0)>$ ซึ่งใช้จุดเริ่มต้นในการปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ GL' เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'$ บางระเบียนไม่มีคุณสมบัติ k-Anonymity และระเบียนเหล่านั้นไม่ซ้ำกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ GL' ลดลงเมื่อข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนซ้ำกับระเบียนบางระเบียนในข้อมูลหลังการแปลง D'

จากรูปที่ 3.6 เป็นตัวอย่างการเพิ่มเข้ามาของระเบียนที่เป็นกรณีค่าข้อมูลใหม่และมีคุณสมบัติ k-Anonymity ในรูปที่ 3.6 (1) เป็นชุดข้อมูลเริ่มต้น D เมื่อนำชุดข้อมูลเริ่มต้น D ไปแปลงค่าข้อมูลตามขั้นตอนวิธี MCCRT โดยกำหนดลำดับให้เป็นไปตามทิศทาง <A, B, C, D, E> โดยขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของแต่ละคอลัมน์มีลักษณะดังตัวอย่างขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นของคอลัมน์ A ดังรูปที่ 3.6 (4) จะได้ข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.6 (2) โดยมีระดับการเจนเนอรัลไลเซชัน GL คือ <(A, 2), (B, 2), (C, 1), (D,0), (E,0)>

ต่อมาเมื่อมีการกำหนดลำดับให้เป็นไปตามทิศทาง <A, B, C, E, D> จะได้ข้อมูลหลังการแปลงค่าข้อมูลตามระดับการเจนเนอรัลไลเซชัน D' ดังรูปที่ 3.6 (3) โดยมีระดับการเจนเนอรัลไลเซชัน GL คือ <(A, 2), (B, 2), (C, 1), (D,0), (E,0)> ซึ่งแสดงให้เห็นว่าการสลับลำดับของคู่คอลัมน์กับระดับระหว่าง D และ E ยังคงให้ผลลัพธ์ในการแปลงคงเดิมและยังให้ค่าระดับในคู่ลำดับคอลัมน์กับระดับเป็นค่าเดิม

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Id	A	B	C	D	E	Class
1	A1	B1	C12	D1	E3	1
2	A2	B2	C12	D1	E3	1
3	A2	B1	C34	D2	E1	0
4	A1	B2	C34	D2	E1	0

กำหนดให้ k=2

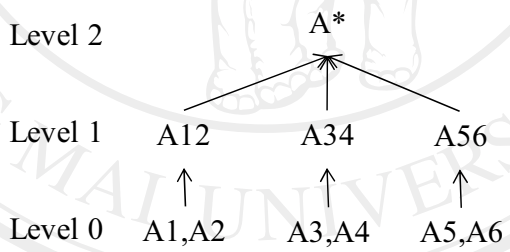
(1) ชุดข้อมูลเริ่มต้น

Id	A	B	C	D	E	Class
1	A*	B*	C12	D1	E3	1
2	A*	B*	C12	D1	E3	1
3	A*	B*	C34	D2	E1	0
4	A*	B*	C34	D2	E1	0

(2) ชุดข้อมูลที่ถูกแปลงค่าโดยระดับการเจนเนอรัลไลเซชันเท่ากับ
<(A, 2), (B, 2), (C, 1), (D, 0), (E, 0)>

Id	A	B	C	D	E	Class
1	A*	B*	C12	D1	E3	1
2	A*	B*	C12	D1	E3	1
3	A*	B*	C34	D2	E1	0
4	A*	B*	C34	D2	E1	0

(3) ชุดข้อมูลที่ถูกแปลงค่าโดยระดับการเจนเนอรัลไลเซชันเท่ากับ
<(A, 2), (B, 2), (C, 1), (E, 0), (D, 0)>



(4) ขั้นตอนการแปลงข้อมูลตามลำดับขั้นของคอลัมน์ A

รูปที่ 3.6 ตัวอย่างการสลับลำดับหลัง *CGP*

จากการนำเสนอหัวข้อ 3.4.1, 3.4.2.1 และ 3.4.2.2 ไปแล้วจะสามารถทำการสรุปค่าความซับซ้อนเชิงคำนวณจากหัวข้อดังกล่าวได้ดังต่อไปนี้

การคำนวณค่าความซับซ้อนเชิงคำนวณ เมื่อระดับการเจนเนอรัลไลเซชันใหม่ GL' ไม่มีการเปลี่ยนลำดับหรือเปลี่ยนตำแหน่งของคู่คอลัมน์กับระดับทุกคู่จากระดับการเจนเนอรัลไลเซชันเดิม GL และทุกระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลเหมือนกับระเบียนใดๆในข้อมูลหลังการแปลง D' หรือมีบางระเบียนเป็นระเบียนที่มีค่าข้อมูลไม่เหมือนกับระเบียนใดๆในข้อมูลหลังหารแปลง D' แต่กลุ่มระเบียนเหล่านั้นมีคุณสมบัติ k-Anonymity จะได้ว่าค่า (Complexity)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ความซับซ้อนเชิงคำนวณในการแปลงค่าข้อมูลให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ มีคุณสมบัติ k-Anonymity และผลลัพธ์ที่ได้เหมือนกับขั้นตอนวิธี MCCRT มีค่าเท่ากับสมการ (1)

$$\text{Complexity} = START_TO_CGP \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n))+k\times(n+\Delta n) \quad (1)$$

โดย $START_TO_CGP$ คือจำนวนระดับจากจุดเริ่มต้น (จุดเริ่มต้นหมายถึงทุกคู่คอลัมน์กับระดับใน GL มีค่าระดับเท่ากับ 0) ถึง CGP สามารถคำนวณได้จากสมการ (2)

$$START_TO_CGP = \sum_{\exists l_j, A_j \in Q_D where A_j \prec = CGP} l_j \quad (2)$$

ค่า n คือจำนวนระเบียนใน D และค่า Δn คือจำนวนระเบียนใน ΔD โดยค่าความซับซ้อนเชิงคำนวณของการคำนวณค่าอัตราความแม่นยำในการจำแนก CCR ในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $k\times(n+\Delta n)$ และค่าความซับซ้อนเชิงคำนวณของการทดสอบข้อมูลว่ามีคุณสมบัติ k-Anonymity หรือไม่ของแต่ละระดับในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $n+\Delta n+(n+\Delta n)\log(n+\Delta n)$ แต่จำนวนระดับที่เป็นไปได้นั้นจะน้อยกว่าคือเท่ากับ $START_TO_CGP$ ซึ่งน้อยกว่า $FULL$ แสดงว่าค่าความซับซ้อนเชิงคำนวณของกระบวนการนี้น้อยกว่าค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT

การคำนวณค่าความซับซ้อนเชิงคำนวณ เมื่อระดับการเจนเนอรัลไลเซชันใหม่ GL' ไม่มีการเปลี่ยนลำดับหรือเปลี่ยนตำแหน่งของคู่คอลัมน์กับระดับทุกคู่จากระดับการเจนเนอรัลไลเซชันเดิม GL และทุกระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลไม่เหมือนกับระเบียนใดๆ ในข้อมูลหลังการแปลง D' แต่กลุ่มระเบียนเหล่านั้นมีคุณสมบัติ k-Anonymity จะได้ว่าค่าความซับซ้อนเชิงคำนวณในการแปลงค่าข้อมูลให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ มีคุณสมบัติ k-Anonymity และผลลัพธ์ที่ได้เหมือนกับขั้นตอนวิธี MCCRT มีค่าความซับซ้อนเชิงคำนวณเท่ากับ Δn ซึ่งมาจากการการแปลงค่าตามระดับเจนเนอรัลไลเซชันเดิม GL และการเพิ่มข้อมูลให้ข้อมูลหลังการแปลงเป็นจำนวน Δn ระเบียนนั่นเอง

การคำนวณค่าความซับซ้อนเชิงคำนวณ เมื่อระดับการเจนเนอรัลไลเซชันใหม่ GL' ไม่มีการเปลี่ยนลำดับหรือเปลี่ยนตำแหน่งของคู่คอลัมน์กับระดับทุกคู่จากระดับการเจนเนอรัลไลเซชันเดิม GL และทุกระเบียนในข้อมูลหลังการแปลง $\Delta D'$ มีค่าข้อมูลไม่เหมือนกับระเบียนใดๆใน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright © by Chiang Mai University
All rights reserved

ข้อมูลหลังการแปลง D' บางกลุ่มระเบียนเหล่านั้นไม่มีคุณสมบัติ k-Anonymity จะได้ว่าค่าความซับซ้อนเชิงคำนวณในการแปลงค่าข้อมูลให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ มีคุณสมบัติ k-Anonymity และผลลัพธ์ที่ได้เหมือนกับขั้นตอนวิธี MCCRT มีค่าเท่ากับสมการ (3)

$$\text{Complexity} = CGP_TO_FULL \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n))+k\times(n+\Delta n) \quad (3)$$

โดย CGP_TO_FULL คือจำนวนระดับจาก CGP ถึง $FULL$ สามารถคำนวณได้จากสมการ (4)

$$CGP_TO_FULL = \sum_{\exists l_j, A_j \in Q_D where A_j >= CGP \& A_j <= FULL} l_j \quad (4)$$

ค่า n คือจำนวนระเบียนใน D และค่า Δn คือจำนวนระเบียนใน ΔD โดยค่าความซับซ้อนเชิงคำนวณของการคำนวณค่าอัตราความแม่นยำในการจำแนก CCR ในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $k\times(n+\Delta n)$ และค่าความซับซ้อนเชิงคำนวณของการทดสอบข้อมูลว่ามีคุณสมบัติ k-Anonymity หรือไม่ของแต่ละระดับในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $n+\Delta n+(n+\Delta n)$ $\log (n+\Delta n)$ แต่จำนวนระดับที่เป็นไปได้นั้นจะน้อยกว่าคือเท่ากับ CGP_TO_FULL ซึ่งน้อยกว่า $FULL$ แสดงว่าค่าความซับซ้อนเชิงคำนวณของกระบวนการนี้น้อยกว่าค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT

3.4.2.3 *CGP* ถอยหลัง (*CGP* move backward)

เมื่อระดับการเจนเนอรัลไลเซชันใหม่ GL' มีการเปลี่ยนลำดับหรือตำแหน่งของคู่คอลัมน์กับระดับจากระดับการเจนเนอรัลไลเซชันเดิม GL โดยคู่คอลัมน์กับระดับมีการเปลี่ยนตำแหน่งคือ CGP ในระดับการเจนเนอรัลไลเซชันเดิม GL ถอยหลังกล่าวคือการย้ายไปทางซ้ายกลายเป็น CGP' ใหม่ในระดับการเจนเนอรัลไลเซชันใหม่ GL' ค่าระดับในคู่คอลัมน์กับระดับใดๆที่อยู่ด้านหลัง CGP ใหม่นี้เพื่อให้เป็นไปตามขั้นตอนวิธี MCCRT คู่คอลัมน์กับระดับเหล่านั้นจะลดค่าระดับลงมาเป็น 0 จากจุดนี้สามารถใช้ระดับการเจนเนอรัลไลเซชันใหม่ GL' ที่ได้นี้เป็นจุดเริ่มต้นในการทดสอบการมีคุณสมบัติ k-Anonymity ใหม่อีกครั้งโดยการปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ GL' เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ไม่มีคุณสมบัติ k-Anonymity หรือไม่ปรับค่าเมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันใหม่ GL' มี

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

คุณสมบัติ *k*-Anonymity แต่ข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันที่ลดลงแล้วไม่มีคุณสมบัติ *k*-Anonymity หรือปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ลดลง เมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันที่ลดลงแล้วมีคุณสมบัติ *k*-Anonymity ตัวอย่าง ให้ระดับการเจนเนอรัลไลเซชันเดิม *GL* คือ $<(A_1,2), (A_2,3), (A_3,2), (A_4,0), (A_5,0)>$ ฉะนั้นลำดับการแปลงค่าข้อมูลคือ $<A_1, A_2, A_3, A_4, A_5>$ และมี *CGP* คือ $(A_3,2)$ ถ้าลำดับใหม่คือ $<A_1, A_3, A_2, A_4, A_5>$ ดังนั้นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จะเท่ากับ $<(A_1,2), (A_3,2), (A_2,0), (A_4,0), (A_5,0)>$ จากตัวอย่างจะเห็นว่าเมื่อ *CGP* ถอยหลังทุกคู่คอลัมน์กับระดับที่อยู่ถัดจาก *CGP′* ใหม่นี้จะถูกลดระดับลงมาที่ระดับ 0 ทำให้ไม่สามารถรู้ได้ว่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* นี้ยังคงคุณสมบัติ *k*-Anonymity หรือไม่ อย่างไรก็ตามระดับการเจนเนอรัลไลเซชันก่อนหน้า *GL′* ซึ่งก็คือ $<(A_1,2), (A_3,1), (A_2,0), (A_4,0), (A_5,0)>$ นั้นไม่มีคุณสมบัติ *k*-Anonymity เพราะระดับการเจนเนอรัลไลเซชัน $<(A_1,2), (A_2,3), (A_3,1), (A_4,0), (A_5,0)>$ ได้ทำการทดสอบแล้วว่าไม่มีคุณสมบัติ *k*-Anonymity เช่นกัน ดังนั้นสามารถใช้คู่คอลัมน์กับระดับที่ *CGP′* ใหม่นี้ซึ่งคือ $<(A_1,2), (A_3,2), (A_2,0), (A_4,0), (A_5,0)>$ เป็นจุดเริ่มต้นที่ใช้ปรับค่าระดับของ *GL′* ต่อไป

การคำนวณค่าความซับซ้อนเชิงคำนวณ เมื่อระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีการเปลี่ยนลำดับหรือตำแหน่งของคู่คอลัมน์กับระดับจากระดับการเจนเนอรัลไลเซชันเดิม *GL* ตามกรณี *CGP* ถอยหลัง จะได้ว่าค่าความซับซ้อนเชิงคำนวณในการแปลงค่าข้อมูลให้ $(D+\Delta D)'$ มีคุณสมบัติ *k*-Anonymity และผลลัพธ์ที่ได้เหมือนกับขั้นตอนวิธี MCCRT มีค่าเท่ากับสมการ (5)

$$\text{Complexity} = NEW_CGP_TO_FULL \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n)) \quad (5)$$

โดยค่า *NEW_CGP_TO_FULL* คือค่าจำนวนระดับจากจุดที่ *CGP′* ใหม่อยู่ถึงจุดสุดท้ายหรือ *FULL* ในกรณีที่การปรับค่าระดับการเจนเนอรัลไลเซชันหลังการได้ *CGP′* ใหม่เป็นการปรับค่าเพิ่ม และค่าความซับซ้อนเชิงคำนวณจะเท่ากับสมการ (6)

$$\text{Complexity} = START_TO_NEW_CGP \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n))+k\times(n+\Delta n) \quad (6)$$

โดยค่า *START_TO_NEW_CGP* คือค่าจำนวนระดับจากจุดเริ่มต้นไปถึง *CGP′* ในกรณีการปรับค่าระดับการเจนเนอรัลไลเซชันหลังการได้ *CGP′* ใหม่เป็นการปรับค่าลดลง ซึ่งทั้งสองกรณีมีค่าความซับซ้อนเชิงคำนวณของการคำนวณค่าอัตราความแม่นยำในการจำแนก *CCR* ในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $k\times(n+\Delta n)$ และค่า

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ความซับซ้อนเชิงคำนวณของการทดสอบข้อมูลว่ามีคุณสมบัติ *k*-Anonymity หรือไม่ของแต่ละระดับในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $n+\Delta n+(n+\Delta n)\log(n+\Delta n)$ แต่จำนวนระดับที่เป็นไปได้นั้นจะน้อยกว่าคือเท่ากับ *NEW_CGP_TO_FULL* กับ *START_TO_NEW_CGP* ซึ่งทั้งสองค่าน้อยกว่า *FULL* แสดงว่าค่าความซับซ้อนเชิงคำนวณของกระบวนการนี้น้อยกว่าค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT

3.4.2.4 *CGP* เดินหน้า (*CGP* move forward)

เมื่อระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีการเปลี่ยนลำดับหรือตำแหน่งของคู่คอลัมน์กับระดับจากระดับการเจนเนอรัลไลเซชันเดิม *GL* โดยคู่คอลัมน์กับระดับที่เป็น *CGP* ในระดับการเจนเนอรัลไลเซชันเดิม *GL* เดินหน้ากล่าวคือการย้ายไปทางขวาจะทำให้คู่คอลัมน์กับระดับที่อยู่ก่อนหน้ากลายเป็น *CGP′* ใหม่ในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ค่าระดับในคู่คอลัมน์กับระดับที่อยู่ทางขวาหรือหลัง *CGP′* ใหม่เพื่อให้เป็นไปตามขั้นตอนวิธี MCCRT คู่คอลัมน์กับระดับนั้นจะลดค่าระดับลงมาเป็น 0 จากจุดนี้สามารถใช้ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ที่ได้นี้เป็นจุดเริ่มต้นในการทดสอบการมีคุณสมบัติ *k*-Anonymity ใหม่อีกครั้งโดยการปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพิ่มขึ้นเมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ไม่มีคุณสมบัติ *k*-Anonymity หรือไม่ปรับค่าเมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีคุณสมบัติ *k*-Anonymity แต่ข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันที่ลดลงแล้วไม่มีคุณสมบัติ *k*-Anonymity หรือปรับค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ลดลงเมื่อข้อมูลหลังการแปลง $\Delta D'+D'$ ของระดับการเจนเนอรัลไลเซชันที่ลดลงแล้วมีคุณสมบัติ *k*-Anonymity ตัวอย่าง ให้ระดับการเจนเนอรัลไลเซชันเดิม *GL* คือ $<(A_1,2), (A_2,3), (A_3,2), (A_4,0), (A_5,0)>$ ฉะนั้นลำดับการแปลงค่าข้อมูลคือ $<A_1, A_2, A_3, A_4, A_5>$ และมี *CGP* คือ $(A_3,2)$ ถ้าลำดับใหม่คือ $<A_1, A_2, A_4, A_3, A_5>$ ดังนั้นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จะเท่ากับ $<(A_1,2), (A_2,3), (A_4,0), (A_3,0), (A_5,0)>$ จากตัวอย่างจะเห็นว่าเมื่อ *CGP* เดิม $(A_3.0)$ เดินหน้า *CGP* เดิมนี้จะถูกลดค่าระดับลงมาเป็น 0 และคู่คอลัมน์กับค่าระดับก่อนหน้าคือ $(A_2, 3)$ จะกลายเป็น *CGP′* ใหม่ ทำให้ระดับการเจนเนอรัลไลเซชัน *GL′* นี้ไม่มีคุณสมบัติ *k*-Anonymity เพราะระดับการเจนเนอรัลไลเซชัน $<(A_1,2), (A_2,3), (A_3,1), (A_4,0), (A_5,0)>$ ได้ทำการทดสอบแล้วว่าไม่มีคุณสมบัติ *k*-Anonymity เช่นกันจึงสามารถให้ระดับการเจนเนอรัลไลเซชันที่เพิ่มระดับจาก *CGP′* ใหม่นี้ 1 ระดับซึ่งก็คือ $<(A_1,2), (A_2,3), (A_4,1), (A_3,0), (A_5,0)>$ เป็นเป็นจุดเริ่มต้นในการค้นหาค่าระดับใน *GL′* ต่อไป

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

การคำนวณค่าความซับซ้อนเชิงคำนวณ เมื่อระดับการเจนเนอรัลไลเซชันใหม่ GL' มีการเปลี่ยนลำดับหรือตำแหน่งของคู่คอลัมน์กับระดับจากระดับการเจนเนอรัลไลเซชันเดิม GL ตามกรณี CGP เดินหน้า จะได้ว่าค่าความซับซ้อนเชิงคำนวณในการแปลงค่าข้อมูลให้ $(D+\Delta D)'$ มีคุณสมบัติ k-Anonymity และผลลัพธ์ที่ได้เหมือนกับขั้นตอนวิธี MCCRT มีค่าเท่ากับสมการ (7)

$$\text{Complexity} = NEW_CGP_TO_FULL \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n)) \quad (7)$$

โดยค่า $NEW_CGP_TO_FULL$ คือค่าจำนวนระดับจากจุดที่ CGP' ใหม่อยู่ถึงจุดสุดท้ายหรือ $FULL$ ในกรณีที่การปรับค่าระดับการเจนเนอรัลไลเซชันหลังการได้ CGP' ใหม่เป็นการปรับค่าเพิ่ม และค่าความซับซ้อนเชิงคำนวณจะเท่ากับสมการ (8)

$$\text{Complexity} = START_TO_NEW_CGP \times (n+\Delta n+(n+\Delta n)\log(n+\Delta n))+k\times(n+\Delta n) \quad (8)$$

โดยค่า $START_TO_NEW_CGP$ คือค่าจำนวนระดับจากจุดเริ่มต้นไปถึง CGP' ในกรณีการปรับค่าระดับการเจนเนอรัลไลเซชันหลังการได้ CGP' ใหม่เป็นการปรับค่าลดลง ซึ่งทั้งสองกรณีมีค่าความซับซ้อนเชิงคำนวณของการคำนวณค่าอัตราความแม่นยำในการจำแนก CCR ในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $k\times(n+\Delta n)$ และค่าความซับซ้อนเชิงคำนวณของการทดสอบข้อมูลว่ามีคุณสมบัติ k-Anonymity หรือไม่ของแต่ละระดับในกระบวนการนี้เท่ากับค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT คือ $n+\Delta n+(n+\Delta n)\log(n+\Delta n)$ แต่จำนวนระดับที่เป็นไปได้นั้นจะน้อยกว่าคือเท่ากับ $NEW_CGP_TO_FULL$ กับ $START_TO_NEW_CGP$ ซึ่งทั้งสองค่าน้อยกว่า $FULL$ แสดงว่าค่าความซับซ้อนเชิงคำนวณของกระบวนการนี้น้อยกว่าค่าความซับซ้อนเชิงคำนวณของขั้นตอนวิธี MCCRT

3.4.2.5 กรณีผสม (Mixed cases)

ในความเป็นจริงการที่ระดับการเจนเนอรัลไลเซชันใหม่ GL' มีการเปลี่ยนลำดับหรือตำแหน่งของคู่คอลัมน์กับระดับจากระดับการเจนเนอรัลไลเซชันเดิม GL สามารถเกิดจากการรวมกันของ 4 กรณีที่กล่าวมาข้างต้นได้แก่ การสลับลำดับก่อน CGP , การสลับลำดับหลัง CGP , CGP ถอยหลังและ CGP เดินหน้าเช่น ให้ระดับการเจนเนอรัลไลเซชันเดิม GL เท่ากับ $<(A_1,2), (A_2,3), (A_3,2), (A_4,0), (A_5,0)>$ จะได้ว่า CGP คือคู่คอลัมน์กับระดับ $(A_3, 2)$ แล้วมีการเปลี่ยนลำดับเป็น $< A_2, A_4, A_3, A_1, A_5>$ ในตัวอย่างนี้สามารถแสดงขั้นตอนการเปลี่ยนลำดับได้ว่า เริ่มจากลำดับ $< A_1, A_2, A_3, A_4, A_5>$ เปลี่ยนไปเป็น $< A_2, A_1, A_3, A_4, A_5>$ ซึ่งเป็นกรณีสลับลำดับ

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ก่อน *CGP* การเปลี่ยนลำดับครั้งที่ 2 เปลี่ยนลำดับจาก < A_2, A_1, A_3, A_4, A_5> ไปเป็น < A_2, A_1, A_4, A_3, A_5> ซึ่งเป็นกรณี *CGP* เดินหน้าทำให้ *CGP* ใหม่ที่มีคอลมันน์เป็นคอลัมน์ A_1 การเปลี่ยนลำดับครั้งที่ 3 เปลี่ยนลำดับจาก < A_2, A_1, A_4, A_3, A_5> ไปเป็น < A_2, A_4, A_1, A_3, A_5> ซึ่งเป็นกรณี *CGP* เดินหน้าทำให้ *CGP* ใหม่ที่มีคอลัมน์เป็น A_2 สุดท้ายเปลี่ยนจาก < A_2, A_4, A_1, A_3, A_5> ไปเป็น < A_2, A_4, A_3, A_1, A_5> ซึ่งจะเป็นการเปลี่ยนลำดับในกรณีเปลี่ยนสลับลำดับหลัง *CGP*

จากตัวอย่างข้างต้นจะแสดงให้เห็นว่าสามารถจัดการกับการรวมกันของกรณีดังกล่าวได้ โดยการหา *CGP′* ใหม่ให้กับระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพื่อใช้ในการปรับค่าต่อไป จากการสังเกตพบว่า *CGP′* ใหม่นั้นจะขึ้นอยู่กับการเปลี่ยนตำแหน่งจากของคู่คอลัมน์กับระดับเพียงสองคู่คือ คู่ที่เป็น *CGP* ในระดับการเจอเนอรัลไลเซชันเดิม *GL* และคู่ที่มีค่าระดับเป็น 0 ตัวแรกเมื่อเปลี่ยนลำดับไปเป็นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* แล้วพบว่า *CGP* ในระดับการเจนเนอรัลไลเซชันเดิม *GL* หรือคู่คอลัมน์กับระดับที่มีค่าระดับเป็น 0 ตัวแรกจากทางซ้าย เกิดการเปลี่ยนที่โดยจากทางซ้ายถ้าพบ *CGP* ก่อนจะทำการปรับค่าระดับของคู่ทางขวาทั้งหมดเป็น 0 แล้วใช้ระดับการเจนเนอรัลไลเซชันนี้เป็นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพื่อใช้ในการปรับค่าตามกรณีเพิ่มหรือลดลงต่อไป แต่ถ้าพบคู่ที่มีค่าระดับเป็น 0 ก่อนให้ปรับค่าระดับของคู่ทางขวาทั้งหมดให้เป็น 0 ใช้ระดับการเจนเนอรัลไลเซชันนี้เป็นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เพื่อใช้ในการปรับค่าตามกรณีเพิ่มหรือลดลงต่อไป

3.5 ขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้น (Incremental Algorithm)

ในรูปที่ 3.2 เป็นรหัสเทียมของขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นซึ่งใช้พื้นฐานจากการสังเกตในกรณีผสม ขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นเริ่มจากการคำนวณค่าอัตราความแม่นยำในการจำแนก *CCR* ในแต่ละคอลัมน์ของข้อมูล $D+\Delta D$ ซึ่งคือข้อมูลเดิมรวมกับข้อมลระเบียนใหม่ที่เพิ่มขึ้น ต่อมาขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นจะทำการเรียงลำดับของระดับการเจนเนอรัลไลเซชันเดิม *GL* ไปเป็นระดับการเจนเนอรัลไลเซชันใหม่ *GL′* โดยเรียงลำดับตามค่าอัตราความแม่นยำในการจำแนก *CCR* จากคู่คอลัมน์กับระดับที่คอลัมน์มี่ค่าอัตราความแม่นยำในการจำแนก *CCR* น้อยไปคู่คอลัมน์กับระดับที่คอลัมน์มีค่าอัตราความแม่นยำในการจำแนก *CCR* มาก จากนั้นขั้นตอนวิธีจะเข้าสู่กระบวนการตรวจสอบกรณีซึ่งแบ่งเป็นสองส่วนโดยส่วนที่หนึ่ง บรรทัดเลขที่ 1 ถึง 16 เป็นการตรวจสอบตามกรณีลำดับไม่เปลี่ยนแปลง หรือเป็นกรณีผสมของการสลับลำดับก่อน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

หน้า *CGP* กับการสลับลำดับหลัง *CGP* ส่วนที่สองเป็นกรณีผสมที่กรณี *CGP* เดินหน้าหรือ กรณี *CGP* ถอยหลังรวมอยู่ด้วย โดยส่วนนี้อยู่ในบรรทัดเลขที่ 17 ถึง 32

ในส่วนที่หนึ่งการตรวจสอบตามกรณีลำดับไม่เปลี่ยนแปลง หรือเป็นกรณีผสมของการสลับลำดับก่อนหน้า *CGP* กับการสลับลำดับหลัง *CGP* โดยตรวจสอบตามบรรทัดเลขที่ 1 กับ 2 ว่าคู่คอลัมน์ที่เป็น *CGP* ในระดับการเจนเนอรัลไลเซชันเดิม *GL* เปลี่ยนตำแหน่งในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* หรือไม่และตำแหน่งแรกที่คู่คอลัมน์กับระดับที่มีค่าระดับเท่ากับ 0 ในระดับการเจนเนอรัลไลเซชันเดิม *GL* เปลี่ยนตำแหน่งในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* หรือไม่ ถ้าเงื่อนไขทั้งสองเป็นจริงแสดงว่าเป็นกรณีที่กล่าวมาข้างต้น ขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นจะทำการปรับค่าระดับการเจนเนอรัลไลเซชันตามบรรทัดเลขที่ 4 ถึง 14 โดยในบรรทัดเลขที่ 4 จะทำการตรวจสอบว่าข้อมูลหลังการแปลง $\Delta D'$ มีระเบียนที่ค่าข้อมูลไม่ซ้ำกับระเบียนใดๆ ของข้อมูลหลังการแปลง D' หรือไม่และกลุ่มระเบียนเหล่านั้นต้องไม่มีคุณสมบัติ *k*-Anonymity ถ้าเงื่อนไขทั้งสองข้อเป็นจริง ขั้นวิธีการแปลงข้อมูลแบบเพิ่มขึ้นจะทำตาม บรรทัดที่ 5 ถึง 8 คือเพิ่มระดับการเจนเนอรัลไลเซชันให้ *GL′* ครั้งละ 1 ระดับจนข้อมูลหลังการแปลงของระดับการเจนเนอรัลไลเซชันให้ *GL′* มีคุณสมบัติ *k*-Anonymity แต่ถ้าเงื่อนไขไม่ตรงกับบรรทัดเลขที่ 4 ขั้นตอนวิธีแปลงค่าข้อมูลแบบเพิ่มขึ้นจะทำการตรวจสอบตามเงื่อนไขบรรทัดเลขที่ 9 คือตรวจสอบว่าข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนเป็นค่าข้อมูลที่ไม่ซ้ำกับระเบียนใดๆของข้อมูลหลังการแปลง D' หรือไม่และกลุ่มระเบียนเหล่านั้นต้องมีคุณสมบัติ *k*-Anonymity ถ้าเงื่อนไขทั้งสองข้อเป็นจริงจะทำตามบรรทัดเลขที่ 10 คือส่งค่าข้อมูลหลังการแปลง $(D+\Delta D)'$ และค่า *GL′* ออกมาเป็นผลลัพธ์ แต่ถ้าเงื่อนไขไม่ตรงกับบรรทัดเลขที่ 4 และ 9 ขั้นตอนวิธีแปลงข้อมูลแบบเพิ่มขึ้นจะทำการตรวจสอบเงื่อนไขตามบรรทัดเลขที่ 11 ว่าข้อมูลหลังการแปลง $\Delta D'$ ทุกระเบียนเป็นระเบียนที่ค่าข้อมูลซ้ำกับระเบียนใดๆของข้อมูลหลังการแปลง D' หรือมีคุณสมบัติ *k*-Anonymity หรือไม่ ถ้าเงื่อนไขเป็นจริง ขั้นวิธีการแปลงข้อมูลแบบเพิ่มขึ้นจะทำตาม บรรทัดที่ 12 ถึง 15 คือทำการทดสอบโดยให้ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ลดลงครั้งละ 1 ระดับจนข้อมูลหลังการแปลง $(D+\Delta D)'$ ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ไม่มีคุณสมบัติ *k*-Anonymity จะให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ และค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* สุดท้ายทีทำให้ข้อมูลหลังการแปลงมีคุณสมบัติ *k*-Anonymity เป็นผลลัพธ์

ในส่วนที่สองตามกรณีผสมที่กรณี *CGP* เดินหน้าหรือ กรณี *CGP* ถอยหลังรวมอยู่ด้วย เมื่อบรรทัดเลขที่ 1 กับ 2 ตรวจสอบแล้วว่าเป็นเท็จกล่าวคือคู่คอลัมน์กับระดับ *CGP* ในระดับการเจนเนอรัลไลเซชันเดิม *GL* เปลี่ยนตำแหน่งในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* หรือตำแหน่ง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

แรกที่คู่คอลัมน์กับระดับที่มีค่าระดับเท่ากับ 0 ในระดับการเจนเนอรัลไลเซชันเดิม *GL* เปลี่ยนตำแหน่งในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ต่อมาขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้นจะข้ามบรรทัดเลขที่ 3 ถึง 17 ซึ่งคือส่วนที่หนึ่งนั่นเอง กระบวนการทำงานเริ่มทำงานที่บรรทัดเลขที่ 18 และ 19 โดยกำหนดให้ค่าระดับของคู่คอลัมน์กับระดับในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เท่ากับ 0 ถ้าคู่คอลัมน์กับระดับนั้นอยู่หลังคู่คอลัมน์กับระดับที่เป็น *CGP* เดิมหรือคู่คอลัมน์กับระดับคู่แรกจากทางซ้ายที่มีค่าระดับเท่ากับ 0 ต่อมาบรรทัดเลขที่ 21 จะทำการตรวจสอบข้อมูลหลังการแปลง $(D+\Delta D)'$ ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ว่ามีคุณสมบัติ *k*-Anonymity หรือไม่ ถ้าข้อมูลหลังการแปลง $(D+\Delta D)'$ ยังไม่มีคุณสมบัติ *k*-Anonymity จะทำตามบรรทัดเลขที่ 22 ถึง 25 คือเพิ่มระดับการเจนเนอรัลไลเซชันใหม่ *GL′* จนข้อมูลหลังการแปลง $(D+\Delta D)'$ ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* มีคุณสมบัติ *k*-Anonymity จะให้ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* และข้อมูลหลังการแปลง $(D+\Delta D)'$ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* เป็นผลลัพธ์ ต่อมาบรรทัดเลขที่ 26 ถ้าข้อมูลหลังการแปลง$(D+\Delta D)'$ มีคุณสมบัติ *k*-Anonymity จะทำตาม บรรทัดที่ 27 ถึง 30 คือทำการทดสอบโดยให้ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ลดลงครั้งละ 1 ระดับจนข้อมูลหลังการแปลง $(D+\Delta D)'$ ของระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ไม่มีคุณสมบัติ *k*-Anonymity จะให้ข้อมูลหลังการแปลง $(D+\Delta D)'$ และค่าระดับการเจนเนอรัลไลเซชันใหม่ *GL′* สุดท้ายทีทำให้ข้อมูลหลังการแปลงมีคุณสมบัติ *k*-Anonymity เป็นผลลัพธ์

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

Input
D : a dataset
D' : a dataset output from MCCRT
minsup: a minimum support threshold
minconf: a minimum confidence threshold
ΔD : an additional dataset
k : a condition of k-Anonymity
GL : a generalization level which makes the D' satisfies the k-Anonymity
$dom^{GL}(D)$: a generalization level D by GL
Output
GL' : new generalization level
$(D + \Delta D)'$: the output dataset, which satisfy the k-Anonymity property,
and which regards to the MCCRT
Method
Calculate CCR of each attribute in GL by $D + \Delta D$
Order GL to GL' by the CCR of each attribute increasingly
$\Delta D' \leftarrow dom^{GL}(\Delta D)$
$(D + \Delta D)' \leftarrow (D' + \Delta D')$

รูปที่ 3.7 รหัสเทียมขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้น

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

```
1  if the position of the CGP attribute in GL does not change in GL′ and
2     the first 0-level attribute in GL is same position with first 0-level attribute
   in GL′
3  then   if ΔD′ have new_value tuples and some of them not satisfied k-
   Anonymity then
4        While (D + ΔD)′ does not satisfy k-Anonymity
5           increase GL′ by 1 level
6           (D + ΔD)′ ← dom^GL′ (D + ΔD)        end while
7     else if ΔD′ is new_value tuples and all of them satisfied k-Anonymity then
8        return (D+ΔD)′, GL′
9     else if(ΔD′ - new_value tuples that satisfy k-Anonymity) is all same_value
   tuples then
10       While (D + ΔD)′ satisfy k-Anonymity
11          Decrease GL′ by 1 level
12          (D + ΔD)′ ← dom^GL′ (D + ΔD)          end while
13    end if
14 else
15 set the level of the attribute after first 0-level attribute in GL′ to 0
16    set the level of the attribute after CGP attribute in GL′ to 0
17    (D + ΔD)′ ← dom^GL′ (D + ΔD)
18    if (D + ΔD)′ does not satisfy k-Anonymity
19       while (D + ΔD)′ does not satisfy k-Anonymity
20          increase GL′ by 1 level
21          (D + ΔD)′ ← dom^GL′ (D + ΔD)
22       end while
23    else if (D + ΔD)′ satisfy k-Anonymity
24       while (D + ΔD)′ satisfy k-Anonymity
25          decrease GL′ by 1 level
26          (D + ΔD)′ ← dom^GL′ (D + ΔD)
27       end while
28    end if
29 end if
```

รูปที่ 3.7 รหัสเทียมขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้น (ต่อ)

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตัวอย่างการทำงานของขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้น

ตัวอย่างนี้จะเป็นการทำงานของขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้นในกรณีที่ข้อมูลที่เพิ่มขึ้นทำให้ลำดับการเจนเนอรัลไลเซชันมีการเปลี่ยนแปลงโดยเป็นกรณี *CGP* ถอยหลัง

เริ่มจากข้อมูล *D* ตามตารางที่ 3.1 ซึ่งเป็นข้อมูลการวินิจฉัยโรคโดยมีคอลัมน์ที่เป็นตัวบ่งชี้บุคคลทางอ้อมคือ คอลัมน์ BirthDate, Sex, Weight, Height และ Career ซึ่งมีขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้น (Hierarchy) ดังรูปที่ 2.1 ถึง 2.5 ตามลำดับ และมีคลาสคือคอลัมน์ Diag เมื่อข้อมูล *D* ผ่านการแปลงข้อมูลตามขั้นตอนวิธี MCCRT โดยกำหนดให้ค่า *k* เท่ากับ 2 ค่าสนับสนุนขั้นต่ำเท่ากับ 2 ระเบียน และค่าความเชื่อมั่นขั้นต่ำเท่ากับ 60%

ผลลัพธ์ของขั้นตอนวิธี MCCRT คือค่าอัตราความถูกต้องในการจำแนก *CCR* ของแต่ละคอลัมน์, ค่าระดับการเจนเนอรัลไลเซชัน *GL* และ ข้อมูลหลังการแปลง *D′* โดยค่าความถูกต้องในการจำแนก *CCR* ของแต่ละคอลัมน์คำนวณดังนี้

คอลัมน์ BirthDate ไม่มีคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งทั้งสองค่า ดังนั้นค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ BirthData เท่ากับ $\frac{0}{5}$ หรือ 0% โดย 0 คือจำนวนระเบียนของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งและ 5 คือจำนวนระเบียนทั้งหมด

คอลัมน์ Sex มีคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งทั้งสองค่า คือ Male $\rightarrow$ Fever ซึ่งมี 3 ระเบียน ดังนั้นค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ Sex เท่ากับ $\frac{3}{5}$ หรือ 60% โดย 3 คือจำนวนระเบียนของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งและ 5 คือจำนวนระเบียนทั้งหมด

คอลัมน์ Weight มีคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งทั้งสองค่า คือ 49 $\rightarrow$ Flu มี 2 ระเบียน กับ 55 $\rightarrow$ Fever มี 2 ระเบียนดังนั้นค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ Weight เท่ากับ $\frac{4}{5}$ หรือ 80% โดย 4 คือจำนวนระเบียนของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งและ 5 คือจำนวนระเบียนทั้งหมด

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

คอลัมน์ Height มีคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งทั้งสองค่า คือ 160 → Flu มี 2 ระเบียน กับ 169 → Fever มี 3 ระเบียนดังนั้นค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ Height เท่ากับ $\frac{5}{5}$ หรือ 100% โดย 5 คือจำนวนระเบียนของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งและ 5 คือจำนวนระเบียนทั้งหมด

คอลัมน์ Career มีคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งทั้งสองค่า คือ B2 → Flu มี 2 ระเบียน กับ B1 → Fever มี 3 ระเบียนดังนั้นค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ Career เท่ากับ $\frac{5}{5}$ หรือ 100% โดย 5 คือจำนวนระเบียนของคู่ระหว่างค่าของคอลัมน์กับค่าคลาสที่ผ่านค่าขีดแบ่งและ 5 คือจำนวนระเบียนทั้งหมด

ตารางที่ 3.1 ตัวอย่างข้อมูล *D*

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	14/2/2520	Female	49	160	B2	Flu
2	28/2/2520	Male	49	160	B2	Flu
3	19/5/2520	Male	53	169	B1	Fever
4	30/5/2520	Male	55	169	B1	Fever
5	31/5/2520	Male	55	169	B1	Fever

จากค่าอัตราความถูกต้องในการจำแนกของแต่ละคอลัมน์จะได้ลำดับของการเจนเนอรัลไลเซชันซึ่งเรียงลำดับตามขั้นตอนวิธี MCCRT เท่ากับ <BirthDate, Sex, Weight, Height, Career> โดยเรียงคอลัมน์จากค่าอัตราความถูกต้องในการจำแนกน้อยไปหามากถ้าอัตราความถูกต้องในการจำแนกเท่ากันให้เรียงจากคอลัมน์ที่มีค่าความสูงของขั้นตอนการแปลงข้อมูลตามลำดับขั้นจากมากไปหาน้อย และได้ค่าระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลตารางที่ 3.1 มีคุณสมบัติ *k*-Anonymity ตามขั้นตอนวิธี MCCRT เท่ากับ <(BirthDate, 2), (Sex, 1), (Weight, 1),(Height, 1), (Career, 0)> และมีคู่คอลัมน์กับคลาสที่เป็น *CGP* คือ (Height, 1) โดยใช้ขั้นตอนการแปลงค่าข้อมูลตามลำดับขั้นตามรูปที่ 2.1 ถึง 2.5 ซึ่งข้อมูลหลังการแปลงข้อมูล *D′* จะได้ดังตารางที่ 3.2

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ตารางที่ 3.2 ตารางข้อมูลหลังการแปลงข้อมูล D' ซึ่งมีคุณสมบัติ 2-Anonymity โดย GL เท่ากับ <(BirthDate, 2), (Sex, 1), (Weight, 1),(Height, 1), (Career, 0)>

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	*/*/2520	*	46-50	160	B2	Flu
2	*/*/2520	*	46-50	160	B2	Flu
3	*/*/2520	*	51-55	169	B1	Fever
4	*/*/2520	*	51-55	169	B1	Fever
5	*/*/2520	*	51-55	169	B1	Fever

ตารางที่ 3.3 ตารางข้อมูลเพิ่มขึ้นจำนวน 2 ระเบียน

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
6	3/6/2520	Female	43	160	B2	Flu
7	5/6/2520	Male	44	160	B2	Flu

ตารางที่ 3.4 ตัวอย่างข้อมูลเดิมรวมกับข้อมูลเพิ่มขึ้น

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	14/2/2520	Female	49	160	B2	Flu
2	28/2/2520	Male	49	160	B2	Flu
3	19/5/2520	Male	53	169	B1	Fever
4	30/5/2520	Male	55	169	B1	Fever
5	31/5/2520	Male	55	169	B1	Fever
6	3/6/2520	Female	43	160	B2	Flu
7	5/6/2520	Male	44	160	B2	Flu

เมื่อมีข้อมูลเพิ่มเข้ามา 2 ระเบียนตามตารางที่ 3.3 จะเริ่มขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้นโดย 4 บรรทัดแรกของขั้นตอนวิธีเป็นส่วนของการเตรียมข้อมูลเริ่มต้น เริ่มจากคำนวณค่าอัตราความถูกต้องในการจำแนก CCR ของแต่ละคอลัมน์ใหม่โดยจะคำนวณจากตารางที่ 3.4 ซึ่งเป็นตารางข้อมูลเดิมรวมกับข้อมูลที่เพิ่มขึ้น จะได้ค่าอัตราความถูกต้องในการจำแนกของคอลัมน์ BirthDate, Sex, Weight, Height, Career เท่ากับ 0%, 42%, 28%, 100%, 100% ตามลำดับ

ต่อมานำระดับการเจนเนอรัลไลเซชันเดิม GL ซึ่งเท่ากับ <(BirthDate, 2), (Sex, 1), (Weight, 1),(Height, 0), (Career, 0)> มาเรียงลำดับตามค่าอัตราความถูกต้องในการจำแนกใหม่ที่คำนวณได้จะได้ระดับการเจนเนอรัลไลเซชันใหม่ GL' เท่ากับ <(BirthDate, 2), (Weight,

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

1), (Sex, 1),(Height, 0), (Career, 0)> นำข้อมูลที่เพิ่มขึ้นมาแปลงค่าข้อมูลโดยใช้ระดับการเจนเนอรัลไลเซชันเดิม *GL* จะได้ข้อมูลเพิ่มขึ้นหลังการแปลงข้อมูล $\Delta D'$ ตามตารางที่ 3.5 นำไปรวมกับข้อมูลหลังการแปลงเดิม D' ตามตารางที่ 3.2 จะได้ข้อมูลรวมหลังการแปลง $(D+\Delta D)'$ ดังตารางที่ 3.6

ตารางที่ 3.5 ตารางข้อมูลเพิ่มขึ้นหลังการแปลงข้อมูล $\Delta D'$ โดยระดับการเจนเนอรัลไลเซชันเดิม *GL* <(BirthDate, 2), (Sex, 1), (Weight, 1),(Height, 0), (Career, 0)>

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
6	*/*/2520	*	41-45	160	B2	Flu
7	*/*/2520	*	41-45	160	B2	Flu

ตารางที่ 3.6 ข้อมูลรวมหลังการแปลง $(D'+\Delta D)'$

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	*/*/2520	*	46-50	160	B2	Flu
2	*/*/2520	*	46-50	160	B2	Flu
3	*/*/2520	*	51-55	169	B1	Fever
4	*/*/2520	*	51-55	169	B1	Fever
5	*/*/2520	*	51-55	169	B1	Fever
6	*/*/2520	*	41-45	160	B2	Flu
7	*/*/2520	*	41-45	160	B2	Flu

เมื่อเตรียมข้อมูลเริ่มต้นทั้งหมดแล้ว ให้พิจารณาระดับการเจนเนอรัลไลเซชันเดิม *GL* เทียบกับระดับการเจนเนอรัลไลเซชันใหม่ GL' โดยพิจารณาค่าคอลัมน์ของ *CGP* และค่าคอลัมน์ของคู่คอลัมน์กับคลาสคู่แรกจากทางซ้ายที่มีค่าระดับเท่ากับ 0 ในระดับการเจนเนอรัลไลเซชันเดิม *GL* ถูกเปลี่ยนตำแหน่งในระดับการเจนเนอรัลไลเซชันใหม่ GL' หรือไม่ ในตัวอย่างนี้ระดับการเจนเนอรัลไลเซชันเดิม *GL* เท่ากับ <(BirthDate, 2), (Sex, 1), (Weight, 1),(Height, 0), (Career, 0)> มี *CGP* คือคู่คอลัมน์กับคลาสที่มีคอลัมน์คือ Weight และคู่คอลัมน์กับคลาสคู่แรกที่มีค่าระดับเท่ากับ 0 คือคู่คอลัมน์กับคลาสที่มีคอลัมน์คือ Height และตัวอย่างนี้มีระดับการเจนเนอรัลไลเซชันใหม่ GL' คือ <(BirthDate, 2), (Weight, 1), (Sex, 1),(Height, 0), (Career, 0)> แสดงว่าตำแหน่งของคอลัมน์ที่เป็น *CGP* ในระดับการเจนเนอรัลไลเซชันเดิม *GL* มีการเปลี่ยนตำแหน่งใน

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

ระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ทำให้การทำงานข้ามจากบรรทัดเลขที่ 1 ของขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้นไปยังบรรทัดเลขที่ 17

การทำงานถัดมาจะทำการเปลี่ยนค่าระดับในระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ของคู่คอลัมน์กับคลาสหลัง (ด้านขวา) คู่คอลัมน์กับคลาสที่เป็น *CGP* ซึ่งในที่นี้คือคู่คอลัมน์กับคลาส Weight และคู่คอลัมน์กับคลาสแรกที่มีค่าระดับเท่ากับ 0 ซึ่งในที่นี้คือคู่คอลัมน์กับคลาส Height โดยเปลี่ยนค่าระดับในคู่คอลัมน์กับคลาสหลังคู่คอลัมน์กับคลาส Weight และคู่คอลัมน์กับคลาส Height ทุกคอลัมน์ให้เท่ากับ 0 ดังนั้นจะทำให้ *GL′* เป็น <(BirthDate, 2), (Weight, 1), (Sex, 0),(Height, 0), (Career, 0)> เมื่อนำระดับเจนเนอรัลไลเซชันใหม่ *GL′* ที่ได้นี้ไปทำการแปลงข้อมูลเดิมรวมกับข้อมูลที่เพิ่มขึ้นตามตารางที่ 3.4 จะได้ข้อมูลรวมหลังการแปลงตามระดับการเจนเนอรัลไลเซชันใหม่ดังตารางที่ 3.7

ตารางที่ 3.7 ข้อมูลรวมหลังการแปลงตามระดับการเจนเนอรัลไลเซชันใหม่ *GL′*
<(BirthDate, 2), (Weight, 1), (Sex, 0),(Height, 0), (Career, 0)>

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	*/*/2520	Female	46-50	160	B2	Flu
2	*/*/2520	Male	46-50	160	B2	Flu
3	*/*/2520	Male	51-55	169	B1	Fever
4	*/*/2520	Male	51-55	169	B1	Fever
5	*/*/2520	Male	51-55	169	B1	Fever
6	*/*/2520	Female	41-45	160	B2	Flu
7	*/*/2520	Male	41-45	160	B2	Flu

เมื่อพิจารณาข้อมูลหลังการแปลงตามระดับการเจนเนอรัลไลเซชันใหม่ *GL′* แล้วจะเป็นได้ว่าข้อมูลที่ได้ยังไม่มีคุณสมบัติของ *2*-Anonymity ดังนั้นจะเป็นไปตามเงื่อนไขในบรรทัดเลขที่ 21 ในรูปที่ 3.2 ทำให้การทำงานเข้าสู่ while loop ในบรรทัดเลขที่ 22 ในรูปเดียวกัน ซึ่งการทำงานในส่วนนี้จะทำการเพิ่มระดับการเจนเนอรัลไลเซชันใหม่ *GL′* ขึ้น 1 ระดับแล้วทำการทดสอบข้อมูลหลังการแปลงตามระดับเจนเนอรัลไลเซชันใหม่ว่ามีคุณสมบัติ *2*-Anonymity

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved

หรือไม่ ถ้าไม่จะทำการเพิ่มระดับครั้งละ 1 ระดับไปจนกว่าจะพบค่าระดับการเจนเนอรัลไลเซชันที่ทำให้ข้อมูลหลังการแปลงมีคุณสมบัติ *2*-Anonymity ซึ่งในตัวอย่างนี้ปรากฏว่าการเพิ่มเพียง 1 ระดับจากระดับการเจนเนอรัลไลเซชันใหม่ *GL'* เท่ากับ <(BirthDate, 2), (Weight, 1), (Sex, 0),(Height, 0), (Career, 0)> ไปเป็น <(BirthDate, 2), (Weight, 2), (Sex, 0),(Height, 0), (Career, 0)> จะทำให้ข้อมูลหลังการแปลงมีคุณสมบัติ *2*-Anonymity โดยข้อมูลหลังการแปลงที่มีคุณสมบัติ *2*-Anonymity เป็นดังตารางที่ 3.8

ตารางที่ 3.8 ข้อมูลรวมหลังการแปลงตามระดับการเจนเนอรัลไลเซชันใหม่ *GL'*
<(BirthDate, 2), (Weight, 2), (Sex, 0),(Height, 0), (Career, 0)>

Tuple-ID	BirthDate	Sex	Weight	Height	Career	Diag
1	*/*/2520	Female	41-50	160	B2	Flu
2	*/*/2520	Male	41-50	160	B2	Flu
3	*/*/2520	Male	51-60	169	B1	Fever
4	*/*/2520	Male	51-60	169	B1	Fever
5	*/*/2520	Male	51-60	169	B1	Fever
6	*/*/2520	Female	41-50	160	B2	Flu
7	*/*/2520	Male	41-50	160	B2	Flu

ตารางที่ 3.8 เป็นตารางข้อมูลหลังการแปลงของข้อมูลเดิมรวมกับข้อมูลใหม่ที่ระดับการเจนเนอรัลไลเซชัน *GL'* เท่ากับ <(BirthDate, 2), (Weight, 2), (Sex, 0),(Height, 0), (Career, 0)> โดยระดับเจนเนอรัลไลเซชัน *GL'* และข้อมูลหลังการแปลงนี้คือผลลัพธ์ขั้นตอนวิธีการแปลงข้อมูลแบบเพิ่มขึ้นนั่นเอง

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved