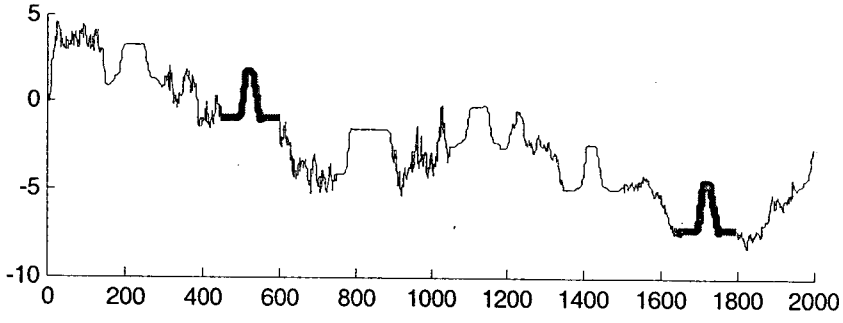
บทที่ 3
การหาโมทีฟความยาวแปรผันของข้อมูลอนุกรมเวลา
Variable-length Time Series Motif Discovery

3.1 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

โมทีฟของข้อมูลอนุกรมเวลา (Time Series Motif) คือลำดับย่อย (Subsequence) ที่คล้ายกันมากที่สุดในข้อมูลอนุกรมเวลา โดยนิยามของงานวิจัยที่ผ่านมาได้ให้คำจำกัดความว่าเป็นข้อมูลที่เกิดขึ้นซ้ำ ๆ กันในระยะ R (Range) ซึ่งเกิดเป็นพารามิเตอร์อีกหนึ่งตัวที่ผู้ใช้ต้องกำหนดขึ้นเองซึ่งมีความคลุมเครือว่าควรให้ค่าเป็นเท่าใด

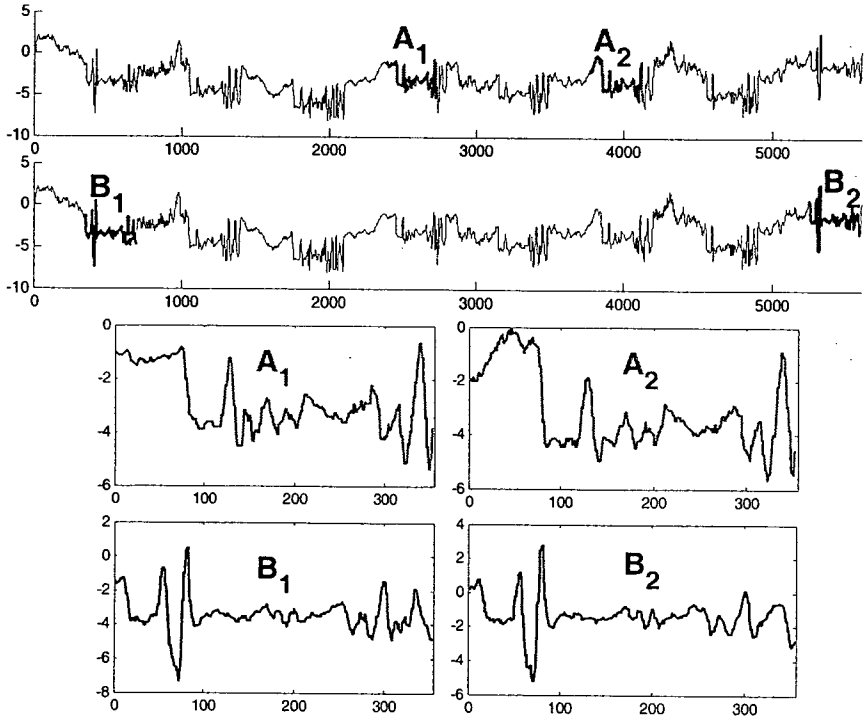


รูปที่ 3.1 โมทีฟของข้อมูล Gun-Point ขนาด 150 จุดถูกค้นพบที่ตำแหน่ง 450 ถึง 599 และ 1,647 ถึง 1,796 ในข้อมูลอนุกรมเวลาขนาด 2,000 จุด

วิธีที่ง่ายที่สุดในการหาโมทีฟคือวิธี Brute force ซึ่งเป็นการเปรียบเทียบรูปร่างที่เป็นไปได้ทั้งหมดในข้อมูลอนุกรมเวลา แล้วหาตัวที่คล้ายกันมากที่สุดออกมาเป็นโมทีฟของข้อมูลอนุกรมเวลานั้น โดยใช้การวัดระยะทางแบบยุคลิดเป็นตัววัดความคล้าย แต่ด้วยวิธีการที่ต้องเปรียบเทียบรูปร่างที่เป็นไปได้ทั้งหมดทำให้เป็นวิธีที่ใช้เวลาในการประมวลผลนานถึง $O(n^2)$ โดย n คือ จำนวนรูปร่างที่เป็นไปได้ทั้งหมดของข้อมูลอนุกรมเวลานั้นที่ขนาดตัวแปรความยาวค่าใดค่าหนึ่ง งานวิจัยในระยะเริ่มแรกของการค้นพบโมทีฟของข้อมูลอนุกรมเวลาจึงเน้นไปที่ การเพิ่มความเร็วของวิธีการหาโมทีฟเพื่อแทนที่วิธี Brute force มีวิธีการหลายวิธีถูกนำเสนอในช่วงทศวรรษที่ผ่านมา [11] [16] โดยส่วนใหญ่เป็นวิธีที่เพิ่มความเร็วในการหาโมทีฟแต่ได้แค่รูปร่างโดยประมาณที่ใกล้เคียงกับโมทีฟเท่านั้น ซึ่งมีอัตราการคลาดเคลื่อนของโมทีฟเพิ่มขึ้นมาแลกกับความเร็วที่เพิ่มขึ้น จนกระทั่งล่าสุดอัลกอริทึม MK (Mueen-Keogh) [18] ซึ่งเป็นอัลกอริทึมที่ดีที่สุดในการหาโมทีฟในขณะนี้ได้ถูกนำเสนอขึ้น วิธีนี้จะได้โมทีฟที่ถูกต้องเสมอคือมีระยะทางยุคลิดที่น้อยที่สุด โดยจะได้รูปร่างที่ตรงกับวิธี Brute force ทุกครั้ง ถึงแม้ว่าจะมีงานวิจัยมากมายเกี่ยวกับการค้นพบโมทีฟของข้อมูลอนุกรมเวลาแต่มีงานจำนวนน้อยมากที่จะกล่าวถึงปัญหาที่เกิดขึ้นทุกครั้งเมื่อเราต้องทำการหาโมทีฟ คือ การกำหนดขนาดตัวแปรความยาวของโมทีฟ ซึ่งเป็นพารามิเตอร์ของทุก ๆ อัลกอริทึมที่ผู้ใช้จะต้องกำหนดขึ้นเองโดยที่ไม่รู้แน่ชัดว่าควรจะกำหนดขนาดเท่าไร ปัญหาของความยาวโมทีฟจึงมีความคลุมเครือในการหาคำตอบที่ถูกต้องชัดเจน เนื่องจากคำตอบจะออกมาหลากหลายตามความคิดของผู้ใช้หลาย ๆ แบบซึ่งไม่เหมือนกัน การกำหนดค่าตัวแปรความยาวของรูปร่างที่แตกต่างกันอาจส่งผลกระทบอย่างใดอย่างหนึ่งต่อโมทีฟของข้อมูลอนุกรมเวลาได้ ในหลาย ๆ งานวิจัยที่ผ่านมาได้กล่าวถึงตัวแปรความยาวของโมทีฟของ

ข้อมูลอนุกรมเวลาไว้ในนิยามว่า เป็นความยาวที่ผู้ใช้กำหนดขึ้นเอง (User-Defined) เท่านั้น [11] [16]

จากปัญหาที่ได้กล่าวไว้ข้างต้น การทดลองเบื้องต้นเพื่อสาธิตการค้นพบรูปร่างโมทีฟของข้อมูลอนุกรมเวลาเมื่อขนาดความยาวเปลี่ยนแปลงจึงได้ถูกจัดทำขึ้น โดยผลการทดลองบางส่วนเป็นดังรูปที่ 9 ซึ่งแสดงให้เห็นว่าการกำหนดความยาวนั้นส่งผลต่อรูปร่างโมทีฟของข้อมูลอนุกรมเวลาที่ค้นพบ ความแตกต่างของความยาวที่กำหนดโดยผู้ใช้งานเพียงเล็กน้อยสามารถส่งผลให้โมทีฟของข้อมูลอนุกรมเวลาที่ค้นพบนั้นแตกต่างกันโดยสิ้นเชิง การกำหนดความยาวของข้อมูลอนุกรมเวลาจึงมีความละเอียดอ่อนและสำคัญควรค่าแก่การทำงานวิจัย ให้ได้ความรู้ที่ลึกซึ้งมากยิ่งขึ้น เพื่อให้คุณสมบัติต่าง ๆ ที่ส่งผลกระทบได้รับการแจกแจงเพื่อประโยชน์ในการใช้งานและการทำวิจัยต่อไปในภายภาคหน้า



รูปที่ 3.2 โมทีฟข้อมูลอนุกรมเวลาจำนวน 2 คู่ของข้อมูล Face-Four ขนาดความยาว 352 และ 353 จุดข้อมูล ตามลำดับ คู่แรก A1 และ A2 พบที่ตำแหน่ง 2,376 และ 3,777 ส่วนคู่ที่สอง B1 และ B2 พบที่ตำแหน่ง 341 และ 5,239

ส่วนหนึ่งของโครงงานวิจัยนี้ จะเป็นการศึกษาลักษณะของโมทีฟที่ถูกค้นพบในข้อมูลอนุกรมเวลาเมื่อขนาดความยาวโมทีฟเปลี่ยนแปลงไป แล้วนำเสนอวิธีการใหม่ในการแก้ปัญหาการกำหนดความยาวของโมทีฟ รวมไปถึงกำหนดนิยามอื่น ๆ ที่เกี่ยวข้องเพื่อประโยชน์ในการใช้งานและการทำวิจัยต่อไปในอนาคต

มาตรวัดระยะทาง (Distance Metric)

ตัววัดระยะทางหรือฟังก์ชันที่เกี่ยวกับการวัดระยะทางที่เป็นตัวกำหนดระยะระหว่างวัตถุต่าง ๆ ที่อยู่ในเซต ซึ่งไม่ใช่ทุกเซตที่มีโครงสร้างเป็น Metric จะมีเฉพาะบางโครงสร้างเท่านั้นที่สามารถอธิบายได้โดยใช้คุณสมบัติ 4 ข้อของ Metric ซึ่งมีดังต่อไปนี้

เซต X บน Metric เป็นฟังก์ชันระยะทาง $d : X \times X \longrightarrow R$ โดย R เป็นเซตของจำนวนจริงและมีวัตถุ x, y, z อยู่ใน X จะมีคุณสมบัติดังนี้

1. $d(x, y) \geq 0$ ระยะทางจาก x ไป y จะไม่มีค่าติดลบ (Non-Negativity)
2. $d(x, y) = 0$ ระยะทางจาก x ไป y จะเป็น 0 ถ้า $x = y$
3. $d(x, y) = d(y, x)$ ระยะทางจาก x ไป y จะเท่ากับ y ไป x (Symmetry)
4. $d(x, z) \leq d(x, y) + d(y, z)$ อสมการความไม่เท่ากันของสามเหลี่ยม (Triangular Inequality)

Euclidean Distance

ระยะทางยุคลิดเป็นตัววัดระยะทางวิธีการหนึ่งที่เป็น Metric คือมีคุณสมบัติของตัววัดระยะทาง Metric ครบ 4 ข้อตามที่กล่าวมาข้างต้น ซึ่งเป็นวิธีการวัดระยะทางระหว่างจุดสองจุดโดยมีนิยามดังต่อไปนี้

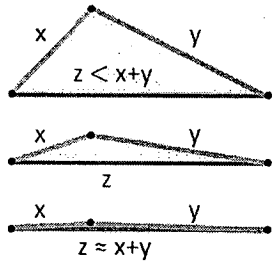
ระยะทางยุคลิดระหว่างจุด p และ q เป็นความยาวของเส้นเชื่อมต่อระหว่างจุด ถ้า $p = (p_1, p_2,..., p_n)$ และ $q = (q_1, q_2,..., q_n)$ เป็นสองจุดที่อยู่บนพื้นที่ยุคลิดแล้วระยะทางระหว่าง p ถึง q คือ

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Triangular Inequality

ทฤษฎีบทความไม่เท่ากันของสามเหลี่ยม เป็นทฤษฎีหนึ่งที่ใช้กันอย่างกว้างขวางในงานวิจัยด้านการค้นพบโมทีฟของข้อมูลอนุกรมเวลา เนื่องจากทฤษฎีความไม่เท่ากันของสามเหลี่ยมนี้บรรจุอยู่ในคุณสมบัติหนึ่งจากทั้งหมดสี่ข้อของตัววัดระยะทาง (Distance Metric) ซึ่งสามารถใช้เป็นค่าขอบเขตล่างเพื่อลดจำนวนการคำนวณหาระยะทางยุคลิดของข้อมูลอนุกรมเวลาได้ โดยมีคุณสมบัติดังนี้

$x + y \geq z$ ความยาวด้าน x รวมกับความยาวด้าน y จะมากกว่าหรือเท่ากับด้าน z เสมอ ดังรูปที่ 3.3 ซึ่งถ้าเรามอง 3 จุดในรูปให้เป็นจุดที่อยู่บนพื้นที่ยุคลิดจะสังเกตได้ว่าถ้าเรารู้ระยะทางของด้าน x และ z เราจะสามารถประมาณค่าของด้าน y ได้ว่า $|x-z| \leq y$ ซึ่งจะนำค่านี้ไปใช้เป็นค่าของขอบเขตล่าง



รูปที่ 3.3 อสมการความไม่เท่ากันของสามเหลี่ยม

งานวิจัยที่เกี่ยวข้อง

ในหัวข้อนี้จะเน้นที่นิยามของแต่ละงานวิจัยที่เกี่ยวข้องกับการค้นพบโมทีฟของข้อมูลอนุกรมเวลาเพื่อเป็นพื้นฐานของการกำหนดนิยามใหม่ที่เกี่ยวข้องกับตัวแปรความยาวของโมทีฟข้อมูลอนุกรมเวลา

Finding Motifs in Time Series [16]

งานวิจัยชิ้นนี้นำเสนอวิธีการค้นพบโมทีฟของข้อมูลอนุกรมเวลาและเสนอนิยามใหม่เนื่องจากมีความคลุมเครือในการเปรียบเทียบความใกล้เคียงของข้อมูลอนุกรมเวลาเมื่อตำแหน่งของคู่เปรียบเทียบต่างเพียงเล็กน้อยซึ่งกรณีนี้จะทำให้การวัดระยะของคู่เปรียบเทียบมีความคล้ายคลึงกันมากซึ่งเป็นสิ่งที่ไม่ควรจะเป็นเพราะจะทำให้โมทีฟที่ได้อยู่ที่ตำแหน่งใกล้เคียงและเกยทับเหมือนเป็นตัวเดียวกัน

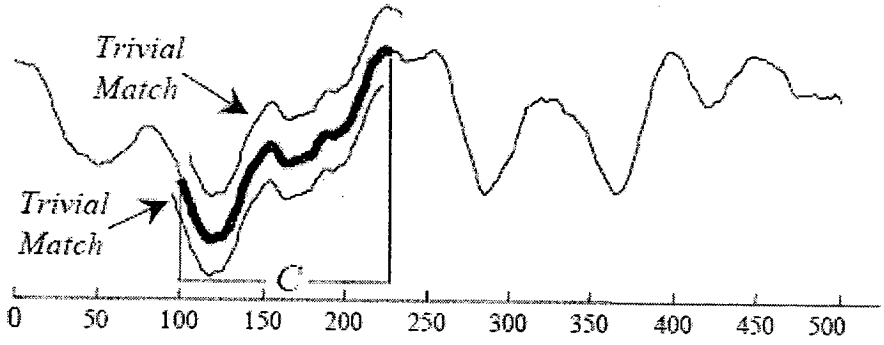
นิยามที่สำคัญต่าง ๆ ถูกกำหนดขึ้นเพื่อขจัดความคลุมเครือของการจัดลำดับย่อยให้อยู่ในชุดของโมทีฟของข้อมูลอนุกรมเวลา มีดังต่อไปนี้

1. Time Series : ข้อมูลอนุกรมเวลา $T = t_1,...,t_m$ คือ ชุดลำดับของจำนวนจริงจำนวน m ตัว

2. Subsequence (ลำดับย่อย) : ถ้ามีลำดับของข้อมูลอนุกรมเวลา T ที่ขนาดความยาว m แล้ว ลำดับย่อย C ใน T ที่ขนาดความยาว $n < m$ คือ $C = t_p,...,t_{p+n-1}$ เมื่อ $1 \le p \le m - n + 1$

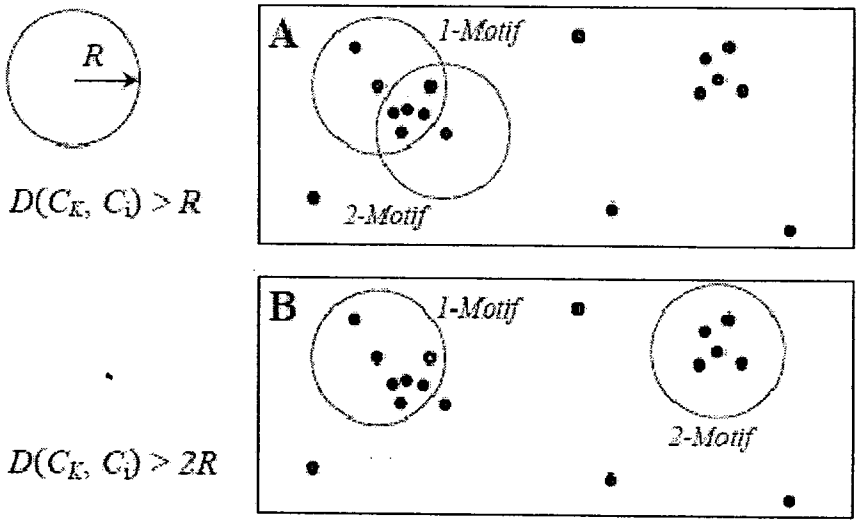
3. Match : เมื่อตัวแปร R เป็นระยะพิสัยที่เป็นจำนวนจริงบวกที่ถูกกำหนดขึ้น และข้อมูลอนุกรมเวลา T มีลำดับย่อย C ที่เริ่มต้นที่ตำแหน่ง p และลำดับย่อย M เริ่มต้นที่ตำแหน่ง q แล้ว C และ M จะมีความสัมพันธ์ตามนิยาม Match ก็ต่อเมื่อ $D(C, M) \le R$ โดย D(C, M) คือระยะทางยุคลิดระหว่างลำดับย่อย C และ M

4. Trivial Match : ถ้ามีข้อมูลอนุกรมเวลา T ซึ่งมีลำดับย่อย C ที่ตำแหน่ง p และมีลำดับย่อย M ที่ตำแหน่ง q ที่มีความสัมพันธ์แบบ Match กับลำดับย่อย C แล้ว เราจะเรียก M กับ C ว่ามีความสัมพันธ์แบบ Trivial Match ก็ต่อเมื่อ p=q หรือ ไม่ปรากฏลำดับย่อย M' ใดใดอีกแล้วที่ $D(C, M') > R$ โดยที่ $q < q' < p$ หรือ $p < q' < q$



รูปที่ 3.4 ลำดับย่อย C เกือบทั้งหมดในข้อมูลอนุกรมเวลาจะเข้าคู่กับลำดับย่อยตำแหน่งไปทางซ้ายและขวาที่ติดกันของ C ในตำแหน่งแรกได้ดีที่สุด [16]

5. K-Motifs : ถ้ามีข้อมูลอนุกรมเวลา T และลำดับย่อยความยาว n และมีระยะพิสัย R จะได้โมทีฟของข้อมูลอนุกรมเวลาที่สำคัญที่สุดที่เรียกว่า 1-Motif คือ ลำดับย่อย C1 ที่มีจำนวนลำดับย่อยอื่น ๆ มามีความสัมพันธ์แบบ Non-Trivial Match กับตัวมันเองสูงสุด ดังนั้น K-Motifs จึงหมายถึงโมทีฟของข้อมูลอนุกรมเวลาที่มีความสำคัญเป็นอันดับที่ K ซึ่งก็คือลำดับย่อย Ck ที่มีจำนวนลำดับย่อยอื่น ๆ มามีความสัมพันธ์แบบ Non-Trivial Match กับตัวมันเองสูงเป็นอันดับที่ K โดยมีระยะทางยุคลิดระหว่างลำดับย่อยแต่ละตัวที่เป็นโมทีฟในแต่ละอันดับ $D(C_K, Ci) > 2R$ โดยที่ $1 \le i < K$



รูปที่ 3.5 การอธิบายนิยามของ K-Motifs ด้วยรูปภาพว่าเหตุใดระยะทางยุคลิดของโมทีฟแต่ละตัวถึงต้องมีระยะห่างจากกันมากกว่า 2R ขึ้นไป ซึ่งมีเหตุผลคือถ้าใช้เพียงระยะทาง R จะทำให้ลำดับย่อยที่เป็นโมทีฟแต่ละตัวใน K-Motifs เกิดการใช้ลำดับย่อยอื่น ๆ ร่วมกันดังรูป A ซึ่งไม่ควรจะเป็น ในทางตรงข้ามถ้าระยะทางมากกว่า 2R จะได้ โมทีฟที่ไม่เกิดการใช้ลำดับย่อยอื่นร่วมกันซึ่งจะได้โมทีฟที่เฉพาะตัวมากกว่า [16]

หลังจากกำหนดนิยามแล้วจึงได้นำเสนออัลกอริทึมในการหาโมทีฟของข้อมูลอนุกรมเวลา โดยวิธีการของอัลกอริทึมนี้จะเริ่มต้นด้วยการลดมิติของข้อมูลอนุกรมเวลาโดยใช้ Piecewise Aggregate Approximation (PAA) ซึ่งเป็นการทำให้ข้อมูลอนุกรมเวลาซึ่งเป็นข้อมูลต่อเนื่องเปลี่ยนเป็นข้อมูลวิยุตซึ่งแทนที่ด้วยข้อมูลที่เป็นตัวอักษร เมื่อลำดับย่อยแต่ละลำดับผ่านกระบวนการลดมิติข้อมูลแล้วจึงนำมาเปรียบเทียบกันเพื่อหาระยะทางของแต่ละลำดับ

Exact Discovery of Time Series Motifs [18]

งานวิจัยชิ้นนี้เป็นงานวิจัยที่เกี่ยวกับการค้นพบโมทีฟของข้อมูลอนุกรมเวลาล่าสุดที่สามารถหาโมทีฟของข้อมูลอนุกรมเวลาที่ถูกต้องแม่นยำโดยไม่ใช่แค่การประมาณค่าใกล้เคียงและสามารถประมวลผลได้เร็วกว่าวิธีอื่นมากซึ่งแสดงให้เห็นในผลการทดลองกับข้อมูลทดสอบต่าง ๆ

นิยามของข้อมูลอนุกรมเวลาที่ใช้ในงานวิจัยชิ้นนี้มีดังต่อไปนี้

1. Time Series : คือ ลำดับของจำนวนจริงที่ต่อเนื่องกันเป็นจำนวน n ตัว T = (t1,...,tn)

2. Time Series Database (D) : เป็นชุดของข้อมูลอนุกรมเวลาจำนวน m ชุดที่อาจมีความยาวที่แตกต่างกัน

3. Time Series Motif ของ Time Series Database (D) : คือ คู่ของข้อมูลอนุกรมเวลา {Ti, Tj} ใน D ที่เป็นคู่ที่คล้ายกันมากที่สุดซึ่งหมายถึงมีระยะทางระหว่างคู่ข้อมูลอนุกรมเวลาน้อยที่สุด

4. kth-Time Series Motif : คู่ของอนุกรมเวลาในฐานข้อมูล D ที่คล้ายกันมากที่สุดเป็นอันดับที่ k โดยคู่ของข้อมูลอนุกรมเวลา {Ti, Tj} เป็นคู่ที่คล้ายกันมากที่สุดเป็นอันดับที่ kth ก็ต่อเมื่อ มีชุดข้อมูล S ของคู่ของข้อมูลอนุกรมเวลาขนาด k-1 โดย $\forall$Td $\in$ D และ {Ti, Td} $\notin$ S และ {Tj, Td} $\notin$ S และ $\forall${Tx,Ty} $\in$ S, {Ta,Tb} $\notin$ S และระยะทาง dist(Tx, Ty) $\leq$ dist(Ti, Tj) $\leq$ dist(Ta, Tb) ตัวอย่างเช่น k = 5 จะมีจำนวนคู่ของข้อมูลอนุกรมเวลาอยู่ในชุด S จำนวน 4 คู่ คือ {Tx,Ty} ซึ่ง {Ti, Tj} จะเป็น 5-Time Series Motif ก็ต่อเมื่อมี {Tx,Ty} จำนวน 4 คู่ที่มีระยะทางน้อยกว่า {Ti, Tj} และมี {Ta,Tb} ที่อยู่นอกชุด S ทั้งหมดที่ระยะทางมากกว่า {Ti, Tj}

5. The Range Motif with Range R : คือ ชุดของข้อมูลอนุกรมเวลาที่มีคุณสมบัติว่าแต่ละข้อมูลในชุดจะมีระยะทางระหว่างพวกมันเองน้อยกว่า 2R ซึ่งเขียนเป็นทางการได้ว่า S เป็น Range Motif ด้วยค่า range R ก็ต่อเมื่อ $\forall Tx,Ty \in S$, $dist(Tx, Ty) \leq 2R$ และ $\forall Td \in D-S$ $dist(Td, Ty) > 2R$ นิยามนี้ใช้เป็นตัววัดความหนาแน่นในแต่ละบริเวณของพื้นที่ของอนุกรมเวลา

นิยามที่กล่าวมาข้างต้นสามารถต่อยอดให้นำไปใช้กับลำดับย่อยของหนึ่งข้อมูลอนุกรมเวลา โดยแทนที่จะพิจารณาข้อมูลอนุกรมเวลาหลายข้อมูลก็เปลี่ยนเป็นพิจารณาลำดับย่อยทั้งหมดของหนึ่งข้อมูลอนุกรมเวลาในฐานข้อมูลอนุกรมเวลา D แทน โดยมีนิยามต่อเนื่องดังนี้

6. Subsequence : ลำดับย่อยขนาดความยาว n ของข้อมูลอนุกรมเวลา T = (t1, t2...,tm) คือข้อมูลอนุกรมเวลา $Ti,n = (ti, ti+1, ...,ti+n-1)$ โดยที่ $1 \leq i \leq m-n+1$

7. Subsequence Motif : เป็นคู่ของลำดับย่อย {Ti,n, Tj,n} ของข้อมูลอนุกรมเวลาขนาดความยาว T ที่มีความคล้ายกันมากที่สุด คือระยะทางของคู่ลำดับย่อยทั้งหมดมีค่าน้อยกว่า ระยะทางของคู่ {Ti,n, Tj,n}

อัลกอริทึมที่งานวิจัยนี้ได้นำเสนอมีชื่อว่า Mueen-Keogh (MK) ซึ่งเป็นวิธีการค้นพบ โมทีฟของข้อมูลอนุกรมเวลาโดยใช้หลักการทางคณิตศาสตร์เรื่องคุณสมบัติของความไม่เท่ากันของสามเหลี่ยมในการข้ามการคำนวณระยะทางที่ไม่สำคัญทิ้งไป โดยวิธีการของอัลกอริทึมจะเริ่มต้นด้วยการกำหนดลำดับย่อยแบบสุ่มขึ้นมาชุดหนึ่งซึ่งเรียกว่าเป็นตัวอ้างอิง กำหนดให้เป็นชุด S = {s1, ..., sn} โดย n เป็นจำนวนตัวอ้างอิงที่สุ่มขึ้นมาจากข้อมูลอนุกรมเวลา T = {t1, ..., tm} โดย m คือจำนวนลำดับย่อยทั้งหมดที่เป็นไปได้ใน T แล้วทำการคำนวณหาระยะทางยุคลิดของตัวอ้างอิงกับลำดับย่อยที่เป็นไปได้ทั้งหมดใน T นั้น เมื่อสังเกตที่ตัวอ้างอิง 1 ตัวที่ s1 เราจะได้ระยะทางจาก s1 ไปยังลำดับย่อย t1 ถึง tm โดยที่ระยะทางที่น้อยที่สุดจะถูกเก็บไว้เป็นตัวแปรที่ชื่อว่า Best-so-far (BSF) สมมติว่าเราพิจารณาที่ระยะทางของคู่ลำดับย่อยโดยให้ A = dist(s1, t1), B = dist(s1, t2) และ C = dist(t1, t2) ค่าของ A และ B นี้เป็นเหมือนค่าของด้าน 2 ด้านของสามเหลี่ยม ซึ่งเราสามารถประมาณค่าของ C ได้จากอสมการความไม่เท่ากันของสามเหลี่ยม คือ $|A-B| \leq C$ ดังนั้น เราจะได้ค่า |A-B| เป็นขอบเขตล่างของ C หมายถึงเป็นค่าประมาณที่ไม่เกินค่าของ C ซึ่งถ้าค่าขอบเขตล่างนี้มากกว่า BSF แล้วแสดงว่า dist(t1, t2) ต้องมากกว่า BSF ด้วยแน่นอน เราจึงสามารถทำการข้ามการคำนวณระยะทางยุคลิดของ t1 และ t2 ไปได้เลยเพราะคู่นี้ไม่มีโอกาสที่จะเป็นคู่โมทีฟอีกแล้วเพราะระยะทางมากกว่าซึ่งหมายถึงความคล้ายกันจึงน้อยกว่าด้วย อัลกอริทึมของงานวิจัยนี้จึงเป็นอัลกอริทึมหนึ่งที่เหมาะสมในการนำมาใช้หาคู่โมทีฟของข้อมูลอนุกรมเวลาเพื่อการศึกษาในการหาความยาวของโมทีฟ

Discovering original motifs with different lengths from time series [21]

งานวิจัยที่เกี่ยวข้องกับการค้นพบตัวแปรความยาวของข้อมูลอนุกรมเวลา โดยในงานชิ้นนี้ได้กล่าวถึงปัญหาของตัวแปรความยาวและได้นำเสนออัลกอริทึมใหม่ในการค้นพบตัวแปรความยาวของโมทีฟข้อมูลอนุกรมเวลาที่เรียกว่า Motif Concatenation

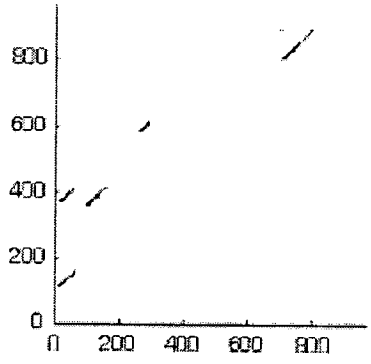
นิยามที่ใช้ในงานวิจัยมีดังต่อไปนี้

1. Time Series : ข้อมูลอนุกรมเวลา T = <t1, ..., tm> เป็นลำดับจำกัดของจำนวนจริง โดย m คือความยาวของข้อมูลอนุกรมเวลา

2. Collision Matrix : เมทริกซ์การชนกัน CM ของข้อมูลอนุกรมเวลา T คือ เมทริกซ์ขนาด q x q โดยสมาชิกแต่ละตัวบนเมทริกซ์แสดงด้วย eij $\in$ M (M น่าจะหมายถึงเมทริกซ์ซึ่งไม่ได้ประกาศว่าเป็นอะไรในงานชิ้นนี้) eij = collision_hit(si, sj) โดย collision_hit คือ ดีกรีความคล้ายกันของ 2 ลำดับย่อยที่ได้มาจากการสร้างเมทริกซ์การชนกันด้วยวิธีการฉายแบบสุ่ม (Random Projection) [11] โดยช่องใดที่มีดีกรีการชนกันสูงสุดจะเรียกว่าเป็น top-k Matches ในที่นี้ลำดับย่อยคู่ใดที่ช่องบนเมทริกซ์มีดีกรีการชนกันสูงสุดจะเป็นโมทีฟของข้อมูลอนุกรมเวลา

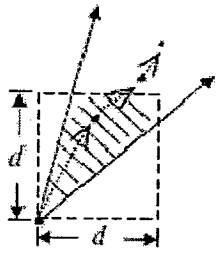
3. Motif : สำหรับลำดับย่อยใดใด 2 ลำดับ si ,sj $\in$ S ด้วยความยาว w ถ้าดีกรีการชนกัน eij = collision_hit(si, sj) มีดีกรีการชนกันสูงสุดเราเรียก คู่ของ M = (si, sj) เป็น k-motif ของข้อมูลอนุกรมเวลา T

อัลกอริทึมของงานวิจัยชิ้นนี้ใช้การสร้างเมทริกซ์การชนกันที่ขนาดตัวแปรความยาวที่แตกต่างกันแล้วจับแต่ละข้อมูลที่เป็นจุดแสดงตำแหน่งของโมทีฟของข้อมูลอนุกรมเวลามาต่อกันซึ่งแสดงได้ดังรูปที่ 3.6



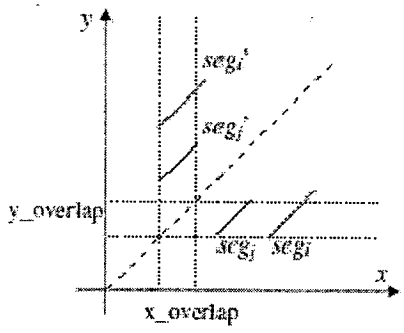
รูปที่ 3.6 เมทริกซ์การชนกันที่แสดงตำแหน่งของโมทีฟของข้อมูลอนุกรมเวลา [21]

ในวิธีการนี้จะแบ่งส่วนของโมทีฟบนเมทริกซ์การชนกันแต่ละจุดออกเป็นกลุ่มๆ เรียกว่า Segment โดยวิธีการแบ่งจุดนั้นทำได้โดยการกำหนดพารามิเตอร์เพิ่มเติมอีก 3 ตัวคือ d, $\alpha 1$ และ $\alpha 2$ โดยที่ d คือ ระยะทางที่จะใช้ต่อจุดของโมทีฟ ส่วน $\alpha 1$ และ $\alpha 2$ คือ ความชันของเส้นตรง 2 เส้นที่เป็นกรอบในการต่อจุดซึ่งแสดงไว้ดังรูปที่ 3.7



รูปที่ 3.7 อัลกอริทึม Motif Concatenation ซึ่งต้องกำหนดพารามิเตอร์ 3 ตัวคือ d คือ ระยะการต่อจุดโมทีฟ $\alpha 1$ และ $\alpha 2$ คือ ความชันของเส้นตรง 2 เส้นที่กำหนดกรอบการต่อจุด [21]

หลังจากการต่อจุดแล้วจะได้ผลดังรูปที่ 3.7 จากนั้นความยาวโมทีฟบนจุดของส่วนที่เกยทับกันของแต่ละ Segment จะนำมาคำนวณค่าเฉลี่ยเพื่อเป็นขนาดของตัวแปรความยาวซึ่งเป็นคำตอบของอัลกอริทึม ซึ่งแสดงการเกยทับกันไว้ดังรูปที่ 3.8



รูปที่ 3.8 ส่วนที่เกยทับกันตามรูปจากแกน x ซึ่งมีการฉายลงบนแกน y เพื่อตรวจสอบดีกรีการเกยทับกันของทั้งสองแกนตามการนิยามในงานวิจัยซึ่งอ่านเพิ่มเติมได้ใน [21]

อัลกอริทึมนี้ยังมีความซับซ้อนของการใช้งานสูงเนื่องจากต้องกำหนดพารามิเตอร์ถึง 4 ตัวคือ d, $\alpha 1$ และ $\alpha 2$ ที่กล่าวไว้ในงานวิจัย ซึ่งต้องพิจารณาหลังจากการสร้างเมทริกซ์การชนกัน CM ว่าควรให้ค่าเป็นเท่าไร อีกทั้งวิธีการต่อจุดของอัลกอริทึม Motif Concatenation ยังต้องการใช้ เมทริกซ์การชนกันเพื่อสร้างพื้นที่และความชันตามพารามิเตอร์ที่ใส่เข้าไป ดังนั้น เมทริกซ์การชนกัน CM จึงต้องเป็นข้อมูลนำเข้าอีกหนึ่งตัว ซึ่งซับซ้อนกว่าการกำหนดพารามิเตอร์ w ซึ่งเป็นความยาวเพียงค่าเดียวมากมายนัก มากไปกว่านั้นการเกยทับกันของจุดโมทีฟบนเมทริกซ์การชนไม่ได้มีโอกาสที่จะเกิดขึ้นเสมอไปในข้อมูลอนุกรมเวลาซึ่งมาจากแหล่งกำเนิดหลายประเภท ซึ่งวิธีการนี้จะใช้ไม่ได้เลยถ้า Segment ไม่เกิดการเกยทับกัน ส่วนวิธีการประเมินการทดลองของงานวิจัยนี้ยังมีความคลุมเครือไม่แน่ชัดว่าโมทีฟที่ได้จากความยาวที่เป็นคำตอบจากอัลกอริทึมคุณภาพเป็นอย่างไร ซึ่งผลการทดลองระบุเพียงสามารถหารูปแบบของโมทีฟของข้อมูลสังเคราะห์ที่สร้างโดยการใส่รูปแบบโมทีฟลงไปในข้อมูลอนุกรมเวลาได้เท่านั้น อัลกอริทึมของงานวิจัยชิ้นนี้จึงยังไม่มีประสิทธิภาพเพียงพอที่จะนำไปใช้งานจริงได้

Approximate Variable-Length Time Series Motif Discovery Using Grammar Inference [15]

งานวิจัยที่มีวัตถุประสงค์ในการหาโมทีฟของข้อมูลอนุกรมเวลาอีกงานหนึ่งที่ใช้หลักการของ Grammar Induction โดยมีนิยามที่ใช้ในงานวิจัย ดังนี้

1. Time Series : ข้อมูลอนุกรมเวลา T = t1, t2...,tm คือชุดของลำดับของจำนวนจริงจำนวน m ตัว

2. Subsequence : ถ้ามีข้อมูลอนุกรมเวลา T ขนาดความยาว m แล้วลำดับย่อย C ใน T คือส่วนแบ่งย่อยความยาว n ≤ m ที่ตำแหน่งที่ติดกันจาก p เป็นต้นไป นั่นคือ C = tp, ...,tp+n-1 โดย 1 ≤ p ≤ m-n+1

3. Sliding Window : ถ้ามีข้อมูลอนุกรมเวลา T และความยาวของลำดับย่อยที่ผู้ใช้งานกำหนดขนาด n แล้วทุก ๆ ลำดับย่อยที่เป็นไปได้ของ T จะถูกแยกออกมาได้โดย sliding window ขนาดความยาว n เลื่อนผ่าน T โดยพิจารณาเฉพาะลำดับย่อย Cp ที่ $1 \le p \le m-n+1$

อัลกอริทึมที่นำเสนอมีชื่อว่า Sequitur ซึ่งเริ่มต้นด้วยการทำให้ข้อมูลอนุกรมเวลาเปลี่ยนเป็นข้อมูลวิยุตโดยใช้ SAX – Symbolic Aggregate approXimation [17] แล้วจึงประมวลผลข้อมูลนั้นด้วย Sequitur เพื่อหาข้อมูลที่ซ้ำกันจากนั้นจึงแปลงจากข้อมูลวิยุตกลับเป็นข้อมูลอนุกรมเวลา

เนื่องจากอัลกอริทึมของงานวิจัยชิ้นนี้อยู่ในการทดลองเบื้องต้นเท่านั้น โดยที่อัลกอริทึม Sequitur มีพารามิเตอร์ที่จะต้องทำการกำหนดเพิ่มเติมอีก 3 ตัว คือ ความยาวเริ่มต้นของโมทีฟโดยจะปรับความยาวเพิ่มขึ้นไปเรื่อย ๆ เพื่อหาความยาวที่เป็นคำตอบ การนำวิธีการของ SAX มาใช้ในอัลกอริทึมจึงต้องกำหนดพารามิเตอร์เพิ่มขึ้นอีก 2 ตัว คือ จำนวน Segment ซึ่งใช้ในการแบ่งส่วนข้อมูลอนุกรมเวลาเพื่อลดจำนวนมิติลง และ จำนวนของตัวอักษรที่จะใช้แทนข้อมูลอนุกรมเวลา ทำให้อัลกอริทึมนี้ยังมีความซับซ้อนในการใช้งาน และยังไม่มีการวัดผลที่ชัดเจนว่ารูปแบบของโมทีฟที่มาจากความยาวที่ค้นพบมีคุณภาพอย่างไร ในงานวิจัยบอกเพียงว่าสามารถหาโมทีฟได้บางข้อมูลเท่านั้น ซึ่งมีความคลุมเครือของคำว่าโมทีฟที่ยังไม่มีนิยามไว้แน่ชัดในงานวิจัยนี้ อัลกอริทึมนี้จึงยังไม่มีประสิทธิภาพเพียงพอในการนำมาใช้งานจริงได้

3.2 การค้นพบโมทีฟความยาวแปรผันสาหรับข้อมูลอนุกรมเวลา

เมื่อกล่าวถึงการพัฒนาวิธีการค้นพบโมทีฟในความยาวที่เหมาะสมนั้น คงหลีกเลี่ยงไม่ได้ที่จะเริ่มต้นการพัฒนาด้วยการหาโมทีฟที่ความยาวแตกต่างกันทั้งหมดที่เป็นไปได้ซึ่งมีหลายค่า เพื่อศึกษาลักษณะของโมทีฟที่เกิดขึ้น ดังนั้น ขั้นตอนแรกจำเป็นต้องมีอัลกอริทึมในการค้นพบโมทีฟที่มีประสิทธิภาพ โดยในงานวิจัยนี้เลือกใช้อัลกอริทึม MK (Mueen-Keogh) [9] ซึ่งเป็นอัลกอริทึมที่ใช้ในการหาโมทีฟที่ดีที่สุดในขณะนี้ จากนั้นในขั้นตอนถัดไปจะเป็นการวิเคราะห์โมทีฟที่ค้นพบในแต่ละความยาวของข้อมูลอนุกรมเวลา แล้วจึงทำการลดจำนวนโมทีฟที่ได้ทั้งหมดลงจนกระทั่งได้เซตของ “โมทีฟที่ดี” เป็นผลลัพธ์ โดยมีขั้นตอนดังต่อไปนี้ ขั้นตอนแรก จะแบ่งกลุ่มโมทีฟโดยใช้เกณฑ์ความคล้ายกันของตำแหน่งที่โมทีฟถูกค้นพบและเกณฑ์ขอบเขตบนของจำนวนโมทีฟ ขั้นตอนที่สองจะทำการเลือกตัวแทนของโมทีฟจากกลุ่มที่เกิดจากการแบ่งกลุ่มนั้น จากนั้นตัวแทนของโมทีฟแต่ละกลุ่มจะนำมาทำการรวมกลุ่มใหม่ แล้วคำนวณคะแนนเพื่อทาการจัดอันดับหากลุ่มที่ดีที่สุด และในขั้นตอนสุดท้ายจะทำการหาตัวแทนของแต่ละกลุ่มนั้นออกมาเป็นคำตอบของอัลกอริทึม

ลำดับการนาเสนอในหัวข้อนี้ จะเริ่มต้นจากขั้นตอนการหาโมทีฟที่ความยาวแตกต่างกันซึ่งเป็นขั้นตอนแรกของวิธีการ จากนั้นจะดำเนินเรื่องตามขั้นตอนของการพัฒนาที่กล่าวไว้เบื้องต้น คือ การแบ่งกลุ่มโมทีฟ การหาตัวแทนของกลุ่มโมทีฟ การคำนวณคะแนนของกลุ่มตัวแทนโมทีฟเพื่อจัดอันดับ “โมทีฟที่ดี” ออกมาเป็นเซตคำตอบของอัลกอริทึม

3.2.1 คำจำกัดความที่ใช้ในงานวิจัย

1. ข้อมูลอนุกรมเวลา(Time Series) คือ ลำดับของจำนวนจริงที่ต่อเนื่องกัน ที่จำนวน n ตัว $T = (t_1, t_2, \ldots, t_n)$

2. ลำดับย่อย(Subsequence) S_i^w คือ ลำดับของจำนวนจริงต่อเนื่องกันที่มีขนาดสั้นกว่าในข้อมูลอนุกรมเวลา T โดย w คือ ขนาดความยาวของลำดับย่อยและ i คือตำแหน่ง เริ่มต้นของลำดับย่อยนั้นโดยลำดับย่อย $S_i^w = (t_i, t_{i+1}, \ldots, t_{i+w-1})$ โดย $w > 1$

3. ระยะทางยุคลิด $EUC(S_i^w, S_j^w)$ คือ ระยะทางระหว่าง 2 ลำดับย่อย S_i^w และ S_j^w ของข้อมูลอนุกรมเวลา T โดย

$$EUC(S_i^w, S_j^w) = \sqrt{\sum_{k=1}^{w} (t_{k+i-1} - t_{k+j-1})^2}$$

$1 \le i \le n\text{-}w$ และ $1 \le j \le n\text{-}w$

4. โมทีฟของข้อมูลอนุกรมเวลา (Time Series Motif) M_w คือ คู่ของลำดับย่อยที่มีรูปร่างคล้ายกันในข้อมูลอนุกรมเวลา T ที่ขนาดความยาว w กำหนดโดย $M_w = (S_{L1}^w, S_{L2}^w, MDist, NDist)$ เมื่อ L_1 และ L_2 คือ ตำแหน่งเริ่มต้นของลำดับย่อย โดยที่ $L_1 < L_2$ และ $MDist = EUC(S_{L1}^w, S_{L2}^w)$ และ NDist คือ ระยะทางยุคลิดในรูปปกติระหว่างลำดับย่อย (S_{L1}^w, S_{L2}^w) ซึ่งคำนวณได้โดย NDist = MDist/w

5. โมทีฟของข้อมูลอนุกรมเวลาอันดับที่k (k^{th} -Time Series Motif) k^{th}-M_w คือโมทีฟที่มีคู่ของลำดับย่อยของอนุกรมเวลา(S_i^w, S_j^w) ที่ขนาดความยาว w ที่คล้ายกันมากที่สุด เป็นอันดับที่ k ในข้อมูลอนุกรมเวลา T โดย k^{th}-Motif และ i^{th}-Motif ไม่ซ้อนทับกันและ $1 \le i < k$ เมื่อการซ้อนทับกันของโมทีฟอาจทำให้เกิดtrivial matches [3][12] ดังนั้น การสกัดโมทีฟทั้งหมดจึงทำโดยการหาโมทีฟที่ไม่ซ้อนทับกัน

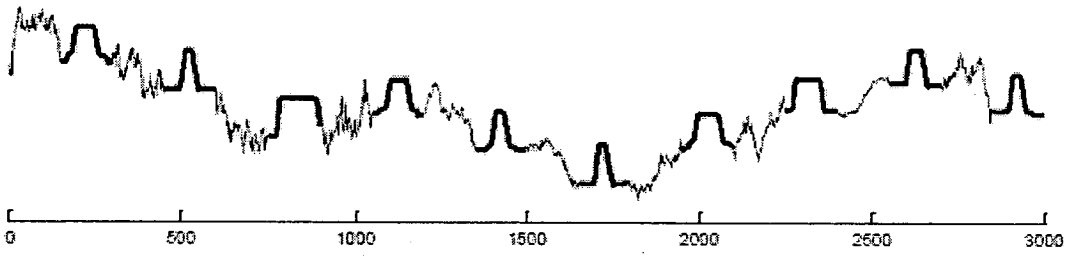
6. โมทีฟซ้อนทับกัน (Overlapped Motifs) โมทีฟสองโมทีฟจะซ้อนทับกันก็ต่อเมื่อ (Mi.L1 ≤ Mj.L1 ≤ Mi,L1+ i หรือMj.L1 ≤ Mi.L1 ≤ Mj,L1+ j) และ (Mi.L2 ≤ Mj.L2 ≤ Mi,L2+ i หรือMj.L2 ≤ Mi.L2 ≤ Mj,L2+ j) โดยที่ i และ j คือความยาวของโมทีฟ

7. โมทีฟที่ดี (Best Motifs) BM_w คือคู่ของลำดับย่อย (S_i^w, S_j^w) ที่มีรูปร่างคล้ายกันที่ยาวที่สุด ที่ถูกค้นพบที่ตำแหน่งใดตำแหน่งหนึ่ง โดยที่ $EUC(S_i^w, S_j^w) \le R$และ R คือ ระยะทางยุคลิดที่มากที่สุดที่ใช้ในการพิจารณาความเป็นโมทีฟที่ดี

8. โมทีฟที่ดีอันดับที่ k (k^{th}-Best Motif) k^{th}- BM_w คือโมทีฟที่มีคะแนนสูงสุด เป็นอันดับที่ k ที่ได้จากฟังก์ชันการให้คะแนนที่นำเสนอในอัลกอริทึม ซึ่งคำนวณโดยมีพื้นฐานจากความคล้ายกันของตำแหน่งการเกิดโมทีฟและความคล้ายกันของคู่ลำดับย่อยของโมทีฟ

3.2.2 การค้นพบโมทีฟของข้อมูลอนุกรมเวลา

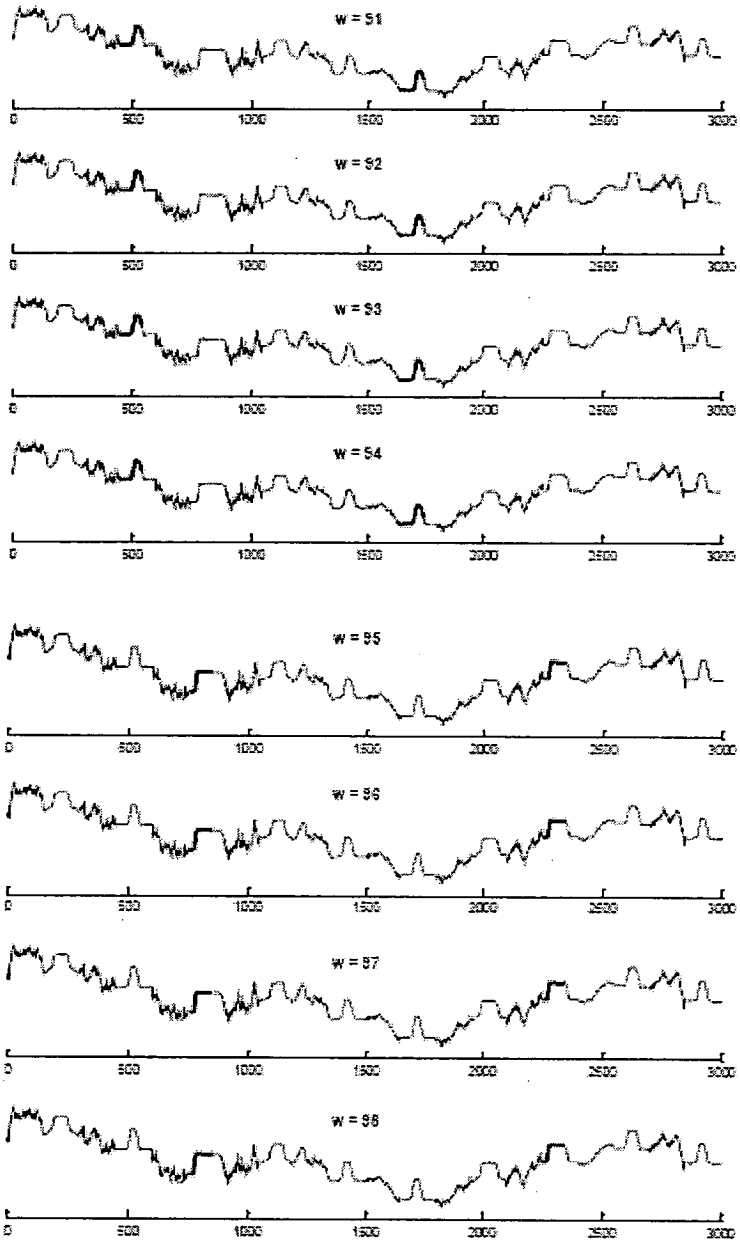
การพัฒนาวิธีการค้นหาโมทีฟในความยาวที่เหมาะสมนั้น จำเป็นต้องทำการวิเคราะห์โมทีฟที่เกิดขึ้นที่ความยาวทั้งหมดที่เป็นไปได้ก่อน เพื่อที่จะนำลักษณะและตำแหน่งการเกิดของโมทีฟมาเปรียบเทียบเพื่อใช้ในการแบ่งกลุ่ม ในงานวิจัยชิ้นนี้ได้เลือกใช้อัลกอริทึม MK (Mueen-Keogh) [9] ซึ่งเป็นอัลกอริทึมที่ใช้ในการค้นพบโมทีฟที่ดีที่สุดในขณะนี้ มาใช้ค้นหาโมทีฟที่ความยาวแตกต่างกัน เพื่อความเข้าใจยิ่งขึ้น จึงทำการสาธิตกับข้อมูลทดลองตัวอย่างซึ่งแสดงดังรูปที่ 3.9



รูปที่ 3.9 ตัวอย่างข้อมูลอนุกรมเวลา Gun-Point [1] ซึ่งสร้างจากการฝังตัวข้อมูล Gun-Point ขนาด 150 จุดข้อมูล ลงไปในข้อมูลแบบสุ่ม (Random Walk) ซึ่งใช้เป็นตำแหน่งของโมทีฟที่ อัลกอริทึมควรจะให้ผลลัพธ์ของโมทีฟมาที่ตำแหน่งและความยาวนี้ โดยโมทีฟที่จะนำมาฝังตัว จะแสดงด้วยเส้นหนา

เนื่องจากข้อมูลอนุกรมเวลา Gun-Point มีขนาด 3000 จุดข้อมูล ดังนั้น ความยาวโมทีฟ (w) ที่เป็นไปได้ทั้งหมด w = คือ 2, 3, 4, ..., 500 จุดข้อมูลโดยที่ความยาวสูงสุดคือ1500 จุดข้อมูลจะได้คู่ของโมทีฟที่แบ่งข้อมูลอนุกรมเวลาออกเป็น 2 ส่วนเท่ากันพอดี ซึ่ง ในงานวิจัยนี้ ได้ทำการหาโมทีฟโดยเปลี่ยนพารามิเตอร์ความยาวของอัลกอริทึม MK โดยให้ค่าตั้งแต่ 2 เป็นต้นไปจนถึง 1500 แล้ววิเคราะห์โมทีฟที่ได้จากอัลกอริทึม ผลจากการหาโมทีฟที่เกิดขึ้นในช่วงระยะความยาวหนึ่ง จะพบว่าโมทีฟที่ถูกค้นพบจะเกิดในตำแหน่งที่ใกล้เคียงกันและเมื่อความยาวเพิ่มขึ้นเรื่อย ๆ จนกระทั่งมีคู่ของลำดับย่อยที่คล้ายกันที่ตำแหน่งอื่นที่คล้ายกันมากกว่าที่ โมทีฟที่เกิดขึ้น จะมีการเปลี่ยนแปลงตำแหน่งโดยมีผลลัพธ์ที่ได้จากการหาโมทีฟในแต่ละความยาว ดังแสดงในภาพที่ 3.10 ในช่วงความยาวของโมทีฟ ตั้งแต่ 91 จุดข้อมูล ไปจนถึง 98 จุดข้อมูล

จากผลลัพธ์โมทีฟดังภาพที่ 3.10 สังเกตได้ว่าโมทีฟที่เกิดขึ้นในแต่ละขนาดความยาวมีทั้งโมทีฟเกิดที่ตำแหน่งใกล้เคียงกันและไม่ใกล้เคียงกัน โดยลักษณะโมทีฟที่เกิดขึ้นที่ตำแหน่ง ใกล้เคียงกันนั้นจะเกิดขึ้นที่ความยาวไล่เรียงกันไปช่วงหนึ่ง คือ โมทีฟที่ความยาว 91 ถึง 94 หลังจากนั้นโมทีฟจะเกิดการเปลี่ยนตำแหน่งไปจากเดิม เป็นโมทีฟที่ความยาว 95 ถึง 98 ดังนั้น จากการสังเกตลักษณะที่เกิดขึ้นเทียบกับข้อมูล Gun-Point ที่ทำการฝังตัวลงไป โมทีฟในตำแหน่งที่ ใกล้เคียงกันนี้จะเป็นส่วนหนึ่งของข้อมูล Gun -Point ซึ่งเป็นข้อสังเกตได้ว่าโมทีฟที่เกิดในตำแหน่งที่ใกล้เคียงกัน จะเป็นส่วนหนึ่งของรูปแบบที่น่าสนใจ ณ ตำแหน่งนั้น โดยจะต้องมีโมทีฟที่มีขนาดความยาวค่าหนึ่งที่ครอบคลุมส่วนที่น่าสนใจทั้งหมด ณ ตำแหน่งนั้น และเนื่องจากในงานวิจัยนี้ ใช้ระยะทางยุคลิดเป็นมาตรวัดความคล้ายกันของคู่ลำดับย่อย ความยาวที่เพิ่มขึ้นของคู่ลำดับย่อย หมายถึงระยะทางยุคลิดที่เพิ่มขึ้น ดังนั้นงานวิจัยนี้จึงต้องหาค่าที่มากสุดของระยะทางยุคลิดของคู่ลำดับย่อยที่เหมาะสมที่บอกว่า “โมทีฟที่ดี” ต้องมีค่าความคล้ายกันไม่เกินค่านี้ซึ่งหมายถึงค่า R ที่เป็นค่าระยะทางยุคลิดที่มากที่สุดสำหรับการพิจารณาว่าเป็น “โมทีฟที่ดี” สำหรับงานวิจัยนี้จึงได้เริ่มต้นโดยทำการแบ่งกลุ่มของโมทีฟในแต่ละความยาวออกเป็นกลุ่ม ๆ โดยใช้เกณฑ์ความคล้ายกันของ ตำแหน่งที่เกิดโมทีฟ ซึ่งจะใช้การซ้อนทับกันของโมทีฟเป็นเกณฑ์ความคล้ายกันในการแบ่งกลุ่ม



ภาพที่3.10 ตัวอย่างผลลัพธ์โมทีฟที่ถูกค้นพบที่ความยาวแตกต่างกัน

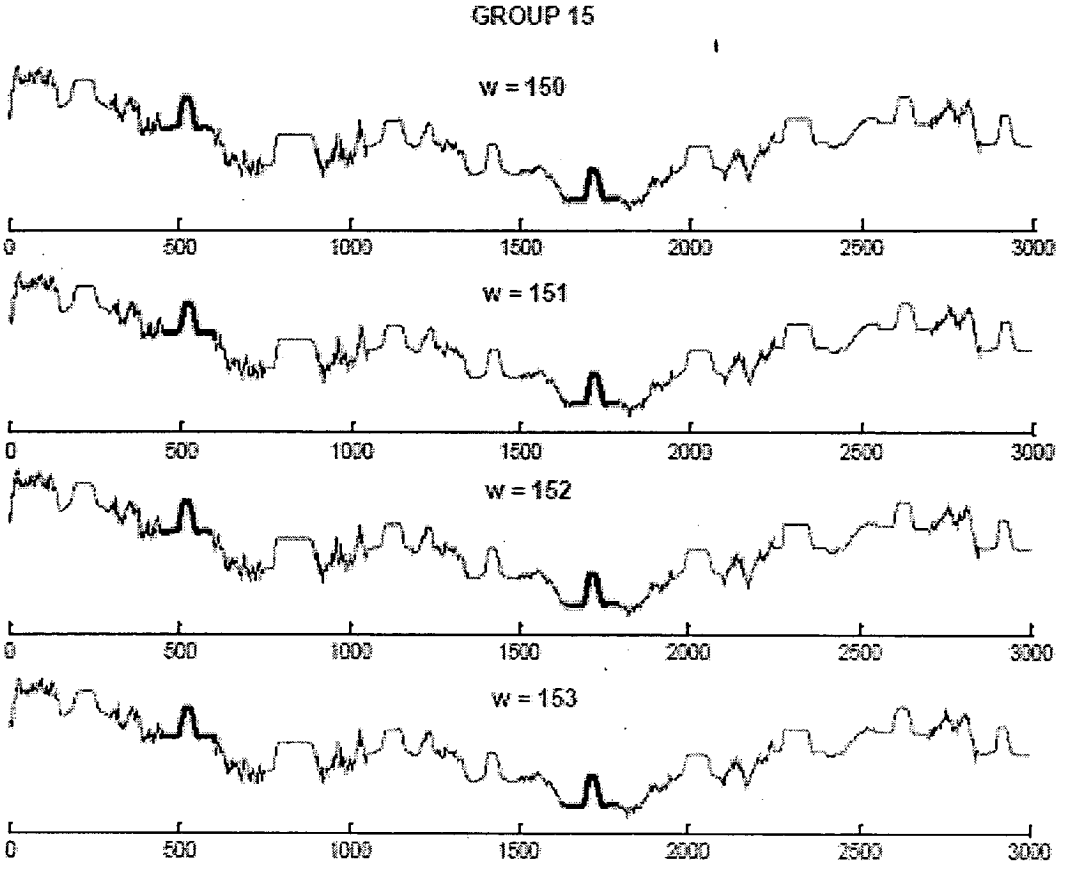
เมื่อได้ผลลัพธ์โมทีฟที่ความยาวทั้งหมดที่เป็นไปได้แล้ว จากนั้นได้ทำการแบ่งกลุ่มของโมทีฟโดยใช้การซ้อนทับของแต่ละโมทีฟเป็นเกณฑ์ โดยโมทีฟที่แต่ละขนาดความยาวจะถูกจัดให้อยู่ในกลุ่มเดียวกันก็ต่อเมื่อคู่ลำดับย่อยของโมทีฟนั้นเกิดการซ้อนทับกัน ดังคำจำกัดความดังนี้

โมทีฟซ้อนทับกัน (Overlapped Motifs) โมทีฟสองโมทีฟจะซ้อนทับกันก็ต่อเมื่อ

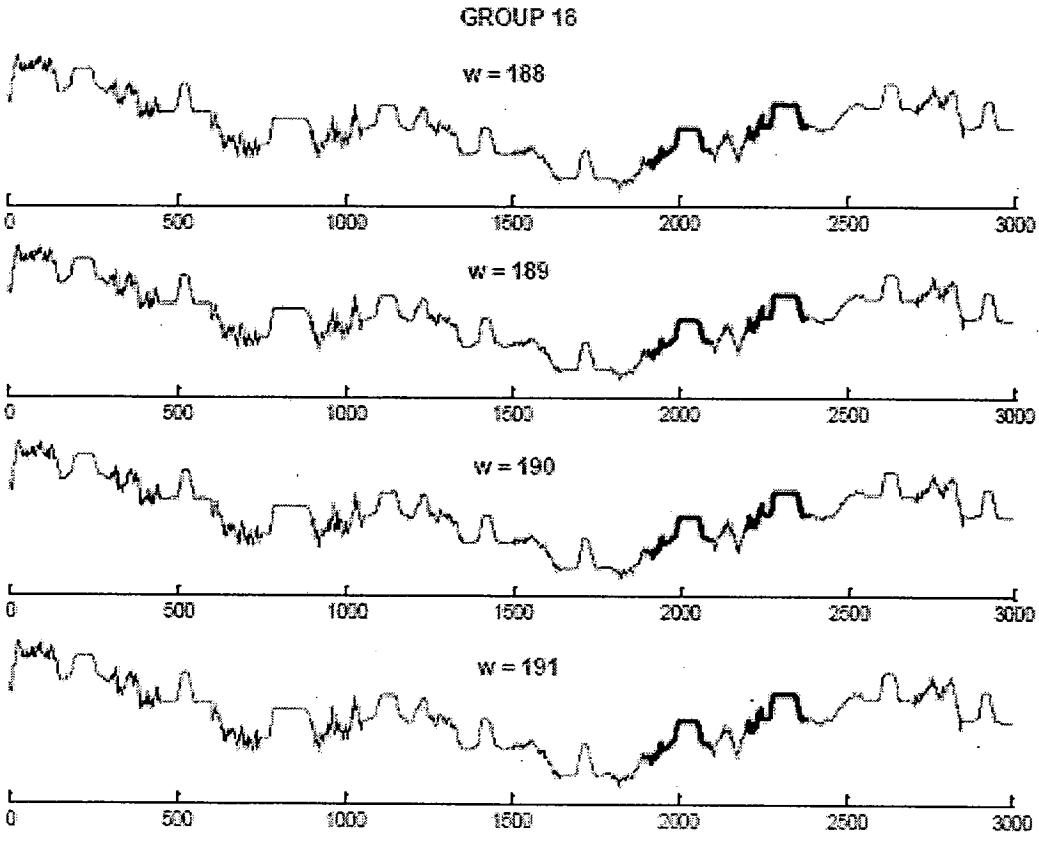
(Mi.L1 ≤ Mj.L1 ≤ Mi,L1+ i หรือMj.L1 ≤ Mi.L1 ≤ Mj,L1+ j) และ

(Mi.L2 ≤ Mj.L2 ≤ Mi,L2+ i หรือMj.L2 ≤ Mi.L2 ≤ Mj,L2+ j) โดยที่ i และ j คือความยาวของโมทีฟ

ซึ่งจะได้ตัวอย่างของกลุ่มโมทีฟ ดังแสดงในภาพที่ 3.11 และภาพที่ 3.12



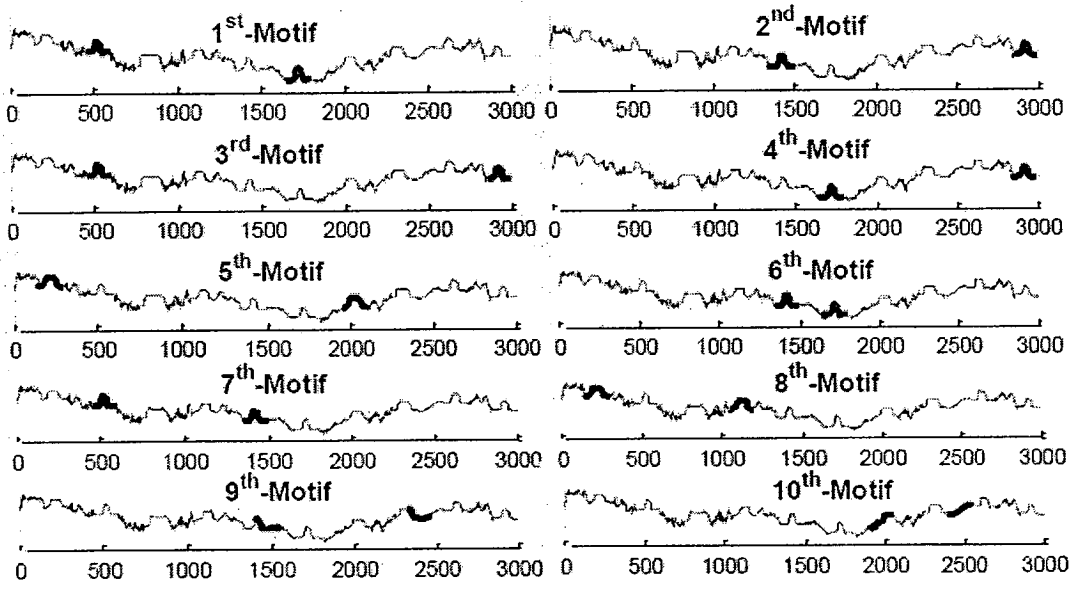
ภาพที่ 3.11 ตัวอย่างกลุ่มโมทีฟกลุ่มที่15 ที่เกิดขึ้นจากการแบ่งกลุ่มแบบใช้เกณฑ์การซ้อนทับกัน โดยในที่นี้โมทีฟที่ความยาว (w) 150, 151, 152 และ 153 ซ้อนทับกัน



ภาพที่3.12 ตัวอย่างกลุ่มโมทีฟกลุ่มที่ 18 ที่เกิดขึ้นจากการแบ่งกลุ่มแบบใช้เกณฑ์การซ้อนทับกัน

สังเกตว่ารูปแบบของโมทีฟที่ได้ในกลุ่มที่ 15 จะมีกลุ่มที่มีโมทีฟความยาว 150 จุดข้อมูล ซึ่งเป็นโมทีฟที่ขนาดความยาวเดียวกับโมทีฟที่ทำการฝังตัว ซึ่งได้โมทีฟออกมารูปแบบตรงกับโมทีฟที่ได้ฝังตัวลงไปในข้อมูลทดลอง ถัดมาในโมทีฟกลุ่มที่ 18 จะพบว่ารูปแบบและตำแหน่งของโมทีฟนั้นใกล้เคียงกับโมทีฟที่ทำการฝังตัวลงไป แต่จะมีข้อมูลแบบสุ่มรวมออกมาด้วย จากการทดลองเบื้องต้นของตัวอย่างนี้แสดงให้เห็นว่าการหาโมทีฟที่ความยาวทั้งหมดที่เป็นไปได้นั้น ไม่เพียงพอที่จะสามารถค้นพบโมทีฟที่ฝังตัวลงไปได้ เนื่องจากโมทีฟที่เป็นโมทีฟที่ดีในตัวอย่างข้อมูลทดลองนี้ มีขนาดความยาวเท่ากัน คือ 150 จุดข้อมูล วิธีการค้นพบเพียงหนึ่งโมทีฟ ต่อหนึ่งความยาวจึงไม่เพียงพอเพราะอาจมีคู่โมทีฟอื่นที่ความยาวเดียวกันเหลืออยู่ แต่ความคล้ายของโมทีฟนั้นเทียบระยะทางยุคลิดแล้วมากกว่าโมทีฟคู่แรกที่ค้นพบ ดังนั้นจึงได้เพิ่มการค้นหาโมทีฟของแต่ละความยาว โดยนำ k^{th}-Time Series Motif มาใช้ในงานวิจัยโดยจะเริ่มต้นหา k^{th}-Motif ที่เป็นไปได้ในแต่ละความยาวโดยวิธีการหา k^{th}-Motif ที่เป็นไปได้นั้น ได้ทำการพัฒนาโดยแยกออกเป็น 2 วิธีการ ดังต่อไปนี้

1. ค้นหาโมทีฟทั้งหมดที่เป็นไปได้โดยเริ่มจากการหา 1^{st}-Motif, 2^{nd}-Motif, 3^{rd}-Motif, ... ตามลำดับ โดยที่ k^{th}-Motif จะต้องไม่ซ้อนทับกับ i^{th}-Motif โดยที่ $1 \le i < k$ ตัวอย่างเช่น 2^{nd}-Motif จะไม่ซ้อนทับกับ 1^{st}-Motif ที่ถูกค้นพบก่อน ซึ่งแปลความโดยสรุปคือ k^{th}-Motif ที่ค้นพบทั้งหมดจะต้องไม่ซ้อนทับกัน โดยผลลัพธ์ของวิธีการนี้แสดงไว้ดังภาพที่ 3.13



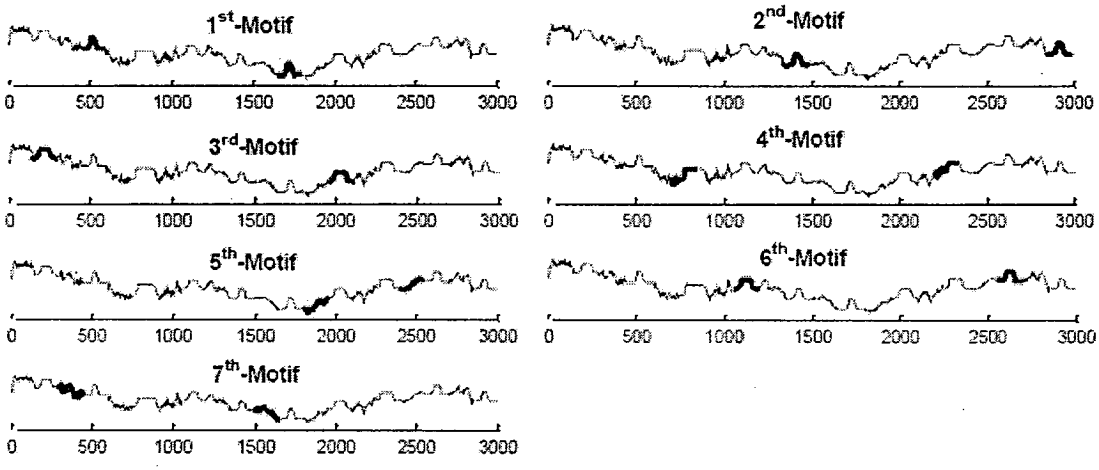
ภาพที่ 3.13 แสดงโมทีฟที่ค้นพบที่ความยาว 150 จุดข้อมูล ซึ่งแสดงตั้งแต่ 1st-Motif ถึง 10th-Motif โดยวิธีการค้นหาโมทีฟทั้งหมดที่เป็นไปได้

วิธีการค้นหาโมทีฟทั้งหมดที่เป็นไปได้ในแต่ละความยาวโมทีฟนี้ ต้องใช้คำนวณหาโมทีฟทั้งสิ้นเป็นจำนวนมากที่สุด โดยคำนวณจาก

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

โดย n คือ จำนวนของลำดับย่อยทั้งหมดที่เป็นไปได้ในข้อมูลอนุกรมเวลา T และจะสามารถหาค่าของ n ได้จาก n = Floor(|T|/w) โดย w คือ ความยาวโมทีฟ ในที่นี้ T ตามภาพที่ 3.13 มีขนาดความยาว3000 จุดข้อมูลและ w=150 ดังนั้น n = Floor(3000/150) = 20 จากนั้น ค่า r คือจำนวนลำดับย่อยที่จะเลือกออกมาเป็นคู่ของโมทีฟ ค่า r จึงมีค่าเท่ากับ 2 เพื่อคำนวณว่าถ้ามีจำนวนลำดับย่อยทั้งหมด 20 จะสามารถเลือกออกมาเป็นคู่โมทีฟได้ทั้งหมดกี่แบบ ดังนั้น จำนวนของโมทีฟที่ต้องค้นหาทั้งหมดที่ความยาวโมทีฟ w=150 คือ $\binom{20}{2}$ = 20!/(2!(20-2)!) = 190 ซึ่งการคำนวณค่านี้เป็นจำนวนที่เป็นขอบเขตบนของ จำนวนโมทีฟที่เป็นไปได้ทั้งหมดในแต่ละความยาวโมทีฟ ซึ่งมีจำนวนมหาศาล ปัญหาที่เกิดขึ้นจากการค้นหาโมทีฟที่เป็นไปได้ ทั้งหมดนี้จึงเป็นปัญหาทางด้านประสิทธิภาพในการประมวลผล ยกตัวอย่างกรณีหนึ่ง เมื่อเริ่มการค้นหาโมทีฟที่ความยาวของโมทีฟที่ค่าน้อย ๆ และชุดข้อมูลทดลองมีขนาดใหญ่ ตัวอย่างจากข้อมูลทดลองเช่น |T|=7000 ซึ่งเป็นขนาดโดยประมาณของข้อมูลทดลองส่วนใหญ่ เมื่ออัลกอริทึมทำการหาโมทีฟทั้งหมดที่ความยาว w=7 และคำนวณค่า n = Floor(7000/7) = 1000 จะมีจำนวนโมทีฟทั้งหมดที่เป็นไปได้ที่เป็นขอบเขตบนที่ต้องมาทำการคำนวณหาทั้งสิ้น $\binom{1000}{2}$= 1000!/(2!(1000-2)!) = 499500 โมทีฟ ซึ่งเป็นจำนวนโมทีฟที่เกิดขึ้นที่ความยาวเดียว ด้วยจำนวนโมทีฟที่มหาศาลนี้ เมื่อต้องค้นหาทุกๆ kth-Motif ที่เป็นไป

ได้ในทุกความยาวโมทีฟ ทำให้เป็นปัญหามากกับการใช้เวลาในการประมวลผลโดย Big-O ของวิธีการนี้ คำนวณจาก $m(_2^n)n^2$ เมื่อ m คือจำนวนค่าความยาวทั้งหมดที่เป็นไปได้ในข้อมูลอนุกรมเวลา T และ n คือจำนวนลำดับย่อยในข้อมูลอนุกรมเวลา T โดยรายละเอียดแล้ว n^2 คือเวลาที่ใช้ในการหาโมทีฟ 1 โมทีฟ ($_2^n$) คือจำนวน k^{th}-Motif ทั้งหมดที่เป็นไปได้ในแต่ละความยาวโมทีฟ ซึ่งจะได้ค่าของ Big-O เป็น$O(mn^4)$ โดยเวลาที่ใช้ในการประมวลผลนี้ยังไม่รวมถึงขั้นตอนถัด ๆ ไปของอัลกอริทึมที่มีทั้งขั้นตอนการจัดกลุ่ม และฟังก์ชันการให้คะแนนซึ่งจากการทำการพัฒนาและทดลองแล้ว ไม่สามารถที่จะทำได้ในระยะเวลาที่จำกัด งานวิจัยนี้จึงจำกัดการหาโมทีฟโดยการใช้แนวทางในวิธีที่ 2 ในการหาโมทีฟ



ภาพที่ 3.14 แสดงโมทีฟที่ค้นพบที่ความยาว 150 จุดข้อมูล ซึ่งแสดงตั้งแต่ 1^{st}-Motif ถึง7^{th}-Motif

วิธีการค้นหาโมทีฟแบบลดจำนวนนี้ ต้องคำนวณหา k^{th}-Motif โดยมีจำนวนโมทีฟที่เป็นขอบเขตบนสูงสุด = Floor(|T|/2w) ถ้าขนาดของข้อมูลอนุกรมเวลา T มีขนาด 3000 ที่ความยาวโมทีฟw=150 จะมีจำนวนโมทีฟที่เป็นขอบเขตบนสูงสุด = Floor(|3000|/2(150)) = 10 โดยมีความหมายว่าจะสามารถหาคู่ของลำดับย่อยจำนวนสูงสุดได้กี่คู่ลำดับ บนข้อมูลอนุกรมเวลาขนาด |T| ดังนั้นจึงต้อง นำขนาดของข้อมูลอนุกรมเวลาหารด้วยความยาวของลำดับย่อยจำนวน 2 ลำดับจึงนำความยาวคูณด้วย 2 ซึ่งเมื่อทำการหาโมทีฟตามการทดลองแล้ว ได้จำนวนทั้งสิ้น 7 โมทีฟ ดังตัวอย่างแสดงในภาพที่ 3.14 จากแนวทางนี้ สามารถลดเวลาในการประมวลผลลงเหลือ $O(mn^2)$ และเมื่อลองเปรียบเทียบโมทีฟที่ได้ในวิธีที่ 2 กับวิธีการแรกดังภาพที่ 3.13 โมทีฟที่ค้นพบในวิธีการแรกที่ 1^{st}-Motif, 2^{nd}-Motif, 3^{rd}-Motif, 4^{th}-Motif, 6^{th}-Motif และ 7^{th}-Motif จะเป็นรูปแบบ GunPoint ที่มีความคล้ายกันจำนวน 4 ลำดับที่ทำการฝังตัวลงไป โดยโมทีฟทั้ง 6 โมทีฟที่ถูกค้นพบนี้ จะสลับคู่ไปมาระหว่างกันอยู่ เมื่อเปรียบเทียบกับโมทีฟที่ได้ในวิธีการที่ 2 1^{st}-Motif และ 2^{nd}-Motif จะเป็นรูปแบบ GunPoint ที่มี ความคล้ายกันจำนวน 4 ลำดับที่ทำการฝังตัวลงไปเช่นกัน ดังนั้น ลำดับทั้งหมดที่ทำการฝังตัวลงไป ก็ยังสามารถถูกค้นพบออกมาในวิธีการที่ 2 อย่างไรก็ตาม ถ้าหากฝังตัวรูปแบบของ GunPoint ที่คล้ายกัน ลงไปทั้งสิ้น 3 ลำดับหรือจำนวนลำดับที่ทำการฝังลงไปเป็นจำนวนคี่ วิธีการที่ 2 นี้ จะสามารถค้นพบลำดับได้เพียงจำนวนคู่เท่านั้น จะมีอยู่ 1 ลำดับที่ไม่

ถูกค้นพบออกมา ซึ่งเป็นประเด็นหนึ่งที่เกิดขึ้นจากการใช้วิธีที่ 2 เพื่อลดจำนวนการค้นพบโมทีฟเพื่อลดเวลาการประมวลผล ลง ทั้งนี้ เนื่องจากหน้าที่ของการค้นพบโมทีฟนั้นเป็นการทำงานส่วนย่อยของกระบวนการหลักของ การทำเหมืองข้อมูลพวกการจัดกลุ่มข้อมูล การแยกประเภทข้อมูล ซึ่งการหาโมทีฟนั้น รับหน้าที่ย่อยเพียงเพื่อค้นหารูปแบบที่น่าสนใจออกมาจากข้อมูลอนุกรมเวลา การค้นพบรูปแบบที่น่าสนใจออกมาได้ จึงถือว่าเพียงพอกับหน้าที่ของการค้นพบโมทีฟ โมทีฟที่เป็นผลลัพธ์ทั้งหมดจากขั้นตอนนี้ จะนำเข้าสู่ขั้นตอนถัดไป ซึ่งจะกล่าวถึงในบทที่ 4 เพื่อจัดกลุ่มลำดับย่อยของข้อมูลอนุกรมเวลาแบบกระแส (Subsequence Clustering for Time Series Data Stream) ต่อไป