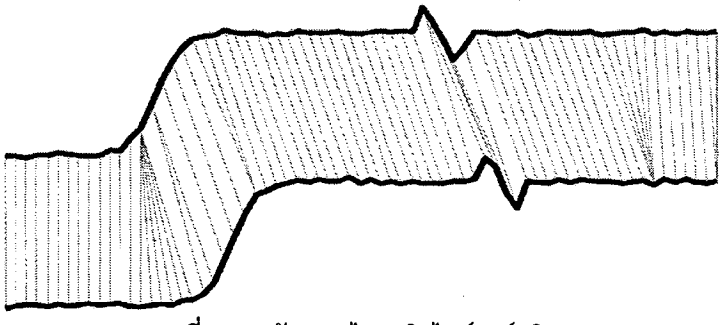
บทที่ 2
การจัดกลุ่มข้อมูลอนุกรมเวลาตามรูปร่าง
Shape-Based Time Series Clustering

2.1 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

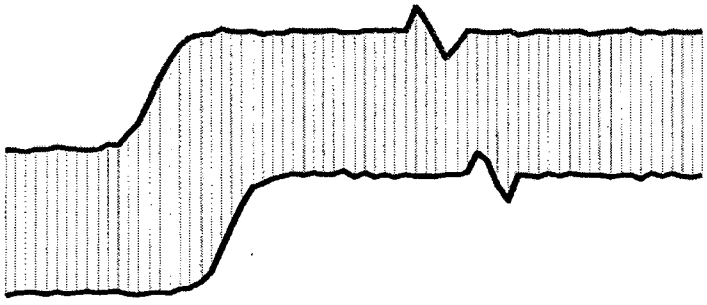
ข้อมูลอนุกรมเวลาได้รับความนิยมมากขึ้นและมีการนำไปใช้ในงานด้านต่าง ๆ อย่างแพร่หลาย เนื่องจากสามารถประยุกต์ใช้กับข้อมูลได้หลายประเภท เช่น ข้อมูลทางชีวภาพ [1], มัลติมีเดีย [2] ด้วยเหตุนี้จึงมีงานวิจัยที่เกี่ยวข้องกับข้อมูลอนุกรมเวลาเกิดขึ้นมากมายไม่ว่าจะเป็น การจำแนก, การจัดกลุ่ม, การหาโมทีฟ (Motif Discovery) และการตรวจสอบความผิดปกติ (Anomaly Detection) โดยงานวิจัยส่วนใหญ่จะใช้ระยะไดนามิกไทม์วอร์ปปิง [3] เป็นมาตรวัดเพราะมีความยืดหยุ่นในการวางแนว (alignment) ระหว่างอนุกรม (ดังรูปที่ 1) จึงมีความแม่นยำสูง



รูปที่ 1 การวัดระยะไดนามิกไทม์วอร์ปปิง

(ที่มา: C. A. Ratanamahatana and E. Keogh, "Making Time-series Classification More Accurate Using Learned Constraints," in SIAM International Conference on Data Mining (SDM), Lake Buena Vista, Florida, 2004, pp. 11-22.)

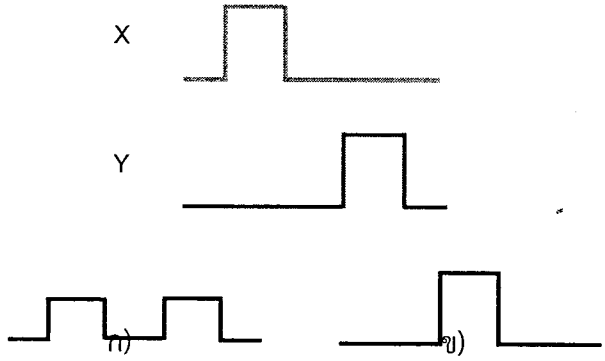
สำหรับข้อมูลอนุกรมเวลา การจัดกลุ่มข้อมูล (Clustering) ถือเป็นหนึ่งในงานที่นักวิจัยให้ความสนใจ โดยอัลกอริทึมที่นิยมใช้ คือ การจัดกลุ่มแบบเคมีนส์ (K-Means Clustering) [4] เนื่องจากอัลกอริทึมที่ไม่ซับซ้อนและมีการทำงานที่รวดเร็ว สำหรับข้อมูลทั่วไป การจัดกลุ่มแบบเคมีนส์จะนิยมใช้การวัดระยะยุคลิด(ดังรูปที่ 2) และการหาตัวแทนของกลุ่มข้อมูลก็จะใช้วิธีการหาค่าเฉลี่ย แต่การวัดระยะยุคลิดซึ่งมีการจำกัดการวางแนวระหว่างอนุกรมให้เป็นแบบหนึ่งต่อหนึ่ง รวมไปถึงการหาตัวแทนกลุ่มข้อมูลโดยการหาค่าเฉลี่ยซึ่งเป็นการเฉลี่ยโดยใช้ระยะยุคลิดหรือการเฉลี่ยแอมพลิจูด (ดังรูปที่ 3 ก.) นั้นไม่เหมาะกับลักษณะของข้อมูลอนุกรมเวลา ด้วยเหตุนี้จึงมีงานวิจัยเกี่ยวกับการหาค่าเฉลี่ยของข้อมูลอนุกรมเวลาโดยอาศัยการวัดระยะไดนามิกไทม์วอร์ปปิง [5] [6] [7] ซึ่งเหมาะกับลักษณะของข้อมูลอนุกรมเวลามากกว่า โดยการเฉลี่ยข้อมูลอนุกรมเวลานั้น นอกจากจะสามารถนำไปใช้ในอัลกอริทึมสำหรับการจัดกลุ่มแล้ว ยังสามารถนำไปประยุกต์ใช้ในงานอื่น ๆ ได้ เช่น ในการจำแนกข้อมูลอนุกรมเวลาที่มีข้อมูลจำนวนมากโดยวิธีเพื่อนบ้านใกล้สุดอันดับหนึ่ง ด้วยการวัดระยะไดนามิกไทม์วอร์ปปิง [7]



รูปที่ 2 การวัดระยะยุคลิด

(ที่มา: C. A. Ratanamahatana and E. Keogh, "Making Time-series Classification More Accurate Using Learned Constraints," in SIAM International Conference on Data Mining (SDM), Lake Buena Vista, Florida, 2004, pp. 11-22.)

ในปี 1996 มีการเสนออัลกอริทึม Non Linear Alignment and Averaging Filter (NLAAF) [5] ซึ่งเป็นการเฉลี่ยข้อมูลอนุกรมเวลาทีละสองอนุกรม โดยอาศัยระยะไดนามิกไทม์วอร์ปปิง แต่อนุกรมที่ได้จากการเฉลี่ยแบบนี้จะมีขนาดยาวขึ้น ต่อมาในปี 2009 ได้มีการเสนอ Prioritized Shape Averaging (PSA) [6] ซึ่งเป็นการเฉลี่ยรูปร่างที่ปรับปรุงเพิ่มจาก NLAAF (ดังรูปที่ 3 ข.) โดยทำการเฉลี่ยข้อมูลทีละสองอนุกรม ตามลำดับที่ได้จากการจัดกลุ่มแบบลำดับขั้น (Hierar-chical Clustering) และทำการปรับขนาดของอนุกรมให้มีความยาวเท่ากับอนุกรมที่นำมาเฉลี่ยโดยใช้การยืดหดแบบยูนิฟอร์ม (Uniform Scaling) แต่การปรับความยาวของอนุกรมด้วยวิธีนี้จะส่งผลต่อรูปร่างของอนุกรม ในปี 2010 จึงมีการนำเสนออัลกอริทึม Cubic-Spline Dynamic Time Warping (CDTW) [8] โดยทำการปรับความยาวของอนุกรมแบบคิวบิคสไปลน์ (Cubic-Spline Interpolation) ซึ่งจะช่วยรักษารูปร่างของอนุกรมไว้ได้



รูปที่ 3 ข้อมูลอนุกรมเวลาที่ได้จากการเฉลี่ยอนุกรม X และ Y
ก) แบบแอมพลิจูด ข) แบบรูปร่าง

สำหรับการเฉลี่ยรูปร่างโดยอาศัยระยะไดนามิกไทม์วอร์ปปิง อนุกรมที่ได้จากการเฉลี่ยจะแตกต่างจากการเฉลี่ยข้อมูลบนพื้นที่ของยุคลิด (Euclidean Space) ซึ่งข้อมูลที่ได้จากการเฉลี่ยบนพื้นที่ของยุคลิดนั้นจะมีตำแหน่งตรงกลางระหว่างข้อมูลที่นำมาเฉลี่ย และผลรวมของระยะระหว่างข้อมูลเฉลี่ยกับข้อมูลที่นำมาเฉลี่ยจะมีค่าน้อยที่สุดเสมอ ในขณะที่อนุกรมที่ได้จากการเฉลี่ยรูปร่างอาจไม่อยู่ตำแหน่งตรงกลางและผลรวมของระยะระหว่างอนุกรมที่ได้จากการเฉลี่ยกับอนุกรมที่นำมาเฉลี่ยไม่เป็นค่าน้อยที่สุดเสมอเมื่อวัดด้วยระยะไดนามิกไทม์วอร์ปปิง [8] จึงมีการเสนออัลกอริทึม Iterative Cubic-Spline Dynamic Time Warping (ICDTW) [7] โดยการปรับปรุง CDTW

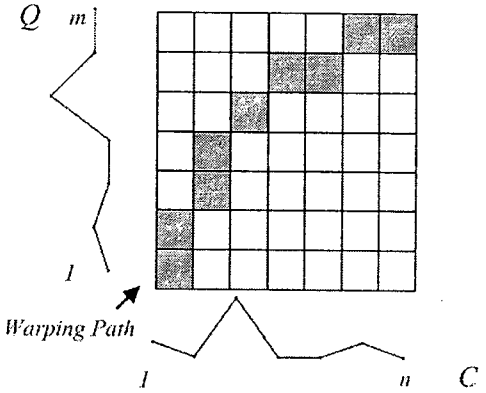
งานวิจัยนี้ มีแนวคิดในการปรับปรุงการจัดกลุ่มข้อมูลอนุกรมเวลาโดยนำวิธีการจัดกลุ่มตามรูปร่างซึ่งนำอัลกอริทึมสำหรับเฉลี่ยข้อมูลอนุกรมเวลามาใช้ในการหาตัวแทนของกลุ่มข้อมูล และนำระยะไดนามิกไทม์วอร์ปปิงมาใช้แทนระยะยุคลิด เพื่อเพิ่มประสิทธิผลและความแม่นยำในการจัดกลุ่มข้อมูลอนุกรมเวลา

ระยะไดนามิกไทม์วอร์ปปิง

ระยะไดนามิกไทม์วอร์ปปิง [3] เป็นมาตรวัดที่นิยมใช้กับข้อมูลอนุกรมเวลา เนื่องจากมีความยืดหยุ่นจึงสามารถปรับการวางแนวระหว่างสองอนุกรมได้เหมาะสมที่สุด การวัดระยะสามารถทำได้โดยใช้สมการที่ (1)

$$dist(q_i, c_j) = dist(q_i, c_j) + min\begin{cases} dist(q_{i-1}, c_j) \\ dist(q_i, c_{j-1}) \\ dist(q_{i-1}, c_{j-1}) \end{cases} \quad (1)$$

สมมติให้ Q และ C เป็นอนุกรมขนาดยาว m และ n ที่ต้องการทำการวัดระยะไดนามิกไทม์วอร์ปปิง โดยมี q_i และ c_j เป็นข้อมูลแต่ละจุดของอนุกรม Q และ C ตามลำดับ จากนั้นทำการสร้างเมทริกซ์ขนาด m x n และทำการคำนวณแบบพลวัต (ดังรูปที่ 4) โดยการสะสมค่าระยะระหว่างแต่ละข้อมูลของอนุกรมโดยใช้สมการที่ (1) ตั้งแต่ช่องแรก (ซ้ายล่าง) ไปจนถึงช่องสุดท้าย (ขวาบน) ซึ่งจะได้ค่าระยะไดนามิกไทม์วอร์ปปิง



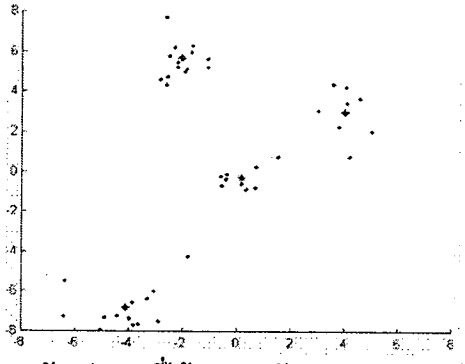
รูปที่ 4 การคำนวณระยะไดนามิกไทม์วอร์ปปิง

การจัดกลุ่มแบบเคมีนส์ (K-means Clustering)

การจัดกลุ่มแบบเคมีนส์ [4] เป็นการจัดกลุ่มข้อมูลที่นิยมนำมาใช้เนื่องจากอัลกอริทึมที่ทำงานรวดเร็วและไม่ซับซ้อน โดยมีขั้นตอนดังนี้

1) สุ่มข้อมูลเริ่มต้นขึ้นมา k ตัวเท่ากับจำนวนกลุ่มที่ต้องการแบ่งเพื่อเป็นตัวแทนของกลุ่ม
2) ทำการจัดกลุ่มข้อมูลโดยการวัดระยะระหว่างข้อมูลที่มีกับข้อมูลตัวแทนและจัดกลุ่มข้อมูลให้อยู่ในกลุ่มเดียวกับข้อมูลตัวแทนที่ใกล้ที่สุด
3) ทำการเฉลี่ยข้อมูลแต่ละกลุ่มเพื่อหาตัวแทนของกลุ่มตัวใหม่

จากนั้นทำซ้ำในขั้นตอนที่ 2 และ 3 ไปเรื่อย ๆ จนกว่าข้อมูลในแต่ละกลุ่มจะไม่มีการเปลี่ยนแปลง ซึ่งตัวอย่างผลการจัดกลุ่มจะเป็นดังรูปที่ 5



รูปที่ 5. ตัวอย่างผลที่ได้จากการจัดกลุ่มแบบเคมีนส์ [4]

งานวิจัยที่เกี่ยวข้อง

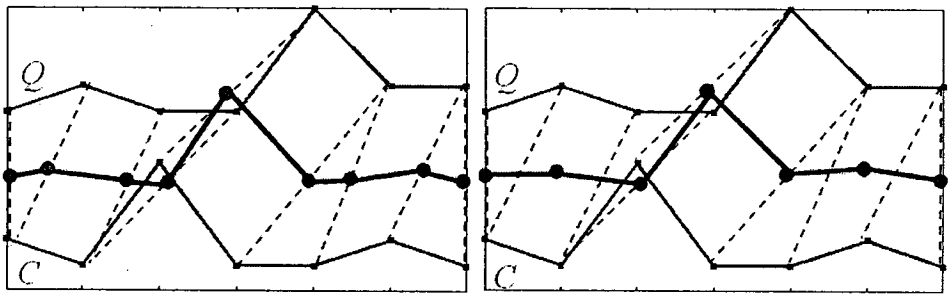
ในปี 1996 ได้มีการเสนออัลกอริทึมสำหรับการหาค่าเฉลี่ยของข้อมูลอนุกรมเวลาทีละสองอนุกรม โดยอาศัยระยะไดนามิกไทม์วอร์ปปิง คือ Non Linear Alignment and Averaging Filter (NLAAF) [5] ซึ่งทำการเฉลี่ยอนุกรมโดยการสุ่มอนุกรมขึ้นมาทีละสองอนุกรม จากนั้นทำการหาเส้นทางวอร์ปปิง (Warping Path) ซึ่งได้จากการคำนวณระยะไดนามิกไทม์วอร์ปปิง แล้วจึงทำการหาค่าเฉลี่ยของค่าในแต่ละตำแหน่งของคู่อันดับที่อยู่ในเส้นทางวอร์ปปิง ซึ่งวิธีนี้อาจทำให้อนุกรมที่ได้จากการเฉลี่ยมีขนาดที่ยาวกว่าอนุกรมที่นำมาเฉลี่ยได้ ต่อมาในปี 2009 มีการเสนอ Prioritized Shape Averaging (PSA) [6] ซึ่งเป็นการเฉลี่ยอนุกรมทีละสองอนุกรมตามลำดับที่ได้จากการจัดกลุ่มตามลำดับขั้น (Hierarchical Clustering) จากนั้นทำการเฉลี่ยโดยใช้ Scaled Dynamic Time Warping (SDTW) ซึ่งเป็นอัลกอริทึมที่ปรับปรุงเพิ่มเติมจาก NLAAF และทำการปรับความยาวของอนุกรม ให้มีขนาดเท่ากับความยาวของอนุกรมที่นำมาเฉลี่ยโดยใช้การยืดหดแบบยูนิฟอร์ม (Uniform Scaling)

ในปี 2010 มีการเสนอ Shape-based Template Matching Framework (STMF) [7] โดยการหาอนุกรมที่คล้ายกันมากที่สุดซึ่งวัดจากระยะไดนามิกไทม์วอร์ปปิงทีละสองอนุกรมจากนั้นทำการเฉลี่ยโดยใช้ Cubic-Spline Dynamic Time Warping (CDTW) ซึ่งเริ่มจากการหาเส้นทางวอร์ปปิง W และ w_i เป็นแต่ละจุดของ W โดยแต่ละ w_i จะประกอบด้วยคู่อันดับ (i, j) ซึ่งเก็บค่าตำแหน่งของอนุกรม Q และ C ที่นำมาเฉลี่ยตามลำดับ จากนั้นทำการเฉลี่ยอนุกรมทีละจุดตามคู่อันดับของเส้นทางวอร์ปปิงโดยใช้สมการที่ (2) และ (3)

$$z_k(x) = \frac{\omega_q q_i + \omega_c c_j}{\omega_q + \omega_c} \qquad (2)$$

$$z_k(y) = \frac{\omega_q q_{i_k} + \omega_c c_{j_k}}{\omega_q + \omega_c} \qquad (3)$$

ค่า ω_q และ ω_c เป็นค่าน้ำหนักของอนุกรม Q และ C ตามลำดับ เนื่องจากอนุกรมที่ได้จากการเฉลี่ย (ดังรูปที่ 6 ก.) จะมีจำนวนจุดที่มากกว่าอนุกรมที่นำมาเฉลี่ย จึงทำการปรับขนาดของอนุกรมแบบคิวบิคสไปลน์ (Cubic-Spline Interpolation) (ดังรูปที่ 6 ข.) ซึ่งวิธีนี้จะช่วยรักษารูปร่างของอนุกรมไว้ได้เมื่อเทียบกับการยืดหดแบบยูนิฟอร์ม (Uniform Scaling)

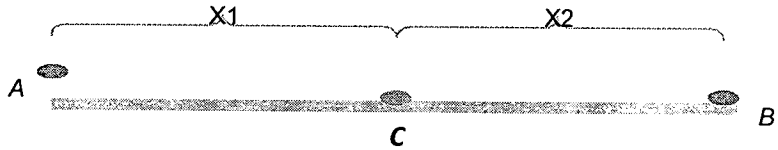


ก) ข)

รูปที่ 6 อนุกรมที่ได้จากการเฉลี่ย ก.) ก่อนการปรับความยาว ข.) หลังการปรับความยาว แบบคิวบิคสไปลน์ (Cubic-Spline interpolation)

โดยทั่วไปเมื่อทำการหาค่าเฉลี่ยของข้อมูลใด ๆ บนพื้นที่ของยุคลิด ข้อมูลเฉลี่ยจะต้องมีตำแหน่งตรงกลางระหว่างข้อมูลที่นำมาเฉลี่ย และในกรณีที่ทำการเฉลี่ยข้อมูลสองข้อมูล เมื่อคำนวณระยะระหว่างข้อมูลเฉลี่ยกับข้อมูลที่นำมาเฉลี่ยทั้งสองข้อมูลจะต้องได้ระยะที่เท่ากัน (ดังรูปที่ 7) และหากมีข้อมูลที่ต้องการเฉลี่ยมากกว่าสองข้อมูลผลรวมของระยะระหว่างข้อมูลเฉลี่ยกับข้อมูลที่นำมาเฉลี่ยจะมีค่าน้อยที่สุดเสมอ [8]

รูปที่ 7 ข้อมูลเฉลี่ย C ซึ่งได้จากการเฉลี่ยข้อมูล A และ B



บนพื้นที่ของยุคลิด ซึ่ง X1 = X2 เสมอ

อย่างไรก็ตามอนุกรมที่ได้จากการเฉลี่ยด้วยวิธี CDTW นั้นไม่ได้มีตำแหน่งตรงกลางระหว่างอนุกรมที่นำมาเฉลี่ยเมื่อวัดด้วยระยะไดนามิกไทม์วอร์ปปิง จึงมีการเสนอ Iterative Cubic-Spline Dynamic Time Warping (ICDTW) [7] ซึ่งเป็นอัลกอริทึมที่ปรับปรุงเพิ่มเติมจาก CDTW เพื่อให้อนุกรมที่ได้จากการเฉลี่ย มีตำแหน่งตรงกลางระหว่างอนุกรมทั้งสองที่นำมาเฉลี่ยเมื่อวัดด้วยระยะไดนามิกไทม์วอร์ปปิง

อัลกอริทึม ICDTW จะทำการหาอนุกรมเฉลี่ยโดยใช้ CDTW ก่อน จากนั้นนำอนุกรมเฉลี่ยที่ได้มาทำการปรับเพื่อให้มีตำแหน่งตรงกลางเมื่อคำนวณด้วยระยะไดนามิกไทม์วอร์ปปิง ในการปรับนั้นทำได้โดยสมมติให้ A และ B ซึ่งเป็นอนุกรมที่ต้องการนำมาเฉลี่ยมีค่าน้ำหนักเป็น β_A และ β_B ตามลำดับ กำหนดให้ค่า β_A เป็นหนึ่ง จากนั้นทำการปรับ β_B จนกระทั่งระยะไดนามิกไทม์วอร์ปปิงระหว่างอนุกรมเฉลี่ยและอนุกรมที่นำมาเฉลี่ยมีค่าเท่ากัน