

RÉPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique



Université 20 Août 1955-Skikda

Faculté des Sciences

Département d'Informatique

THESE

En vue de l'obtention du diplôme de :

Doctorat de 3^o cycle (LMD) en Informatique

Option: *Computation et Cognition des systèmes informatiques*

Présentée par :

Imen Boulnemour

Thème

Caractérisation des séries temporelles pour la recherche et la classification

Présentée Publiquement le : 14/02/2021 ; devant le Jury :

| | | | |
|----------------------|-------------|---------------------------------|--------------------|
| Pr Redjimi Mohamed, | Professeur, | Université 20 Août 1955-Skikda, | Président |
| Dr Brahimi Said, | MCA, | Université 08 Mai 1945-Guelma, | Examineur |
| Pr Mazouzi Smaïne, | Professeur, | Université 20 Août 1955-Skikda, | Examineur |
| Pr Boucheham Bachir, | Professeur, | Université 20 Août 1955-Skikda, | Directeur de thèse |

Remerciements

C'est avec beaucoup de plaisir et de gratitude que j'exprime ici ma reconnaissance aux personnes qui m'ont portées leurs aides, ne serait-ce que des encouragements, pour mener cette thèse à son terme.

Je voudrais tout d'abord remercier grandement mon directeur de thèse, Professeur Bachir Boucheham, Université 20 Août 1955 de Skikda, pour toute son aide. Je suis ravie d'avoir travaillé en sa compagnie, car, outre son appui scientifique, il a consacré de nombreuses heures à me soutenir et à me conseiller. Son investissement dans cette thèse a été mon principal moteur dans mes moments de doutes. C'est aussi lui qui m'a appris ce que rigueur et précision pouvaient apporter.

J'exprime aussi ma gratitude au Pr Redjimi Mohamed, Université 20 Août 1955 de Skikda d'avoir accepté de présider mon Jury, Dr Brahimi Said, Université 08 Mai 1945 de Guelma et Pr Mazouzi Smaine, Université 20 Août 1955 de Skikda pour l'honneur qu'ils m'ont fait d'être dans mon jury de thèse. Qu'ils trouvent tous ici l'expression de ma profonde gratitude.

Je tiens aussi à remercier tous les enseignants à l'université 20 août 1955 Skikda qui ont contribué à ma formation. Qu'ils trouvent ici l'expression mon respect éternel.

Je voudrais aussi étendre mes remerciements particuliers aux chercheurs à travers le monde qui ont montré, d'une façon ou d'une autre, de l'intéressement à certains de mes travaux publiés. Notamment, j'exprime mes sincères remerciements au Professeur Eamonn Keogh, spécialiste en Data Mining des Time Series à l'UCR (University of California at Riverside) qui m'a contactée par email à propos de mon travail sur la détection d'anomalies dans l'Electrocardiogramme [Boulnemour, 2016]. Mes remerciements s'étendent aussi aux chercheurs qui ont cité certains de mes travaux : Andreas Dengel, Professeur à l'université de Kaiserslautern et Directeur scientifique du Centre de recherche allemand pour l'intelligence artificielle, Sheraz Ahmed, Chercheur sénior, Clodoaldo Lima Professeur à l'université de São Paulo et à Xiaowei XU, Chercheur à l'université de Notre Dame à Paris.

Dédicaces

Je dédie cette thèse à mes chers parents. Vous serez toujours un modèle pour moi. Papa, pour ta détermination, ton intelligence et ton honnêteté. Maman pour ta générosité, ta force et ton dévouement pour nous. Vous n'avez jamais douté de moi et vous m'avez offert un cadre idéal pour mener à bien mes travaux. Je vous aime tant !

À mes chers frères et belles sœurs. Particulièrement toi, Yassine parce que, tu m'as appris beaucoup de choses et tu m'as encouragé à poursuivre mes études.

À tous mes oncles et tantes adorés, pour votre amour et votre soutien. Tonton Mouloud, tu es comme un deuxième papa pour moi! Tata Rachida, tu m'as toujours gâté, tout en m'apprenant à être forte et déterminée !

À mes chers cousins et cousines. Je vous aime tous sans exception.

Un grand merci à tous mes collègues et amis pour leurs soutiens et leurs gentillesse.

À mon cher mari et à toute ma belle famille.

À ma petite chipie adorée Malak Anaïs qui ne m'a jamais laissé finir ma thèse à temps mais que j'aime par-dessus tout au monde.

Résumé

L'objet de cette thèse concerne principalement : (a) La problématique de caractérisation des séries temporelles (ST) et (b) L'investigation du domaine obligatoirement lié à cette problématique : Les mesures de distances et les mesures de similarité entre ST, ainsi que les techniques d'appariement (ou alignement) des ST. Ces dernières techniques sont beaucoup plus robustes que les mesures de distance/similarité, mais, elles sont peu nombreuses et difficiles à inventer. La caractérisation des ST a pour but de décrire de façon compacte, mais pertinentes ce type de données (les ST). Compacte au sens de la réduction autant que possible de la redondance des données dans les ST, en n'extrayant que les caractéristiques effectivement discriminantes. Cela nécessite le plus souvent de faire muter les ST du domaine temporel à un domaine beaucoup plus expressif en termes de pouvoir de discrimination. Dans ce contexte, nous allons établir un état de l'art et une catégorisation des techniques de caractérisation des ST. Grossièrement parlant, nous pouvons déjà avancer que ces techniques peuvent être classées en quatre domaines: (1) Le domaine temporel tels que les histogrammes, les motifs et l'approche PLA (Piecewise Linear Approximation), (2) Le domaine fréquentiel, tel que la Transformée de Fourier Discrète (DFT), (3) Le domaine temps-fréquences, tel que la Transformée des Ondelettes Discrètes (DWT : Discrete Wavelet Transform) et (4) La transformée par décomposition en valeurs singulières (SVD : Singular Value Decomposition), qui relève du domaine de l'algèbre linéaire.

Pour le point d'intérêt (b), notre étude montre l'existence de quatre familles de techniques de comparaison des ST. Il y a d'abord (1) Les techniques de comparaison utilisant les données directes des ST combinées à des mesures de distance/similarité, telle que la distance Euclidienne, (2) Les techniques de comparaison basées caractéristiques combinées aussi à une mesure de distance / similarité, telle que la DFT et (3) Les technique d'alignement, telle que la célèbre et très pertinente méthode : DTW (Dynamic Time Warping). Enfin, (4) une autre classe de techniques d'alignement des ST basées signatures (caractéristiques globales) a été récemment proposée par Boucheham: la méthode SEA (Shape Exchange Algorithm). Son auteur la présente comme la rivale de la DTW ; mais, plus précise et plus pertinente.

Comme contributions personnelles dans le cadre de cette thèse, nous avons proposé quatre travaux dont deux ont été publiés au niveau international : (1) L'élaboration d'une technique d'alignement pour l'alignement des ST quasi-périodiques (STQP), telles que les tracés ECG (Electrocardiogramme) et les ST du Capnogram (Qualité de la respiration). Notre méthode, intitulée QPDTW (Quasi-Periodic Dynamic Time Warping), est une fusion entre la méthode DTW et la méthode SEA. L'intérêt de ce travail est que, selon les travaux de Boucheham, l'alignement des STQP serait le cas le plus complexe à résoudre. Brièvement, notre méthode QPDTW permet d'aligner de manière effective les STQP même lorsqu'elles sont déphasées et bruitées de manières significatives. (2) Développement d'un classifieur binaire basé QPDTW pour la détection d'anomalies dans les tracés ECG. Ce travail en particulier a eu un écho remarquable sur Researchgate. (3) Développement d'une méthode accélérée à base de la transformation en ondelettes de Haar et de la méthode QP-DTW pour la recherche de similarité dans l'ECG. (4) Développement d'un classifieur de type KNN basé QPDTW pour la classification des ST générales. La base utilisée dans ce cas étant l'UCR (University of California at Riverside) utilisée mondialement comme benchmark pour évaluation des techniques de classification et de clustering des TS. Les résultats sont prometteurs, que se soit pour la détection d'anomalies, la recherche ou pour la classification des séries temporelles.

Mots-Clés : Séries Temporelles, Caractérisation, Alignement, ECG, DTW, SEA, QP-DTW.

Abstract

This study is mainly concerned with: (a) Time Series (TS) Characterization problematic and (b) Investigation the, necessarily linked domain: Distance and similarity measures within the context of TS. This last point includes also the more complicated problem of TS alignment. These are also TS matching techniques, but known as being at the same time more effective but more hard to derive.

TS characterization aims to express in a compact, yet pertinent fashion TS. Compact means that data redundancy is reduced as much as it can, yet, the extracted features are as discriminatory as possible. To this end, TS are usually moved from the temporal domain to other more expressive domains in terms of power of discrimination. In this context, a state of the art and a categorization of such techniques will be established in this thesis. Generally speaking, we already can announce the existence of four classes of such methods. (1) Temporal domain techniques, such as Histogram, Motifs and PLA (Piece-Wise Linear Approximation) approaches, (2) Frequency domain techniques, such as DFT (Discrete Fourier Transform) and DCT (Discrete Cosine Transform), (3) Time-Frequency techniques, such as DWT (Discrete wavelet Transform) under its various variants and (4) SVD: Singular Value Decomposition techniques that belongs to the algebraic domain.

For aim (b), the study shows existence of four families of such techniques: (1) Techniques that use directly data combined with some kind of distance/similarity, such as the Euclidian distance, (2) Feature-based techniques that use some kind of extracted features, like Fourier coefficients, combined to some kind of distance/similarity measure, (3) Alignment techniques, such as the famous and very appreciated DTW (Dynamic Time Warping) technique. Recently, a new TS alignment technique has been proposed by Boucheham known as SEA (Shape Exchange Algorithm) as a rival to DTW and the sole capable of aligning QPTS (Quasi-Periodic Time Series). SEA is rather based on the novel paradigm of Signature Exchange between the two TS to align.

As personal contributions within the context of this thesis, we have proposed four works, two of which published internationally: (1) Elaboration of a novel TS alignment technique that is capable also to align QPTS, such as ECG (Electrocardiogram) and Capnogram (Quality of respiration signal) traces. Our method is denominated QPDTW for it is a fusion of the two TS rivals methods: DTW and SEA of Boucheham. According to Boucheham, aligning QPTS represents the most complex situation in the context of TS alignment techniques. Briefly, QPDTW can effectively align even significantly noisy and also phase-shifted TS for instance. (2) Development of a binary QPDTW-based classifier for anomaly detection in ECG traces. Perhaps it is worth to mention that this work attracted great attention on Researchgate. (3) Development of a Haar-Transform based accelerated version of QPDTW for similarity search in ECG TS. (4) Development of QPDTW-based, KNN-type of general TS classifier. This classifier was tested on the UCR (University of California at Riverside), universally admitted as the reference Benchmark for TS classification and Clustering.

Briefly, obtained results for QPDTW, Anomaly detection in ECG traces, TS Search by similarity and Classification are promising.

Keywords: Time Series, Characterization, Alignment, ECG, DTW, SEA, QP-DTW.

ملخص :

يتعلق موضوع هذه الأطروحة أساسًا بما يلي: (أ) مشكلة توصيف السلاسل الزمنية (ST) و (ب) البحث و التوسع في المجال المرتبط بالضرورة بهذه المشكلة: مقاييس المسافات و مقاييس التشابه بين ST، وكذلك تقنيات مطابقة (أو محاذاة) ST. هذه التقنيات الأخيرة أقوى بكثير من مقاييس المسافات / التشابه، لكنها قليلة ويصعب اختراعها. الغرض من وصف ST هو وصف هذا النوع من البيانات (ST) بطريقة مدمجة ولكن ذات صلة. مدمجة بمعنى تقليل التكرار قدر الإمكان للبيانات في ST ، من خلال استخراج الخصائص المميزة بشكل فعال فقط. يتطلب هذا عادةً نقل ST من المجال الزمني إلى مجال أكثر تعبيرًا من حيث قوة التمييز. في هذا السياق ، سنؤسس استعراضاً للأدبيات وتصنيف تقنيات توصيف ST. بمعنى عمومي ، يمكننا القول بأن هذه التقنيات يمكن تصنيفها إلى أربعة مجالات: (1) المجال الزمني مثل الرسوم البيانية (Histogram) ونهج (Piece-wise Linear Approximation) PLA ، (2) مجال التردد ، مثل تحويل فورييه المنفصل (DFT(Discrete Fourier Transform)) ، (3) مجال التردد الزمني ، مثل تحويل الموجات المنفصلة (DWT(Discrete wavelet Transform)) و (4) و مجال الجبر الخطي مثل، تحليل القيمة الفردية (SVD (Singular Value Decomposition)).

بالنسبة إلى نقطة الاهتمام (ب)، توضح دراستنا وجود أربع مجموعات من التقنيات. أولاً ، هناك (1) تقنيات مقارنة تستخدم بيانات ST مباشرة مع مقاييس المسافة / التشابه ، مثل المسافة الإقليدية ، (2) تقنيات المقارنة القائمة على الخصائص مع مقاييس المسافة / التشابه ، مثل DFT و (3) تقنيات المحاذاة ، مثل الطريقة الشهيرة: (DTW (Dynamic Time Warping)). أخيرًا ، (4) تم اقتراح فئة أخرى من تقنيات محاذاة ST القائمة على البصمة (الخصائص العامة) من قبل بوالشحم : طريقة SEA (خوارزمية تبادل الأشكال). يقدمها مؤلفها كمنافسة DTW. ولكن ، بأكثر دقة وأكثر صلة. كمساهمات شخصية في هذه الأطروحة، اقترحنا أربعة أعمال، اثنان منها قد تم نشرهما دولياً: (1) تطوير تقنية محاذاة لمواءمة ST شبه الدورية (STQP) ، مثل تخطيط القلب (الكهربائي) ECG وأثار Capnogram (جودة التنفس). تعتبر طريقتنا، التي تحمل اسم (Quasi-Periodic Dynamic Time Warping) QPDTW، عبارة عن اندماج بين طريقة DTW وطريقة SEA . فائدة هذا العمل هي أنه وفقاً لبوالشحم، فإن محاذاة STQP ستكون القضية الأكثر تعقيداً التي يجب حلها. باختصار ، تعمل طريقة QPDTW الخاصة بنا على محاذاة STQP بشكل فعال حتى عندما تكون خارج المرحلة ومشوهة بصفة كبيرة. (2) تطوير QPDTW على أساس مصنف ثنائي للكشف عن الحالات الشاذة في أثار ECG. كان لهذا العمل الخاص صدق ملحوظ على Researchgate. (3) تطوير طريقة تسريع تعتمد على تحويل الموجات Haar وطريقة QP-DTW للبحث عن التشابه في ECG. (4) تطوير مصنف KNN على أساس QPDTW لتصنيف ST العامة. القاعدة المستخدمة في هذه الحالة هي قاعدة UCR (جامعة كاليفورنيا في ريفرسايد) المستخدمة في جميع أنحاء العالم كمعيار لتقييم تقنيات تصنيف وتكتل TS. النتائج واعدة، سواء للكشف عن الحالات الشاذة، البحث أو تصنيف السلاسل الزمنية.

الكلمات المفتاحية: التوصيف، المطابقة، المتتاليات الزمنية، 'DTW' 'SEA' 'QP-DTW'.

Sommaire

| | |
|--|----|
| Chapitre I. Introduction Générale..... | 14 |
| I.1 Contexte du travail | 14 |
| I.2 Problématique..... | 14 |
| I.3 Objectifs de la thèse | 15 |
| I.4 Contributions de la thèse | 15 |
| I.5 Organisation de la thèse | 17 |
| Chapitre II. Etat de l'art des techniques de caractérisation des ST | 18 |
| II.1 Introduction | 18 |
| II.2 Présentation des séries temporelles | 18 |
| II.2.1 Définition formelle d'une série temporelle | 18 |
| II.2.2 Définition d'une séquence temporelle | 19 |
| II.2.3 Domaines d'applications des séries temporelles | 19 |
| II.3 Techniques de caractérisation des séries temporelles | 19 |
| II.3.1 Domaine temporel | 19 |
| II.3.2 Domaine fréquentiel | 32 |
| II.3.3 Domaine temps-fréquence (Ondelettes, Wavelets) | 34 |
| II.3.4 Domaine algébrique..... | 36 |
| II.4 Discussion et conclusion | 37 |
| Chapitre III. Etat de l'art des Techniques de comparaison et d'alignement des ST | 39 |
| III.1 Introduction | 39 |
| III.2 Difficultés associées à l'appariement des séries temporelles..... | 39 |
| III.2.1 Décalages et différence d'échelles entre les séries temporelles..... | 39 |
| III.2.2 Le Bruit | 40 |
| III.2.3 Valeurs aberrantes..... | 41 |
| III.3 Mesures de similarités utilisées pour l'appariement des ST | 41 |
| III.3.1 Techniques de comparaison basées mesures de Distances/Similarités sur les données brutes | 41 |
| III.3.2 Techniques de comparaison basées caractéristiques..... | 45 |
| III.3.3 Techniques d'alignement des ST | 46 |
| III.4 Classification des méthodes d'appariement selon la nature des données | 53 |
| III.5 Techniques d'alignement avancées | 55 |
| III.5.1 Les améliorations de la méthode DTW..... | 55 |
| III.5.2 Améliorations de la méthode SEA :..... | 59 |
| III.6 Conclusion..... | 63 |
| Chapitre IV. Recherche de similarité dans les séries temporelles | 64 |
| IV.1 Introduction | 64 |
| IV.2 Recherche de similarité dans les séries temporelles..... | 64 |
| IV.3 Problème de la recherche de similarité dans les séries temporelles | 65 |
| IV.3.1 Techniques d'optimisation des indexes des BDD des séries temporelles .. | 65 |
| IV.4 Les différents types de requêtes de similarité | 69 |
| IV.4.1 Requêtes de plages de similarité (Range Queries)..... | 69 |

| | | |
|---------------------|--|-----|
| IV.4.2 | Requêtes de similarité des k-plus proches voisins (KNN Query)..... | 70 |
| IV.4.3 | Requête de similarité de jointure | 70 |
| IV.5 | Applications de la recherche de similarité dans les séries temporelles | 71 |
| IV.6 | Conclusion..... | 72 |
| Chapitre V. | Classification des séries temporelles | 73 |
| V.1 | Introduction | 73 |
| V.2 | Méthodes de classification supervisée des séries temporelles | 73 |
| V.2.1 | Méthodes de classification basées sur l'extraction de caractéristiques | 74 |
| V.2.2 | Méthodes de classification basées sur des mesures de similarité..... | 78 |
| V.2.3 | Méthodes de classification basées sur les modèles | 80 |
| V.3 | Conclusion..... | 80 |
| Chapitre VI. | QPDTW : Technique d'alignement des STQP basée DTW et SEA | 82 |
| VI.1 | Introduction | 82 |
| VI.2 | Méthode d'alignement proposée | 82 |
| VI.3 | Tests expérimentaux et discussions..... | 86 |
| VI.3.1 | Comparaison des tracés similaires | 90 |
| VI.3.2 | Comparaison de séries temporelles dissimilaires | 97 |
| VI.4 | Conclusion..... | 101 |
| Chapitre VII. | Méthode proposée pour la détection d'anomalies dans l'ECG | 102 |
| VII.1 | Introduction | 102 |
| VII.2 | ECG normal et ECG Anormal..... | 102 |
| VII.3 | Approches existantes | 104 |
| VII.4 | Méthode proposée..... | 105 |
| VII.5 | Tests expérimentaux | 106 |
| VII.6 | Conclusion | 110 |
| Chapitre VIII. | Méthode proposée A-QPDTW pour la recherche de similarité dans les ST...111 | |
| VIII.1 | Introduction | 111 |
| VIII.2 | Méthode proposée : AQP-DTW | 112 |
| VIII.3 | Expérimentation..... | 113 |
| VIII.4 | Conclusion | 115 |
| Chapitre IX. | Méthode proposée basée 1NN-QPDTW pour la classification des séries temporelles | 116 |
| IX.1 | Introduction | 116 |
| IX.2 | Méthode proposée | 116 |
| IX.3 | Expérimentation | 117 |
| IX.4 | Résultats et discussions | 119 |
| IX.4.1 | Comparaison des techniques d'alignement DTW, QP-DTW et SEA..... | 120 |
| IX.4.2 | Comparaison de 1NN-QP-DTW à d'autres classifieurs basés 1NN..... | 124 |
| IX.5 | Conclusion..... | 128 |
| Conclusion Générale | | 129 |
| Bibliographie | | 130 |

Table des Figures

| | |
|---|----|
| Figure1 : Consommation d'électricité sur toute la France | 20 |
| Figure2 : Nombre de passagers dans les transports aériens..... | 21 |
| Figure3 : Histogramme à une échelle | 21 |
| Figure4 : Algorithme de construction des histogrammes des séries temporelles..... | 22 |
| Figure5 : Deux séries temporelles différentes avec le même histogramme..... | 22 |
| Figure6 : Histogrammes à deux échelles pour les séries temporelles TA et TB..... | 23 |
| Figure7 : Exemple d'approximation d'une série temporelle avec la méthode PAA..... | 24 |
| Figure8 : Exemple d'approximation d'une série temporelle avec la méthode PLA..... | 26 |
| Figure9 : Etapes d'approximation d'une série temporelle avec la méthode SAX..... | 27 |
| Figure10 : Échantillons de feuilles de deux espèces | 28 |
| Figure11 : Une forme peut être convertie en une "série temporelle" unidimensionnelle | 29 |
| Figure12 : Le shapelet dont il est question dans la figure 10 | 29 |
| Figure13 : le classifieur d'arbre de décision pour le problème de la feuille..... | 29 |
| Figure14 : Exemple d'une ST et le graphique associé dérivé de l'algorithme de visibilité..... | 31 |
| Figure15 : Appariement point par point et signatures de X et Y..... | 32 |
| Figure16 : Difficultés associées à l'appariement des séries temporelles | 40 |
| Figure17 : Séries temporelles provenant du Benchmark UCR (Beef) normale v.s bruitée.... | 40 |
| Figure18 : ST bleue normale et ST rouge contenant une valeur aberrante à l'instant 5..... | 41 |
| Figure 19 : Comparaison de deux séries temporelles similaires, mais décalées dans le temps avec la distance Euclidienne | 43 |
| Figure20 : Illustration de l'alignement effectuée par DTW | 47 |
| Figure 21: Chemin de déformation optimal de la DTW | 48 |
| Figure 22 : La matrice des distances cumulées entre les séquences X et Y | 49 |
| Figure 23 : Comparaison de deux séries temporelles à l'aide de la méthode LCSS | 51 |
| Figure 24 : Difficultés au niveau de l'appariement des séries temporelles quasi-périodiques. | 55 |
| Figure 25: Classification des méthodes d'appariement des ST selon la nature des données... | 55 |
| Figure 26: Illustration de la technique Early Abandon Dynamic Time Warping (EA_DTW). | 56 |
| Figure 27 : Illustration du problème de singularité de la DTW..... | 57 |
| Figure28: Création d'une série temporelle à partir de l'image du squelette d'un crâne humain avec le balayage radial..... | 58 |
| Figure29 : Les quatre résolutions évaluées durant l'exécution complète de Fast-DTW..... | 59 |

| | |
|---|----|
| Figure30: Forme du signal quasi-périodique du Capnographe | 60 |
| Figure 31 : Série temporelle compressée par la méthode FAN | 61 |
| Figure 32 : Une intuition visuelle de la LB_Yi..... | 67 |
| Figure 33 : Une intuition visuelle de la LB_Kim | 67 |
| Figure 34 : Une intuition visuelle de la LB_Keogh..... | 68 |
| Figure 35 : Les deux contraintes de chemins optimale les plus populaires, utilisées dans l'implémentation de DTW..... | 68 |
| Figure 36 : Plage de requêtes de similarité (Range Queries)..... | 69 |
| Figure 37 : Requête de similarité des k- plus proches voisins. K= 4..... | 70 |
| Figure 38 : Groupement hiérarchique de huit crânes de primates basés sur la vue latérale, en utilisant la distance euclidienne..... | 72 |
| Figure 39 : Arbre de décision pour étudier la possibilité d'aller à la plage | 77 |
| Figure 40 : Illustration de la méthode KNN..... | 79 |
| Figure 41 : Un segment d'ECG typique avec trois périodes (cycles)..... | 82 |
| Figure 42 : Illustration du problème de déphasage significatif de DTW..... | 83 |
| Figure 43 : Illustration de l'efficacité d'alignement de la méthode SEA par rapport A la méthode DTW..... | 83 |
| Figure 44 : Illustration du problème de bruit que peut rencontrer la méthode SEA..... | 84 |
| Figure 45 : Diagramme illustratif de la méthode d'alignement proposée QP-DTW..... | 86 |
| Figure 46 : Signatures des tracés X et Y établies par les méthodes SEA (a) et QP-DTW..... | 87 |
| Figure 47: Comparaison des méthodes SEA, QP-DTW et DTW pour le cas de séries temporelles quasi périodiques à plusieurs périodes avec déphasage | 89 |
| Figure 48 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (113,113) avec un décalage de 30 secondes pour Y..... | 92 |
| Figure 49 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (215,215) avec un décalage de 35 secondes pour Y..... | 93 |
| Figure 50 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (104,104) avec un décalage de 20 secondes pour Y..... | 94 |
| Figure 51 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (118,118) avec un décalage de 35 secondes pour Y..... | 95 |
| Figure 52 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (115,115) avec un décalage de 25 secondes pour X..... | 96 |

| | |
|---|-----|
| Figure 53 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus de deux enregistrements ECG différents (103,123) avec un décalage de 40 secondes pour Y..... | 98 |
| Figure 54 : Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus de deux enregistrements ECG différents (100,112) avec un décalage de 50 secondes pour Y..... | 99 |
| Figure 55 : Facteur de corrélation en fonction de la taille de la série temporelle rec.118 (MITBIH) pour les méthodes DTW, SEA et QP-DTW | 100 |
| Figure 56 : Facteur de PRD en fonction de la taille de la série temporelle rec.118 (MITBIH) pour les méthodes DTW, SEA et QP-DTW | 101 |
| Figure 57 : Tracé d'un cycle ECG normal | 103 |
| Figure 58 : Extrasystole ventriculaire (ESV)..... | 103 |
| Figure 59 : Fibrillation atriale (Tracé Haut) et ECG normal..... | 104 |
| Figure 60 : Tachycardie supra-ventriculaire paroxystique..... | 104 |
| Figure 61 : Détection des changements morphologiques en comparant deux segments ECG de la même personne (contractions auriculaires prématurées)..... | 107 |
| Figure 62 : Détection d'un changement morphologique en comparant deux segments ECG de la même personne (contraction ventriculaire prématurée)..... | 107 |
| Figure 63 : Détection de changements morphologiques en comparant deux segments ECG de la même personne (QRS isolé)..... | 108 |
| Figure 64 : Détection de changements morphologiques en comparant deux segments ECG de la même personne (Bloc de branche)..... | 108 |
| Figure 65 : Détection de changements morphologiques en comparant deux segments ECG de la même personne (contraction ventriculaire prématurée)..... | 109 |
| Figure 66 : Diagramme illustratif de la méthode AQP-DTW proposée pour la recherche de similarité dans les ST..... | 112 |
| Figure 67 : Temps d'exécution de la méthode A-QPDTW par rapport à QP-DTW..... | 114 |
| Figure 68 : Précision de la méthode AQP-DTW. La ST en rouge représente l'objet requête et la ST bleue représente les 20 objets retournées qui sont similaires à la requête..... | 115 |
| Figure 69 : Algorithme de classification des séries temporelles "One Nearest Neighbor Quasi Periodic Dynamic Time Warping (1NN-QPDTW)"..... | 117 |
| Figure 70 : Comparaison de la précision de DTW et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN..... | 122 |
| Figure 71 : Comparaison de la précision de SEA et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN..... | 122 |

| | |
|---|-----|
| Figure 72 : Comparaison du taux d'erreurs des méthodes SEA et QP-DTW pour la classification de 85 ensembles de données de l'UCR en utilisant le classifieur 1-NN..... | 123 |
| Figure 73 : Comparaison du taux d'erreurs des méthodes SEA et QP-DTW pour la classification de 85 ensembles de données de l'UCR en utilisant le classifieur 1-NN..... | 123 |
| Figure 74 : Comparaison du taux d'erreurs des mesures de distances/similarité avec QP-DTW pour la classification de 85 ensembles de données de l'UCR avec le classifieur 1NN..... | 127 |

Liste des tableaux

| | |
|---|-----|
| Tableau1 : Les étapes de décomposition de la méthode DWT..... | 35 |
| Tableau2 : Techniques de caractérisation des séries temporelles..... | 38 |
| Tableau3 : Comparaison de QP-DTW avec DTW et SEA pour la série temporelle rec.215... | 90 |
| Tableau4.a : Comparaison des méthodes DTW, SEA et QP-DTW pour les ST similaires..... | 97 |
| Tableau4.b : Comparaison de DTW, SEA et QP-DTW pour les ST dissimilaires..... | 100 |
| Tableau 5 : Résultats de la méthode QP-DTW pour la détection d'anomalies dans l'ECG.. | 109 |
| Tableau 6 : Description de la base de données ECG utilisée..... | 113 |
| Tableau 7 : Précision de A-QPDTW V.S QP-DTW..... | 115 |
| Table8.a : Ensembles des séries temporelles..... | 118 |
| Table8.b : Ensembles des séries temporelles..... | 119 |
| Table9 : Comparaison de QP-DTW avec DTW et SEA pour la classification 1NN | 121 |
| Table10.a : Comparaison des techniques de comparaison avec QP-DTW en utilisant 1NN. | 125 |
| Table10.b : Comparaison des techniques d'alignement avec QP-DTW en utilisant 1NN.... | 126 |

Chapitre I. Introduction Générale

I.1 Contexte du travail

Une série temporelle est une collection de données réelles ordonnées chronologiquement ou logiquement. Elles sont le reflet d'activités et de phénomènes de diverses natures; naturelles (le signal ECG ou EEG, etc.), économiques (valeurs des échanges de stocks boursiers, prix des biens, etc.), technologiques (valeurs renvoyées par des senseurs, densité du trafic dans un réseau informatique, etc.).

Les séries temporelles sont des données hiérarchiques. Une série temporelle peut être analysée par sa hiérarchie temporelle, par exemple horaire, hebdomadaire, mensuelle ou annuelle. Elles peuvent aussi être de natures multivariées. Par exemple, les données météorologiques comprennent des mesures bien connues telles que la température, la pression, l'humidité, etc.

Les séries temporelles sont présentes dans de nombreuses applications telles que la reconnaissance biométrique des chiffres manuscrits [Serfass, 2017], la requête de tranches de musique [Tavenard, 2005], la détection d'anomalies ou de nouveautés dans les systèmes physiologiques [Boucheham, 2011] [Boulnemour, 2016], le clustering et la compression des séries temporelles [Aghabozorgi, 2014], la reconnaissance vocale [Graves, 2013], la classification des arythmies cardiaques [Singha, 2018], L'alignement et la classification des expressions génétiques [Smith, 2009], la reconnaissance du visage [Qi, 2018], etc.

Dans ce cadre contexte, nous nous intéressons à des points importants relatifs à ce type de données, essentiellement, le duo (a) caractérisation des ST et (b) mesures de distances/similarités et techniques d'alignement. Des cas d'études comprenant (a) la recherche et (b) la classification des ST sont par la suite envisagés.

I.2 Problématique

Comme mentionné ci-haut, la problématique principale envisagée ici est double. D'une part, nous nous intéressons aux techniques de caractérisation des ST. D'autre part, nous nous

intéressons aussi aux mesures de similarité/distances et aux techniques d'alignement des ST. Dans le cadre de l'exploitation et fouille des ST, ces deux points, en plus d'être essentiels, sont intimement liés, voir inséparables. En effet, l'application des mesures de distance/similarité fait généralement suite à l'étape de caractérisation pour beaucoup, pour ne pas dire toutes, les tâches d'exploration et fouille des ST. Notamment, la phase de caractérisation permet d'extraire les caractéristiques (dites aussi attributs) discriminantes dans les séries temporelles. L'objectif ici est aussi double. (a) d'une part, cela permettra de réduire leur dimensionnalité et, d'autre part, (b) permettra d'améliorer l'efficacité et la performante des différentes tâches associées aux séries temporelles, entre autres, (a) la recherche et (b) la classification des séries temporelles, tâches qui seront considérées dans cette étude.

I.3 Objectifs de la thèse

Pour résumer, deux grands objectifs sont tracés dans cette thèse :

1. Réalisation d'études critiques des techniques existantes (état de l'art) en matière de caractérisation, comparaison/appariement, recherche et classification des séries temporelles.
2. Suite à l'étude théorique, nous proposons des contributions dans le cadre de la problématique traitée.

I.4 Contributions de la thèse

Dans le cadre de cette thèse, nous avons réalisé les objectifs tracés suivants :

1. L'étude théorique et critique a fait émerger les techniques les plus significatives suivantes en matière de :
 - a. Caractérisation des ST: Nous pouvons classer les techniques de caractérisation en quatre domaines : *Le domaine temporel* tels que les histogrammes [Chen, 2003], les motifs [Agrawal, 1995-b] et l'approche PLA (Piecewise Linear Approximation) [Boucheham, 2010], *le domaine fréquentiel*, tel que la Transformée de Fourier Discrète (DFT) [Agrawal, 1993], *le domaine temps-fréquences*, tel que la Transformée des Ondelettes Discrètes (DWT : Discrete Wavelet Transform) [Chan, 1999] et la transformée par décomposition en valeurs singulières (SVD : Singular Value Decomposition) [Wu, 1996], qui relève du *domaine de l'algèbre linéaire*.
 - b. Appariement des ST: Pour ce point, notre étude montre l'existence de quatre familles de techniques de comparaison des ST. Il y a d'abord (1) Les techniques de comparaison

utilisant les données directes des ST combinées à des mesures de distance/similarité, telle que la distance Euclidienne, (2) Les techniques de comparaison basées caractéristiques combinées aussi à une mesure de distance / similarité, telle que la DFT [Agrawal, 1993-a] et (3) Les technique d'alignement, telle que la célèbre et très pertinente méthode : DTW (Dynamic Time Warping) [Sakoe, 1978], [Berndt 1994]. Enfin, (4) une autre classe de techniques d'alignement des ST basées signatures (caractéristiques globales) a été récemment proposée par Boucheham [Boucheham, 2008] : la méthode SEA (Shape Exchange Algorithm). Son auteur la présente comme la rivale de la DTW ; mais, plus précise et plus pertinente. Cette technique est présentée par son auteur comme étant, avec ses dérivées ASEAL [Boucheham, 2010] et FANSEA [Boucheham, 2013], les seuls capables d'aligner les séries temporelles quasi-périodiques, telles que les tracés de l'électrocardiogramme (activité électrique cardiaque, ECG) et du Capnogram (Signal de mesure de la qualité de respiration, notamment chez les asthmatiques et les personnes en détresse de respiration suite à l'inhalation de fumée et autres). De plus, l'auteur avance que l'alignement de ces séries serait le cas le plus complexe à résoudre [Boucheham, 2013]. La difficulté de résolution de ce cas particulier est récompensée par l'avantage exclusif de cette approche à pouvoir localiser dans le temps les évènements. En effet, que se soit pour la DTW, SEA ou ses dérivées, ces méthodes comparent deux ST effectuant un appariement points-à-points. Ce qui permet de localiser dans le temps les points, en particulier, de similarité quasi-parfaite des points dits de discorde.

2. Dans la partie contributions personnelles, nous avons proposé quatre travaux dont deux ont été publiés au niveau international.

- (a) L'élaboration d'une technique d'alignement basée caractéristique pour l'alignement des séries temporelles quasi périodiques. Notre méthode proposée «QP-DTW» (Quasi-Periodic Dynamic Time Warping) [Boulnemour, 2015], [Boulnemour, 2018], est une fusion entre la méthode DTW et la méthode SEA. Elle a été comparée avec les méthodes SEA et DTW sur la base des ECG de l'MIT-BIH [MITBIH, 2000]. Les résultats montrent que QP-DTW effectue de très bons alignements qui sont meilleurs que ceux de SEA et ceux de DTW.
- (b) Développement d'un classifieur binaire basé QP-DTW pour la détection d'anomalies dans les tracés ECG [Boulnemour, 2016]. Notre classifieur est capable de détecter les anomalies en alignant un tracé ECG normal et un tracé anormal sans avoir besoin de segmenter l'ECG.

- (c) Développement d'une version accélérée de la méthode QP-DTW à base de la technique de caractérisation « Haar Wavelets Transform » et de la méthode QP-DTW pour rechercher la similarité à l'aide de l'approche KNN dans les tracés ECG de l'UCR (Université de Californie à Riverside) [Chen, 2015], utilisée mondialement comme benchmark pour la recherche et la classification des ST. Les résultats obtenus montrent que la méthode proposée est environ trois fois plus rapide que QP-DTW, sans pratiquement aucune perte de précision.
- (d) Un classifieur de type KNN basé QP-DTW pour la classification des ST générales. Les résultats basés sur les expériences effectuées sur la base UCR comprenant des séries temporelles réelles provenant de différents domaines d'application montrent que notre méthode 1NN-QPDTW est une méthode prometteuse et qu'elle est en moyenne meilleure que les méthodes 1NN-DTW et 1NN-SEA en termes de qualité de classification.

I.5 Organisation de la thèse

Le reste de ce manuscrit se déroule comme suit :

Dans le deuxième chapitre, nous faisons une étude exhaustive des techniques existantes de caractérisation des ST. Dans le troisième chapitre, nous dresserons l'état de l'art des techniques d'appariement des séries temporelles. Dans le quatrième chapitre, nous allons présenter le problème de la recherche de similarité dans les séries temporelles, ainsi que les différents types de requêtes utilisées dans ce problème. Dans le cinquième chapitre, nous effectuons une étude exhaustive des méthodes de classification des séries temporelles. Dans le sixième chapitre, nous présentons notre méthode QP-DTW pour l'appariement des séries temporelles. Dans le septième chapitre, nous présentons notre contribution dans le cadre de la classification binaire basée QPDTW pour la détection d'anomalies dans les tracés ECG. Dans le huitième chapitre, nous présentons notre contribution dans le cadre de la caractérisation et de la recherche dans les ST, en l'occurrence, une technique de recherche de similarité dans les ST de type plus proche voisin (PPV, KNN) basée Haar-QPDTW. Dans le neuvième chapitre, nous présentons notre dernière contribution dans le cadre de la classification des séries temporelles générales. Nous proposons d'utiliser notre méthode QP-DTW comme mesure de similarité pour le classifieur 1NN. Ce travail se termine par une conclusion générale dans laquelle nous résumons nos contributions et présentons nos différentes perspectives.

Chapitre II. Etat de l'art des techniques de caractérisation des ST

II.1 Introduction

Dans ce chapitre, nous allons dans un premier temps faire une présentation des séries temporelles (ST) et des différents domaines d'application des ST.

Dans un deuxième temps, nous allons présenter les techniques de caractérisation des ST qui, comme déjà avancé, sont aussi utilisées pour extraire les caractéristiques pertinentes en termes de pouvoir de discrimination dans les ST. Du coup, deux avantages sont comptabilisés, suite à cette opération : (a) Réduction de la dimensionnalité des données et (b) amélioration de l'efficacité et la performante des différentes tâches associées aux ST. Particulièrement dans cette thèse, nous visons les tâches de recherche de similarité et de classification des ST.

II.2 Présentation des séries temporelles

II.2.1 Définition formelle d'une série temporelle

Une série temporelle ou time series est une collection de données obtenues de manière séquentielle au cours du temps. Plus formellement, une série temporelle est une suite de couples $\langle (v_1, t_1), (v_2, t_2), \dots, (v_i, t_i), \dots \rangle$ où v_i est une valeur ou un vecteur de valeurs prises à un instant t_i . Cette notation peut être abrégée en $\langle v_1, v_2, \dots, v_i, \dots \rangle$ lorsque la référence au temps n'a pas besoin d'être explicitée. Ceci n'est possible que lorsque la fréquence d'échantillonnage est constante, soit $f = \text{Nombre d'Echantillons}/UT$, UT étant l'intervalle de temps entre deux mesures et s'exprime par seconde, minute, ..., jour, mois, année, etc. A noter que les valeurs v_i d'une séquence sont souvent des nombres réels. Mais, par extension de cette notion, le type de la valeur peut s'étendre à d'autres types : caractère (Texte ou ADN), Image (Vidéo), etc.

En particulier aussi, lorsqu'on réalise la saisie simultanée sur plusieurs capteurs, par exemple, de même, ou de différents types, on a naturellement des séquences multidimensionnelles ou multi-variées, v_i étant, dans ce cas, un vecteur à k dimensions [Agrawal, 1993].

Pour le cas des séries usuelles, la plupart du temps, il est bien utile de représenter la série temporelle sur un graphe construit de la manière suivante : En abscisse le temps, en ordonnée

la valeur de l'observation à chaque instant. Pour des questions de lisibilité, les points ainsi obtenus sont reliés par des segments de droite. Le graphe apparaît donc comme une ligne brisée [Viano, 2004]. La Figure 1 illustre ce principe.

II.2.2 Définition d'une séquence temporelle

Une séquence temporelle est une partie d'une série temporelle observée pendant un intervalle de temps. On l'utilise, entre autres, pour la recherche en comparant une (ou des) séquence(s) temporelle(s) avec une (ou des) séquence(s) de référence.

II.2.3 Domaines d'applications des séries temporelles

Les séries temporelles sont intensivement utilisées en finances et en économie où l'évolution des indices boursiers et des prix des données économiques des compagnies est vitale [Agrawal, 1993, 1995], [Faloutsos, 1994]. Les ST sont aussi intensivement utilisées dans l'analyse des tendances météorologiques [Egondi, 2012], en traitement du signal (exemple : signaux de communications, parole) [Lindgren, 2003], en médecine, pour le suivi des évolutions des pathologies et, en particulier, pour l'analyse d'électro-encéphalogrammes (signal de l'activité cérébrale, EEG) et de l'électrocardiogramme (signal de l'activité électrique cardiaque, ECG) [Ye, 2012].

II.3 Techniques de caractérisation des séries temporelles

La revue de la littérature montre qu'en général, les techniques de caractérisation relèvent de quatre domaines suivants : Le domaine temporel, le domaine fréquentiel, le domaine temps-fréquence et, le domaine algébrique. Ces domaines sont présentés comme suit :

II.3.1 Domaine temporel

Dans le domaine temporel, nous recensons en premier lieu les approches statistiques telles que les histogrammes, le facteur de corrélation, etc. Nous avons aussi les approches dites géométriques, telles que : PAA (Piecewise Aggregate Approximation), PLA (Piecewise Linear Approximation) et SAX [Symbolic Aggregate Approximation]. Nous recensons aussi

dans cette catégorie les approches structurelles, telles que les Motifs, les graphes, les shapelets, les bornes inférieurs (Lower Bound) et les approches basées sur l'extraction de signatures. Toutes ces approches sont présentées et référencées dans ce qui suit.

II.3.1.1 Approches statistiques

II.3.1.1.a. Facteur de corrélation

Une caractéristique intéressante qui peut être utile pour décrire les séries temporelles est la valeur de corrélation (Eq (1)) calculée entre les séries temporelles. Elle est décrite dans [Asuero, 2006] comme le degré d'association entre deux variables. La corrélation des séries temporelles peut être affectée par divers facteurs, tels que le bruit et les valeurs aberrantes [Asuero, 2006].

$$\text{Corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}} \quad (1)$$

$$\text{cov}(X, Y) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y}) \quad (2)$$

$$\text{var}(X) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{X})^2 \quad (3)$$

II.3.1.1.b. Saisonnalité (Périodicité)

La saisonnalité d'une série temporelle est une composante cyclique plus ou moins régulière, résultante de phénomènes divers, revenant périodiquement [Walas, 2013]. La Figure 1 montre un exemple d'une série temporelle avec une composante saisonnière (périodique).

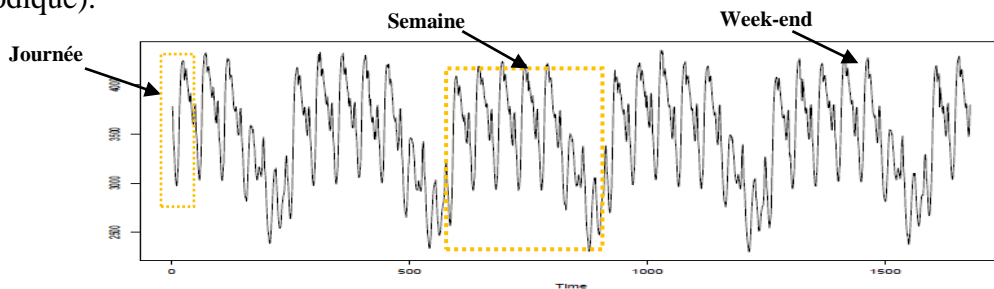


Fig.1. Consommation d'électricité sur toute la France. Une donnée toutes les demi-heures, pendant 35 jours en juin-juillet 1991. Soit 1680 données. On y voit bien deux composantes périodiques correspondant à la journée (48 unités de temps) et à la semaine, ainsi qu'un effet week-end massif. [Viano, 2004].

II.3.1.1.c. Tendance

La tendance d'une série temporelle représente le mouvement de fond d'un phénomène particulier. Elle est décrite parfois comme la tendance globale d'un phénomène à long terme [Walas, 2013]. La Figure 2 montre un exemple d'une série temporelle avec une tendance linéaire croissante qui peut être représentée par une droite.

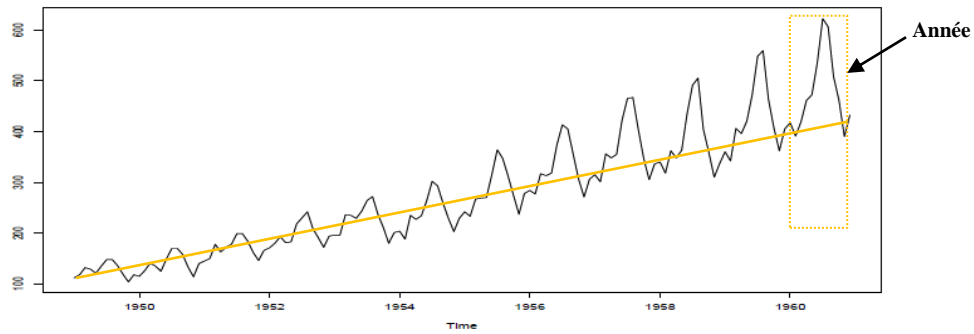


Fig.2. Nombre de passagers (en milliers) dans les transports aériens à paris. Une donnée par mois de 1949 à 1960. Constatations visuelles : tendance générale linéaire, forte augmentation de la variabilité et une composante périodique correspondant à l'année [Viano, 2004].

II.3.1.1.d. Histogramme à une échelle

L'histogramme [Chen, 2003] est, schématiquement, la figure constituée de rectangles dont les bases sont les classes et dont les aires sont égales aux fréquences de ces classes [Gaudoin, 2012].

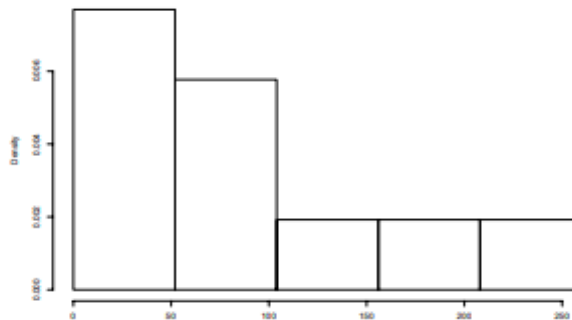


Fig.3. Histogramme à une échelle [Gaudoin, 2012].

Analytiquement parlant, étant donné les valeurs maximales et minimales des séries temporelles, telles que \max_T et \min_T , nous divisons la plage $[\min_T, \max_T]$ en C sous-régions de taille égales disjointes, appelées classes d'histogrammes. Étant donné une séquence temporelle T , nous pouvons calculer l'histogramme H de T en comptant le nombre de points de données h_i ($1 \leq i \leq n$) qui sont localisés dans chaque classe d'histogramme : $H=[h_1, \dots, h_c]$.

La Figure 4 décrit les étapes pour construire l'histogramme d'une série temporelle T .

Algorithme de construction de l'histogramme d'une série temporelle

```

/* entrées: Un ensemble de séries temporelles  $T$  et le nombre de classes d'histogrammes  $C^*$ 
/* sortie : un ensemble d'histogrammes */
1: trouver  $\max_T$  et  $\min_T$  de l'ensemble des données  $T$ 
2: diviser  $[\min_T, \max_T]$  en  $C$  classes d'histogrammes de tailles égales disjointes  $h_i$ 
3: pour chaque série temporelle  $T_i$  de  $T$  faire
4:   pour chaque point de données  $x_i$  de  $T_i$  faire
5:     pour chaque classe d'histogramme  $h_i$  faire
6:       si  $h_{i\text{borne inférieure}} \leq x_i < h_{i\text{borne supérieure}}$ , alors
7:          $h_i = h_i + 1$ ;
8:       fin si
9:     fin pour
10:  fin pour
11: insérer l'histogramme  $H_i$  de chaque  $T_i$  à l'ensemble des résultats
12: fin pour
13: retourner l'ensemble des résultats.

```

Fig.4.Algorithme de construction de l'histogramme d'une série temporelle.

II.3.1.1.e. Histogramme multi échelles

Les histogrammes multi-échelles [Chen, 2003] viennent palier les problèmes des histogrammes à une échelle (globaux) qui ne considèrent pas l'ordre d'affichage temporel des valeurs. Par exemple, sur la Figure 5, deux séries temporelles, T_A et T_B , diffèrent considérablement en termes d'ordre d'apparence. Cependant, ils ont les mêmes histogrammes à une échelle illustrés par

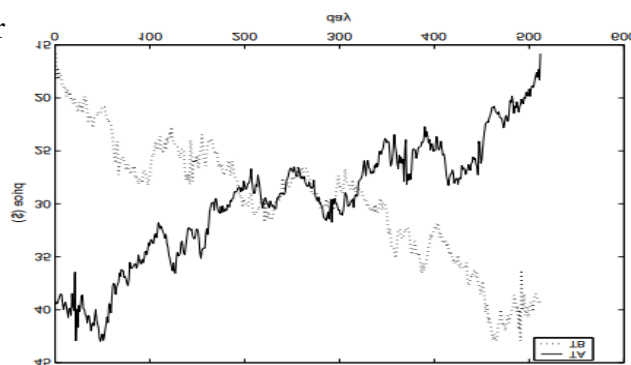
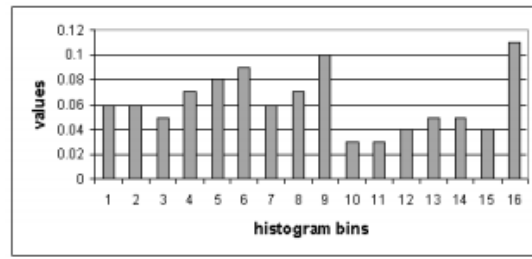
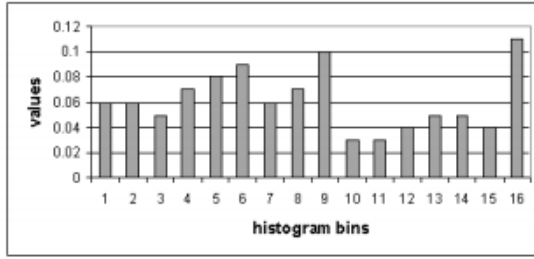


Fig.5. Deux séries temporelles différentes avec le même histogramme [Chen, 2003].

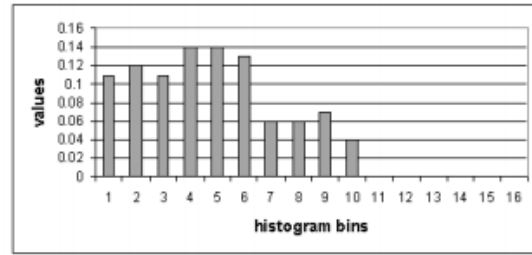
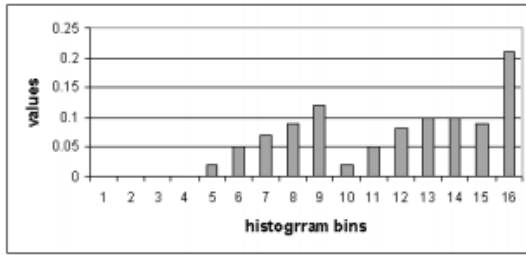
La Figure 6.a et la Figure 6.b montrent les histogrammes à une échelle des séries temporelles T_A et T_B de la figure 5. On voit que les histogrammes sont identiques malgré que les séries temporelles soient différentes. La Figure 6.c et la Figure 6.d montrent les

histogrammes à deux échelles des séries temporelles T_A et T_B . On voit que les histogrammes des deux séries sont inversés parce que les séries temporelles ont des tendances inversées.

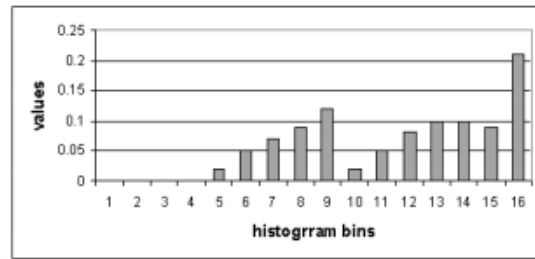
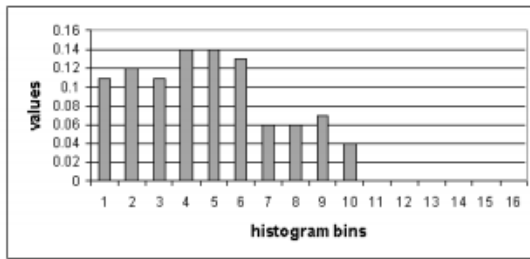


(a) Histogramme de T_A à l'échelle 1

(b) Histogramme de T_B à l'échelle 1



(c) Deux histogrammes de T_A à l'échelle 2



(d) Deux histogrammes de T_B à l'échelle 2

Fig.6. Histogrammes à deux échelles pour les séries temporelles T_A et T_B .

Une représentation multi échelle de l'histogramme de chaque série temporelle est conçue pour une meilleure discrimination des données en fonction de leurs détails temporels. Une série temporelle T_A de longueur n peut être divisée de manière égale en segments δ ($1 \leq \delta \leq n$). Pour chaque segment, nous pouvons calculer son histogramme de série temporelle. Si $\delta = 1$, nous obtenons l'histogramme de la série temporelle complète telle que définie dans la section précédente. Si une précision supplémentaire est requise pour comparer deux séries temporelles, nous pouvons définir $\delta > 1$, créer un histogramme pour chaque segment (pour les deux séries temporelles) et comparer les histogrammes correspondants. Ainsi, des histogrammes multi-échelles peuvent être développés en spécifiant une valeur δ qui définit le niveau de précision (c'est-à-dire l'échelle) de l'histogramme [Chen, 2003].

Les histogrammes multi-échelles sont invariants par rapport au décalage et à la mise à l'échelle sur les axes du temps et de l'amplitude. En outre, ils peuvent supprimer les perturbations introduites par le bruit [Chen, 2003].

Les histogrammes multi-échelles offrent aux utilisateurs la possibilité d'interroger les données des séries temporelles sur différents niveaux de précision. En outre, des histogrammes à plus petite échelle peuvent être utilisés pour éliminer les fausses alarmes de la base de données avant d'interroger des histogrammes à plus grande échelle [Chen, 2003].

La difficulté des histogrammes multi-échelles est la fixation du nombre de segments pour déterminer le nombre d'échelles.

II.3.1.2 Approches géométriques

II.3.1.2.a. Méthode PAA (Piecewise Aggregate Approximation) et APCA(AdaptivePAA)

La méthode Piecewise Aggregate Approximation (PAA) a été proposée par Keogh et al [Keogh, 2001, a] pour la réduction du nombre de données. Elle est basée sur la simple observation que pour la plupart des séries temporelles, il est possible de faire une approximation des données en divisant les séquences en sections de même longueur et en sauvegardant la valeur moyenne de ces sections. Ces valeurs moyennes peuvent être indexées dans un espace de dimension inférieur.

Une série temporelle de longueur n est représentée dans l'espace réduit N par un vecteur $\bar{X} = \bar{x}_1, \dots, \bar{x}_N$, tel que N est la dimensionnalité de l'espace à indexer. Un élément de X est calculé avec la formule suivante :

$$\bar{x}_i = \frac{N}{n} \sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i} x_j \quad (4)$$

L'équation 1 calcule la valeur moyenne des données dans un segment et le vecteur de ces valeurs donnera la représentation réduite des données.

Keogh et al [Keogh, 2001-b] proposent une version plus étendue de PAA appelée Adaptive Piecewise Constant Approximation (APCA) dans laquelle la longueur de chaque segment n'est pas fixée, mais adaptée à la forme de la série. Selon Keogh *et al* [Keogh, 2001-b], cette méthode a une complexité de $O(n)$. Elle peut être utilisée pour un traitement en temps réel puisqu'à chaque réception de la nouvelle valeur correspondante à un segment, on peut calculer la moyenne [Keogh, 2001-b].

La Figure 7 présente une représentation réduite d'une série temporelle avec la méthode PAA. La série temporelle est divisée en segments et la moyenne de chaque segment est calculée. Le vecteur de ces moyennes donnera la représentation réduite des données

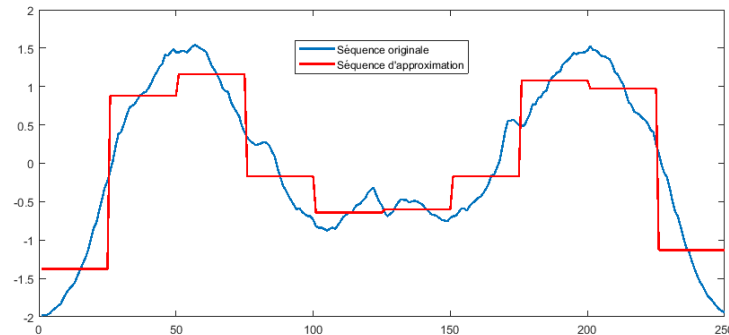


Fig.7. Exemple d'approximation d'une série temporelle avec la méthode PAA.

II.3.1.2.b. Méthode PLA (Piecewise Linear Approximation)

La méthode Piecewise Linear Approximation (PLA) ou Approximation linéaire par morceaux a été utilisée par plusieurs auteurs [Shatkay, 1996],[Korn, 1997], [Li, 1997], [Nygaard, 1997], [Keogh, 1998], [Boucheham, 2006] pour la caractérisation et la réduction des séries temporelles. La clé de la méthode PLA est la segmentation optimale de la série temporelle en sous séquences qui seront ensuite approximeés avec des fonctions linéaire [Chen, 2005-b].

Contrairement aux premiers auteurs qui utilisent une stratégie séquentielle pour segmenter les séries temporelles, Boucheham [Boucheham, 2006] utilise une méthode de segmentation récursive inspirée de l'algorithme de simplification de lignes de Douglas et Peucker [Douglas, 1973]. Il utilise ensuite une mesure d'erreur inspirée de la distance de Hausdorff pour exprimer l'approximation linéaire par morceaux de l'ECG et il illustre que la stratégie récursive est meilleure que la stratégie séquentielle en termes de complexité temporelle et qualité de l'approximation pour la même erreur maximale tolérée ou le même budget de points de ruptures (BreakPoints) utilisés.

La figure 8montre un exemple d'utilisation de cette méthode.

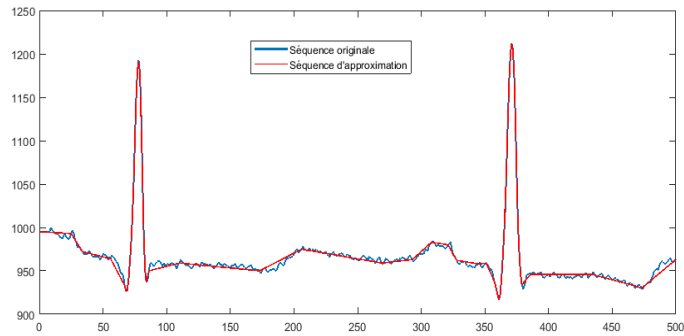


Fig.8. Exemple d'approximation d'une série temporelle avec la méthode PLA.

II.3.1.2.c. Méthode SAX (Symbolic Aggregate approxImation)

La méthode d'approximation symbolique (SAX) a été proposée dans [Lin, 2007] par Lin et al. Elle permet à une série temporelle de longueur arbitraire n d'être réduite en une chaîne symbolique de longueur arbitraire w ($w < n$). Cette approche transforme les données d'origine des séries temporelles en chaînes symboliques [Lin, 2007] en passant par deux étapes: durant la première étape, une transformation des données avec la méthode d'approximation PAA est effectuée. La deuxième étape consiste à transformer cette représentation en une chaîne de caractères discrète. Une technique de discrétisation [Apostolico, 2002], [Lonardi, 2001] est appliquée pour obtenir des symboles avec équiprobabilité.

La méthode SAX est basée sur le fait que les séries temporelles normalisées ont une distribution gaussienne. En déterminant les points d'arrêt « break points » qui correspondent à la taille de l'alphabet utilisé, on peut obtenir des régions de taille équitables [Larsen, 1986].

L'intervalle entre deux breaks points successifs est assigné à un symbole de l'alphabet, et chaque segment de PAA qui est dans cet intervalle est discrétisé par ce symbole.

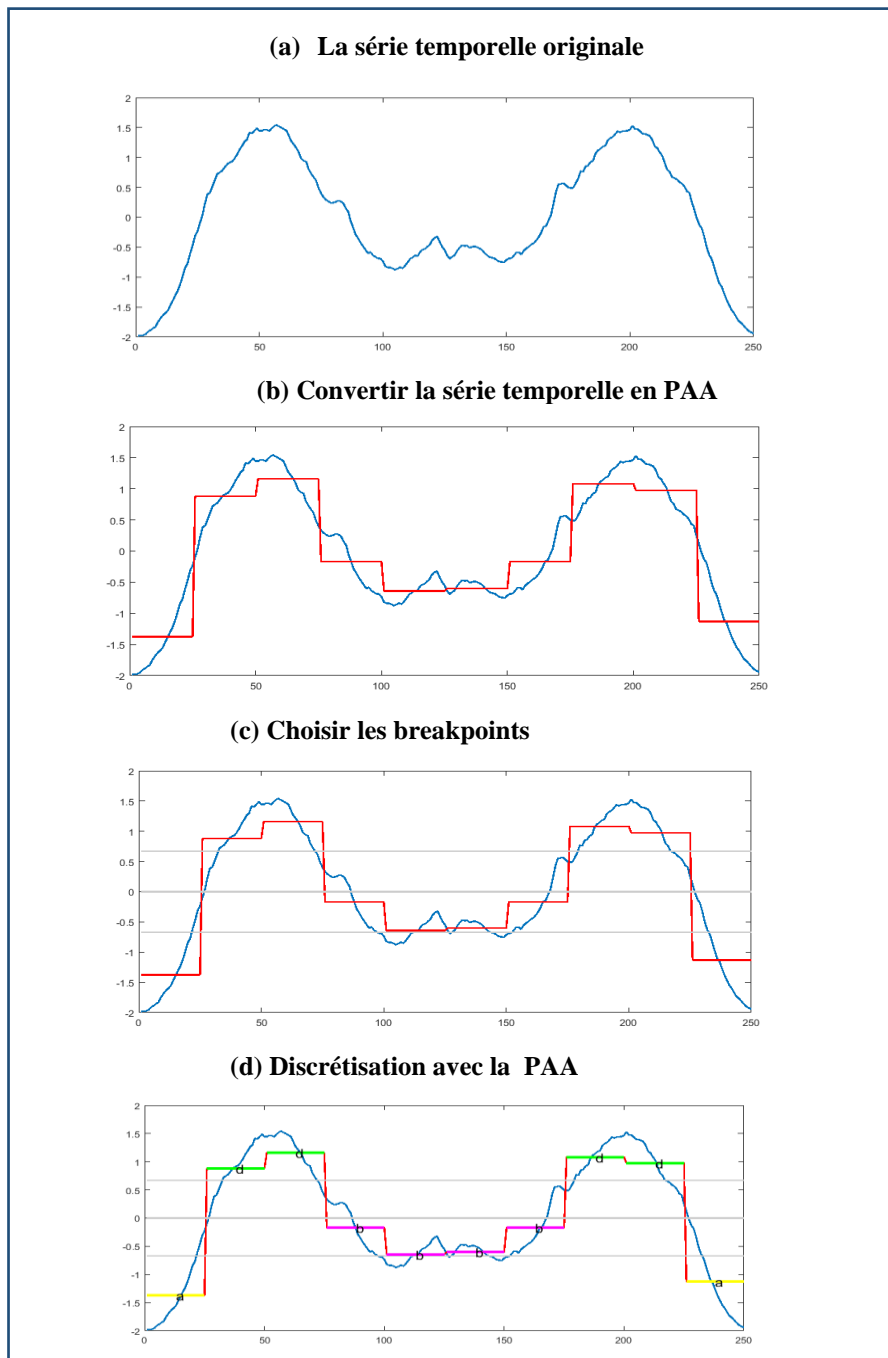


Fig.9. Etapes d'approximation d'une série temporelle avec la méthode SAX.

Figure 9 illustre les étapes de la SAX. On obtient d'abord un PAA de la série temporelle. Tous les coefficients PAA qui sont en dessous du plus petit breakpoint sont mappés sur le symbole "a", tous les coefficients supérieur ou égal au plus petit breakpoint et moins que le deuxième plus petit point d'arrêt sont mappés sur le symbole «b», etc. Finalement, on transforme une série temporelle de longueur 250 en une chaîne de caractère : « addbbbdada » de longueur 10.

La représentation symbolique utilisée par la méthode SAX lui ouvre la porte aux algorithmes qui manipulent les chaînes de caractères tels que les arbres de suffixes, la bioinformatique, etc [Lin, 2007].

L'inconvénient des techniques PLA, PAA et SAX selon Bettaiah et al [Bettaiah, 2014] est qu'elles ne capturent pas les informations relatives à la forme des séries temporelles et leurs tendances puisqu'elles les découpent en segments. De notre part, nous excluons de cette liste la PLA, puisque, dans cette technique, les segments de lignes sont typiquement des tendances, plus ou moins locales.

II.3.1.3 Approches structurelles

II.3.1.3.a. Extraction de sous-sections de formes (Shapelets)

Les shapelets sont des sous-séquences de séries temporelles qui sont, dans un certain sens, très représentatives d'une classe de ST. Elles ont été proposées par Ye et Keogh [Ye, 2009] pour la classification des séries temporelles.

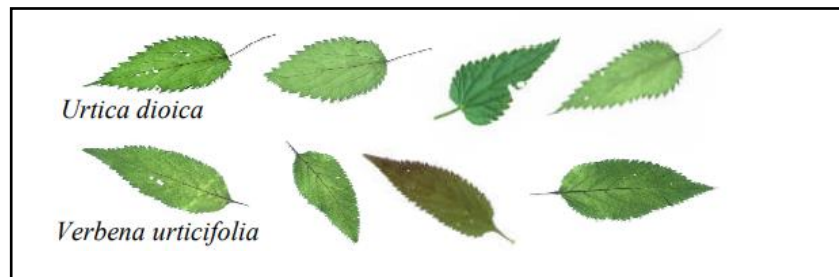


Fig.10.Échantillons de feuilles de deux espèces. Notez que plusieurs feuilles sont endommagées par des piqûres d'insectes [Ye, 2009].

Dans la Figure 10 [Ye, 2009], quelques exemples de feuilles d'orties (*Urtica dioica*) et de verveine (*Verbena urticifolia*) sont présentés. Ces deux plantes sont communément confondues, car elles ont la même couleur et presque la même forme. Les auteurs de la méthode essaient d'abord de transformer la forme de la feuille entière en une série temporelle unidimensionnelle mais ils ont trouvé que la différence entre les formes des séries temporelles est subtile à cause du bruit contenu dans les données (c.-à-d. les morsures d'insectes et les différentes longueurs de tiges) (voir Figure 10). Dans la Figure 11 et la Figure 12, nous pouvons voir que le shapelet a "découvert" que la différence déterminante entre les deux espèces est que *Urtica dioica* a une tige reliée à la feuille à presque 90 degrés, alors que la tige de

Verbena urticifolia est reliée à la feuille à un angle beaucoup plus plat. Les auteurs de [Ye, 2009] vont ensuite comparer que les sous-sections discriminantes « shapelet » des formes des deux classes au lieu de comparer les formes entières.

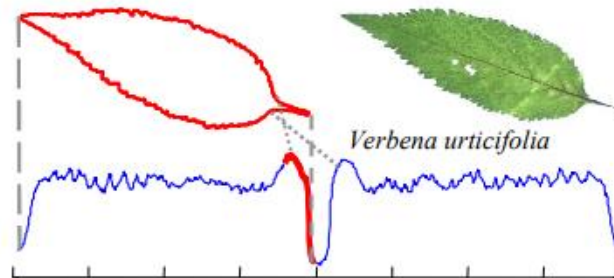


Fig.11. Une forme peut être convertie en une "série temporelle" unidimensionnelle. La partie mise en évidence de la série temporelle sera expliquée dans la figure suivante [Ye, 2009].

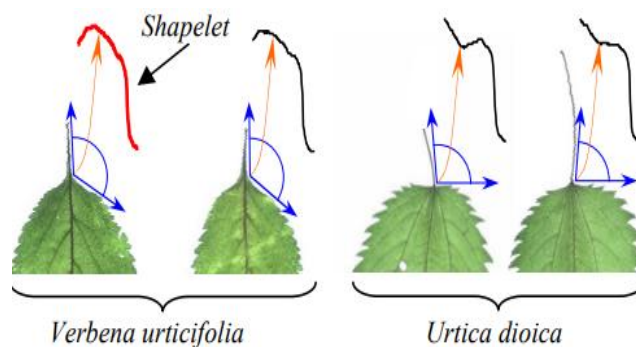


Fig.12. Le shapelet de la Figure 10 (dans les deux cas représentés par une ligne en gras), est la sous-séquence qui fait la distinction entre les deux classes [Ye, 2009].

Après avoir trouvé le shapelet et enregistré sa distance par rapport à la sous-séquence la plus proche dans tous les objets de la base de données, les auteurs construisent un classificateur d'arbre de décision simple comme il est démontré dans la figure 13.

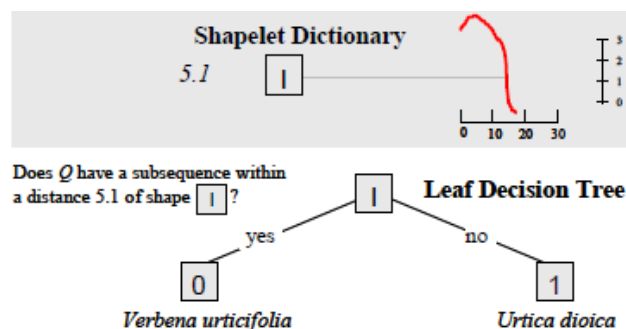


Fig.13. Un classifieur d'arbre de décision pour le problème de la feuille. Les sous-séquences de l'objet à classier sont comparées au shapelet. Si une sous-séquence est inférieure à la valeur déterminée empiriquement de 5.1, elle est classée comme *Verbena urticifolia* [Ye, 2009].

L'avantage de cette méthode par rapport aux autres méthodes de réduction des données est qu'elle extrait des caractéristiques locales et donc elle permet de résoudre les problèmes de bruit lors de la recherche et de la classification.

II.3.1.3.b. Découverte des Motifs des ST

Les motifs d'une série temporelle sont des modèles de « sous séquences » répétés approximativement à l'intérieur d'une série temporelle plus longue. [Nurjahan, 2014]. La présence de motifs n'est pas seulement due à la chance ; les motifs contiennent généralement des informations importantes sur les ST et fournissent des informations précieuses sur la problématique étudiée [Torkamani, 2017].

Le problème de la découverte de motifs a été initialement centré sur la découverte d'ensemble d'items (itemsets) fréquents [Agrawal, 1993]. Il a été proposé dans le cadre de l'analyse du panier de la ménagère (Market-Basket Analysis), dont le but fut de rechercher la collection de produits corrélés souvent achetés ensemble, en examinant les tickets de caisse enregistrés par les supermarchés. Cette connaissance des co-occurrences existantes entre produits est ensuite exploitée pour générer ce qui est appelé les règles d'association qui sont des formes d'implications entre ensembles de produits [Oulad-Naoui, 2018].

Les travaux sur l'extraction de motifs ont ensuite été étendus pour extraire des motifs structurels comme les séquences [Agrawal, 1995-b], les graphes [Inokuchi, 2003], les arbres [Chi, 2004] ou les arbres attribués [Pasquier, 2013]. Ces derniers sont des arbres dans lesquels les nœuds sont associés à des itemsets.

Les problèmes courants des algorithmes de découverte de motifs sont la détermination de la longueur des motifs, la gestion des données en continu, la détection des motifs de longueur variable, l'énorme complexité temporelle et la découverte de motifs mal connus. Selon l'application, les algorithmes de découverte de motifs diffèrent dans leur façon d'aborder les problèmes. Ces méthodes peuvent être adaptées à des motifs exacts ou approximatifs. Ils peuvent aussi avoir besoin de détecter des motifs de longueur variable ou fixe [Torkamani, 2017].

II.3.1.3.c. Extraction des caractéristiques des graphiques

Dans le travail de Li *et al* [Li, 2018], une nouvelle approche pour la caractérisation et la classification des séries temporelle a été proposée. L'approche est fondée sur le calcul des distributions de probabilités des motifs répétés dans les graphiques de visibilité illustré par la

Figure 14 [Lacasa, 2008], [Lacasa, 2009] et d'autres caractéristiques statistiques comme la densité, la variabilité et la corrélation. Ce mécanisme d'extraction de caractéristiques est sans paramètres, de sorte qu'il peut être facile à utiliser et permet d'obtenir des résultats reproductibles [Li, 2018].

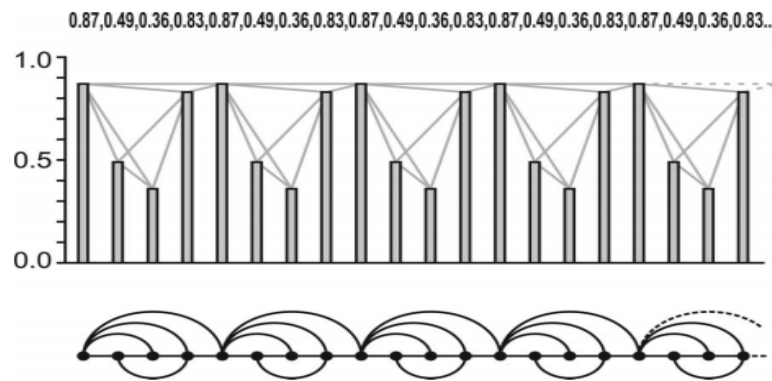


Fig.14.Exemple d'une série temporelle (20 valeurs de données) et le graphique associé dérivé de l'algorithme de visibilité. Dans le graphique, chaque nœud correspond à une donnée de la série temporelle, dans le même ordre. Les rayons de visibilité entre les données définissent les liens reliant les nœuds dans le graphique [Lacasa, 2009].

La Figure 14, un exemple tiré de [Lacasa, 2009], illustre la façon dont une série temporelle se transforme en un graphique de visibilité à l'aide de l'algorithme de visibilité. Une analyse préliminaire a montré que la structure des séries est héritée dans le graphique de visibilité [Lacasa, 2008]. En conséquence, les séries périodiques se transforment en graphiques réguliers, les séries aléatoires en graphiques aléatoires et les séries fractales en graphiques sans échelle [Albert, 2002].

Les auteurs de la méthode considèrent les séries temporelles comme des graphiques/réseaux complexes et extraient de ces réseaux des caractéristiques importantes dans des vecteurs de caractéristiques non ordonnés, qui sont utilisées par la suite par les classifieurs telles que les SVM (Support Vector Machine) pour une sélection et une classification efficaces des caractéristiques [Lacasa, 2008].

II.3.1.4 Approches basées extraction de signatures

La méthode SEA (Sahpe Exchange Algorithm) [Boucheham, 2008] et ses dérivées (ASEAL [Boucheham, 2010], FANSEA [Boucheham, 2013] et Fuzzy-SEA [Betancourt, 2013]) sont des méthodes d'alignement et de caractérisation qui s'inspirent du principe du

bloc-sorting de Burrows et Wheeler [Burrows, 1994]. Elles procèdent à l'extraction de signatures discriminantes des séries temporelles en les triant d'abord sur leurs amplitudes. A signaler que les signatures extraites sont de nature globale (voir Figure15). La méthode LMDS-SEA [Lahreche, 2018] qui appartient partiellement à cette classe, quant à elle, utilise des signatures globales et des signatures locales. Elle consiste à subdiviser les séries temporelles en segments (sous-séquences) pour extraire des informations locales (signatures locales) des séries temporelles et à appliquer une étape d'apprentissage pour sélectionner la bonne mesure de similarité entre la distance Euclidienne, SEA ou Local-Matching-SEA à utiliser avec le classifieur 1NN, selon la nature et le nombre de segments des séries temporelles.

La méthode SEA et ses dérivées sont décrites plus en détail dans le chapitre III de cette thèse.

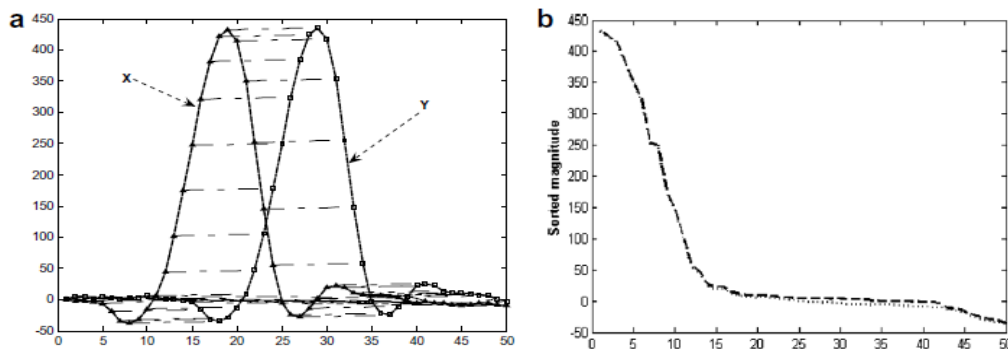


Fig.15. (a)Appariement point par point (noter ici que l'appariement est effectué lors d'un balayage d'amplitude) et (b) Signatures de X et Y presque identiques malgré le décalage temporel entre X et Y [Boucheham, 2008].

II.3.2 Domaine fréquentiel

II.3.2.1 Transformée de Fourier Discrète DFT (Discrete Fourier Transform)

La DFT [Agrawal, 1993] est une technique de réduction de données dans les séries temporelles. Elle est aussi considérée comme une technique de caractérisation dans le domaine fréquentielle, utile dans la comparaison des séries temporelles. Elle est basée sur le principe du théorème de Parseval [Oppenheim, 1999], qui dit que l'énergie des séquences dans le domaine temporel est la même dans le domaine fréquentiel. Agrawal *et al* déduisent ainsi, que la distance euclidienne entre deux séries temporelles dans le domaine temporel est la même dans le domaine fréquentiel [Agrawal, 1993]. En outre, Agrawal *et al* ont remarqué

que les fréquences des amplitudes ne changent pas lors des décalages temporels [Agrawal, 1993].

La distance entre deux séries temporelles X, Y dans [Agrawal, 1993] est définie par la racine carrée de la somme des différences carrées entre les séquences fréquentielles (Eq. 5). Spécifiquement, la distance entre deux séquences x_t et y_t est la racine de l'énergie E des différences. L'énergie E d'une séquence X est définie comme la somme des énergies à chaque point de la séquence (Eq. 6).

$$D(X, Y) = \sqrt{\sum_{t=0}^k |x_t - y_t|^2} = \sqrt{E(X - Y)} \quad (5)$$

$$E(X) \equiv \|X\|^2 \equiv \sum_{i=0}^{n-1} |x_i|^2 \quad (6)$$

Il s'agit de la distance dans l'espace temporelle. En utilisant seulement les quelques premiers coefficients de DFT pour calculer la distance euclidienne, on obtient une distance qui est minorant à la distance euclidienne présentée dans la section II.4.2.1. Selon [Muhammad Fuad, 2011], une série temporelle de 256 dimensions, par exemple, peut être représentée par 128 coefficients de Fourier complexes qui sont des nombres complexes.

La complexité de calcul des coefficients de DFT est de l'ordre de $O(n^2)$, mais un algorithme connu sous le nom de Fast Fourier Transform (FFT) a une complexité inférieure de $O(n \log n)$, [Shatkey, 1995].

Suite à son établissement, si cette distance est inférieure à un seuil ε prédéfini par l'utilisateur, les deux séquences seront dites similaires.

Pour chacune des séries temporelles X et Y de longueur n, les auteurs proposent d'extraire uniquement les k ($k < n$) premiers coefficients de la DFT (fréquences) [Muhammad Fuad, 2011].

Le $j^{\text{ème}}$ coefficient de Fourier d'une séquence X est calculé par la formule suivante:

$$X_f = 1/\sqrt{(n)} \sum_{t=0}^{n-1} \exp(-j2\pi ft/n), \quad t = 0, 1, \dots, n-1, \quad j = \sqrt{-1} \quad (7)$$

Et la transformée inverse est définie par:

$$x_t = 1/\sqrt{(n)} \sum_{f=0}^{n-1} X_f \exp(j2\pi ft/n), \quad t = 0, 1, \dots, n-1 \quad (8)$$

En ce qui concerne la contrainte de l'égalité des longueurs, Faloutsos et al. [Faloutsos, 1994] ont généralisé le travail d'Agrawal et al [Agrawal, 1993] pour permettre l'appariement des séquences. Les séquences de données peuvent être de longueurs différentes,

mais la séquence faisant office de requête doit être plus petite que n'importe quelle autre séquence [Faloutsos, 1994].

II.3.3 Domaine temps-fréquence (Ondelettes, Wavelets)

Dans ce cadre, nous présentons d'abord la transformée en ondelettes DWT (Discrete Wavelet Transform), puis une méthode récente de mesure de similarités entre TSs basée Wavelets.

II.3.3.1 Transformée en ondelettes DWT (Discret Wavelet Transform)

La DWT ou la méthode de la transformée en ondelettes (Discrete Wavelet Transform) [Chan, 1999] est un outil mathématique basé sur une fonction de décomposition hiérarchique des signaux. La transformée en ondelettes est aussi utilisée comme une technique efficace de caractérisation et de réduction de la dimensionnelle pour la recherche par similarité dans les séries temporelles.

La transformée en ondelettes permet d'effectuer la recherche par similarité à différentes échelles [Chan, 1999]. La mesure de similarité pourrait alors être définie de manière adaptative et interactive et être utilisée dans plusieurs applications de fouille des données [Chaovalit, 2011].

En particulier, la transformée de Haar en ondelettes est la forme la plus simple des transformées en ondelettes. C'est une série d'opérations de rapprochement et de différenciation. Pour avoir une idée de comment fonctionne la transformée de Haar en ondelettes, considérons la série temporelle à 4 dimensions suivante: $X = [10, 6, 5, 7]$. La première étape consiste à extraire la médiane de chaque couple de valeurs successives. On obtient une série temporelle à deux dimensions: $X' = [8, 6]$.

Pour revenir à la série temporelle originale à 4 dimensions, on a besoin d'utiliser les *Coefficients DWT*, qui sont dans ce cas: $2 = \frac{10-6}{2}$ et $-1 = \frac{5-7}{2}$.

En répétant successivement ce processus, on obtient la décomposition complète de X. La table1 illustre les étapes de décomposition :

Table1. Les étapes de décomposition de la méthode DWT

| Résolution | Moyenne | DWT coefficient |
|------------|------------|-----------------|
| 4 | [10,6,5,7] | |
| 2 | [8, 6] | [2,-1] |
| 1 | [7] | [1] |

La transformation en ondelettes (Wavelets) de X est alors [7, 1, 2, -1].

L'idée de base de DWT est similaire à celle de DFT dans le sens où une série temporelle peut être uniquement représentée par une transformée en ondelettes, mais en ne conservant que les K premiers coefficients, on peut réduire la dimensionnalité et conserver une grande partie de l'information qui est dans la série temporelle originelle.

Une comparaison entre DFT et DWT pour l'appariement des séries temporelles a été présentée dans [Wu, 2000]. Leurs résultats montrent que la DWT ne réduit pas l'erreur relative d'appariement par rapport à celle de la DFT et n'augmente pas la précision de la recherche par similarité dans les séries temporelles. Cependant, elle présente plusieurs avantages comme sa capacité de recherche à différentes échelles et sa complexité temporelle qui est de $O(n)$, par rapport à celle de DFT qui est de $O(n^2)$. L'inconvénient de DWT est qu'elle exige que la longueur d'une série temporelle soit une puissance de 2.

II.3.3.2 Technique SIM Shape

Une nouvelle mesure de similarité SimShape basée sur la forme est introduite dans [He, 2014] pour traiter les multiples distorsions des séries temporelles. Le nombre de types de distorsions et le degré de distorsions sont utilisés pour mesurer la robustesse de la méthode.

Pour extraire les caractéristiques les plus intéressantes de la série temporelle originelles, une nouvelle représentation multi-échelles de la méthode Haar Wavelet Transform est introduite dans SIMshape.

Il a été démontré dans [He, 2014] que contrairement aux méthodes d'approximation évoquées précédemment, telles que la DFT, la SAX et la PAA, la méthode SIMshape réduit considérablement la dimensionnalité des données et n'a pas besoin de réglage de paramètres.

He et al [He, 2014] ont aussi constaté que SIMshape est plus résistante que la distance Euclidienne, la LB-Keogh [Keogh, 2005], CID (Complexity-Invariant Distance) [Batista, 2011] et ASEAL [Boucheham, 2010] par rapport aux bruits et aux décalages temporelle, et que SIMshape et ASEAL peuvent tolérer le plus grand degré de décalage d'amplitude et de

différence d'échelle d'amplitude par rapport aux trois autres mesures de similarités, parce que le décalage d'amplitude affecte des valeurs spécifiques plutôt que l'ensemble des informations de la forme [He, 2014]. Les auteurs de SIMshape expliquent aussi que ASEAL utilise l'échange de toutes les valeurs d'amplitude pour reconstruire deux nouvelles séries temporelle, et SIMshape est basée sur la forme de base des séries temporelles. Il est donc difficile d'avoir un impact sur les informations de forme de base par le biais du décalage et de la différence d'échelle d'amplitude. Cependant, les trois autres mesures de similarité sont basées sur des informations détaillées alors avec un degré de distorsion élevé, ED, LB_keogh et CID perdront leur efficacité, tandis que ASEAL et SIMshape resteront inchangées.

II.3.4 Domaine algébrique

II.3.4.1 Décomposition en valeurs singulières (SVD)

Globalement, la SVD [Wu, 1996] utilise, comme la DFT et la DWT, le principe de décomposition des ST. Seulement, elle représente la série temporelle en tant que combinaison linéaire de valeurs propres, mais ne conserve que les premiers K coefficients pour faire une représentation de dimension inférieure à celle de la série temporelle d'origine [Muhammad Fuad, 2011].

L'intuition derrière la SVD est qu'un ensemble de m points se regroupera en un groupe sur quelques dimensions. Chaque groupe représente un « composant principal ». Ce composant a un haut pouvoir discriminatoire et est orthogonal aux autres groupes. SVD trouve ces dimensions et projette les points sur ces dimensions.

Mathématiquement, SVD peut être exprimé comme suit : Étant donné une base de données des séries temporelles qui contient m séries temporelles dont la dimension est n , la composition SVD de cette base de données est une matrice X [Muhammad Fuad, 2011]:

$$X = U \times \Lambda \times V^t \quad (9)$$

Où U est la colonne ortho-normale $N \times r$ de la matrice (r est un rang de la matrice X), Λ la matrice diagonale $r \times r$ des valeurs propres λ_i de X et V la colonne ortho-normale $M \times r$ de la matrice qui est essentiellement la décomposition spectrale de la matrice X ;

$$X = \lambda_1 \mu_1 \times V_1^t + \lambda_2 \mu_2 \times V_2^t + \dots + \lambda_r \mu_r \times V_r^t \quad (10)$$

Où μ_1 et v_1 sont les vecteurs de colonne de U et V et λ_i est l'élément diagonal de Λ .

Quand X est décomposée en forme d'un spectre, les auteurs suggèrent de tronquer la somme en équation des k premiers termes, $k \leq r \leq M$:

$$\hat{X} = \sum_{i=1}^k \lambda_i \mu_i \times V_i \quad (11)$$

Où k dépend de l'espace permis. Ces k termes sont aussi connus comme les "principales composantes" dans l'analyse PCA.

Il est démontré que le taux de compression après l'application de la SVD est :

$$s = \frac{n * k + k + k + k * M}{N * M} \approx \frac{k}{M} \quad (12)$$

La SVD surpasse la DFT et la DWT, mais sa complexité est très élevée $O(mn^2)$ [Muhammad Fuad, 2011].

II.4 Discussion et conclusion

Dans ce chapitre, nous avons présenté les séries temporelles et des techniques permettant de les caractériser. En guise de récapitulatif et rapprochement entre les différentes techniques recensées, nous rapportons cette comparaison et discussion dans la Table 2.

Table2. Techniques de caractérisation des séries temporelles.

| | | |
|-------------------------|---|--|
| Domaine temporel | Approches statistiques (histogramme, corrélation,...) | Avantage/Inconvénients : Approche simples et rapides mais globales sauf pour l’histogramme multi-échelles qui retient l’information locale de chaque segment de la série temporelle. |
| | Approches géométriques (PLA, PAA, SAX) | Avantage/Inconvénients : Approches faciles à mettre en œuvre. Cependant, elles ne capturent pas les informations relatives à la forme des séries temporelles et leurs tendances puisqu’elles les découpent en segments. Ces descripteurs sont inefficaces pour la représentation des données avec une grande variabilité, car il y a une suppression de caractéristiques importantes [Garcia, 2014]. |
| | Approches structurelles (motif, shapelets, graphe) | Avantage/Inconvénients : Approches qui prennent en considération la forme des séries temporelle mais qui sont trop dépendantes de la forme des séries temporelles. |
| | Approches basées extraction de signatures (SEA, ASEAL, FANSEA, LMDS-SEA, ...) | Avantage/Inconvénients : La méthode SEA, ASEAL, FANSEA et Fuzzy-SEA prennent en considération la forme globale des séries temporelles sans avoir besoin de paramètres de réglage. Cependant, elles sont sensibles au bruit [Boulnemour, 2018] car elles ne capturent que les informations globales. La méthode LMDS-SEA palie les problèmes du bruit car elle capture les informations locales et globales [Lahreche, 2018]. Le seul désavantage de cette méthode est qu’elle est moins rapide que SEA puisque elle a besoin d’une étape d’apprentissage pour connaitre le bon nombre de segments de la série temporelle. |
| Domaine fréquentiel | DFT, FFT | Avantage/Inconvénients : Ces approches permettent de résoudre le problème de décalage temporel car les fréquences des amplitudes ne changent pas lors des décalages temporels. Elles permettent aussi de réduire le bruit. Cependant, elles ont besoin de faire un apprentissage des données pour spécifier le bon nombre K de coefficients. |
| Domaine temps-fréquence | DWT, SIM-Shape | Avantage/Inconvénients : L’idée de base de DWT est similaire à celle de DFT mais sa complexité temporelle est de $O(n)$, par rapport à celle de DFT qui est de $O(n^2)$. L’inconvénient de DWT est qu’elle exige que la longueur d’une série temporelle soit une puissance de 2. L’avantage de SIMshape et qu’elle réduit considérablement la dimensionnalité des données et qu’elle n’a pas besoin de réglage de paramètres [He, 2014]. |
| Domaine algébrique | SVD | Avantage/Inconvénients : SVD surpasse DFT et DWT, mais elle est très couteuse en termes de temps de calcul $O(mn^2)$ [Muhammad Fuad, 2011] et pour la représentation de grands intervalles, la réduction de la dimensionnalité présente des pertes [Garcia, 2014]. |

Chapitre III. Etat de l'art des Techniques de comparaison et d'alignement des ST

III.1 Introduction

L'appariement des séries temporelles (Time Series Matching) inclue la *comparaison* et l'*alignement* de ce type de données. Précisons dans ce contexte que, la comparaison permet d'obtenir le degré de similarité entre deux séries temporelles X et Y alors que l'alignement réalise une correspondance entre les points de X et ceux de Y [Boucheham, 2013].

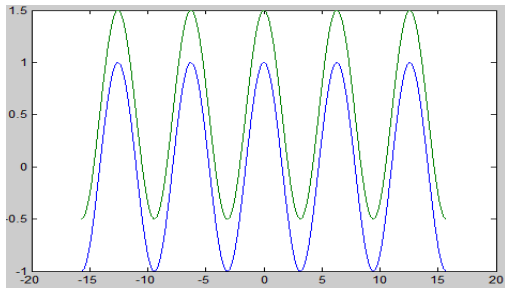
Nous présentons dans ce chapitre une classification des mesures de similarité utilisées pour la comparaison et l'alignement des séries temporelles.

III.2 Difficultés associées à l'appariement des séries temporelles

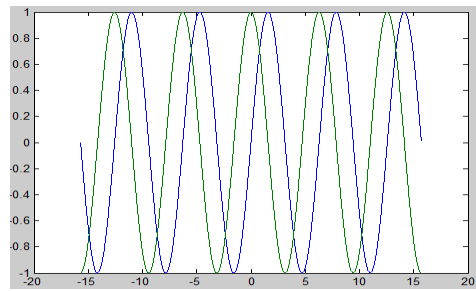
Plusieurs difficultés viennent rendre difficile le problème de l'appariement des ST [Argyros, 2003]. Nous les rapportons comme suit :

III.2.1 Décalages et différence d'échelles entre les séries temporelles

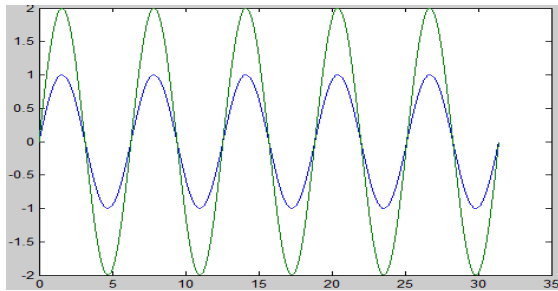
Les séries temporelles ne sont souvent pas alignées sur les axes de l'amplitude ou celui du temps (Fig16.a-b). Elles peuvent aussi avoir des différences d'échelles d'amplitude ou de temps (Fig16.c, d). Les techniques d'appariement ne sont pas toutes capables de faire face à ces types de difficultés, même si les séries temporelles sont visuellement similaires, comme dans la Figure16.



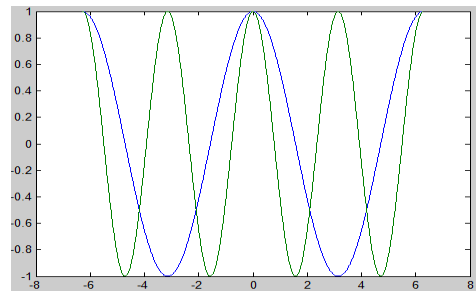
(a) Décalage sur l'axe d'amplitude



(b) Décalage sur l'axe du temps



(c) Différences d'échelles d'amplitudes



(d) Différences d'échelles de temps

Fig. 16. Difficultés associées à l'appariement des séries temporelles.

III.2.2 Le Bruit

La quantité de bruit dans les séries temporelles peut affecter particulièrement les techniques d'appariement (voir Figure 17). Certaines techniques ont cependant, un comportement plus robuste que d'autres en présence du bruit [Esling, 2012].

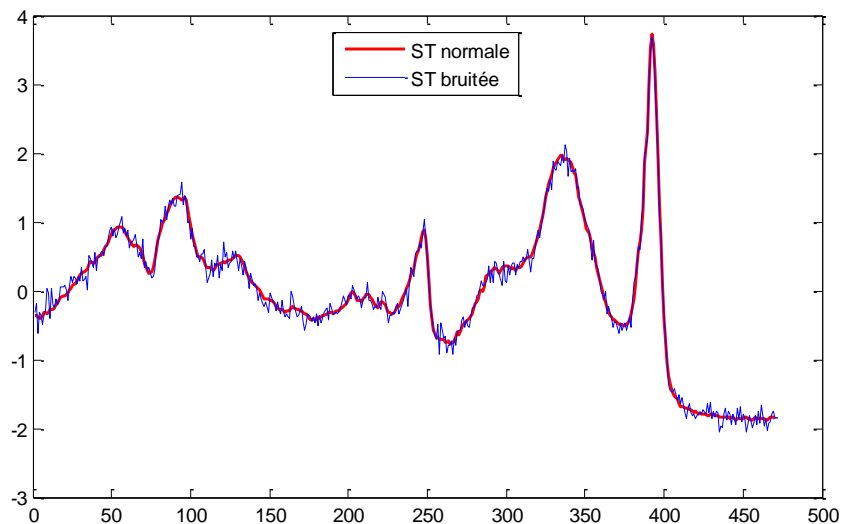


Fig.17. Séries temporelles provenant du Benchmark UCR (Beef) normale v.s bruitée.

III.2.3 Valeurs aberrantes

Les valeurs aberrantes sont les points de données qui s'écartent particulièrement du reste des données (Figure 18). Ils affectent beaucoup les techniques d'appariement des séries temporelles [Esling, 2012].

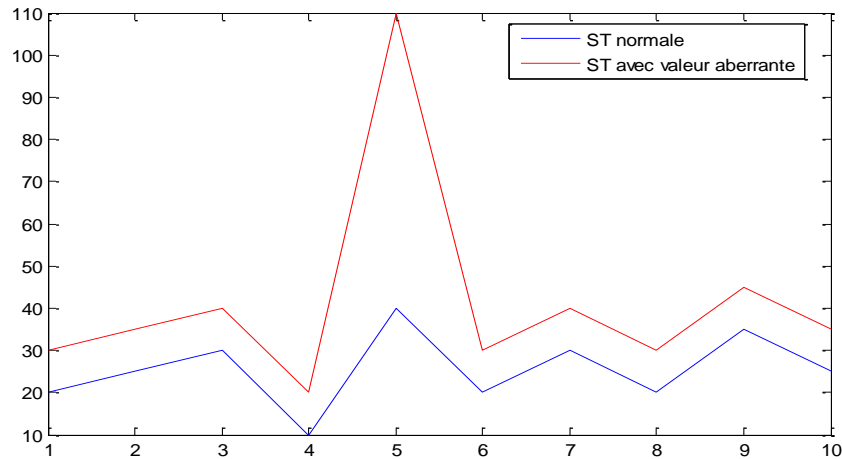


Fig.18. Séries temporelles bleue normale et série temporelle rouge contenant une valeur aberrante à l'instant 5.

III.3 Mesures de similarités utilisées pour l'appariement des ST

Nous pouvons classer les mesures de similarité utilisées pour l'appariement des séries temporelles, comme suit : Les techniques de comparaison basées sur les mesures de distances/similarités utilisant les données brutes (raw data), les techniques de comparaison utilisant des caractéristiques extraites et des mesures de distance/similarité sur ces caractéristiques, les techniques d'alignement utilisant les données brutes et les techniques d'alignement basées caractéristiques. Il va de soit aussi que les techniques d'alignement renvoient en fin de leurs applications, sous une forme ou une autre, une mesure de distance/similarité (Boucheham, 2013).

III.3.1 Techniques de comparaison basées mesures de Distances/Similarités sur les données brutes

Ces techniques mettent en œuvre des mesures de distances/similarités qui respectent une ou plusieurs contraintes posées sur les espaces des valeurs. Nous recensons alors les mesures suivantes : Distance, Quasi-distance, Métrique, Quasi-Métrique, Divergence et Similarité.

Dans le langage courant la distance entre deux points est la longueur du plus court chemin pour aller du point de départ au point d'arrivée. C'est la distance "à vol d'oiseau".

D'un point de vue mathématique, étant donné deux séries temporelles x et y , la fonction de distance $d(x, y)$ entre les deux séries est considérée comme *métrique* si elle respecte un certain nombre de propriétés [Deza, 2009] qui sont les suivantes :

1. Non négativité: La distance entre x et y doit toujours être une valeur supérieure ou égale à zéro : $d(x, y) \geq 0$. Cette assertion exprime le fait que la distance entre un point et un autre ne peut pas être négative.
2. Identité des indiscernables : La distance entre x et y est égale à 0 si et seulement si x est égale à y : $d(x, y) = 0$ ssi $x = y$. Cette assertion exprime le fait que la distance entre un point et lui-même est nulle.
3. Symétrie : La distance entre x et y est égale à la distance entre y et x : $d(x, y) = d(y, x)$. Cette assertion exprime que la distance entre le départ et l'arrivée est la même que celle entre l'arrivée et le départ.
4. Inégalité triangulaire : Considérons la présence d'un troisième point z , la distance entre x et y est toujours inférieure ou égale à la somme des distances entre x et z et la distance entre y et z : $d(x, y) \leq d(x, z) + d(z, y)$. Cette assertion peut s'interpréter dans certains cas par la maxime suivante: "le plus court chemin est la ligne droite".

Si la condition (3) n'est pas satisfaite, d est une *quasi-métrique* et Si la condition (4) n'est pas satisfaite, d est une *semi-métrique* [Huriot, 1990].

Dans la classification utilisée par Cha et al [Cha, 2007] et de [Prasath, 2017], huit familles de mesures de similarité/ distances sont considérées. Nous rappelons les différentes familles avec des exemples de chacune d'elles.

III.3.1.1 L_p : Mesures de distance de Minkowski

Cette famille de distances inclut trois distances métriques célèbres, selon les différentes valeurs de la puissance des distances p

Où p est une valeur entière positive. Lorsque $p=2$, nous obtenons la *distance Euclidienne* et lorsque $p=1$, il s'agit de la distance de *Manhattan*. La *distance de Chebyshev* est une variante de la distance Minkowski quand $p=\infty$. En notant x_i est la $i^{\text{ème}}$ valeur du vecteur x et y_i est la $i^{\text{ème}}$ valeur du vecteur y , nous avons alors les équations suivantes :

$$D_{Mink}(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^r \right)^{1/r} \quad (13)$$

$$D_{Euc}(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (14)$$

$$D_{Manh}(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (15)$$

$$D_{Cheb}(x, y) = \max_{i=1}^n |x_i - y_i| \quad (16)$$

La distance Euclidienne a couramment été utilisée pour la recherche de similarité dans les séries temporelles, car elle a le mérite d'être simple à mettre en œuvre et d'être calculée en seulement $O(n)$ où n est le nombre de valeurs dans la ST. Cependant elle est très fragile lorsque les séries temporelles ont une différence dans l'échelle du temps ou de l'amplitude ou lorsqu'elles sont bruitées.

La Figure19 montre deux séries temporelles comparées point par point avec la distance Euclidienne. Notons ici que les deux pics des deux ST ne sont pas appariés correctement avec la distance Euclidienne car les ST sont prises à des instants différents (décalage sur l'axe temporel).

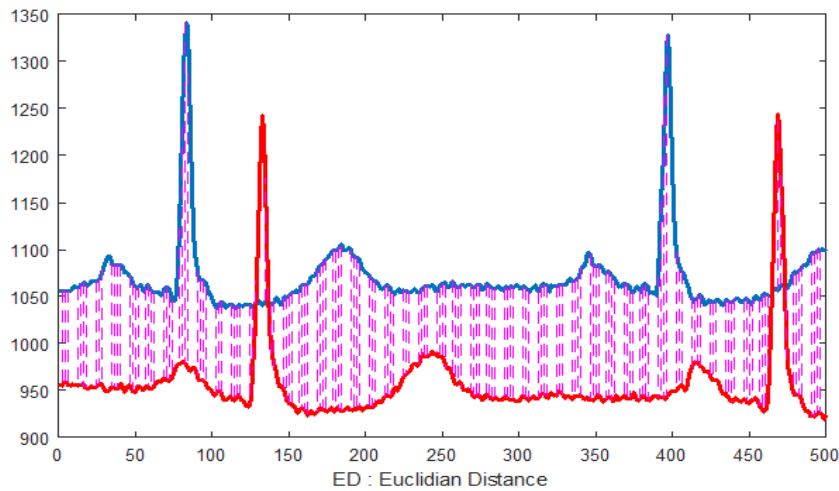


Fig. 19. Comparaison de deux séries temporelles similaires, mais décalées dans le temps avec la distance Euclidienne.

III.3.1.2 Mesures de distance L1

Cette famille de distances/mesures est très sensible aux petits changements [Prasatha, 2017]. Nous donnons une brève présentation de deux de ces distances que nous allons utiliser pour la classification des séries temporelles dans le chapitre VIII.

a. Distance de Lorentzian (Lortz):

$$Lortz(x, y) = \sum_{i=1}^d \ln(1 + |P_i - Q_i|) \quad [17]$$

b. Mesure de similarité Canberra(Canb) :

$$Canb(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad [18]$$

III.3.1.3 Mesures de distance/ similarité L2 au carré:

Dans la famille des mesures de distance L2, le carré de la différence en chaque point le long des deux vecteurs est considéré pour la distance totale. Nous donnons une brève présentation de trois de ces mesures que nous allons utiliser pour la classification des séries temporelles dans le chapitre VIII.

a. Distance Euclidienne au carré (SED) :

$$SED(x, y) = \sum_{i=1}^n (x_i - y_i)^2 \quad (19)$$

b. Mesure de similarité Squared Chi-Squared (SCSD) :

$$SCSD(x, y) = \sum_{i=1}^n \frac{(x_i - y_i)^2}{|x_i + y_i|} \quad (20)$$

c. Mesure de Divergence (DivD) :

$$DivD(x, y) = 2 \sum_{i=1}^n \frac{(x_i - y_i)^2}{(x_i + y_i)^2} \quad (21)$$

III.3.1.4 Autres mesures de distance:

Dans ce qui suit, nous rapportons les définitions des distances que nous allons utiliser pour la classification des séries temporelles dans le chapitre VIII.

- a. **Distance Average(L_1, L_∞)** : La distance Average(L_1, L_∞) est la moyenne des distances de Manhattan et de Tchebychev.

$$AvgD(x, y) = \frac{\sum_{i=1}^n |x_i - y_i| + \max |x_i - y_i|}{2} \quad (22)$$

- b. **Distance Invariable en fonction de la Complexité (CID)** : La distance Invariable en fonction de la Complexité a été introduite par Batista et al [Batista, 2014]. Elle est définie comme suit :

$$CID(x, y) = ED(x, y) \times CF(x, y) \quad (23)$$

où CF est un facteur de correction de complexité défini comme suit :

$$CF(x, y) = \frac{\max(CE(x), CE(y))}{\min(CE(x), CE(y))} \quad (24)$$

et CE est une estimation de la complexité d'une série temporelle.

$$CE(x) = \sqrt{\sum_{i=1}^{n-1} (x_i - x_{i+1})^2} \quad (25)$$

- c. **Distance de Hassanat(HasD)** : La distance Hassanat a été introduite aussi assez récemment par [Hassanet, 2014] dans la l'intervalle [0, 1].

$$HasD(x, y) = \sum_{i=1}^n D(x_i, y_i) \quad (26)$$

$$\text{Où } D(x, y) = \begin{cases} 1 - \frac{1 + \min(x_i, y_i)}{\max(x_i, y_i)}, & \min(x_i, y_i) \geq 0 \\ 1 - \frac{1 + \min(x_i, y_i) + |\min(x_i, y_i)|}{\max(x_i, y_i) + |\min(x_i, y_i)|}, & \min(x_i, y_i) < 0 \end{cases}$$

III.3.2 Techniques de comparaison basées caractéristiques

Pour les techniques de comparaison basées caractéristiques, nous recensons les techniques les plus significatives suivantes : Les histogrammes[Chen, 2003], la DFT (Discrete Fourier Transform) [Agrawal, 1993], la SVD (Singular Value Decomposition) [Wu, 1996], la DWT (Discrete Wavelet Transform) [Chan, 1999], la PAA (Piecewise Aggregate Approximation) [Keogh, 2001, a], la PLA (Piecewise Linear Approximation), [Korn, 1997], [Li, 1997], [Nygaard, 1997] et la SAX (Symbolic Aggregate approXimation) [Lin, 2007]. Ces techniques sont décrites en détail dans le chapitre II de cette thèse dans le cadre de la caractérisation des séries temporelles.

III.3.3 Techniques d'alignement des ST

Les techniques d'alignement mettent en correspondance un ou plusieurs points d'une série temporelle avec un ou plusieurs points de l'autre série. Elles ont l'avantage de préciser où les deux séries correspondent contrairement aux techniques de comparaison précédentes qui donnent seulement des degrés de ressemblance globale entre les séries. Parmi les techniques d'alignement dans la littérature, nous recensons comme techniques les plus significatives : la méthode DTW (Dynamic Time Warping) [Sakoe, 1978] et [Berndt 1994], la méthode LCSS (Longest Common Subsequence) [Crochemore, 1994], la méthode SEA (Shape Exchange Algorithm) [Boucheham, 2008]. Ces méthodes sont présentées dans ce qui suit.

III.3.3.1 Méthode DTW (Dynamic Time Warping)

La méthode de déformation temporelle dynamique (DTW) a été proposée pour appairer des signaux dans le domaine de la reconnaissance de la parole [Sakoe, 1978]. Berndt et Clifford [Berndt, 1994] ont ensuite proposé d'employer cette technique pour mesurer la similarité des séries temporelles dans le domaine de la fouille des séries temporelles. D'autres travaux récents ont également employé cette mesure de similarité dans la fouille des séries temporelles [Keogh 2000, Park 2000b], la vérification des signatures en ligne [Sharma, 2017], l'analyse de similarité des séquences de protéines [Wenbing, 2017], la classification des séries temporelles [Petitjean, 2016], la reconnaissance des gestes humains (Gesture recognition) [Jambhale, 2014] et aussi la reconnaissance de l'écriture [Li, 2013].

La DTW est connue pour être la meilleure mesure de similarité pour de nombreuses applications dans le domaine des séries temporelles [Silva, 2018], [Movchan, 2015]. Elle a la capacité de gérer les décalages temporels qui peuvent éventuellement exister entre deux séries. Au lieu de comparer chaque point d'une série avec celui d'une autre série qui intervient au même instant t , elle permet de comparer chaque point d'une série avec un ou plusieurs points de l'autre, ceux-ci pouvant être décalés dans le temps comme dans la Figure 20.

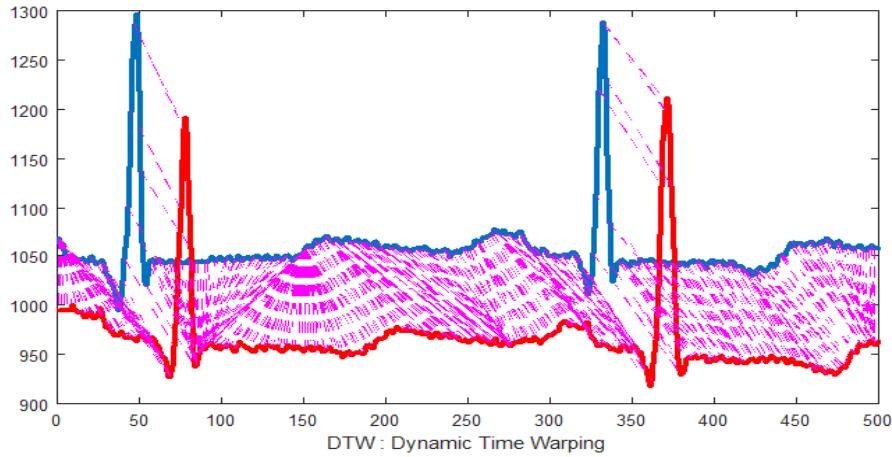


Fig. 20. Illustration de l'alignement effectuée par DTW. Dans ce cas, la DTW fait un alignement presque parfait malgré le décalage temporel.

Le fait que la distance DTW soit élastique et autorise des dilatations et des compressions dans l'axe du temps, permet à une série temporelle requête de correspondre à plusieurs autres séries temporelles de différentes tailles. Cette caractéristique permet d'utiliser la distance DTW dans la reconnaissance de formes où l'objectif est de conclure si une série donnée est similaire à un prototype bien défini.

Pour calculer la distance DTW entre deux séries temporelles X et Y , deux étapes principales sont nécessaires. Premièrement, une matrice de distances aussi appelée matrice des coûts ou matrice cumulative doit être calculée. Pour cela des distances prédéfinies (le plus souvent la distance Euclidienne) entre tous les éléments de X et Y doivent être calculées. Chaque entrée de la matrice de distances $D(i, j)$ correspond à la distance Euclidienne entre une paire de points (x_i, y_i) .

Soit $D(i, j)$ la matrice de distance entre les séquences x_1, x_2, \dots, x_i de $X(i)$ et y_1, y_2, \dots, y_i de $Y(j)$ avec $1 \leq i \leq m$ et $1 \leq j \leq n$ alors :

$$D(i, j) = \begin{cases} |x_i - y_i| & \text{Si } i=j=1 \\ |x_i - y_j| + \min \begin{pmatrix} D(i-1, j) \\ D(i-1, j-1) \\ D(i, j-1) \end{pmatrix} & \text{Si non} \end{cases} \quad (27)$$

Ensuite le chemin de déformation optimale doit être trouvé. Le chemin de déformation W doit satisfaire plusieurs contraintes:

- *Borne.* $w_1 = (1, 1)$ et $w_K = (m, n)$, pour forcer les points terminaux de chaque série à correspondre.

- *Continuité.* Etant donné $w_k = (a, b)$ alors $w_{k-1} = (a', b')$ où $a-a' \leq 1$ et $b-b' \leq 1$, pour obliger le chemin de déformation à aller vers les cellules adjacentes.

- *Monotonie.* Etant donné $w_k = (a, b)$ alors $w_{k-1} = (a', b')$ où $a-a' \geq 0$ et $b-b' \geq 0$ pour forcer les points de W à être espacés de façon monotone dans le temps.

Il y'a plusieurs Chemins de déformation qui peuvent satisfaire cette condition, mais on prend celui qui minimise la distance cumulative (DTW) comme un appariement entre les deux séries temporelles (voir Figure 21). L'alignement optimal correspond à la distance cumulative minimale :

$$DTW(X, Y) = \min \left\{ \sqrt{\frac{\sum_{k=1}^K w_k}{K}} \right\} \quad (28)$$

Le $k^{ème}$ élément de W est défini par :

$$W = w_1, w_2, \dots, w_k, \dots, w_K \quad (29)$$

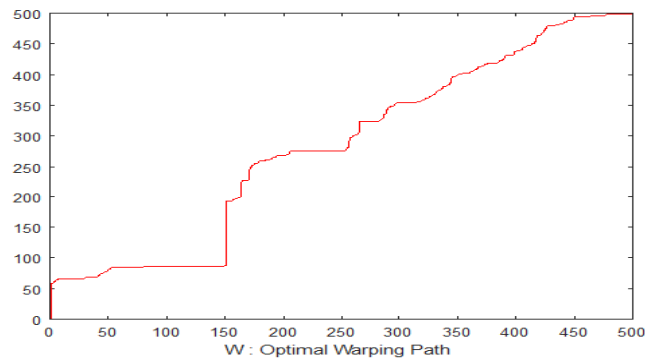


Fig. 21. Chemin de déformation optimal de la DTW.

Dans l'exemple représenté par la Figure 22, la matrice cumulative M est présentée. Les séquences à comparer $X = (5, 8, 9, 7)$ et $Y = (7, 5, 8, 7, 8)$ sont présentées en dessous de la matrice. Après le calcul de la matrice cumulative, on calcule le chemin de déformation optimal W et on obtient les séquences X' et Y' . Les valeurs des cellules sont calculées comme suit:

1. Première cellule : $M[1,1] = [x_1 - y_1]$
2. Première ligne : $M[1, j] = [x_j - y_1] + M[1, j - 1]$
3. Première colonne: $M[i, 1] = [x_1 - y_i] + M[i - 1, 1]$
4. Tous les autres éléments ($i, j > 1$)
 $M[i, j] = [x_j - y_i] + \min(M([i - 1, j - 1], M([i - 1, j), M([i, j - 1)$
5. L'appariement choisi pour la DTW est celui qui minimise la distance Euclidienne.

| X/Y | 5 | 8 | 9 | 7 |
|-----|----|---|---|---|
| 7 | 2 | 3 | 5 | 5 |
| 5 | 2 | 5 | 7 | 7 |
| 8 | 5 | 2 | 3 | 4 |
| 7 | 7 | 3 | 4 | 3 |
| 8 | 10 | 3 | 4 | 4 |

$X = (5, 8, 9, 7)$ et $Y = (7, 5, 8, 7, 8)$

Fig. 22. La matrice des distances cumulées entre les séquences X et Y.

Dans cet exemple, $DTW(X, Y) = 4$. Elle est représentée par la valeur de la dernière cellule. Les cellules en couleurs grises représentent le chemin de déformation optimal "W". Ils indiquent les valeurs répliquées dans les deux séquences et l'appariement entre les deux séquences. Les valeurs proches dans les deux séquences sont répliquées, comme la valeur 5 de la séquence X qui est répliquée deux fois dans la séquence X' et mise en correspondance avec les valeurs 5 et 7 de la séquence Y'

On peut aussi paramétrer l'écart temporel maximum permis à la mesure pour comparer deux Points. On définit ainsi une fenêtre temporelle (appelée δ) que l'on fera "glisser" sur chacune des séries à comparer. Par exemple, si on fixe la taille de la fenêtre $\delta = 3$, chaque point d'une série qui intervient à un instant t ne pourra être comparé qu'avec les points de l'autre série qui interviennent aux instants $t \in [t-3, t+3]$. Le choix de la taille de cette fenêtre peut avoir une influence sur le résultat et il convient de la fixer avec soin, une fenêtre temporelle trop large ou trop fine peut nuire à la qualité de la classification [Gaudin, 2005].

La distance par déformation temporelle (DTW) permet de résoudre un certain nombre de problèmes posés par l'utilisation des distances de comparaison. Cependant, elle souffre principalement de la complexité temporelle qui est quadratique.

Sur le plan de la pertinence, La DTW est sensible aux valeurs extrêmes (outliers) puisqu'elle met en correspondance toutes les valeurs des séquences comparées.

Un autre inconvénient de la DTW est qu'elle ne peut pas traiter les problèmes de déphasage significatifs dans les séries temporelles quasi-périodiques.

III.3.3.2 Distances d'édition

La première distance d'édition a été établie entre deux séquences alphanumériques (dites aussi textuelles). Elle a été définie par Wagner et Fischer [Wagner, 1972] comme étant le nombre d'opérations minimum nécessaire pour transformer une requête donnée en un motif prédéfini.

Plus formellement, étant donné deux séquences (chaînes) $X = x_1x_2x_3 \dots x_n$ et $Y = y_1y_2y_3\dots y_m$, le problème de la distance d'édition est de trouver une série d'opérations d'édition avec un coût minimal pour transformer X en Y . Les opérations d'édition autorisées incluent : *insertion* de caractère, *suppression* de caractères et *remplacement* de caractères.

Soit $D(i, j)$ la distance minimale entre les deux séquences $X(i)$ et $Y(j)$. On définit récursivement cette distance comme suit :

$$D(i, j) = \begin{cases} j & \text{si } i = 0 \\ i & \text{si } j = 0 \\ D(i-1, j-1) & \text{si } i, j > 0 \text{ et } x[i] = y[j] \\ \min \left\{ \begin{array}{l} D(i-1, j-1) + 1, \\ D(i-1, j) + 1, D(i, j-1) + 1 \end{array} \right\} & \text{si non} \end{cases} \quad (30)$$

La formule (30) induit un algorithme de recherche de la distance d'édition entre deux séquences de texte X et Y de longueur N et M avec une complexité d'ordre $O(mn)$, donc, quadratique.

Chen et Ng [Chen, 2004] combinent des mesures de similarité basées sur l'édition et des normes L_p et nomment la mesure de similarité résultante : *distance d'édition avec pénalité réelle* (ERP).

La Distance d'édition avec pénalité réelle (**ERP**) est robuste contre les imperfections de données et correspond au nombre d'opérations d'insertion, de suppression et de remplacement nécessaires pour changer X en Y [Chen, 2004]. De plus, il existe des combinaisons et des extensions de mesures de similarité basées sur l'édition avec d'autres mesures de similarité. Marteau [Marteau, 2009] étend les mesures de similarité basées sur l'édition et présente la distance d'édition du Time Warp (**TWED**) qui est un algorithme de programmation dynamique pour les opérations d'édition.

III.3.3.3 Méthode LCSS (Longest Common Subsequence)

Une forme particulière de l'algorithme de la distance d'édition renvoie la longueur de la plus longue sous séquence commune entre deux séquences (LCSS) ou (PLSC) [Crochemore, 1994]. Cette forme considère que $D(i, j)$ est le nombre minimal d'effacements, et d'insertions, mais pas d'échanges nécessaires pour transformer un texte $X[i]$ en $Y[j]$. C'est une version restreinte de la distance d'édition où les échanges ne sont pas permis.

La mesure de distance de la plus longue sous-séquence commune (LCSS) utilise la programmation dynamique pour rechercher la plus longue sous-séquence commune entre des préfixes qu'elle allonge tant qu'une similarité est reconnue.

Elle a été proposée par Agrawal *et al* [Agrawal, 1995] pour mesurer la similarité entre les séquences temporelles. Agrawal *et al* se basent sur le nombre de paires de sous-séquences temporelles ordonnées dans le temps, qui sont semblables, mais qui ne se chevauchent pas. S'il y en a assez les séquences seront considérées comme similaires. La Figure23 illustre cette méthode.

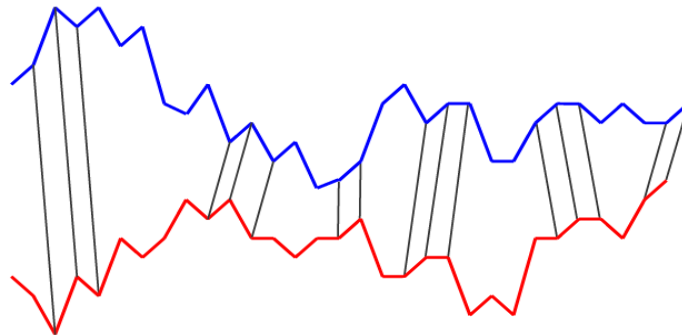


Fig.23. Comparaison de deux séries temporelles à l'aide de la méthode LCSS.

Pour affiner la sélection des sous-séquences communes, on peut paramétrer une fenêtre temporelle de la même manière qu'avec la distance DTW. On doit aussi fixer une fenêtre spatiale (appelée epsilon ϵ) qui servira à définir quel est l'écart maximum toléré pour pouvoir considérer deux sous séquences comme similaires. La méthode compare ensuite chaque point d'une série avec 0 ou 1 point de l'autre, contrairement à la DTW qui compare chaque point d'une série avec k points de l'autre série.

La LCSS est donc une distance élastique qui permet à deux séquences de s'étirer afin de faire correspondre les éléments identiques entre eux. Sa complexité temporelle est de $O(n^2)$ pour des séquences de tailles n .

Plus formellement, soit $D(i, j)$ la distance entre la sous-séquence $X=x_{i1}, x_{i2}, \dots, x_{in}$ et $Y=y_{j1}, y_{j2}, \dots, y_{jm}$ avec $1 \leq i \leq m$ et $1 \leq j \leq n$.

$$D(i, j) = \begin{cases} 0 & \text{Si } i=0 \text{ ou } j=0 \\ 1 + d(i-1, j-1) & \text{Si } |p_i - f_j| < \varepsilon \\ \max \left(\begin{matrix} D(i-1, j) \\ D(i, j-1) \end{matrix} \right) & \text{Sinon} \end{cases} \quad (31)$$

Yazdani et al. [Yazdani, 1996] ont proposé un algorithme LCSS spécialement pour les séquences d'images, qui réduit la complexité temporelle en $O(n+m)$. Ils l'ont ensuite généralisé aux valeurs réelles des signaux pour les appliquer sur les séries temporelles qui représentent des ensembles de caractéristiques comme les coefficients de Fourier. Dans [Yazdani, 1996], chaque image est représentée par un ensemble de coefficients de Fourier.

La méthode LCSS autorise les changements d'échelles et les décalages. Elle est principalement utilisée pour traiter les données bruitées, car elle ignore les valeurs extrêmes en comptabilisant seulement les sous-séquences communes, contrairement aux autres méthodes qui comparent directement les séquences temporelles. Cependant, la méthode présente les inconvénients suivants : le nombre de ces sous-séquences n'est pas évident à fixer, elle ne respecte pas l'inégalité triangulaire et elle souffre d'une complexité temporelle assez élevée.

Des extensions ou des modifications existent. Par exemple, Morse et Patel [Morse, 2007] présentent la méthode d'évaluation rapide des séries temporelles utilisée pour évaluer les valeurs seuils de LCSS et ses modifications. Wang et al [Wang, 2013] soulignent que les mesures élastiques telles que DTW, LCSS, EDR et ERP sont significativement plus précises que les autres mesures. Ils constatent aussi que DTW est plus précise que les autres méthodes, mais pour de plus petites séries temporelles.

III.3.3.4 Méthode SEA (Shape Exchange Algorithm)

La méthode SEA (Shape Exchange Algorithm) proposée dans [Boucheham, 2008] serait, selon son auteur une technique d'alignement pertinente surtout pour les séries temporelles quasi-périodiques [Boucheham, 2008]. Ces séries temporelles sont des concaténations de motifs quasi périodiques appelés pseudo-périodes, comme les tracés de l'électrocardiogramme (ECG) et ceux du signal Capnogram. L'auteur prouve que la méthode

SEA est plus précise, plus rapide et qu'elle consomme moins de mémoire que la méthode DTW pour l'alignement des séries temporelles quasi-périodiques [Boucheham, 2008]. Elle est aussi simple à mettre en œuvre et ne nécessite aucun paramètre. La clé de cet algorithme se résume dans l'échange de signatures entre les séries temporelles. Ces signatures sont établies en triant les séries sur leurs amplitudes. Cette étape permet de prendre en charge les variations d'amplitudes contrairement à la DTW qui ne résout que les problèmes de distorsion de l'axe temporelle. La méthode procède par la suite au tri des signatures échangées sur leurs coordonnées temporelles respectives. Finalement la série temporelle d'origine est comparée point par point avec la série reconstruite obtenue par le processus inverse en utilisant la signature de l'autre série temporelle. L'alignement est tel que les périodes d'une série temporelle sont mises en correspondance avec les périodes de l'autre série temporelle, même si les séries contiennent un nombre différent de périodes [Boucheham, 2008].

Pour les séries de longueurs différentes, une autre étape est nécessaire avant l'échange de signatures. Une technique de correspondance linéaire a été utilisée par Boucheham [Boucheham, 2008] pour forcer la plus longue série temporelle à se rétrécir en sautant le pas de parcours de l'axe temporel jusqu'à ce qu'elle puisse correspondre à la série la plus courte ; et vice versa, la série la plus courte va s'étendre pour correspondre à la série la plus longue.

Selon Boucheham [Boucheham, 2008], la complexité temporelle globale de SEA est de $O(2(n^2 + m^2) + 2(n + m))$, où n et m sont les tailles respectives des séries temporelles, donc d'ordre quadratique et sa complexité spatiale est de $O(2(n + m))$, donc d'ordre linéaire. La méthode SEA s'avère plus avantageuse que la méthode DTW, car elle consomme moins d'espace mémoire (la complexité spatiale de DTW étant quadratique).

III.4 Classification des méthodes d'appariement selon la nature des données

Boucheham [Boucheham, 2013] propose une catégorisation des techniques d'alignement des ST existantes en fonction de la complexité des données dans ces mêmes ST. Ainsi, les classes des méthodes d'appariement des séries temporelles sont rapportées dans l'ordre ascendant de cette complexité comme suit :

Classe (a): une seule période sans déphasage (ou non-périodique)

Les méthodes dans cette classe peuvent traiter les séries temporelles ayant une seule période avec la condition que les séries temporelles ne soient pas significativement déphasées les unes par rapport aux autres (e.g., Figure. 24 (a)). Ces méthodes incluent, à titre d'exemple, la DFT, LCSS, ED et DTW.

Classe (b): une seule période avec déphasage:

Les méthodes dans cette classe peuvent traiter les séries temporelles quasi-périodiques ayant une seule période avec la possibilité que les séries soient significativement déphasées les unes par rapport aux autres (e.g., Figure. 24. (b)). La méthode Phase Independent DTW (PI-DTW) [Keogh, 2006] est dans cette classe.

Classe (c): Périodique avec plusieurs périodes déphasées

Les méthodes dans cette classe peuvent traiter les séries temporelles quasi-périodiques composées de plusieurs périodes chacune, mais pas nécessairement de même nombre de périodes. En effet, les séries temporelles qui contiennent le même nombre de périodes ramènent les données à la classe (a) s'il n'y a pas de déphasage (significatif) et à la classe (b) s'il y a un déphasage (significatif). Ainsi, les types de données spécifiques à cette classe sont des séries temporelles qualifiée par Boucheham [Boucheham, 2013] de complexes, car elles peuvent être déphasées, en plus d'être composées d'un nombre différent de périodes chacune, sans parler des autres difficultés citées ci-dessus (e.g., Figure 24. (c)). Selon Boucheham [Boucheham, 2013], les méthodes de cette classe seraient exclusivement : SEA, ASEAL et FANSEA, proposées par Boucheham dans [Boucheham, 2008, 2010, 2013]. FastSEA et LMDS-SEA proposées par Lahreche et Boucheham dans [Lahreche, 2017] et [Lahreche, 2018], la méthode proposée par Betancourt et al qui est à base d'un Fuzzy-SEA [Betancourt, 2013] seraient aussi dans cette classe. Dans ce cadre d'analyse, nous suggérons dans ce travail de thèse une méthode d'alignement des séries temporelles quasi-périodique dénommée «QP-DTW» (Quasi-Periodic Dynamic Time Warping) [Boulnemour, 2018] (Chapitre VI de cette thèse) qui serait elle aussi dans cette classe.

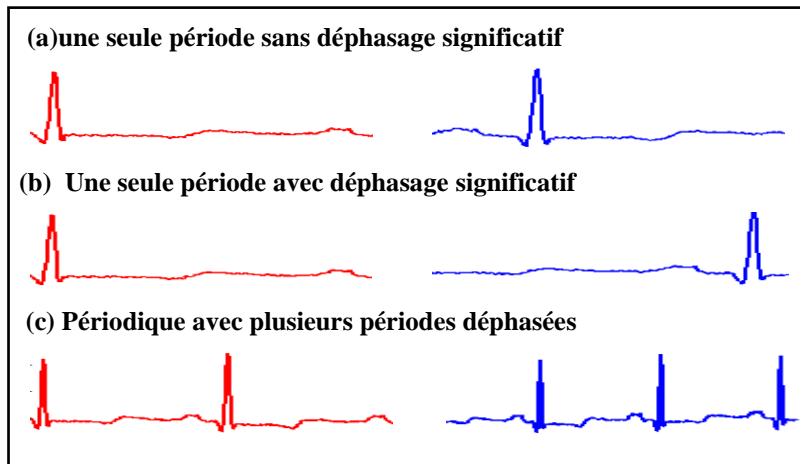


Fig. 24. Difficultés au niveau de l'appariement des séries temporelles quasi-périodiques.

La Figure 25 [Boucheham, 2013] illustre les classes des méthodes d'appariement des séries temporelles organisées en ordre d'inclusion. On voit que DTW ne peut pas traiter les classes b et c, alors que SEA et QP-DTW peuvent traiter tous les cas.

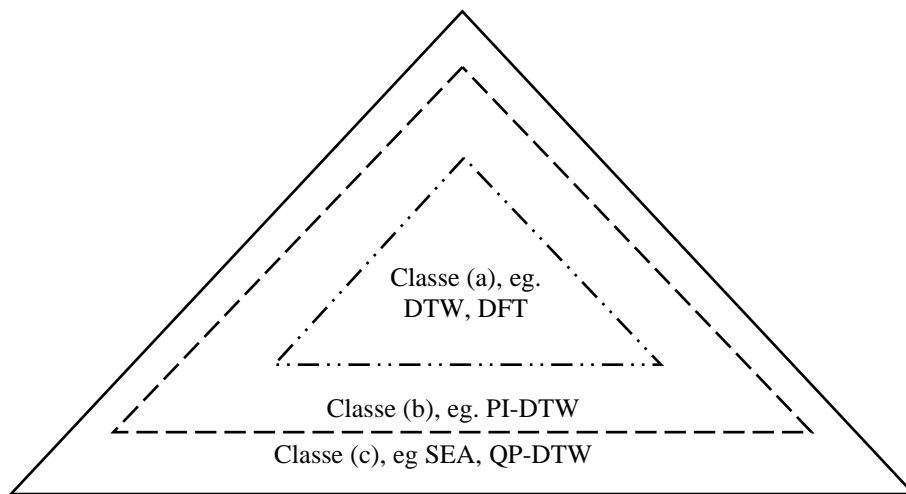


Fig. 25. Classification des méthodes d'appariement des séries temporelles selon la nature des données. Notez que les méthodes dans la classe (c) peuvent traiter les données dans les classes (a), (b) et (c), les méthodes dans (b) peuvent traiter les données dans (a) et (b) et les méthodes dans (a) peuvent traiter les données dans (a) seulement [Boucheham, 2013].

III.5 Techniques d'alignement avancées

III.5.1 Les améliorations de la méthode DTW

Sachant qu'il est admis que l'inconvénient majeur de la DTW est sa complexité temporelle, une multitude de méthodes pour accélérer la DTW ont été proposées, mais peu

d'entre elles essayent d'améliorer sa précision d'alignement [Keogh, 2001-c]. Nous présentons dans ce qui suit, les améliorations les plus récentes de la DTW.

III.5.1.1 Méthode EA_DTW (Early abandon dynamic time warping)

La méthode Early abandon dynamic time warping (EA_DTW) a été proposée par [Jun, 2009] pour accélérer le calcul de la DTW. La méthode admet une contrainte sur les valeurs des cellules dans la matrice des distances cumulée. La distance d'une cellule quelconque ne doit pas dépasser un certain seuil, sinon, le calcul se termine dans cette cellule. Plus formellement, pour $\epsilon > 0$, si la valeur de la cellule $\gamma(i, j) > \epsilon$, on appelle cette cellule débordement et on note #.

Les résultats d'analyse de [Jun, 2009] montrent que si une cellule fait un débordement, il faudrait remplacer sa valeur par une valeur quelconque au-dessus du seuil et omettre le calcul de la distance de base sans affecter le résultat final. En outre, si les cellules dans une ligne ou une colonne débordent, le calcul de DTW doit se terminer, et retourner un non-alignement. L'idée de la méthode EA_DTW est illustrée dans la Figure 26, où la valeur de ϵ est égale à 9. EA_DTW retourne « faux » si les deux séquences ne sont pas appariées, c'est à dire, la distance dépasse le seuil donné, et retourne « vrai » si elles sont mises en correspondance. Le signe ## indique que le résultat de calcul est terminé.

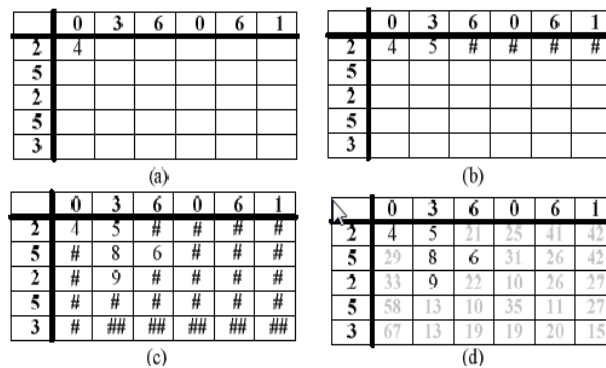


Fig.26. Illustration de la technique Early Abandon Dynamic Time Warping(EA_DTW) ; (a) Calculer le premier point et vérifier s'il déborde; (b) Calculer les marges de la matrice; (c) Calculer le corps de la matrice avec EA-DTW; (d) Les résultats des calculs comparatifs sans EA-DTW. ("#" = Dépassement, "##" = arrêter le calcul.)[Jun, 2009].

La complexité temporelle de EA_DTW est de $O(mn)$, où m et n sont égaux à la longueur des deux séquences, respectivement. Jun et al démontrent que EA_DTW est plus rapide que la DTW et ceux pour toutes les configurations du seuil, même si ce seuil excède la distance exacte.

III.5.1.2 Méthode de recherche exacte de sous-séquences avec la DTW

Le calcul de la DTW est trop lent pour être utilisé pour des bases de données massives. Mais Rakthanmanon et al [Rakthanmanon, 2012] ont proposé une nouvelle technique de recherche de sous-séquences exactes avec la DTW basée sur quatre nouvelles idées qu'ils appellent la suite UCR (aussi, UCR-DTW). Ils normalisent les sous-séquences des séries temporelles, ils appliquent ensuite la technique Early abandon dynamic time warping en réorganisant son ordre et finalement, ils implémentent une cascade des techniques de Lower Bound comme la LB-KIM [Kim, 2001] et la LB-Keogh [Keogh, 2005] et ils choisissent la plus fine et la plus rapide d'entre elles. Selon Rakthanmanon et al [Rakthanmanon, 2012], cette accélération de la DTW serait la plus efficace et la plus rapide pour la recherche dans les séries temporelles massives qui vont jusqu'à un trillion (un million de millions). Cette capacité de recherche permettrait à la méthode de traiter les problèmes en temps réel.

III.5.1.3 Méthode DDTW (Derivative Dynamic Time Warping)

La méthode DTW est capable d'aligner deux séquences similaires ayant des accélérations et des décélérations locales significativement différentes dans l'axe du temps. Cependant, elle a des problèmes lorsque les deux séquences diffèrent également dans l'axe des Y. Dans l'exemple de Keogh et al [Keogh, 2001-c] illustré par la Figure 27, deux séquences identiques produisent clairement un alignement parfait. Mais quand il y a une légère modification d'une caractéristique locale, en l'occurrence la profondeur d'une vallée, DTW tente d'expliquer la différence en termes d'axe du temps et produit deux singularités.

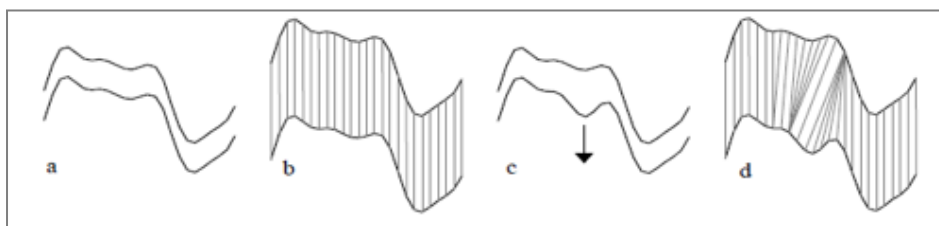


Fig.27. Illustration du problème de singularité de la DTW. En utilisant DTW, deux séquences identiques (a) produisent clairement un alignement parfait (b). Cependant, avec la légère modification de la profondeur d'une vallée (c), DTW tente d'expliquer la différence en termes d'axe de temps et produit deux singularités (d) [Keogh, 2001-c].

La méthode DDTW [Keogh, 2001-c] permet d'éviter ce problème en ne tenant pas compte des valeurs des points de données de Y, mais plutôt de la plus haute valeur de la forme. Ils obtiennent cette valeur en considérant la première dérivée des séquences.

Comme dans la DTW, une matrice n par m est construite où le $i^{\text{ème}}$ et le $j^{\text{ème}}$ élément de la matrice contiennent la distance $d(q_i, c_j)$ entre les deux points q_i et c_j . La différence entre DTW et DDTW est que la distance mesurée $d(q_i, c_j)$ n'est pas Euclidienne, mais c'est plutôt la différence au carré des dérivées estimées (D_x) de q_i et c_j .

$$D_x[q] = \frac{(q_i - q_{i-1}) + ((q_{i+1} - q_{i-1})/2)}{2} \quad 1 < i < m \quad (32)$$

Selon Keogh et al [Keogh, 2001-c], DDTW produit des alignements meilleurs que ceux de DTW et elle est plus robuste aux valeurs aberrantes que DTW.

III.5.1.4 La méthode Phase independant DTW(PI-DTW)

Un problème assez intéressant dans le cadre de la requête d'images d'objets par leurs formes (Shape Based Image Retrieval), formulés comme problème d'appariement des objets dans les images par leurs contours peut être résolu avec la DTW en transformant d'abord les formes 2D en signal 1D (voir Figure 28). Cependant, si une rotation est appliquée à l'un des objets, la tâche devient impossible à réaliser avec la DTW.

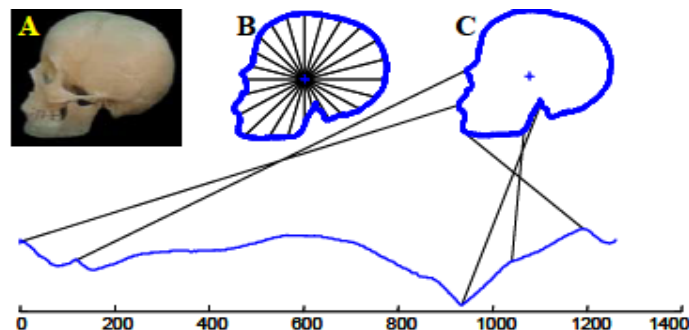


Fig.28.Création d'une série temporelle à partir de l'image du squelette d'un crâne humain avec le balayage radial. (VLDB '06, September 12–15, 2006, Seoul, Korea. Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09).

Pour résoudre ce problème, Keogh et al proposèrent la méthode Phase Independent DTW (PI-DTW) [Keogh, 2006]. Cependant, il est évident que leur méthode est très coûteuse, car elle met en œuvre des techniques d'indexation pour permettre la vérification de toutes les rotations possibles de formes. Cette méthode traite également les séries temporelles quasi périodiques; mais ces séries temporelles ne peuvent pas avoir un nombre différent de

périodes. Ils peuvent avoir soit une seule période, ou le même nombre de périodes chacune (classe b, selon la classification de Boucheham [Boucheham, 2013]).

III.5.1.5 Méthode Fast_DTW (Fast Dynamic Time Warping)

Salvador et Chan font aussi l'approximation de la DTW [Salvador, 2007]. Leur algorithme Fast_DTW utilise une approche multi-niveau qui, récursivement, projette une solution à partir d'une résolution grossière de la série temporelle à une résolution plus élevée, pour l'affiner encore plus par la suite (Voir Figure 29). L'algorithme a une complexité spatiale et temporelle linéaires.

L'algorithme Fast_DTW utilise une approche multi-niveau avec trois opérations clés:

- 1) Résolution grossière: Réduire une série temporelle en une série temporelle plus petite qui représente la même courbe possible avec moins de points de données.
- 2) Projection: Trouver un chemin d'alignement optimal à une résolution inférieure et l'utiliser comme point initial pour obtenir un chemin optimal à une résolution élevée.
- 3) Raffinement : Affiner le chemin de la chaîne projeté à partir d'une résolution plus faible grâce à des ajustements locaux du chemin de la chaîne.

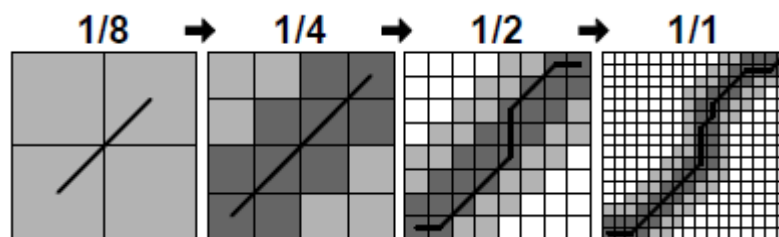


Figure. 29. Les quatre résolutions évaluées durant l'exécution complète de Fast-DTW

III.5.2 Améliorations de la méthode SEA :

III.5.2.1 La méthode Fuzzy-SEA

La technique proposée par Betancourt et al [Betancourt, 2013] est utilisée pour classifier le signal du Capnogramme et le signal de l'électrocardiogramme (ECG). Le signal ECG représente les variations de l'activité électrique du cœur en fonction du temps [Benali, 2013] et le signal du capnogramme représente la courbe de la concentration du CO₂ mesurée au niveau de l'air expiré par une personne. Si la personne respire bien, la courbe du Capnogramme est plutôt une suite de formes rectangulaires, mais si elle respire moins bien, le signal du Capnogramme perd sa forme normale [Betancourt, 2013] (Figure30).

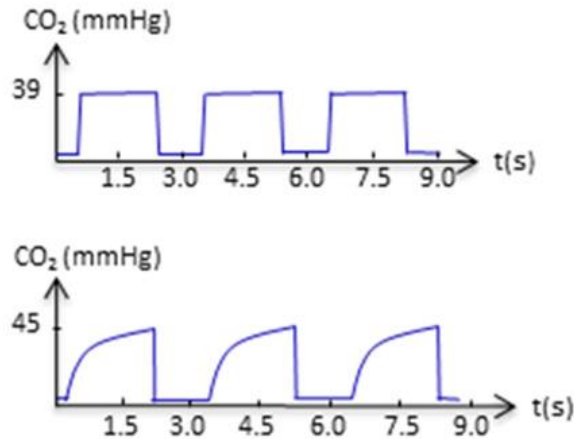


Fig.30.Forme du signal quasi-périodique du Capnogramme, forme normale (haut) et forme anormale (bas) [Betancourt, 2013].

Le signal à classifier est premièrement comparé avec deux groupes de signaux à l'aide de la méthode SEA de Boucheham [Boucheham, 2008]. Le premier groupe est composé de signaux normaux et le deuxième est composé de signaux anormaux. Les résultats de cette comparaison sont concrétisés par le *facteur de corrélation* et la *mesure de distorsion PRD (Percent Root Difference)*. Ensuite, en appliquant un système à inférence floue (*Fuzzy*) aux résultats de SEA, le signal entré est classifié en un signal normal ou anormal.

D'après Betancourt *et al* [Betancourt, 2013], la combinaison de SEA avec le système à inférence floue serait utile pour classifier le signal ECG et sa complexité temporelle serait quadratique [Betancourt, 2013].

III.5.2.2 Méthode ASEAL (Approximate Shape Exchange ALgorithm)

La méthode ASEAL est basée sur la combinaison de deux algorithmes : une version modifiée par Boucheham et al [Boucheham, 2007] de l'algorithme de simplification de lignes de Douglas-Peucker [Douglas, 1973] pour la réduction des données dans les séries temporelles et de l'algorithme SEA pour l'alignement des séries temporelles quasi périodiques. Selon son auteur, la méthode ASEAL aurait obtenue des résultats d'alignements comparables à ceux de la méthode SEA tout en réduisant drastiquement sa complexité temporelle.

III.5.2.3 Méthode FANSEA

La méthode FANSEA est une hybridation entre l'algorithme de simplification de ligne FAN [Gardenhire, 1964] et l'algorithme SEA (Shape Exchange Algorithm). Elle a été proposée par Boucheham [Boucheham, 2013]. La méthode FANSEA est une autre version accélérée de la méthode SEA.

Alors que la méthode SEA effectue l'alignement sur les séries temporelles entières, FANSEA procède d'abord à l'extraction du plus petit nombre de points, les plus significatifs des deux séries temporelles extraites par FAN respectant la contrainte d'erreur maximale tolérée, sur lesquelles la mise en correspondance est réalisée. Par conséquent, la méthode FANSEA est composée de deux phases principales. Dans la première phase, il y'a l'extraction des points dominants (Dominant Points, DPs) des deux séries temporelles. La Figure 31 montre l'exemple d'une série temporelle compressée à l'aide de la méthode FAN. Dans la deuxième phase, la méthode d'appariement SEA est appliquée aux deux ensembles de points dominants extraits à partir des deux séries temporelles à aligner.

L'auteur compare cette méthode à la méthode SEA, sur des tracés ECG qui représentent des données quasi-périodiques bruitées et présentant des variabilités locales significatives. Les résultats obtenus montrent que la méthode est aussi efficace qu'effective pour l'alignement des séries temporelles, même très longues (*long Time Series*), car elle consomme beaucoup moins de temps que SEA, pour des alignements presque identiques.

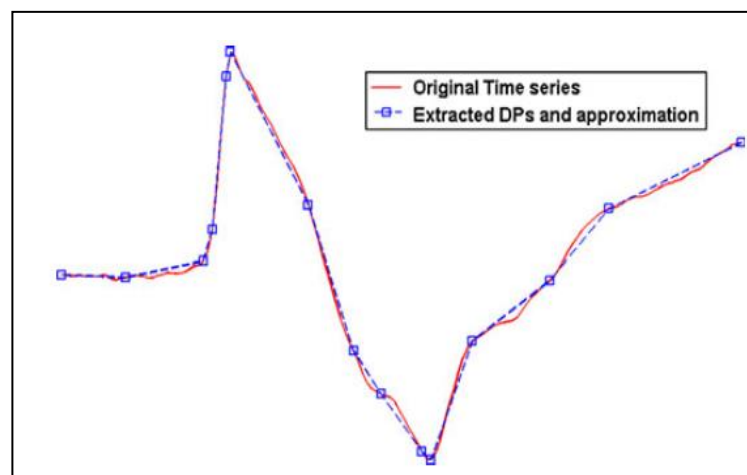


Fig.31.Exemple de série temporelle compressée par la méthode FAN avec 15 points dominants (DP) calculés sur sa courbe. Les points dominants sont des points perceptivement attractifs sur la courbe des séries temporelles [Boucheham, 2013].

III.5.2.4 Méthode FastSEA (Fast Shape Exchange Algorithm)

La méthode FastSEA [Lahreche, 2017] est une accélération de la méthode SEA sans perte de précision contrairement aux méthodes ASEAL et FANSEA. Elle permet d'avoir un temps de calcul linéaire en appliquant un algorithme de tri plus efficace dans l'implémentation de l'algorithme SEA. L'algorithme de tri utilisé est l'algorithme de tri par comptage (counting sort) qui est beaucoup plus rapide ($O(n)$) et plus simple que l'algorithme de tri par sélection utilisé dans SEA. La méthode FastSEA vise à aligner de très longues séries temporelles en de très courts temps d'exécution.

III.5.2.5 Méthode LMDS-SEA

La méthode LMDS-SEA [Lahreche, 2018] est une autre extension de la technique SEA qui se base sur le principe de la *Selection de Distance (Distance Selection)* [Mosbah, 2017]. Elle consiste à diviser les séries temporelles en segments (sous-séquences). Le but étant d'extraire des informations locales (*signatures locales*) des séries temporelles. La longueur des segments appropriée est celle qui donne les meilleurs résultats en termes de précision de classification à l'étape d'apprentissage. Par conséquent, la longueur des segments peut différer d'un ensemble de données (classe) à l'autre. Elle est comprise entre 1 et n (n étant la longueur de toute la série temporelle).

Après l'étape d'apprentissage, une étape de sélection de distance est effectuée. Trois cas sont distingués à ce niveau:

1) la longueur de chaque segment est 1 : cela signifie qu'il existe une segmentation maximale de la série temporelle. Dans ce cas, les deux séries temporelles sont comparées en utilisant la distance Euclidienne.

2) La longueur du segment est n : cela signifie qu'il n'y a qu'un seul segment de la série temporelle. Dans ce cas, la méthode SEA « globale » est utilisée.

3) $1 < l < n$: dans ce cas, la mesure de similarité est sélectionnée automatiquement entre la distance Euclidienne, la méthode SEA « globale » et la méthode SEA « locale » calculée pour chaque segment.

La dernière étape consiste à utiliser les mesures de similarités choisies pour la classification 1NN des séries temporelles.

III.6 Conclusion

Dans ce chapitre, nous avons proposé une revue des techniques d'appariement des séries temporelles. Ensuite, nous avons présenté une classification de ces techniques, selon la nature des données, comme proposée dans [Boucheham, 2013].

Enfin, nous avons présenté les améliorations des techniques d'appariement DTW et SEA, qui sont utilisées pour rendre la recherche par similarité de séquences temporelles et la classification des séries temporelles plus rapide et/ou plus pertinentes.

Chapitre IV. Recherche de similarité dans les séries temporelles

IV.1 Introduction

Etant donné un ensemble d'éléments et un critère de similarité, la recherche par similarité/proximité revient à déterminer dans cet ensemble les éléments similaires/assez proches d'une requête donnée. Ce problème est alors une extension naturelle de la recherche exacte [Peredes, 2009].

La recherche par similarité dans les séries temporelles a plusieurs applications, telles que la recherche dans les données multimédias [Zhu, 2003], la recherche de séquences d'ADN [Noé, 2005], etc. Dans ce type d'applications, la réponse à la requête donnée n'est pas exacte. Car, simplement, la probabilité que deux images soient exactement similaires est presque nulle, sauf si elles sont des copies numériques de la même source, également dans la fouille des données musicales, les musiciens jouent les mélodies différemment, donc rechercher des correspondances exactes est peu utile.

Dans ce chapitre, nous allons nous intéresser au problème de la recherche par similarité dans les bases de séries temporelles. Nous présentons, en particulier, les différents types de requêtes de similarité et les applications de la recherche de similarité dans les séries temporelles.

IV.2 Recherche de similarité dans les séries temporelles

Selon Faloutsos et al [Faloutsos, 1994], Terano et al [Terano, 2000], Keogh et al [Keogh, 2001-a], il y a essentiellement deux façons de rechercher la similarité dans les séries temporelles : *L'appariement complet (Whole sequence Matching)* et *l'appariement de sous-séquences (Sub-Sequence Matching)*. L'appariement complet nécessite que les séquences à comparer X et Y aient la même longueur [Keogh, 2001-a]. L'appariement de sous-séquences exige que la requête X soit comparée à chaque décalage possible avec la longue séquence Y [Terano, 2000].

Une troisième catégorie basée intervalles pour la recherche de similarité a été proposée par Aßfalg et al [Aßfalg, 2007]. Les auteurs spécifient des intervalles de temps pertinents comme paramètre supplémentaire et comparent les objets de la série temporelle uniquement à l'intérieur de cet intervalle de temps défini par l'utilisateur.

IV.3 Problème de la recherche de similarité dans les séries temporelles

La recherche de similarité est limitée à cause de la grande taille des BDD des séries temporelles. La meilleure solution, selon [Keogh, 2001-a] serait d'utiliser une technique de réduction de la dimensionnalité sur les données telle que la DFT, DWT, SVD, PAA, PLA, ou SAX, puis à indexer les données réduites avec une méthode d'accès spatiale (structure multidimensionnelle) telles que les arbres R (R-Trees). Après application de la méthode d'accès spatiale aussi appelée structure multidimensionnelle, les données sont transférées de leur espace originelle vers un espace d'index.

Selon Keogh et al [Keogh, 2001-a], la plupart des techniques d'indexation pourrait souffrir de deux problèmes : les fausses alarmes (False alarms) et les faux rejets (False dismissals). Les fausses alarmes se produisent lorsque des objets qui semblent proches dans l'espace indexé sont en réalité distants. Parce que les fausses alarmes peuvent être supprimées au cours d'une étape de post-traitement, elles peuvent être tolérées dans la mesure où elles sont relativement peu fréquentes. En revanche, les faux rejets sont généralement inacceptables, lorsque des objets qui répondent aux critères choisis sont ignorés parce qu'ils semblent être éloignés dans l'espace indexé [Keogh, 2001-a]. Selon Faloutsos et al [Faloutsos, 1994], la solution pour garantir qu'il n'y ait pas de faux rejets, serait que, la mesure de distance soit métrique et satisfasse la condition suivante dans l'espace d'indexation:

$$D_{\text{espace d'index}}(A, B) \leq D_{\text{espace originelle}}(A, B) \quad [33]$$

IV.3.1 Techniques d'optimisation des indexes des BDD des séries temporelles

Ces structures sont utilisées pour réduire la taille des séries temporelles. Nous en rapportons quelques une dans ce qui suit :

IV.3.1.1 Structure de données multidimensionnelle (Arbre R, B, ...)

L'arbre R (R-Tree) est une structure de données hiérarchique qui permet d'indexer les objets [Guttman, 1984]. L'arbre R a ensuite été utilisé par les chercheurs pour offrir un

traitement efficace des requêtes dans des ensembles de données multidimensionnelles telles que les séries temporelles. L'arbre R est capable de traiter divers types d'objets, en utilisant des approximations. Cela signifie qu'un objet est approximé par son rectangle (R) de délimitation minimum afin de fournir une étape de filtrage efficace. Les objets qui survivent à l'étape de filtrage sont inspectés de façon plus approfondie pour vérifier leur pertinence lors de l'étape de raffinement [Papadopoulos, 2009].

L'arbre B appelé également B-tree ou arbre de Bayer [Bayer, 1972] est un arbre équilibré (Balanced tree) dont tous les nœuds feuilles sont au même niveau et dont tous les chemins de la racine aux nœuds feuilles ont la même longueur.

IV.3.1.2 Bornes inférieures (Lower Bound)

Les mesures de bornes inférieures (Lower Bound, LB) sont des fonctions qui contribuent à accélérer le calcul des mesures de distance et à optimiser l'indexation des BDD des ST.

Selon Ratanamahatana et Keogh [Ratanamahatana, 2004-b] Il n'y aurait que deux propriétés souhaitables pour avoir une LB d'une distance donnée:

- La LB doit être rapide à calculer. Une mesure dont le calcul prend autant de temps que la mesure initiale est peu utile.
- La LB doit être relativement serrée et elle doit se rapprocher le plus possible de la distance réelle.

Les mesures de LB sont nombreuses mais les plus connues d'entre elles sont : la LB_Kim [Kim, 2001], la LB_Yi [Yi 1998] et la LB_Keogh [Keogh, 2002].

La **LB_Yi** [Yi 1998], illustrée par la figure 32 est définie par la formule suivante :

$$LB_Yi(Q, C) = \sum_{q_i > \max(C)} |q_i - \max(C)| + \sum_{c_j < \min(Q)} |c_j - \min(Q)| \quad [34]$$

Si les longueurs des séries temporelles sont m et n, respectivement, La fonction LB_Yi obtient leur LB en O (m+n). Elle a la capacité à traiter des séries temporelles de longueurs différentes. Cependant, malgré que LB_Yi soit une fonction de borne inférieure pour plusieurs distances, selon [Bartos, 2012], elle ne l'est pas toujours pour la distance DTW [Sakoe, 1978], [Berndt, 1994] car elle donne une distance supérieure à celle de DTW.

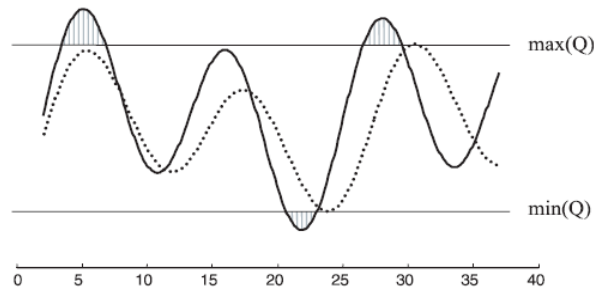


Figure.32. Une intuition visuelle de la LB_Yi. La somme de la longueur carrée des lignes grises représente la mesure de borne inférieure [Keogh, 2002].

La fonction **LB_Kim** [Kim, 2001], illustrée par la Figure33 et définie par l'équation (32), quant à elle, extrait un vecteur de caractéristique à quatre dimensions de chaque séquence. Les caractéristiques sont le premier et le dernier élément de la séquence, avec le maximum et le minimum de celle-ci. La différence carrée maximale des caractéristiques correspondantes est présentée comme la borne inférieure. Elle est défini par la formule suivante :

$$LB_Kim(Q, C) = \max\left(|q_{premier} - c_{premier}|, |q_{dernier} - c_{dernier}|, |q_{min} - c_{min}|, |q_{max} - c_{max}|\right) \quad [35]$$

La LB_Kim est selon son auteur une fonction de borne inférieure pour plusieurs distances et elle fonctionne avec des séries temporelles de longueurs différentes.

Selon Bartos et al, la LB_Kim ne réduit pas toujours la mesure de similarité DTW [Bartos, 2012].

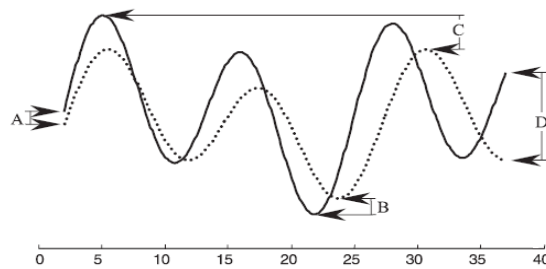


Figure. 33. Une intuition visuelle de la LB_Kim. La différence carrée maximale entre les deux premières séquences (A), la dernière (D), minimale (B) et le maximum de points (C) est retournée comme la mesure de borne inférieure [Keogh, 2002].

La fonction de borne inférieure **LB_Keogh** [Keogh, 2005] a été introduite par Keogh et al [Keogh, 2005]. Les auteurs prouvent à travers plusieurs expériences que la LB_Keogh est la meilleure LB pour la DTW et pour plusieurs distances. Elle est en particulier très utile pour l'indexation des formes présentées sous différents angles de rotation [Keogh, 2006]. Elle est

aussi utilisée pour accélérer la classification des séries temporelles [Ratanamahatana, 2004, b], et pour les séries temporelles en streaming [Wei, 2005].

Ci-dessous une intuition visuelle de la LB_Keogh. Une enveloppe englobante est faite autour de la série temporelle rouge. La distance Euclidienne entre la série temporelle bleue et la plus proche partie de l'enveloppe englobant est la distance de borne inférieure de DTW.

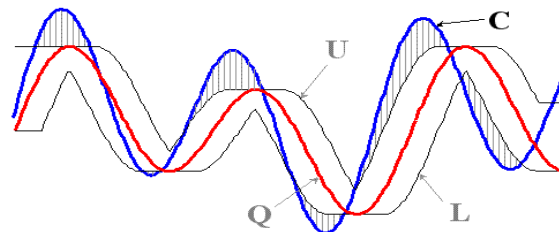


Fig. 34. Une intuition visuelle de la LB_Keogh [Keogh, 2005].

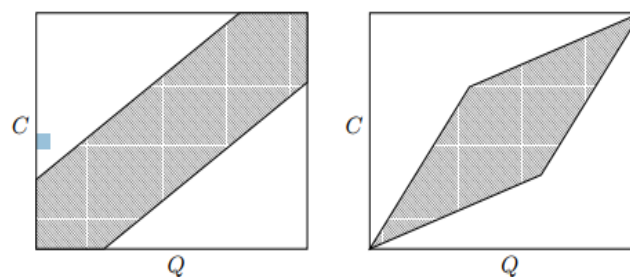
Soit Q , une série temporelle requête et U et L respectivement la borne supérieure et la borne inférieure qui sont construites selon le problème posé, comme suit :

$$U_i = \max(q_{i-r} : q_{i+r}) \quad (36)$$

$$L_i = \min(q_{i-r} : q_{i+r}) \quad (37)$$

Où r est un terme qui définit la plage de déformation autorisée. Il peut être fixé soit avec la « Sakoe-Chiba band » [Sakoe, 1978] ou le « Itakura parallelogram » [Itakura, 1975].

Sakoe et Chiba placent des bornes symétriques autour de la diagonale principale de la matrice des distances cumulées de la DTW (voir figure 35.a). Le chemin d'alignement optimal est forcé à rester à l'intérieur de ces bornes. Le parallélogramme d'Itakura est une autre contrainte de chemin globale très fréquemment utilisée. Itakura place un parallélogramme autour de la diagonale principale de la matrice des distances cumulées qui borne la plage de distorsion (voir figure 35.b).



(a) Sakoe and Chiba band

(b) Itakura Parallelogram

Fig. 35. Les deux contraintes de chemins optimale les plus populaires, utilisées dans l'implémentation de DTW [Rath, 2002].

La mesure de borne inférieure de la DTW est définie comme suit :

$$LB_Keogh(Q, C) = \sqrt{\sum_{i=1}^n \begin{cases} (c_i - U_i)^2 & \text{si } c_i > U_i \\ (c_i - L_i)^2 & \text{si } c_i < U_i \\ 0 & \text{si non} \end{cases}} \quad (38)$$

IV.4 Les différents types de requêtes de similarité

Fondamentalement, il existe trois types de requêtes de similarité : Les requêtes de plages de similarité (*Range Queries*), les requêtes de similarité des k-plus proches voisins, (k -Nearest Neighbor Queries, KNN) et les requêtes de similarité de jointure (Similarity Join).

IV.4.1 Requêtes de plages de similarité (Range Queries)

Dans ce type de requête, il faut trouver l'ensemble U des séries temporelles qui sont à une distance r de q [Cassisi, 2012]. Où r représente la distance tolérée (threshold) et q , la requête. Ce problème peut être défini par la formule suivante :

$$R(q, r) = \{u \in U; d(q, u) \leq r\} \quad [39]$$

La figure 36 donne un exemple de Range query.

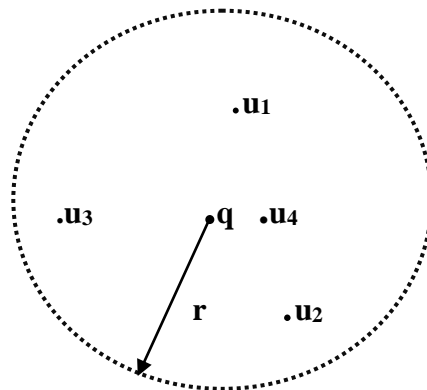


Fig.36. Plage de requêtes de similarité (Range Queries)

L'inconvénient principal de la requête de plage de similarité (Range queries) est que la fixation de la distance maximum r entre les objets de la base et la requête peut être très difficile, si la base de données n'est pas connue. Dans ce cas, une valeur inappropriée de r

peut donner lieu à deux situations indésirables ; retourner un ensemble de réponses qui est trop grand, ou retourner un ensemble de réponses vide, [Muhammad Fuad, 2011].

IV.4.2 Requêtes de similarité des k-plus proches voisins (KNN Query)

Dans ce genre de recherche et étant donné une série temporelle requête q , la recherche de similarité des k-plus proches voisins se fait par la recherche de l'ensemble des premières k séries temporelles dans la base de données les plus similaires à q [Cassisi, 2012]. La figure 37 illustre un exemple de la requête de similarité des k- plus proches voisins.

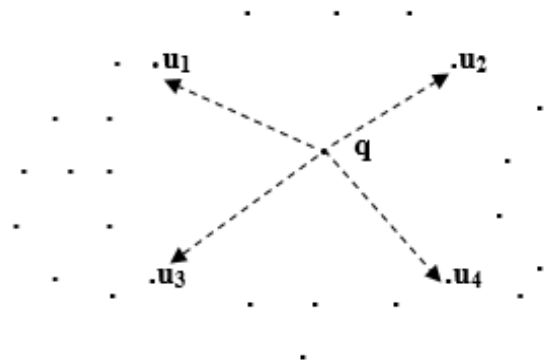


Fig.37. Requête de similarité des k- plus proches voisins. $K=4$

Contrairement à la requête de plage de similarité, l'ensemble des réponses n'est jamais vide. Cependant, il doit être fixé au préalable par l'utilisateur [Muhammad Fuad, 2011].

IV.4.3 Requête de similarité de jointure

Étant donné deux jeux de données, une requête de similarité par jointure consiste alors à trouver des paires d'objets (un de chaque jeu) satisfaisant une distance de similarité. Si les deux jeux de donnée coïncident, il s'agit de similarité auto jointure [Paredes, 2009]. Le problème est largement utilisé pour le nettoyage et l'intégration des données.

Le coût de cette méthode est supérieur à celui des k plus proches voisins et de range queries. En effet, il est d'ordre quadratique, ce qui l'empêche d'être utilisé dans de grandes bases de données [Dohnal, 2003].

IV.5 Applications de la recherche de similarité dans les séries temporelles

La recherche de similarité dans les séries séquences temporelles est utilisée dans de nombreuses applications de data mining telles que la reconnaissance vocale [Obin, 2016], la reconnaissance de l'écriture manuscrite [Schimke, 2007], la reconnaissance du visage [Qi, 2018], La recherche d'images basée sur le contenu [Keogh, 2006], la recherche par similarité de tranches de musiques [Tavenard, 2005] et bien d'autres applications. Cependant aucune application ne cherche à trouver, par exemple, une séquence audio exactement similaire à une séquence de référence ou une image dont les pixels sont parfaitement alignés à l'image de référence [Chávez, 2006].

Dans le domaine des séries temporelles, les images sont souvent représentées par leurs contours (formes) et ces contours sont convertis en signaux unidimensionnels représentant les séries temporelles. Plusieurs techniques font la conversion des images en séries temporelles [Adamek, 2004], [Attalla, 2005], [Cardone, 2003], [Vlachos, 2005] (Figure 28). Les séries temporelles extraites sont ensuite comparées avec une mesure de similarité. La distance Euclidienne, la distance de déformation temporelle (DTW) et la distance de la plus longue sous-séquence commune (LCSS) sont les plus utilisées pour la recherche d'images sous forme de séries temporelles [Adamek, 2003, Bhanu, 2004, Ratanamahatana, 2005, Rath, 2002, Vlachos, 2002]. La méthode Phase Independent DTW (PI-DTW) proposée par Keogh et al. [Keogh, 2006] consiste à extraire une série temporelle à partir de la forme d'un crâne de primate et à chaque rotation de la forme, la distance est calculée entre la série temporelle extraite et les autres séries temporelles de la base.

Dans ce cadre, la Figure 38 [Keogh, 2006], présente les crânes de primates et leurs séries temporelles classés hiérarchiquement en quatre groupes. Cette hiérarchie respecte l'évolution de la forme du crâne humain à travers le temps. La différence dans les séries temporelles entre chaque groupe n'est pas visible à l'œil nu. D'où l'utilité d'une mesure de distance.

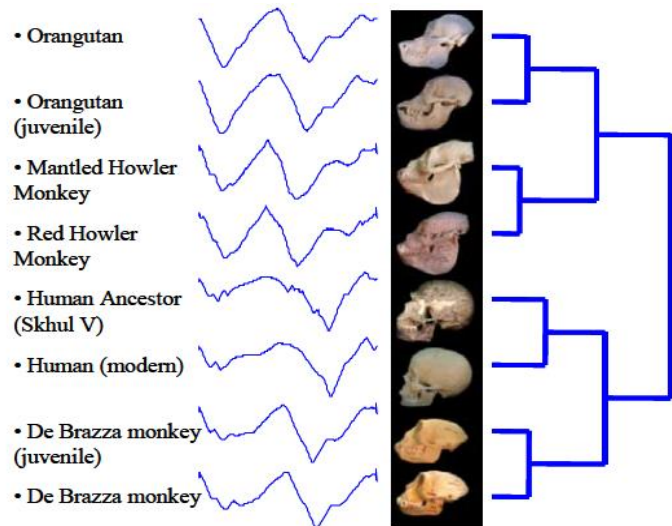


Fig.38.Groupement hiérarchique de huit crânes de primates basés sur la vue latérale, en utilisant la distance Euclidienne [Keogh, 2006]

IV.6 Conclusion

Dans ce chapitre, nous avons présenté les différents problèmes liés à la recherche de similarité dans les séries temporelles. Ensuite, nous avons présenté les trois types de requêtes les plus utilisés pour la recherche de similarité dans les séries temporelles et finalement nous avons donné quelques exemples d'applications de ce problème.

Chapitre V. Classification des séries temporelles

V.1 Introduction

La classification de séries temporelles consiste à assigner une série temporelle à l'une des classes présentes dans un ensemble prédéfini [Ziat, 2017]. Elle trouve des applications dans de nombreux domaines, tel que, la classification des espèces de crickets selon les sons qu'ils émettent [Dietrich, 2004], la détection des crises d'épilepsie humaine [Ramgopal, 2017] et beaucoup d'autres cas.

Plusieurs approches ont été proposées dans la littérature pour résoudre le problème de la classification des séries temporelles, telles que les approches de Support Vector Machine (SVM) [Huerta, 2012], les Réseaux de Neurones à Convolution (CNN) [Liu, 2019], les arbres de décision [Salperwyck, 2015], les approches des sacs de mots (BoW, Bag of Words) [Schäfer, 2015] entre autres.

Dans ce chapitre, nous allons présenter les méthodes les plus utilisées pour la classification des séries temporelles.

V.2 Méthodes de classification supervisée des séries temporelles

Plusieurs méthodes ont été proposées pour la classification des séries temporelles. Nous avons choisi de suivre la catégorisation établie par [Amr, 2012] qui se présente comme suit : les approches fondées sur l'extraction de caractéristiques, les approches fondées sur les mesures de similarité et les approches fondées sur les modèles. Les approches fondées sur l'extraction de caractéristiques discrétisent les données en représentations symboliques ou extraient directement les caractéristiques descriptives telles que la moyenne ou l'écart type pour les injecter en entrée à un classifieur, tel que les réseaux de neurones et les arbres de décision [Nguyen, 2017]. Les approches fondées sur les mesures de similarité utilisent les séries temporelles numériques avec des mesures de similarité telles que la distance euclidienne et la DTW et généralement un classifieur de One Nearest Neighbor (1-NN). La troisième famille concerne les approches fondées sur les modèles tels que les modèles de Markov cachés (HMM). Selon [Liao, 2005]. Ces approches construisent un modèle pour les

séries temporelles au sein d'un cluster (classe dans notre cas) et classifient les nouvelles séries selon le modèle qui les caractérisent le mieux.

V.2.1 Méthodes de classification basées sur l'extraction de caractéristiques

V.2.1.1 Approches des sacs de mots (Bag of Words, BoW)

Les modèles de classification des séries temporelles basés sur l'extraction des caractéristiques descriptives rassemblent un ensemble de caractéristiques de chaque série temporelle pour former des "sacs de mots" (BoW pour "Bag of Words"), qui sont ensuite donnés en entrée à un classifieur [Ziat, 2017]. Ces modèles diffèrent dans les caractéristiques extraites. Le modèle de représentation BoP (Bag of Patterns) proposé par [Lin, 2012] convertit une série temporelle en une chaîne de caractères en utilisant la méthode SAX (Symbolic Aggregate approxImation). L'ordre temporel des segments locaux, d'une série temporelle est ignoré et tous les segments locaux d'une série temporelle sont représentés dans des histogrammes pour construire un sac de motifs. Selon son auteur, cette représentation serait effective pour trouver la similarité entre les séries temporelles et elle permettrait un appariement invariable à la rotation de formes mais sa dimensionnalité pourrait être très grande.

Le modèle TSBF (Time-Series Bag of Features) [Baydogan, 2013], par exemple extrait de multiples sous-séquences au hasard et les divise en intervalles plus courts pour capturer les informations locales. Par conséquent, les caractéristiques calculées à partir de ces sous-séquences mesurent les propriétés à différents endroits et les dilatations à partir de la série originelle. Il s'agit d'une approche basée sur les caractéristiques qui peut, selon son auteur, gérer les décalages temporelles. Chaque série temporelle est ensuite transformée en une chaîne de caractères en utilisant la méthode SAX. Ensuite, un classifieur supervisé intègre l'information de localisation dans un livre de codes grâce à des estimations de probabilité de classe.

Une autre approche récente est notée BOSS (Bag-of-SFA-Symbols) [Schäfer, 2015]. Elle consiste à combiner la tolérance au bruit de la méthode SFA (Symbolic Fourier Approximation) [Schäfer, 2012] avec la représentation structurée du modèle des sacs de mots [Lin, 2012].

V.2.1.2 Les Réseaux de neurones

Les réseaux de neurones (RNA) sont inspirés de la structure biologique des réseaux de neurones naturels. L'unité de base du RNA est appelée cellule. Elle prend en entrée un nombre de caractéristiques et donne en sortie un signal qui peut lui-même être donné en entrée à une autre cellule. Une seule cellule est capable de modéliser un simple signal mais plusieurs cellules combinées entre elles dans une seule structure peuvent modéliser des signaux plus complexes.

Les réseaux de neurones les plus utilisés sont les réseaux neuronaux récurrents (RNN) et les réseaux neuronaux à convolution (CNN). Les réseaux de neurones récurrents RNN comprennent plusieurs boucles de rétroactions sur les cellules [Mefteh, 2011]. Selon l'auteur de [Mefteh, 2011] ces réseaux possèdent une représentation implicite du temps c.à.d. que l'ordre temporelle n'est pas représenté explicitement en tant qu'ordre des composantes de la série temporelle d'entrée mais de manière implicite, par les effets qu'il a sur les traitements.

Les réseaux de neurones à convolution (CNN) sont des réseaux neuronaux artificiels à action directe avec une ou plusieurs couches convolutives contenant des filtres d'apprentissage. Les CNN sont de puissants classificateurs en raison de leur capacité à apprendre automatiquement des caractéristiques discriminantes à partir des données d'entrée [Lines, 2016].

Par rapport aux RNN qui conviennent à l'extraction de données séquentielles et qui fonctionnent bien pour les tâches de reconnaissance vocale [Graves, 2013], les CNN sont généralement bien adaptés aux tâches de reconnaissance d'images. Par exemple, les réseaux résiduels profonds (ResNets) [He, 2015] seraient selon leur auteur très appréciés pour la reconnaissance d'images.

Les CNN ont récemment été utilisés pour la classification des séries temporelles par Fawas et al [Fawas, 2019]. Les auteurs de [Fawas, 2019] ont introduit un réseau de neurones à convolution profond (CNN), appelé *InceptionTime*, qui surpasserait les classifieurs de pointe tels que les ensemble de classifieurs Flat-COTE [Line, 2015] et HIVE-COTE [Lines, 2016] avec un gain /égalité/ perte sur le benchmark UCR (85 bases) de 40/6/39 en faveur de *InceptionTime* et qui serait également beaucoup plus rapide [Fawas, 2019].

V.2.1.3 Approches fondées sur les noyaux (SVM)

L'idée de base du SVM pour l'approximation des fonctions est de projeter les données x dans un espace de caractéristiques à haute dimension par une projection non linéaire, puis d'effectuer une régression linéaire dans l'espace des caractéristiques [Samsudin, 2010].

Plusieurs travaux ont été menés pour étendre les travaux sur les noyaux au cadre des séries temporelles. On peut trouver une comparaison des approches fondées sur plusieurs noyaux appliqués aux séries temporelles (Linear Kernel, Radial Basis Kernel, Fourier Kernel) dans [Rüping, 2001]. Les expériences de Rüping *et al* ont montré que Radial Basis Kernel est le mieux adapté à différents types de séries temporelles.

Eads et al [Eads, 2005] ont proposé un algorithme génétique basé sur l'extraction de caractéristiques et un SVM a été entraîné avec les caractéristiques extraites. Eads *et al*, ont testé leur algorithme « ZEUS » par rapport aux classifieurs conventionnels tels que linear Kernel, Radial Basis Kernel et Fisher Linear Discriminant (FLD) [Duda, 2001] sur sept ensembles de données et ont démontré que « ZEUS » surpasserait ces derniers classifieurs en terme de précision de classification. Rodríguez et al [Rodríguez, 2005] quant à eux ont utilisé des intervalles de séries temporelles pour extraire les caractéristiques sur lesquelles un SVM a été entraîné. Les expériences de Rodríguez et al on montré la validité de leur méthode proposée.

V.2.1.4 Les arbres de décision

Les arbres de décision sont connus pour leurs simplicité et leurs rapidité d'exécution. « Il s'agit d'une méthode itérative de partitionnement des données. En effet, la méthode construit des classes d'individus, les plus homogènes possible, en posant une succession de questions binaires (de type oui/non) sur les attributs de chaque individu » [Santos, 2015]. Rodriguez et Alonso, par exemple, appliquent des arbres de décision basés sur la DTW pour la classification des séries temporelles [Rodriguez, 2004].

La représentation graphique des arbres de décision est facile à lire et les décisions sont aisément interprétables. L'arbre de décision présenté dans la Figure 39 est constitué de feuilles terminales représentant les classes d'individus (gagné, perdu) obtenues en suivant un chemin le long des nœuds. Chaque nœud correspond à une question binaire (Aller à la plage, Mauvaises conditions climatiques). L'ensemble d'apprentissage sur la base duquel cet arbre est construit est le suivant (Aller à la plage, Mauvaises conditions climatiques, gagné, perdu)

(Si Aller à la plage = « Non » alors classe = « Perdu »)

ou (Si Aller à la plage = « Oui » & Mauvaises conditions climatiques = « Non » alors class=« Gagné»)

ou (Si Aller à la plage = « Oui » & Mauvaises conditions climatiques = « Oui» alors class=« Perdu »)

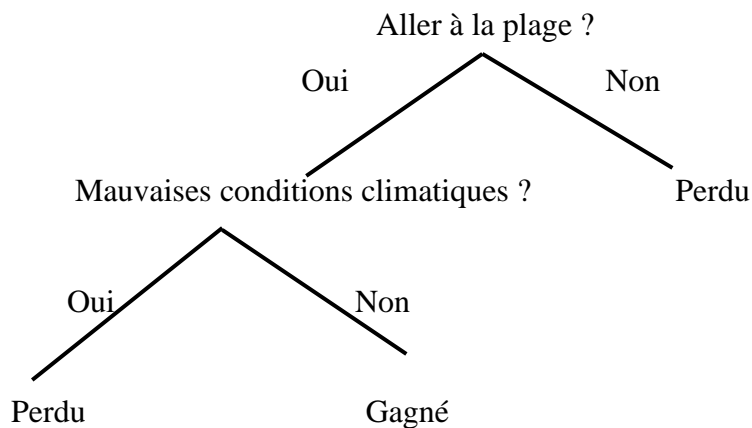


Fig.39.Arbre de décision pour étudier la possibilité d’aller à la plage.

Un certain nombre d’approches des arbres de décision a été élaboré pour la classification des séries temporelles. L’approche de la Forêt des séries temporelle (Time Series Forest-TSF) [Deng, 2013] dérive d’abord un résumé des caractéristiques pour toutes les séries temporelles en les divisant en intervalles et en résumant chaque intervalle avec sa moyenne, son écart-type et son gradient. Ensuite, elle choisit aléatoirement un sous-ensemble de ces caractéristiques à chaque nœud de chacun des arbres d’un ensemble.

Lucas et al [Lucas, 2018] introduisent la Forêt de proximité, un algorithme qui permet d’apprendre de façon précise à partir d’ensembles de données contenant des millions de séries temporelles, et classe une série temporelle en quelques millisecondes. Les modèles sont des ensembles d’arbres de proximité hautement randomisés.

Alors que les arbres de décision classiques se basent sur les valeurs des caractéristiques (et donnent généralement de mauvais résultats pour les séries temporelles), les arbres de proximité se base sur la proximité d’une série temporelle par rapport à une autre série temporelle, en sélectionnant de manière stochastique les mesures de similarité utilisés ce qui leur permet de gagner en efficacité et en précision.

V.2.1.5 Les méthodes d’ensemble de classifieurs ou combinaison de classifieurs

Les méthodes d’ensemble combinent les résultats de plusieurs classifieurs et sont connues pour avoir de bons résultats de classification, mais elles ont souvent des temps d’exécution très longs [Bagnal, 2015].

La Collection des ensembles de transformation (COTE) [Bagnal, 2015] est un ensemble de 35 classificateurs répartis en quatre domaines de transformation des séries temporelles :

temps, fréquence, changement et transformation de forme. Pour le domaine temporel, COTE utilise 11 mesures de similarité [Lines, 2015], tandis que dans les trois autres domaines, les classificateurs sont sélectionnés parmi les arbres de décision, les forêts aléatoires, les SVM, et autres.

Le classifieur des Ensemble Elastique (EE) [Lines, 2015] est un ensemble de classifieurs 1-NN, chacun utilisant une mesure de similarité élastique différente comme la DTW, la LCSS, la distance d'édition, etc. La stratégie d'assemblage la plus efficace selon Lines et al est de combiner tous les classificateurs en une hiérarchie plate (Flat-COTE). Selon Lines et al [Lines, 2016], Flat-COTE serait plus précis que les classifieurs de pointe tels que TSF (Time Series Forest) [Deng, 2013], ST (Shapelet Transform Ensemble) [Hills, 2014], Boss (Bag-of-SFA-Symbols) [Schafer, 2015] et TSBF(Time-Series Bag of Features) [Baydogan, 2013].

HIVE-COTE (Hierarchical Vote Collective of Transformation-based Ensembles) [Lines, 2016], est une version améliorée de Flat-COTE. Il contient des modules qui capturent la similarité avec des séries entières (Ensemble élastique (EE)), des sous-séries indépendantes de la phase (Shapelet Transform Ensemble (ST)), un ensemble basé sur les intervalles (Time Series Forest (TSF)), un ensemble de dictionnaires (Bag-of-SFA-Symbols (BOSS)) et un ensemble spectral. HIVE-COTE capture davantage de caractéristiques discriminatoires possibles dans les séries temporelles et possède une structure plus modulaire et intuitive. Plus important encore, HIVE-COTE serait selon Lines *et al* [Lines, 2016] nettement plus précis que Flat-COTE et représenterait un nouvel état de l'art pour la classification des séries temporelles.

V.2.2 Méthodes de classification basées sur des mesures de similarité

Ces méthodes utilisent la technique du plus proche voisin 1NN (One-Nearest Neighbour 1NN) couplée à une mesure de similarité et s'appliquent généralement sur les données brutes des ST.

V.2.2.1 Approche du plus proche voisin 1NN

L'algorithme du plus proche voisin, ou KNN (k Nearest Neighbor) a été introduit par Fix et Hodges [Fix, 1951] pour la classification des données. Il consiste à affecter une nouvelle série dont l'étiquette est inconnue, à la classe à laquelle appartient la majorité des séries qui lui sont

les plus proches au sens de la métrique, i.e., qui appartiennent aux k séries les plus proches de la série requête. Cette méthode, de par sa simplicité et son efficacité, est encore très répandue.

Dans l'exemple de la Figure 40, nous avons deux classes d'éléments (carrés bleus et cercles rouges) et nous essayons de placer un nouvel élément (représenté par le trapèze marron) dans l'une de ces classes. Nous regardons les voisins "les plus proches" et votons sur ce que l'élément le plus récent est. Dans la Figure 40, ce qui se passe si nous prenons $k = 1$, le nouvel exemple serait mis dans la classe 1, pour $k = 3$, le nouvel exemple serait mis dans la classe 2.

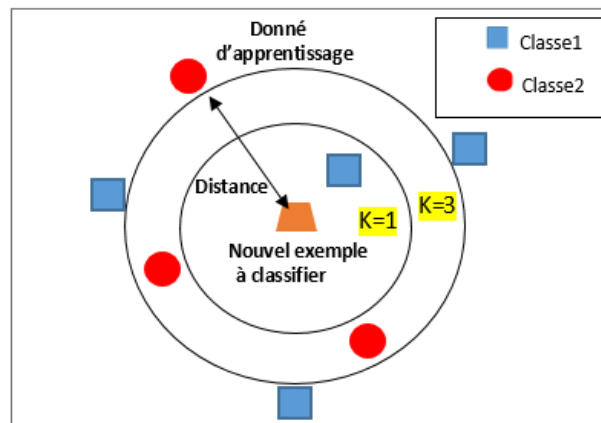


Fig.40. Illustration de la méthode KNN : La méthode classe le trapèze dans la classe 1 en se basant sur $K=1$ voisin.

Dans le contexte de la classification des ST, le classifieur basé 1NN serait le plus efficace selon les travaux récents de l'état de l'art en la matière. En particulier, il est rapporté que 1-NN avec DTW est difficile à battre [Xi, 2006], [Ding, 2008], [Guisti, 2013], [Bagnall, 2014], [Petitjean, 2016], [Dau, 2018], [Boulnemour, 2020].

KNN est le classifieur de base pour de nombreux problèmes de classification tels que la reconnaissance des formes [Xu, 2008], la catégorisation des textes [Manne, 2012], la reconnaissance des objets [Bajramovic, 2006], et les applications de reconnaissance d'événements [Yang, 2000]. Toutefois, KNN a besoin d'une mesure de similarité appropriée pour trouver les voisins les plus proches.

Plusieurs travaux ont été réalisés pour chercher les meilleures mesures de similarité pour le classifieur KNN, La distance Euclidienne (ED) et la DTW sont les plus utilisées. D'autres mesures de distance telles que Time Warp Edit Distance (TWED) [Marteau, 2009], Edit Distance on Real sequence (EDR) [Chen, 2005-a], Minimum Jump Cost (MJC) [Serra, 2012], Move-Split-Merge (MSM) [Stefan, 2013] Complexity-Invariant Distance (CID) [Batista,

2014], distance de Hassanat [Hassanat, 2014], [Alkasassbeh, 2015], ont récemment été utilisé pour le classifieur KNN.

Le principal inconvénient de KNN est son temps de calcul qui devient très élevé lorsqu'il traite un grand ensemble de données d'apprentissage, puisque chaque échantillon de l'ensemble de test doit être comparé à chaque échantillon de l'ensemble d'apprentissage.

V.2.3 Méthodes de classification basées sur les modèles

V.2.3.1 Les modèles de Markov cachés (HMM)

Selon [Liao, 2005], cette catégorie d'approches considère que chaque série temporelle est générée par une sorte de modèle ou par un mélange de distributions de probabilités sous-jacentes. Les séries temporelles sont considérées comme similaires lorsque les modèles caractérisant les séries individuelles après ajustement du modèle soient similaires.

Laxman et Sastry [Laxman, 2006] classent les modèles selon qu'ils soient prédictifs ou descriptifs. Les modèles prédictifs tels que les modèles de Markov cachés (Hidden Markov Models – HMM), sont utilisés dans les applications de prévision et de classification, tandis que les modèles descriptifs sont utiles pour résumer les données [Laxman, 2006].

Dans [Antonucci, 2015] une nouvelle approche de la classification précoce des séries temporelles par les HMM avec des paramètres à valeur fixe a été proposée. Selon Antonucci et al, les résultats sont prometteurs et l'approche à valeur fixe est plus performante que l'approche standard pour la reconnaissance de la parole.

D'un autre coté, la difficulté de l'apprentissage des états initiaux et des probabilités de transition peuvent rendre les modèles HMM moins compétitifs que la dtw [Ravinder, 2010]. Plusieurs auteurs ont ainsi cherché à coupler les méthodes HMM avec la DTW [Oates, 1999, 2001], [Hu, 2006].

V.3 Conclusion

Dans ce chapitre, nous avons présenté une revue des méthodes de classification des séries temporelles. Les deux familles les plus connues dans la fouille des données en général sont les méthodes de classification supervisées et les méthodes de classification non supervisée. Dans le domaine des séries temporelles, la famille des méthodes de classification supervisée est

elle-même classifiée en trois familles : Les approches fondées sur l'extraction de caractéristiques, Les approches fondées sur les mesures de similarité et les approches fondées sur les modèles.

le classifieur basé 1NN serait le plus efficace selon les travaux récents de l'état de l'art en la matière. D'un autre côté, Les classifieurs tels que les arbres de décision et les réseaux de neurones ont besoin de données d'entraînement et donc d'un point de vue d'experts d'où l'utilité du classifieur KNN que nous allons utiliser pour la classification des séries temporelles dans le Chapitre VIII.

Chapitre VI. QPDTW : Technique d'alignement des STQP basée DTW et SEA

VI.1 Introduction

Dans ce chapitre nous allons proposer une technique d'appariement qui compare et aligne spécialement les séries temporelles Quasi-périodiques (STQP ou Quasi-Periodic Time Series : QPTS). Comme déjà avancé, ce type de séries temporelles serait le plus complexe selon la classification établie par Boucheham et al [Boucheham, 2013]. Comme déjà annoncé, notre méthode proposée dénommée QPDTW est obtenue par fusion des deux méthodes les plus en vue en matière d'alignement des ST : DTW et SEA de Boucheham [Boucheham, 2008]. Du coup, les deux méthodes initialement rivales passent du domaine de la rivalité au stade de la coopération grâce à QPDTW. Nous visons ainsi gagner les avantages des deux méthodes : La Quasi-périodicité de SEA et le pouvoir d'alignement optimal de la DTW. Nous appliquons notre technique sur des ST de l'électrocardiogramme (ECG), les ST de l'ECG étant typiquement Quasi-périodique. C'est dire que les cycles d'un tracé ECG normaux sont quasi-similaires, mais avec quelques changements morphologiques locaux, dus parfois aux bruits causés par la variabilité de mesure et/ou les variations physiologiques naturelles.

La Figure41 présente les caractéristiques d'un électrocardiogramme composé de trois périodes quasi-périodiques, représentant trois cycles cardiaques. Chaque cycle (période) est lui-même composé de trois motifs consécutifs : l'onde P, le complexe QRS et l'onde T.

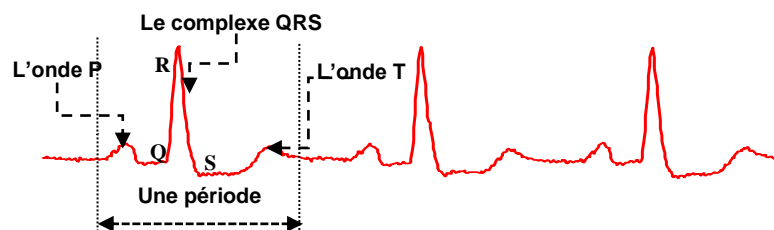


Fig.41. Un segment d'ECG typique avec trois périodes (cycles).

VI.2 Méthode d'alignement proposée

La DTW (Dynamic Time Warping) [Sakoe, 78] est généralement reconnue par plusieurs chercheurs [Ratanamahatana, 2005], [Keogh, 2006], [Ding, 2008], [Rakthanmanon, 2012]

comme étant la méthode d'alignement la plus précise pour l'alignement des séries temporelles. Cependant, elle ne peut pas traiter les problèmes de déphasage significatifs dans les séries temporelles quasi-périodiques [Boucheham, 2013]. Dans l'exemple présenté par la Figure 42, les deux tracés ECGs sont similaires puisqu'ils qu'ils proviennent de la même personne, mais ils ne commencent pas en même temps. Dans ce cas, la DTW fait un mauvais alignement, car seul le complexe QRS a été correctement aligné, les autres composantes des deux ECG (ondes P et T) viennent d'être appariées avec leurs équivalents séquentiels dans le temps.

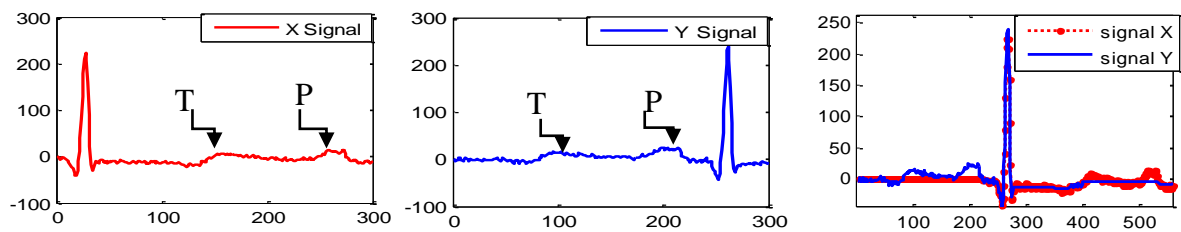


Fig.42. Illustration du problème de déphasage significatif de DTW. Signal Y est significativement déphasé par rapport au signal X. Les ondes P et T n'ont pas été correctement alignées avec la DTW.

D'un autre côté, la méthode SEA est la seule méthode qui traite les séries temporelles quasi-périodiques ayant plusieurs périodes différentes et déphasées [Boucheham, 2013]. Dans l'exemple issu de [Boucheham, 2008] présenté par la Figure 43, SEA aligne efficacement les séries quasi-périodiques similaires et déphasées par rapport à la DTW qui fait la correspondance entre les pics R seulement.

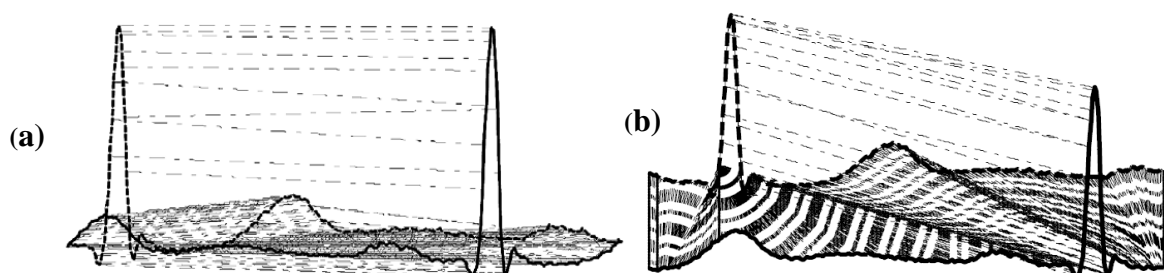


Fig.43. Illustration de l'efficacité d'alignement de la méthode SEA(a) par rapport à la méthode DTW (b) [Boucheham, 2008].

L'inconvénient de la méthode SEA est qu'elle utilise une technique de correspondance linéaire pour égaliser les longueurs des séries temporelles. Cette technique de correspondance n'est pas très précise, car la série la plus longue perd parfois des valeurs importantes en se

rétrécissant. SEA a aussi des difficultés lorsque les séries temporelles sont significativement bruitées comme le montre la Figure 44.

La méthode QP-DTW a été envisagée dans l'optique de palier aux faiblesses des deux méthodes, tout en profitant de leurs forces respectives. Elle est alors décrite dans ce qui suit :

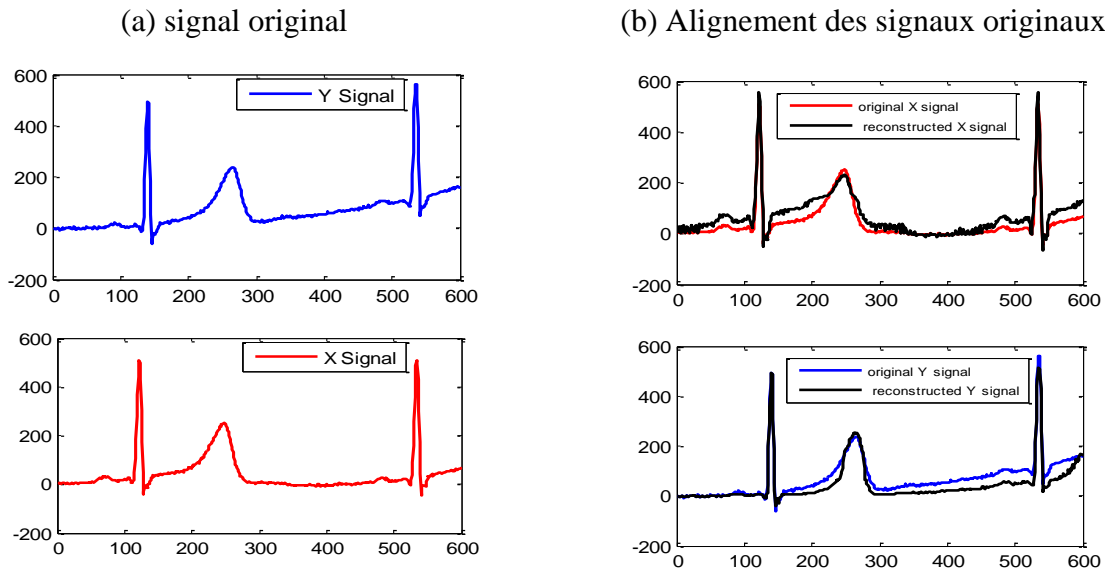


Fig.44. Illustration du problème de bruit que peut rencontrer la méthode SEA. Ici, le signal le plus haut (bleu) est significativement contaminé par le bruit de basse fréquence (Signal respiratoire).

Algorithme QP-DTW (Quasi-Periodic Dynamic Time Warping):

Étape1 : Trouver le chemin optimal avec la DTW

- L'algorithme DTW est exécuté pour étirer les deux séries temporelles et faire une première mise en correspondance entre elles par la réplique de certaines valeurs qui sont proches dans les deux séries temporelles.

Étape2 : Tri sur l'amplitude et comparaison

- Le tri des séries temporelles étirées sur les coordonnées de leurs amplitudes est établi pour leur donner une signature stable. Le résultat de cette étape est une matrice pour chaque série temporelle contenant les amplitudes triées par ordre ascendant ou descendant avec leurs coordonnées temporels équivalents non triés. De cette manière, les séries temporelles similaires, mais déphasées et de longueur différentes ont toujours quasiment le même tracé qui représente leur signature. Un exemple des signatures extraites est présenté dans la Figure 46.

Etape3 : Échange de signatures

- Cette étape se divise en deux parties. La première consiste à faire l'échange de signature entre les deux séries temporelles. La première série temporelle va recevoir les coordonnées de amplitudes (signature) de la deuxième série temporelle, mais elle va garder ses propres indexes temporels non triés et vice et versa pour la deuxième série. Cette étape permet de prendre en charge les variations d'amplitudes contrairement à la DTW qui ne résout que les problèmes de distorsion de l'axe temporelle.

- La deuxième partie consiste à rétablir la taille normale des séries temporelles après l'échange de signatures. Ainsi, les indexes temporelles répliquée dans chaque nouvelle série doivent être supprimés avec leurs amplitudes équivalentes. Cette partie n'affecte pas les résultats de l'alignement, elle permet juste d'avoir une inspection visuelle plus claire des tracés.

Etape4 : Tri sur le temps et alignement

- Les signatures représentent bien les caractéristiques des séries temporelles, mais elles ne considèrent pas les décalages et les différences d'échelles du temps. Le tri sur les indexes temporels respectifs des séries temporelles dont les signatures ont été échangées permet de les reconstruire.

- Chaque série temporelle va ensuite être alignée avec sa série temporelle reconstruite. Suite à cela, les facteurs de corrélation Corr et de PRD (Percent Root Difference) sont utilisés comme des critères objectifs d'évaluation du matching, en parallèle avec l'inspection visuelle comme un critère subjectif.

Le diagramme présenté dans la Figure 45 décrit l'algorithme proposé.

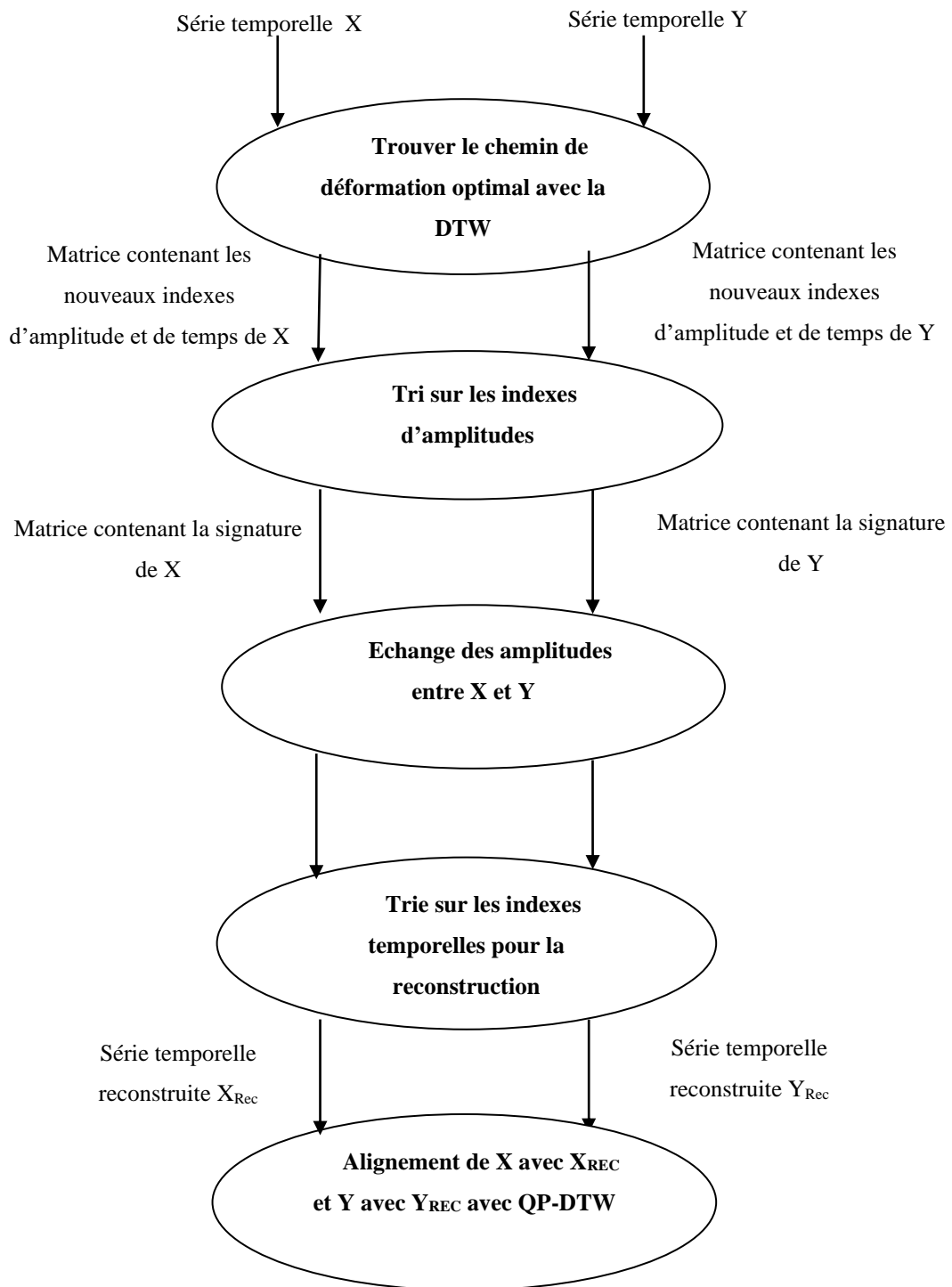


Fig.45. Diagramme illustratif de la méthode d’alignement proposée QP-DTW

VI.3 Tests expérimentaux et discussions

Dans cette section, nous allons tout d'abord présenter un exemple de comparaison numérique et visuelle entre QP-DTW, DTW et SEA. La comparaison numérique est effectuée avec le PRD (Eq (40)) et le facteur de corrélation Corr (Eq (1)). La valeur de la corrélation est dans l'intervalle [0,1] et la valeur du PRD est un pourcentage dans l'intervalle [0, 100].

$$PRD(X, Y) = \sqrt{\frac{\sum_{i=1}^{i=N} |x_i - y_i|^2}{\text{Max}(\sum_{i=1}^{i=N} |x_i - \bar{x}|^2, \sum_{i=1}^{i=N} |y_i - \bar{y}|^2)}} \times 100\% \quad (40)$$

Nous commençons la comparaison en introduisant un exemple de séries temporelles quasi-périodiques de la classe (c), selon la classification de Boucheham [Boucheham, 2013]. Dans l'exemple présenté par la Figure 47.a et la Figure 47.b, les séries temporelles X et Y sont extraites du même enregistrement « rec. 215 » provenant de la base de données publique des ECG "Massachusetts Institute of Technology-Beth Israel Hospital (MIT-BIH)". [MITBIH, 2000]. Tous les enregistrements de la base ont été numérisés à 360 échantillons par seconde et par canal avec une résolution de 11 bits sur une plage de 10 mV.

Les Figures 46.a et Fig46.b présentent les signatures des séries temporelles X et Y pour les méthodes SEA et QP-DTW. Nous constatons que les signatures établies par QP-DTW sont jointes, contrairement aux signatures de SEA qui sont détachées. Ce qui confirme un alignement de bonne qualité effectué par QP-DTW.

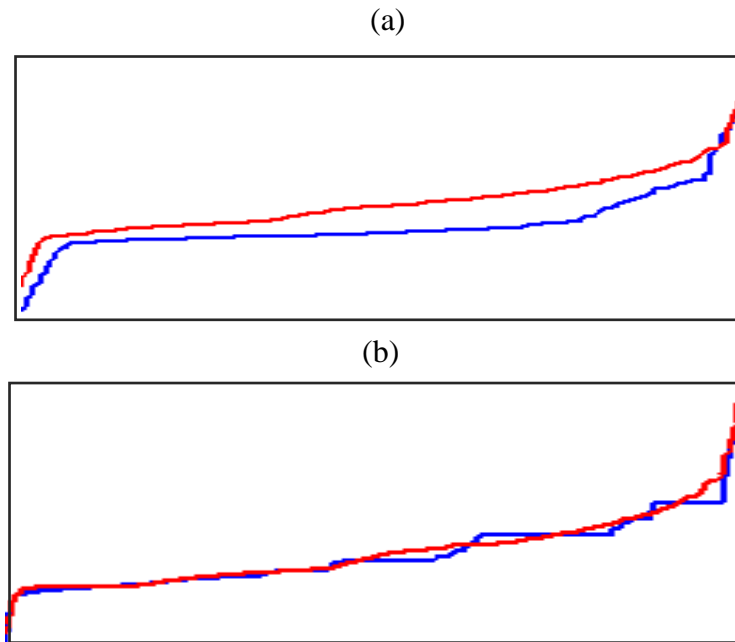


Fig.46. Signatures des tracés X et Y établies par les méthodes SEA (a) et QP-DTW (b).

La Figure 47.I montre les tracés originaux X et Y. La figure 47.II et la figure 47.III illustrent l'alignement effectué par les méthodes SEA et QP-DTW respectivement, pour la série temporelle X et sa série temporelle reconstruite et pour la série temporelle Y et sa série temporelle reconstruite. La figure.47. IV présente l'alignement effectué par la méthode DTW pour les séries temporelles X et Y. Notez que les deux tracés ECG proviennent de la même série temporelle, mais qu'ils ont des longueurs différentes. Ils ont aussi des pics d'amplitudes de longueurs légèrement différentes et n'ont pas été pris en même temps (décalage de 35 secondes pour Y). Le segment (X) a été pris au début de l'enregistrement (ECG) et le segment (Y) à la fin de l'enregistrement pour illustrer le problème de déphasage significatif dans des séries temporelles quasi périodiques.

Les formules sont appliquées aux deux tracés (X et Y) avec leurs tracés reconstruits respectifs (X_{Rec} et Y_{Rec}). Nous présenterons également dans la Table 3 les PRD (X, X_{rec}), Corr (X, X_{rec}) et PRD (Y, Y_{rec}), Corr (Y, Y_{rec}) pour les méthodes SEA et QP-DTW et le PRD (X, Y), Corr (X, Y) pour la méthode DTW

L'inspection visuelle de la Figure 47 et les résultats numériques rapportés dans Table 3 montrent que malgré tous les problèmes précédemment cités, QP-DTW arrive à aligner ces séries temporelles très complexes mieux que SEA et DTW. Nous remarquons aussi que DTW fait un très mauvais alignement car le deuxième QRS de la série temporelle Y n'a pas été aligné avec les QRS de la série X.

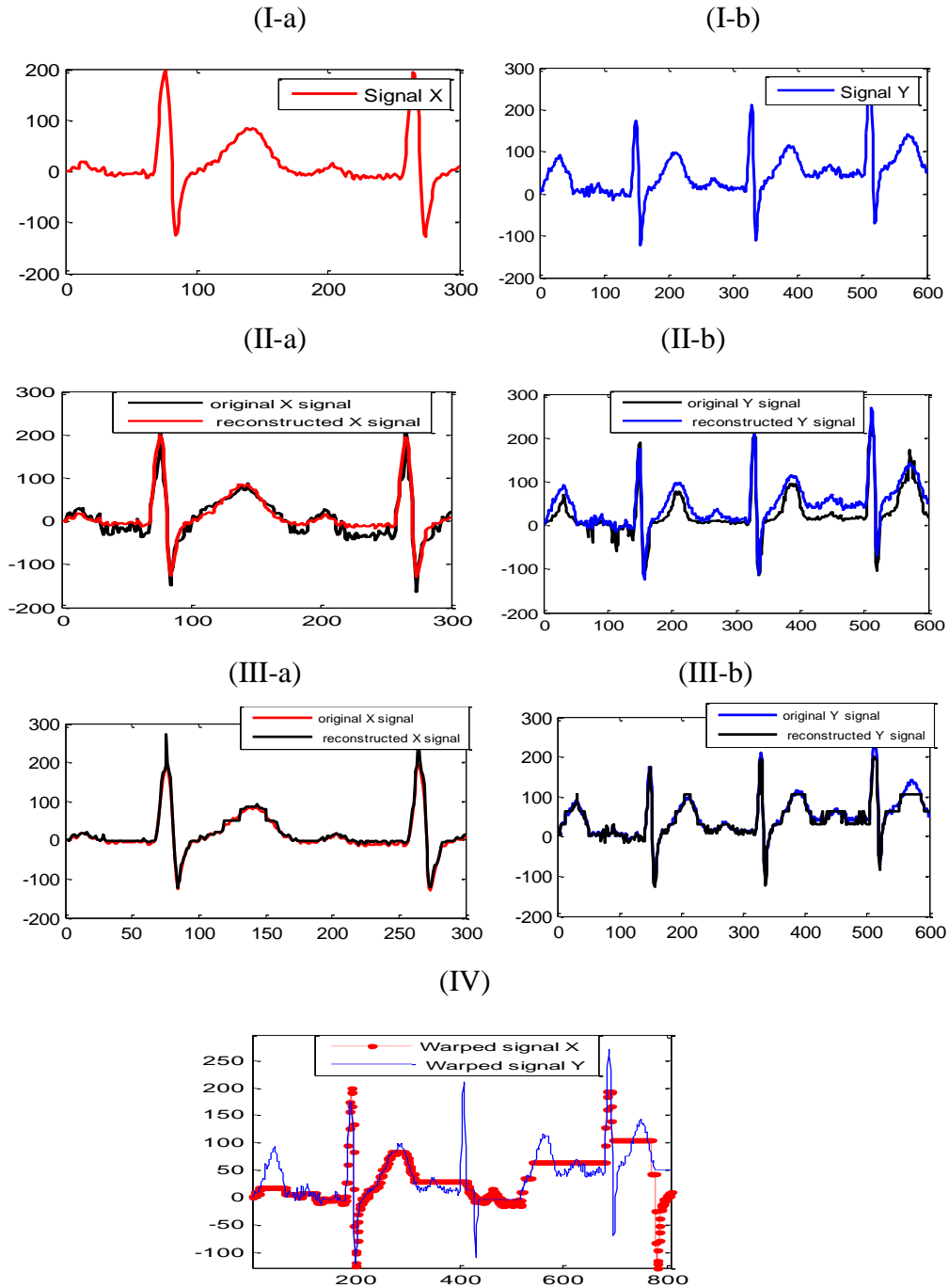


Fig.47. Comparaison des méthodes SEA, QP-DTW et DTW pour le cas de séries temporelles quasi périodiques à plusieurs périodes avec déphasage (rec.215). (I) Signaux originaux. (II) Signaux originaux vs signaux reconstruits et alignés par SEA. (III) Signaux originaux vs signaux reconstruits et alignés par QP-DTW. (IV) Signaux originaux alignés par DTW.

Table 3. Comparaison de QP-DTW avec DTW et SEA pour la série temporelle rec.215 et la série reconstruite Xrec (resp. Yrec).

| Method | PRD(X,Y) | PRD(X,X _{rec}) | PRD(Y,Y _{rec}) | Corr(X,Y) | corr(X,X _{rec}) | Corr(Y,Y _{rec}) |
|--------|----------|--------------------------|--------------------------|-----------|---------------------------|---------------------------|
| SEA | / | 32.0207 | 50.3374 | / | 0.9566 | 0.9561 |
| QP-DTW | / | 16.8858 | 21.6276 | / | 0.9906 | 0.9817 |
| DTW | 68.0843 | / | / | 0.7582 | / | / |

VI.3.1 Comparaison des tracés similaires

La Figure 48 montre le cas de deux tracés du même enregistrement avec 1000 échantillons de la série temporelle rec.113 pour X et 1200 échantillons de la même série temporelle avec un décalage de 30 secondes pour Y. La figure 48.I présente les séries temporelles originales. Elles ont quelques pics d'amplitudes avec des hauteurs légèrement différentes dues au bruit. La figure 48.II montre l'alignement moyen effectué par SEA car les séries temporelles originales et reconstruites sont presque détachées à cause du problème du bruit. La figure 48.III montre que malgré le problème de bruit, QP-DTW a effectué une reconstruction presque parfaite, au point qu'il est difficile de faire la distinction entre les séries temporelles originales et reconstruites. La figure 48.IV montre le désalignement entre X et Y effectué par DTW à cause du problème de quasi-périodicité des séries. Les résultats numériques présentés dans le tableau 4.a confirment ce résultat avec un très faible taux d'erreur (PRD) et une très forte corrélation en faveur de QP-DTW. Au contraire, ils indiquent un taux d'erreur élevé (PRD) pour DTW et SEA, une corrélation moyenne pour SEA et une faible corrélation pour DTW.

La Figure 49 montre le cas de deux tracés du même enregistrement avec 1000 échantillons de la série temporelle rec.215 pour X et 2000 échantillons de la même série temporelle avec un décalage de 35 secondes pour Y. La figure 49.I présente les séries temporelles originales. Le signal X a été pris au début de l'enregistrement et le signal Y a été pris à la fin de l'enregistrement pour illustrer le problème de déphasage significatif. La figure 49.II montre l'alignement moyen effectué par SEA. On voit que la méthode SEA échoue face au problème du bruit de basse fréquence qui est présent dans le signal Y. La figure 49.III montre l'alignement presque parfait effectué par QP-DTW. La figure 49.IV présente le mauvais alignement effectué par DTW. Nous constatant que notre méthode résout efficacement les problèmes de déphasage et de bruit. Les facteurs de similarité (Corr) et de dissimilarité (PRD) dans la Table 4.a indiquent clairement la supériorité de notre méthode par rapport à DTW et

SEA. La corrélation de QP-DTW est plus élevée que celle de DTW et de SEA et son PRD est le plus bas. Il représente la moitié de la valeur du PRD de DTW.

La Figure 50 montre le cas de deux tracés du même enregistrement ECG avec 1000 échantillons de la série temporelle rec.104 pour X et 2000 échantillons de la même série temporelle avec un décalage de 20 secondes pour Y. La figure 50.I présente les séries temporelles originales. Chaque série est composée d'un nombre différent de périodes qui sont déphasées. Le signal Y présente également le problème du bruit. L'alignement effectué par SEA est présenté dans la figure 50.II. La figure 50.III présente l'alignement effectué par QP-DTW. Il montre que notre méthode fait face à tous les problèmes précédemment cités. L'alignement effectué par DTW est présenté dans la figure 50.IV. On voit que DTW et SEA échouent face au problème du déphasage significatif et du bruit. Les résultats numériques confirment les résultats visuels avec une corrélation de 0,98 pour QP-DTW, 0,87 pour SEA et 0,93 pour DTW et un PRD de 20.51 pour QP-DTW, 51.83 pour SEA et 34.48 pour DTW.

La Figure 51 montre le cas de deux courbes du même enregistrement avec 1000 échantillons de la série temporelle rec.118 pour X et 2000 échantillons de la même série temporelle avec un décalage de 35 secondes pour Y. Les séries temporelles sont déphasées. L'inspection visuelle et les résultats numériques indiquent que l'alignement effectué par QP-DTW est meilleur que l'alignement effectué par SEA et DTW avec une corrélation de 0,99 pour QP-DTW, 0,96 pour SEA et 0,90 pour DTW et un PRD de 23,11 pour QP-DTW, 34,55 pour SEA et 43,91 pour DTW.

La Figure 52 montre le cas de deux tracés du même enregistrement avec 1000 échantillons de la série temporelle rec.115 pour X et 1200 échantillons de la même série temporelle pour Y avec un décalage de 25 secondes pour X. L'inspection visuelle de la figure montre que QP-DTW fait un alignement presque parfait contrairement à DTW. Les résultats numériques indiquent une forte corrélation et un faible PRD pour QP-DTW, un PRD et une corrélation moyens pour SEA, une corrélation faible et un PRD élevé pour DTW.

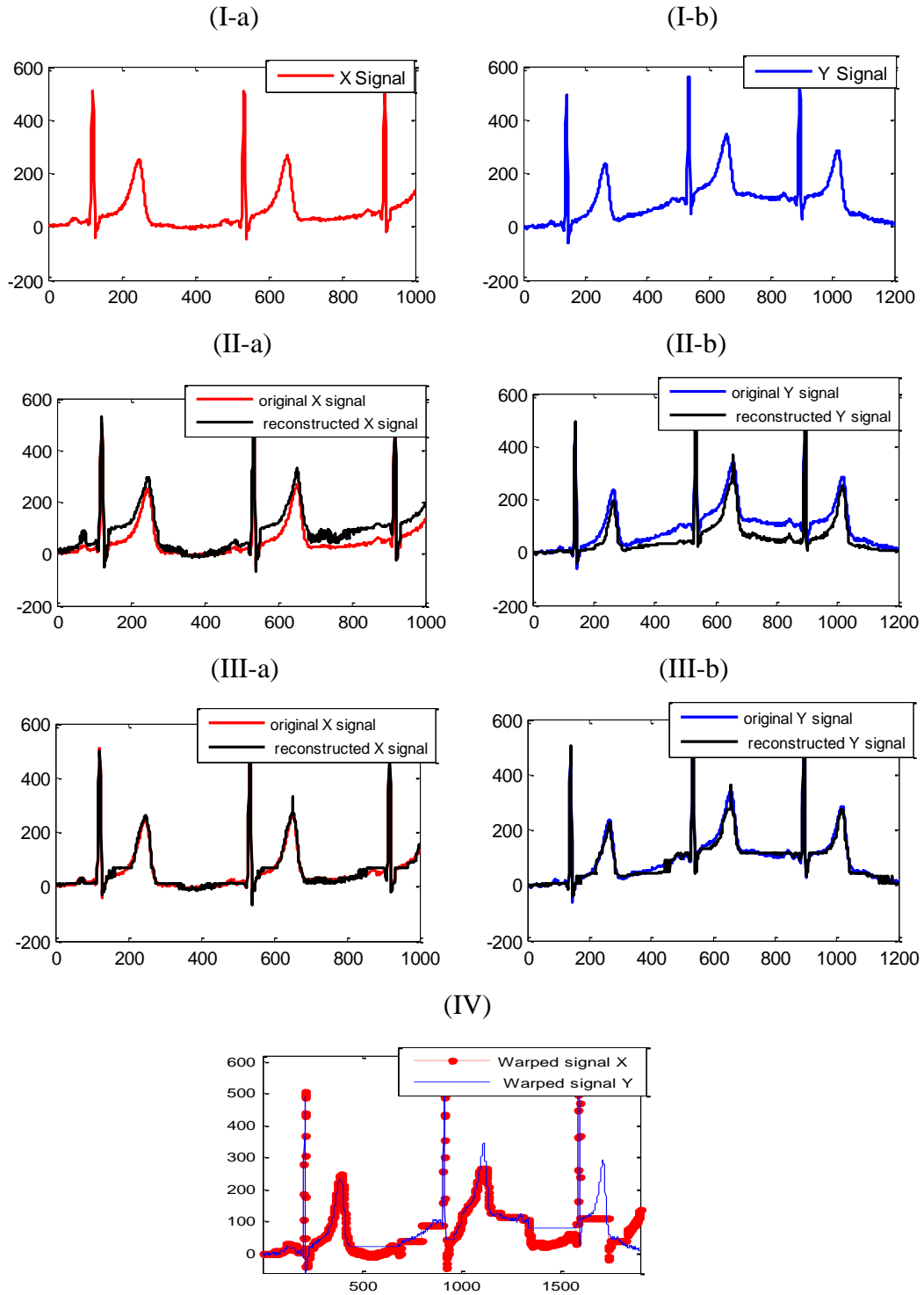


Fig.48. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (113,113) avec un décalage de 30 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

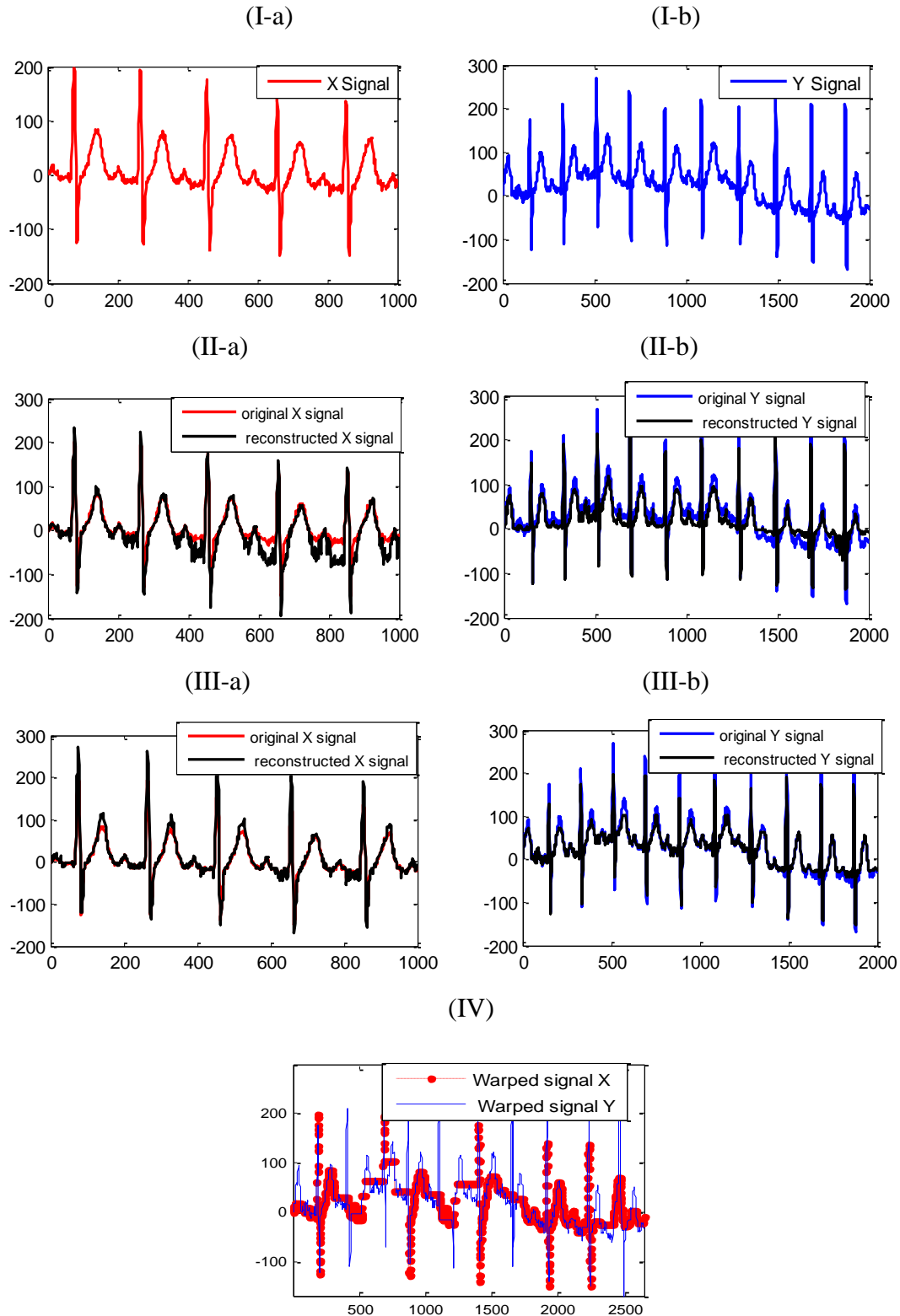


Fig.49. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (215,215) avec un décalage de 35 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

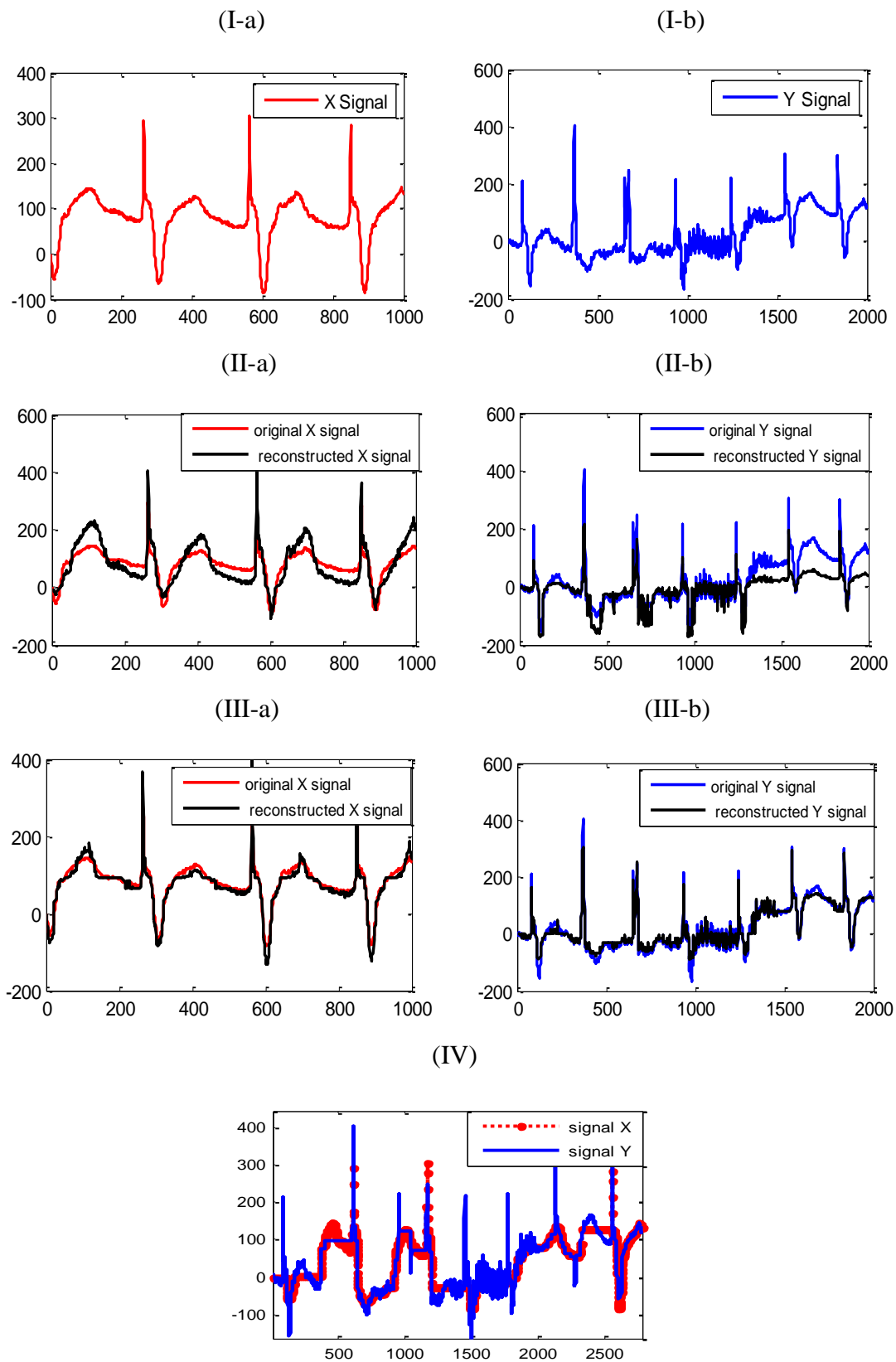


Fig.50. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (104,104) avec un décalage de 20 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

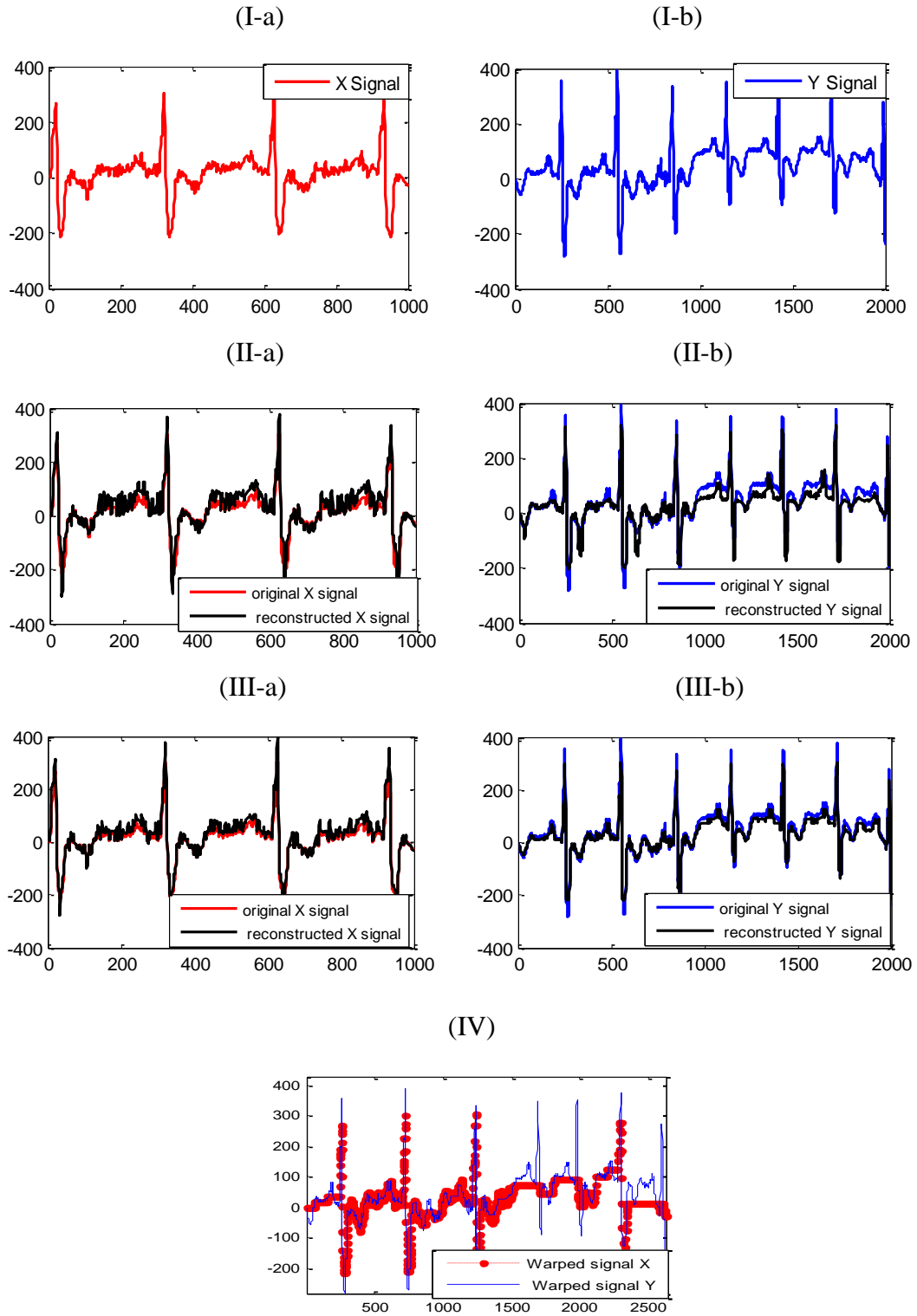


Fig.51. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (118,118) avec un décalage de 35 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

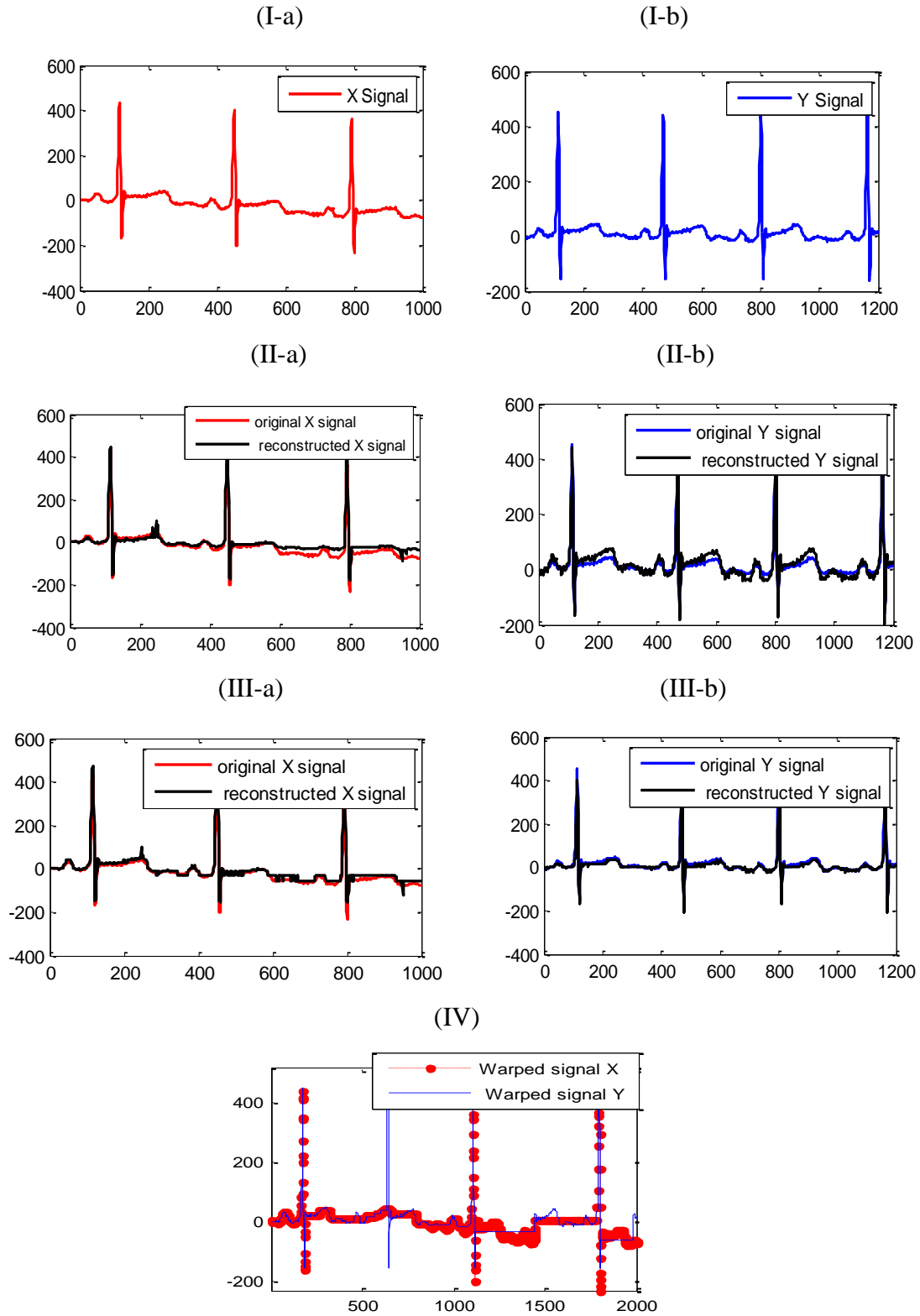


Fig.52. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus du même enregistrement ECG (115,115) avec un décalage de 25 secondes pour X. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

Table 4.a. Comparaison des méthodes DTW, SEA et QP-DTW pour les séries temporelles similaires présentées dans Fig.44, Fig. 45, Fig.46, Fig.47 et Fig.48.

| ECG Segments | DTW | | SEA | | QP-DTW | |
|------------------|-------|--------|-----------|-----------|--------------|---------------|
| | PRD% | Corr | Mean PRD% | Mean Corr | Mean PRD% | Mean Corr |
| (215,215) | 51.97 | 0.8554 | 35.32 | 0.9680 | 23.82 | 0.9887 |
| (113,113) | 40.48 | 0.9225 | 46.85 | 0.9689 | 14.55 | 0.9914 |
| (104,104) | 34.48 | 0.9389 | 51.83 | 0.8734 | 20.51 | 0.9849 |
| (118,118) | 43.91 | 0.9050 | 34.55 | 0.9609 | 23.11 | 0.9922 |
| (115,115) | 47.06 | 0.8827 | 35.75 | 0.9590 | 25.14 | 0.9810 |

VI.3.2 Comparaison de séries temporelles dissimilaires

La Figure 53 montre le cas de deux tracés issu de deux enregistrements différents avec 2000 échantillons de la série temporelle rec.103 pour X et 2000 échantillons de la série temporelle rec.123 pour Y et un décalage de 40 secondes pour Y. L'inspection visuelle de la Figure 53 montre que les séries originales et les séries reconstruites par QP-DTW sont détachées. Les résultats numériques dans le tableau.4.b indiquent une très faible corrélation et un PRD très élevé en faveur de QP-DTW. Cela prouve que notre méthode est capable de différencier des séries temporelles dissemblables.

La Figure 54 montre deux tracés issus de deux enregistrements différents avec 1000 échantillons de la série temporelle rec.100 pour X et 2000 échantillons de la série temporelle rec.112 pour Y et un décalage de 50 secondes pour Y. L'inspection visuelle et les résultats numériques dans le tableau4.b indiquent que QP-DTW a fait le meilleur désalignement.

Les Figures 55 et 56 montrent respectivement les corrélations et les PRDs des trois méthodes en fonction de la taille des séries temporelles. La Figure55 indique que la corrélation de QP-DTW est la meilleure et qu'elle est toujours proche de 1. La figure 56 indique que le PRD de QP-DTW est toujours le plus bas malgré l'augmentation de la taille des données.

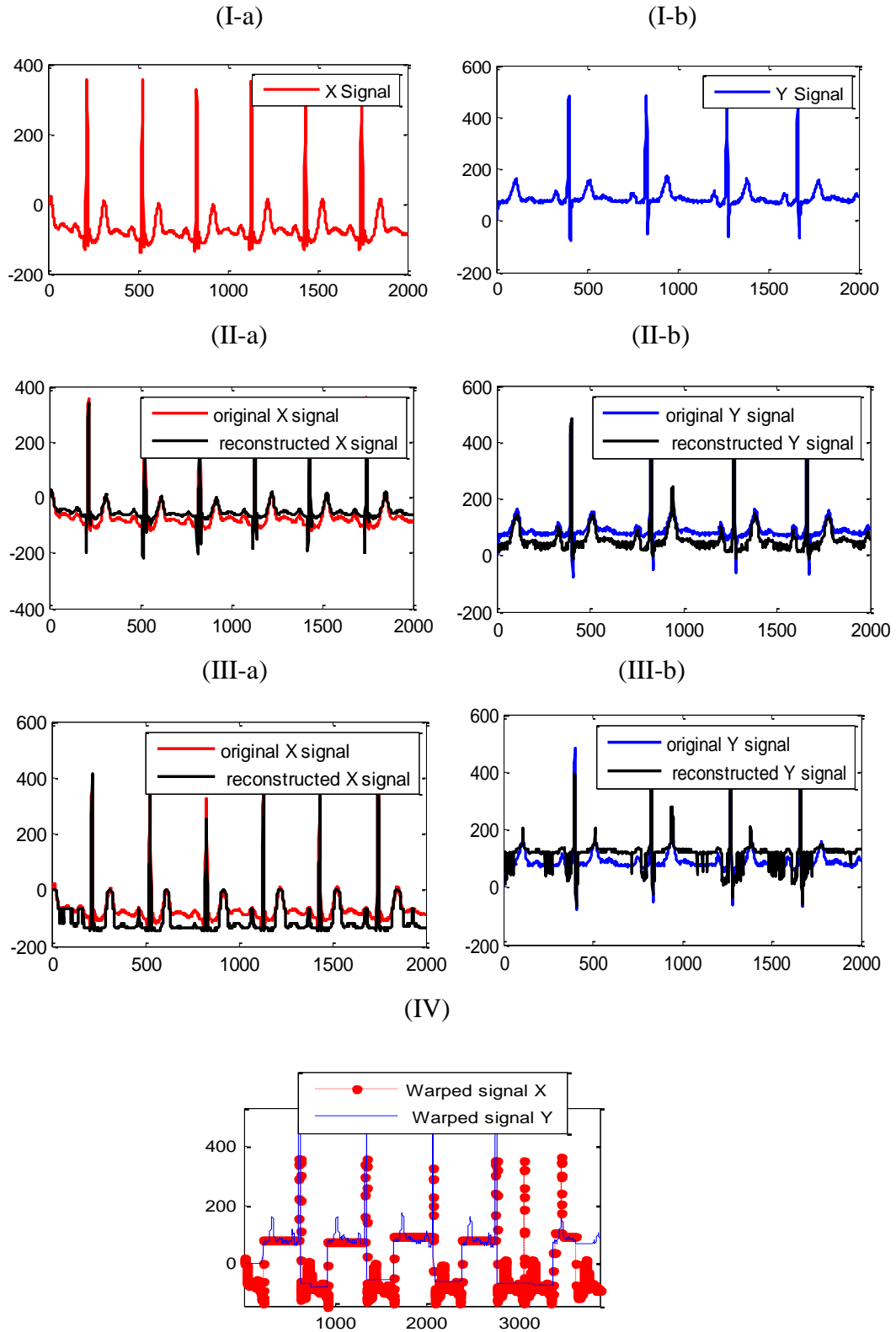


Fig.53. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus de deux enregistrements ECG différents (103,123) avec un décalage de 40 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

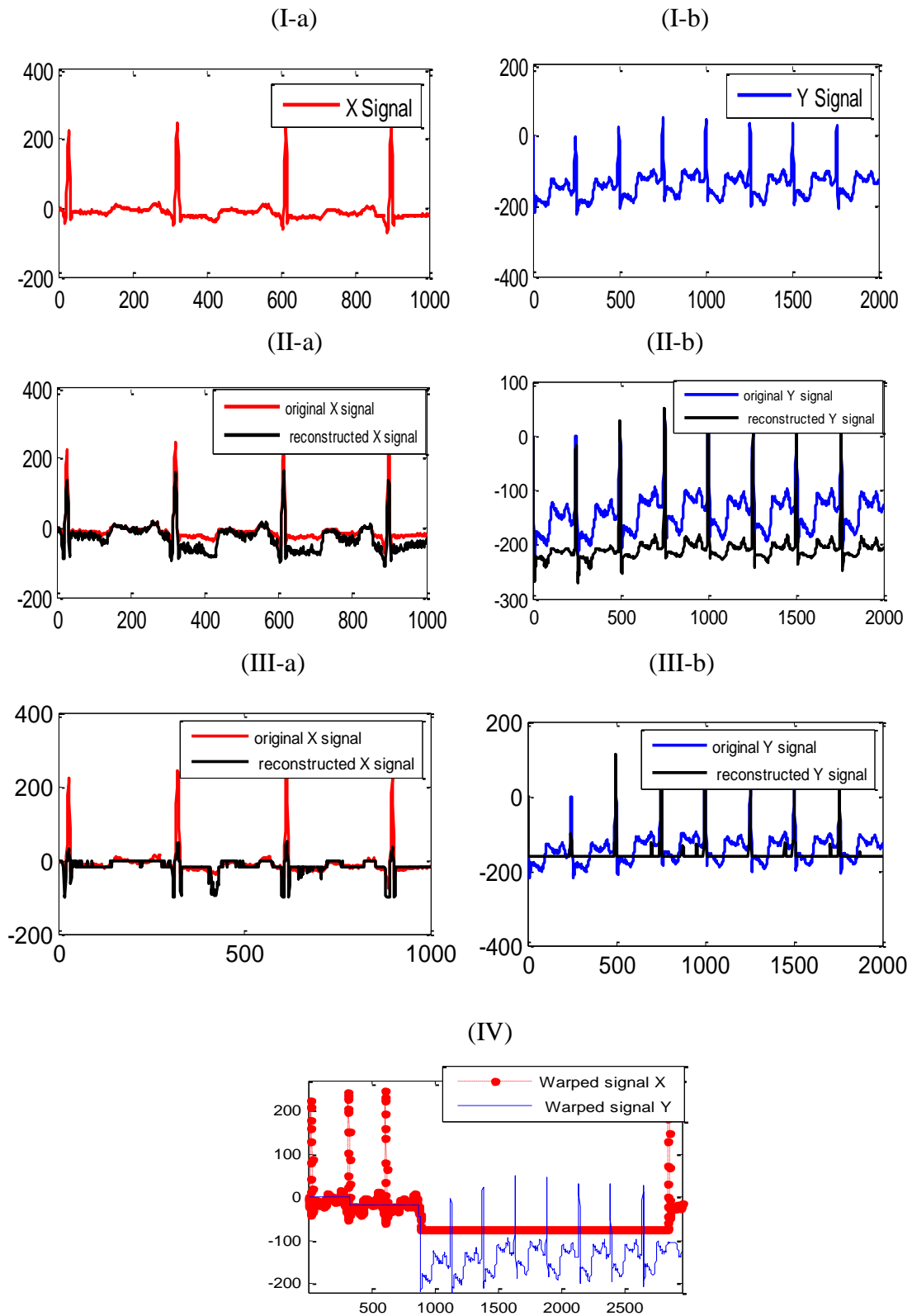


Fig.54. Illustration de l'alignement entre les tracés X (I-a) et Y (I-b) issus de deux enregistrements ECG différents (100,112) avec un décalage de 50 secondes pour Y. L'alignement est effectué par SEA(II), QP-DTW(III) et DTW(IV).

Table 4.b. Comparaison des méthodes DTW, SEA et QP-DTW pour les séries temporelles dissimilaires présentées dans Fig.49, Fig.50.

| ECG Segments | DTW | | SEA | | QP-DTW | |
|--------------|-------|---------------|-----------|-----------|--------------|---------------|
| | PRD% | Corr | Mean PRD% | Mean Corr | Mean PRD% | Mean Corr |
| (103,123) | 57.24 | 0.8472 | 51.67 | 0.9673 | 69.01 | 0.8953 |
| (100,112) | 80.25 | 0.7034 | 62.58 | 0.8709 | 88.38 | 0.5866 |

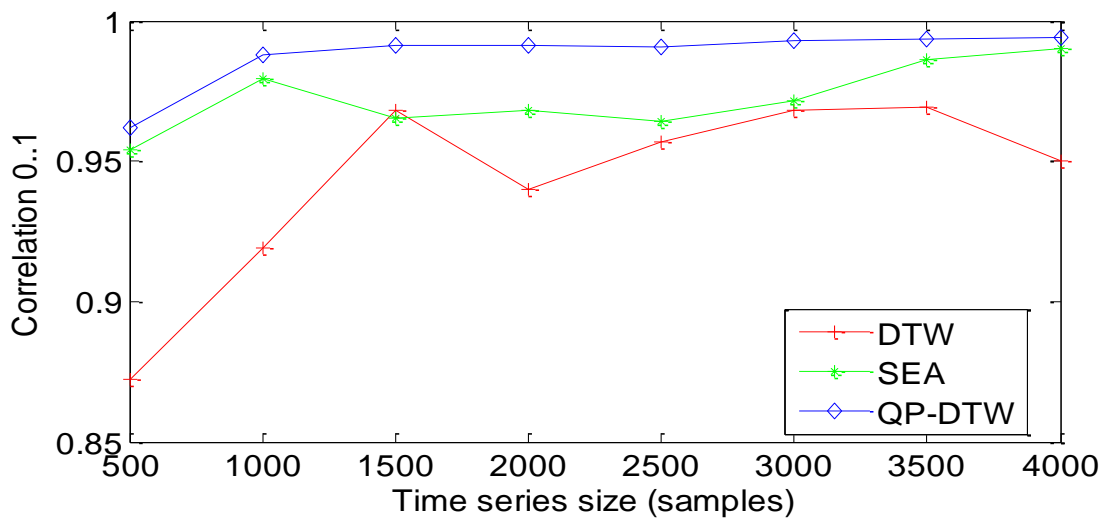


Fig.55. Facteur de corrélation en fonction de la taille de la série temporelle rec.118 (MITBIH) pour les méthodes DTW, SEA et QP-DTW.

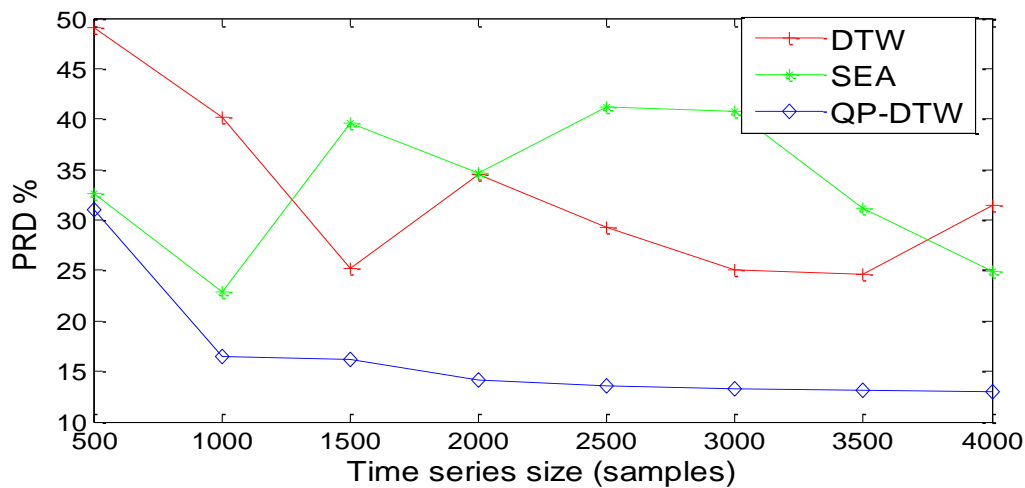


Fig.56.Facteur de PRD en fonction de la taille de la série temporelle rec.118 (MITBIH) pour les méthodes DTW, SEA et QP-DTW.

VI.4 Conclusion

Dans ce chapitre, nous avons proposé une technique d'appariement qui compare et aligne de façon plus pertinente les séries temporelles quasi-périodiques, tels que les tracés ECG.

La méthode proposée QP-DTW, peut être vue comme une amélioration des méthodes DTW et SEA qui peut aligner les séries quasi périodiques mieux que les méthodes SEA et DTW qualitativement et quantitativement. Cependant le temps d'exécution de QP-DTW est significativement supérieur à celui de SEA et de DTW.

Chapitre VII. Méthode proposée basée QPDTW pour la détection d'anomalies dans l'ECG

VII.1 Introduction

Dans ce chapitre, nous allons développer un classifieur binaire basé sur notre méthode QP-DTW pour la détection d'anomalies dans les tracés ECG. Là où la démarche adoptée jusqu'à ce moment consiste à segmenter d'abord ces ST en cycles de base pour pouvoir ensuite les classifier, notre classifieur est capable de classifier un tracé ECG en tracé *normal* ou *anormal* sans avoir besoin de le segmenter. Notre contribution est principalement inspirée du travail de Boucheham [Boucheham, 2011] qui utilise sa méthode SEA. Notre méthode par contre est basée sur la QPDTW.

Dans la première partie de ce chapitre, nous allons tout d'abord expliquer la différence entre un ECG normal et un ECG anormal. Dans la seconde partie, nous allons essayer de nous situer par rapport aux autres travaux qui font la détection d'anomalies dans l'ECG et nous proposerons notre propre méthode par la suite. Dans la dernière partie de ce chapitre, nous ferons des tests expérimentaux sur la base de données des arythmies MIT-BIH.

VII.2 ECG normal et ECG Anormal

L'étude du rythme cardiaque se fait à partir du repérage des ondes R. Ce rythme est caractérisé par deux propriétés : La fréquence des ondes R, exprimée en nombre de battements par minute (bpm), et leur régularité [Dubois, 2004]. La Figure 57 montre les caractéristiques d'un cycle ECG normal.

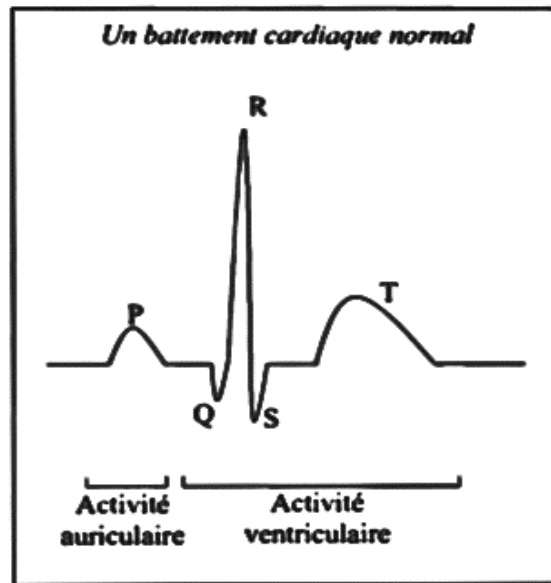


Fig.57. Tracé d'un cycle ECG normal [Dubois, 2004].

Les troubles du rythme cardiaque désignent des battements de cœur trop lents (bradycardie), trop rapides (tachycardie) ou irréguliers (arythmie). Ils sont d'origines variées et non dénuées de conséquences en absence de prise en charge rapide et efficace : des extrasystoles ventriculaires en passant par une tachycardie sinusale (suite à un effort physique par exemple) jusqu'à la fibrillation ventriculaire (arrêt cardiaque) [Rutishauser, 2004].

Dans la Figure 58, le rythme n'est pas normal, car le complexe QRS n'est pas précédé de l'onde P. Ce type d'anomalie s'appelle Extrasystole Ventriculaire (ESV) ou Contraction Ventriculaires Prématuurée (PVC).

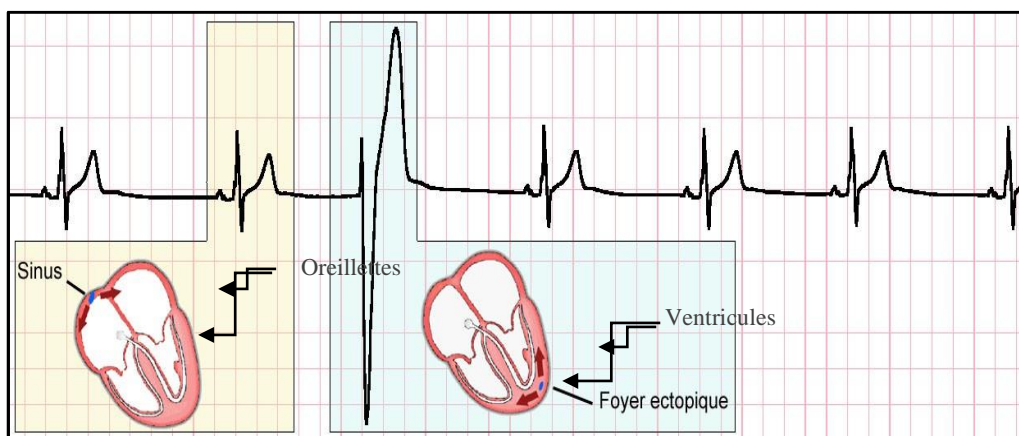


Figure 58. Extrasystole ventriculaire (ESV). Ce type de battement n'est pas précédé d'une onde P, et il a un QRS plus large qu'un QRS normal [Dubois, 2004].

L'exemple suivant illustré par la Figure 59 présente le problème des troubles du rythme atriaux ou auriculaires, ceux-ci se manifestent dans l'oreillette [Farhi 2015].

On parle d'ACFA (arythmie complète par fibrillation auriculaire) quand la fréquence cardiaque est > 90 battements par minute, [Farhi 2015].

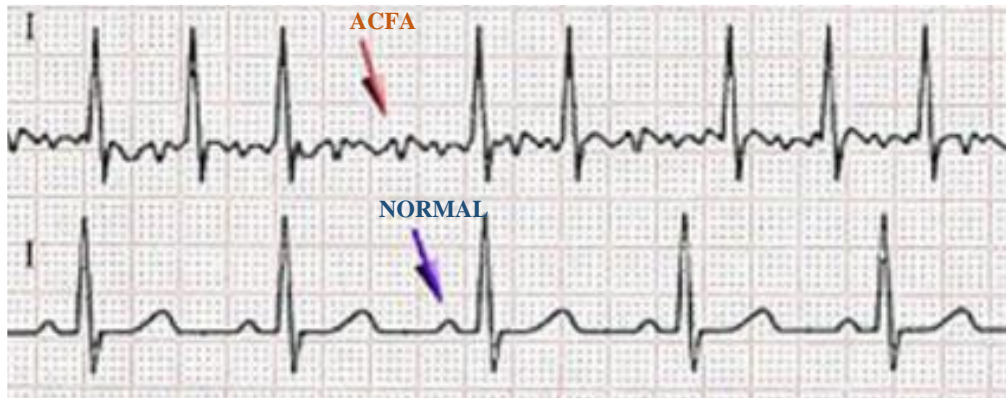


Fig. 59. Fibrillation atriale (Tracé Haut) et ECG normal (Tracé Bas) [Farhi 2015]

Le prochain exemple illustré par la Figure 60 présente le problème de la tachycardie supra-ventriculaire paroxystique. Elle correspond à des crises de tachycardie (accélération du rythme cardiaque), se manifestant par des palpitations à débuts et à fins brusques et dont le mécanisme est bien individualisé.

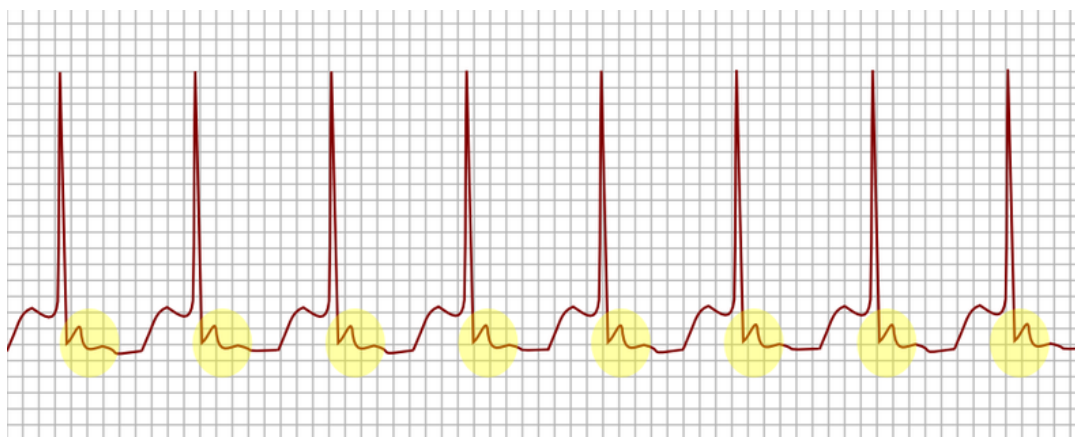


Figure 60. Tachycardie supra-ventriculaire paroxystique. En jaune les ondes P sont juste derrière les complexes QRS [Wikipedia.Maladie de Bouveret]

VII.3 Approches existantes

La détection d'anomalies dans les séries temporelles représentant l'ECG est une application utile pour les cardiologues, car elle les aide à la détection précoce d'anomalies cardiaques dans les tracés ECG qui peuvent parfois durer plusieurs heures. De plus, il est possible de trouver une variabilité dans l'inspection manuelle de l'ECG et le suivi par des experts.

Les principales méthodes de détection d'anomalies reposent sur l'extraction des caractéristiques de l'ECG avec des algorithmes complexes ainsi que sur la segmentation, de l'ECG en cycles pour pouvoir classifier chaque battement cardiaque séparément. Dans le travail de Mi et al [Mi, 2010], un battement cardiaque est intercepté en 3 segments (onde P, intervalle QRS, segment ST) avec une méthode d'analyse en composantes indépendantes qu'ils ont développées. Ensuite, l'algorithme SVM (Support Vector Machine) est utilisé pour la classification des ECG. Dans le travail de Chang et al [Chang, 2012], les tracés ECG ont été transformés en un modèle de densité et un nouveau système hybride avec HMM et GMM a été présenté pour le diagnostic et la classification de l'infarctus du myocarde. Yeh et al [Yeh, 2010] ont utilisé la méthode de clustering c-means pour le clustering et la détection d'anomalies dans l'ECG. Annam et al [Annam, 2011] ont utilisé l'algorithme de clustering K-medoids avec DTW pour identifier les anomalies dans l'ECG. La méthode proposée par Boucheham [Boucheham, 2011] n'a pas besoin de segmenter l'ECG en cycles. Elle s'est avérée efficace pour détecter les changements de morphologie dans les ECG avec une grande précision. L'idée de base de cette méthode est d'effectuer des comparaisons par segments de l'ECG où un segment représente le comportement de référence (normal) et l'autre le comportement inconnu. La comparaison entre les deux segments se fait ensuite avec la méthode SEA. Dans [Karadayi, 2020], une nouvelle approche pour la détection d'anomalies dans un ensemble de données spatio-temporelles à plusieurs variables est proposée en utilisant un modèle hybride d'apprentissage profond.

VII.4 Méthode proposée

Comme pour la méthode proposée par Boucheham [Boucheham, 2011], notre méthode n'a pas besoin de segmenter l'ECG en cycles et pour extraire les caractéristiques de l'ECG. Elle choisie un segment ECG complet et le compare avec un autre segment. Tout simplement. Ceci est rendu possible par le pouvoir d'alignement des STQP dont jouissent les méthodes SEA et QPDTW (à titre exclusif, cela dit).

Les étapes principales de notre méthode se présentent comme suit :

Algorithme QP-DTW pour la détection d'anomalie dans l'ECG :

Etape1 : Extraire les ECG

Dans cette étape, un segment ECG normal N de taille n et un segment ECG en entré R de taille m sont extrait du même enregistrement (même personne) à partir de la base de données des ECG MIT-BIH. Nous prenons en compte l'annotation manuelle des ECG pris par les cardiologues pour extraire un segment normal N et préciser le type d'anomalie (si elle existe) contenu dans le segment R.

Etape2 : Appliquer l'algorithme QP-DTW

- L'algorithme DTW est exécuté pour étirer le segment normal N et le segment R et pour faire une première mise en correspondance entre eux.
- Le tri des segments sur leurs amplitudes est établi pour extraire leurs signatures.
- Echange de signature entre les segments N et R et rétablissement de leurs tailles normales.
- Le tri sur les index temporels respectifs des segments N et R pour les reconstruire. Nous obtenons les segments reconstruits N_{Rec} et R_{Rec} .

Etape3 : Comparaison et décision

Dans cette étape, le facteur de corrélation est calculé pour le couple (R, R_{Rec}) . Nous utilisons une valeur de seuil T sur le facteur de corrélation pour décider de l'existence d'une anomalie :

- Si $Corr(R, R_{Rec}) < T$, alors 'il y a une anomalie' Sinon 'aucune anomalie n'est détectée'.
- L'inspection visuelle est utilisée comme un critère subjectif.

VII.5 Tests expérimentaux

Pour tester notre méthode de détection d'anomalies, nous utilisons la base de données des ECG MIT-BIH [MITBIH, 2000].

Nous sélectionnons cinq enregistrements numérotés dans la base de données comme les enregistrements 223, 219, 218, 104, 223. Les Figures [61-65] illustrent la capacité de notre méthode à détecter le changement de morphologie dans l'ECG de chaque personne en comparant son segment ECG sain et son segment ECG anormal. Le tableau5 démontre numériquement la précision d'alignement de notre méthode. Grâce à des expériences intensives, nous avons constaté que les facteurs de corrélation inférieurs à 0,950 indiquent qu'il y a un changement de morphologie entre les segments comparés, alors la valeur de seuil $T=0,950$ a été retenue pour cette base.

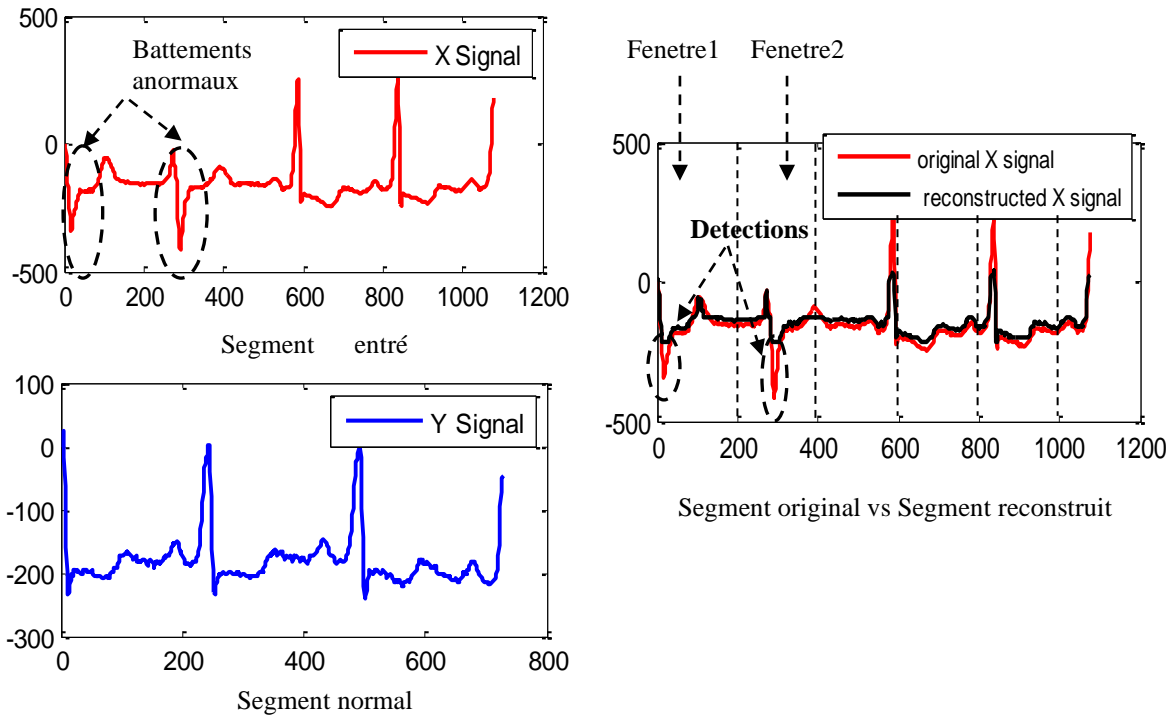


Fig.61.Détection de changements morphologiques (fenêtre 1 et 2) en comparant deux segments ECG de la même personne; Dans ce cas-là c'est à cause des contractions auriculaires prématurées (PAC) rencontré dans l'enregistrement 223. Taux d'erreur : 0.884.

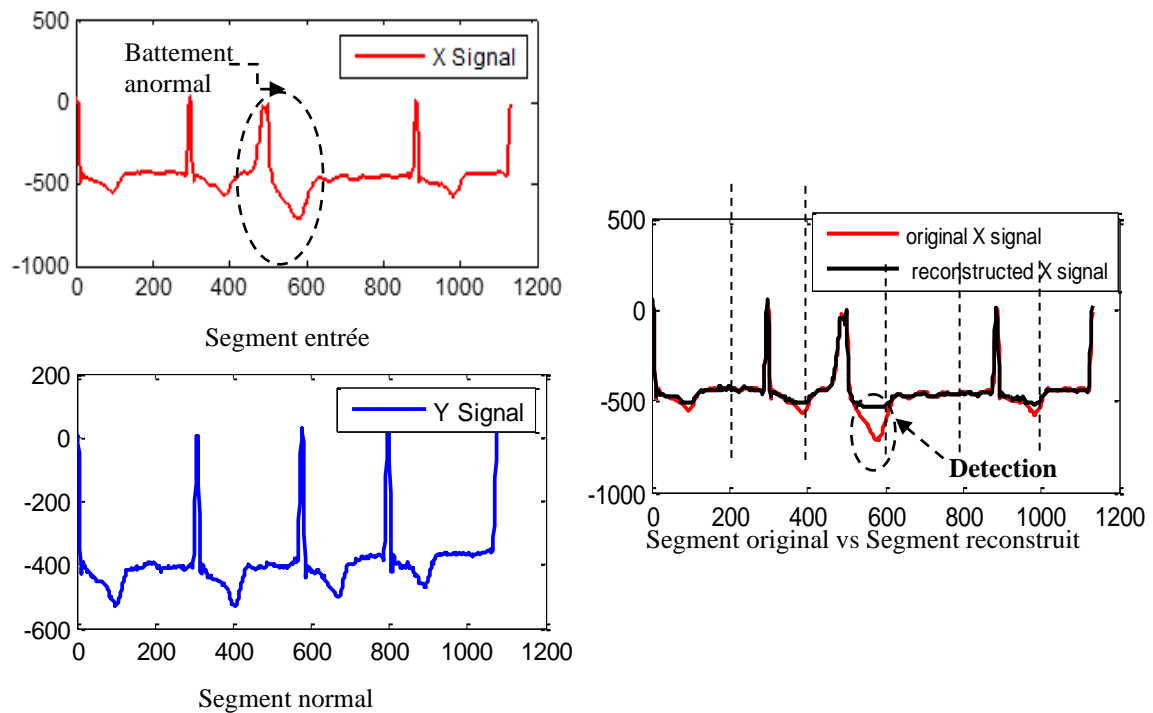


Fig.62.Détection d'un changement morphologique (Fenêtre 3) en comparant deux segments ECG de la même personne; Dans ce cas là, c'est à cause d'une contraction ventriculaire prématurée (PVC) rencontrée dans l'enregistrement 219. Taux d'erreur : 0.947.

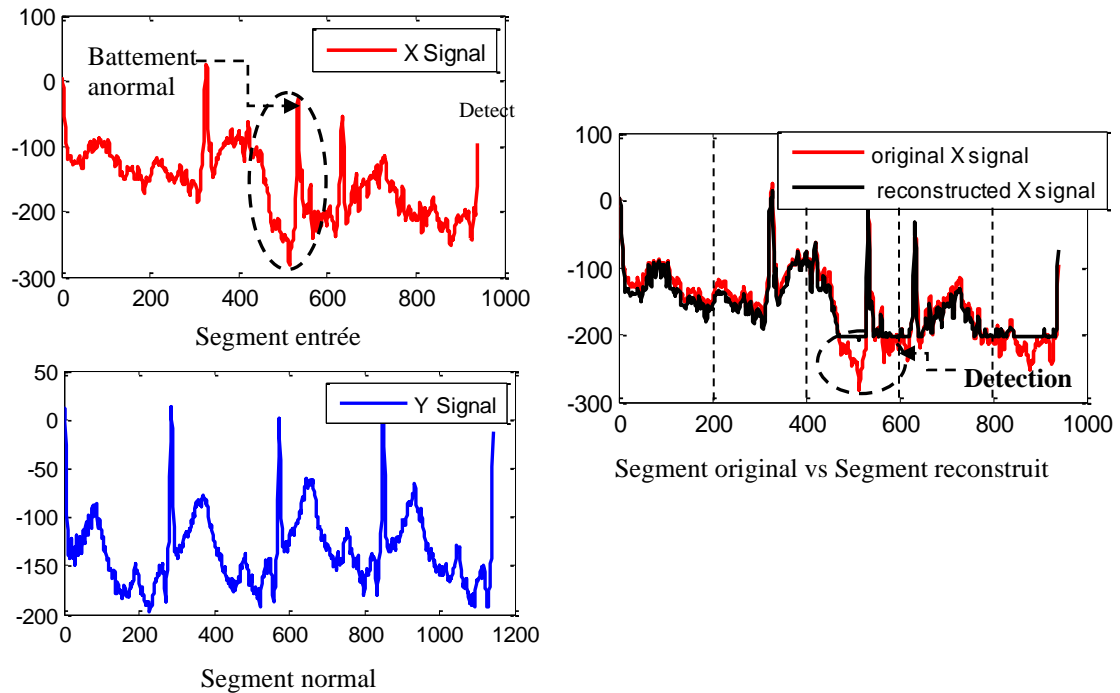


Fig.63. Détection de changements morphologiques (Fenêtre 3) en comparant deux segments ECG de la même personne; Dans ce cas-là c'est d'un QRS isolé à cause du bruit rencontré dans l'enregistrement 228. Taux d'erreur : 0.944.

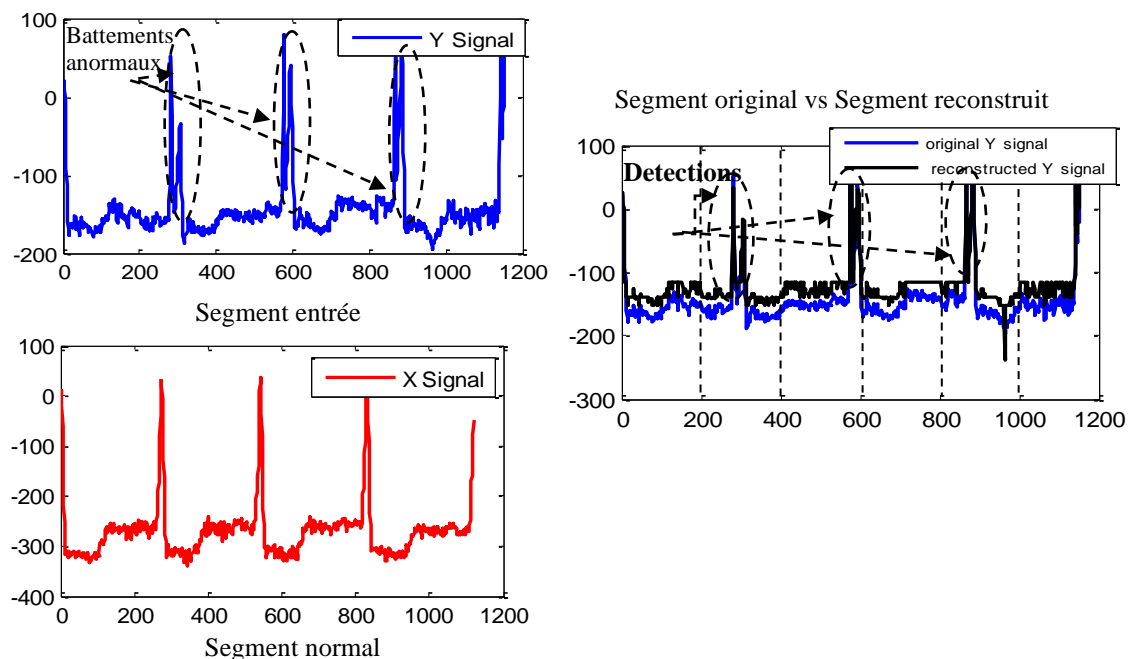


Fig.64. Détection de changements morphologiques (Fenêtre 2, 3, 5) en comparant deux segments ECG de la même personne; Dans ce cas-là c'est un Bloc de branche observé dans l'enregistrement 104. La contraction des deux ventricules n'est pas parfaitement simultanée, ce qui s'observe par deux pics successifs correspondant aux deux ventricules. Taux d'erreur : 0.903.

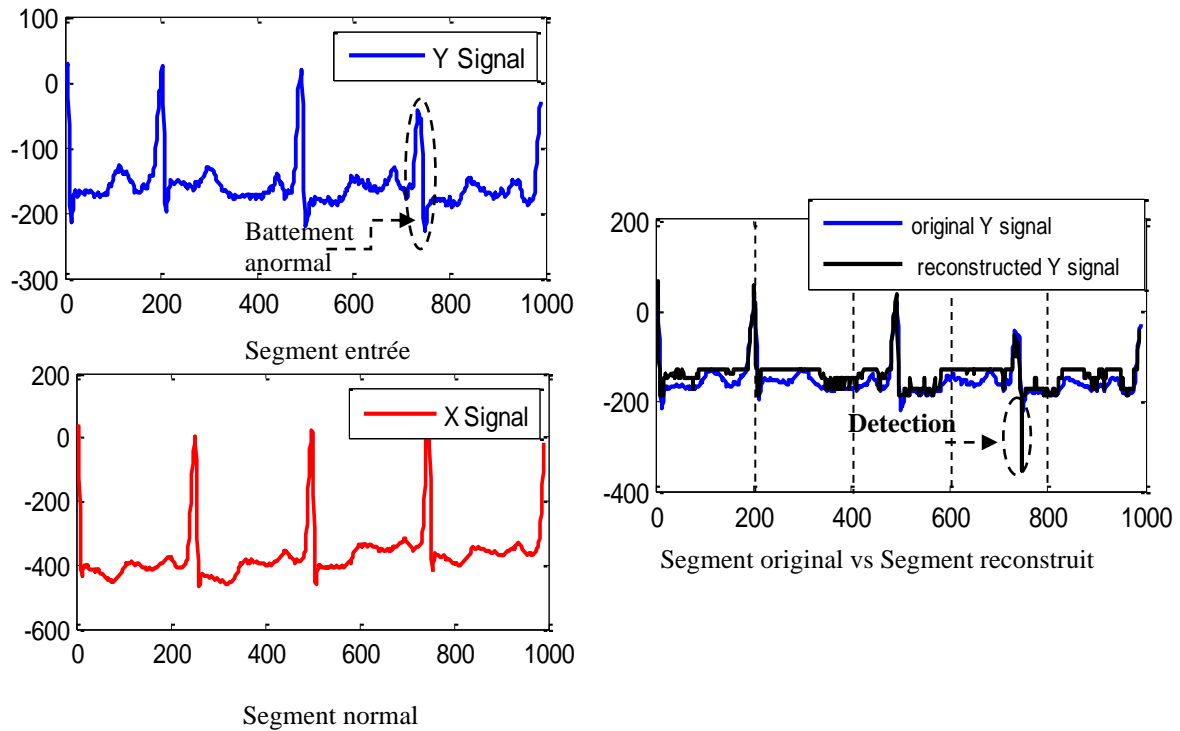


Fig.65.Détection de changements morphologiques (Fenêtre 4) en comparant deux segments ECG de la même personne; Dans ce cas-là c'est à cause d'une contraction ventriculaire prématurée (PVC) rencontré dans l'enregistrement 223. Taux d'erreur : 0.906.

Tableau5.Résultats de la méthode QP-DTW pour la détection d'anomalies dans l'ECG

| Segments ECG | Anomalies | Corrélations |
|--------------|---------------------------------------|--------------|
| (223, 223) | Contractions auriculaires prématurées | 0.884 |
| (219, 219) | Contraction ventriculaire prématurée | 0.947 |
| (228, 228) | QRS isolés à cause du bruit | 0.944 |
| (104, 104) | Bloc de branche | 0.903 |
| (223, 223) | Contraction ventriculaire prématurée | 0.906 |

La Table 5 indique les résultats de la comparaison des cinq segments utilisés avec leurs segments reconstruits. Il indique une corrélation très faible pour le segment 223, ce qui implique une différence entre le segment entré et le segment normal. La figure 61 illustre les changements significatifs de morphologies dans le segment entré 223, ceci étant dû au "Complexe Prématuré auriculaire"(PAC).

La Table 5 indique une corrélation moyenne, inférieure à la valeur T (0,95) pour le segment 219. D'autre part, la Figure 62 montre une différence significative dans X_{Rec} par rapport à X. Ces résultats indiquent qu'il y a un changement de morphologie dans

l'enregistrement 219. Dans ce cas, l'anomalie est une "contraction ventriculaire prématurée" (PVC).

Dans la Figure 63, nous détectons un changement de morphologie en comparant le segment 228 avec sa reconstruction. L'anomalie observée est un QRS isolé à cause du bruit rencontré lors de l'enregistrement. Le tableau 5 confirme ce résultat avec une corrélation inférieure au seuil T.

La Figure 64 illustre les changements de morphologie entre l'enregistrement 104 et sa reconstruction. L'anomalie est un Bloc de branche. Le tableau 5 indique une corrélation très faible, ce qui confirme la différence entre les deux segments.

La Figure 65 montre une différence entre Y et Y_{Rec} dans le troisième QRS causée par un PVC. Les résultats numériques du tableau 5 indiquent une très faible corrélation, ce qui confirme le changement de morphologie dans l'enregistrement 223.

VII.6 Conclusion

Dans ce chapitre, nous avons proposé une méthode pour la détection d'anomalies dans les tracés ECG. Notre méthode est simple à mettre en œuvre, elle ne nécessite pas la segmentation de l'ECG en cycles et n'a besoin que d'un seul paramètre, qui est le seuil T posé sur le facteur de corrélation.

Chapitre VIII. Méthode proposée A-QPDTW pour la recherche de similarité dans les séries temporelles

VIII.1 Introduction

Plusieurs méthodes ont été proposées pour la recherche de similarité dans les séries temporelles (ST). Notre méthode QP-DTW (Quasi Periodic Dynamic Time Warping), présentée précédemment, a été proposée comme une méthode concurrente à DTW pour l'alignement des ST quasi périodique. Cependant, la méthode est coûteuse en termes de complexité temporelle. Par conséquent, elle n'est pas appropriée pour la recherche de similarité dans les ST, où les méthodes avec une complexité temporelle moindre sont nécessaires. Dans ce chapitre, nous proposons une version accélérée de la méthode QP-DTW et nous l'utilisons pour résoudre le problème de la recherche de séries temporelles ECG (électrocardiogramme) similaires dans une base de données. Nous nous sommes concentrés sur l'ECG pour son caractère morphologique Quasi-Périodique. La méthode accélérée de QP-DTW (A-QPDTW) utilise la transformée de Haar en ondelettes comme technique pour réduire la dimensionnalité des ST et l'approche KNN (K nearest neighbor) pour rechercher des ST qui sont similaires à une requête donnée. Les résultats expérimentaux montrent que la méthode proposée est plus efficace que la méthode QPDTW avec une précision comparative. Tous les résultats ont été obtenus en utilisant la Base de séries temporelles ECG de l'UCR (Université de Californie à Riverside), universellement admis comme premier point de référence dans les analyses de séries temporelles et comptant au fait 85 bases de ST de sources variées [Chen, 2015].

VIII.2 Méthode proposée : AQP-DTW

Dans cette section, nous décrivons la méthode proposée: Accelerated Quasi-Periodic Dynamic Time Warping (AQP-DTW), [Boulnemour, 2020-a]. A-QPDTW est principalement composée de deux étapes.

A. Extraction de caractéristiques

Dans la littérature, il existe plusieurs types de transformations en ondelettes. Nous avons utilisé les transformées de Haar en ondelettes à un seul niveau en raison de sa simplicité, de son efficacité (ordre linéaire de complexité), de sa capacité à extraire les caractéristiques discriminantes et à réduire la dimension des séries temporelles quasi périodiques.

B. Mesure de similarité

La deuxième étape consiste à appliquer la mesure de similarité QP-DTW sur les données réduites et à rechercher les ST similaire à une requête donnée dans une base de données, en utilisant l'approche KNN. La Figure 66 illustre le diagramme de la méthode proposée A-QPDTW.

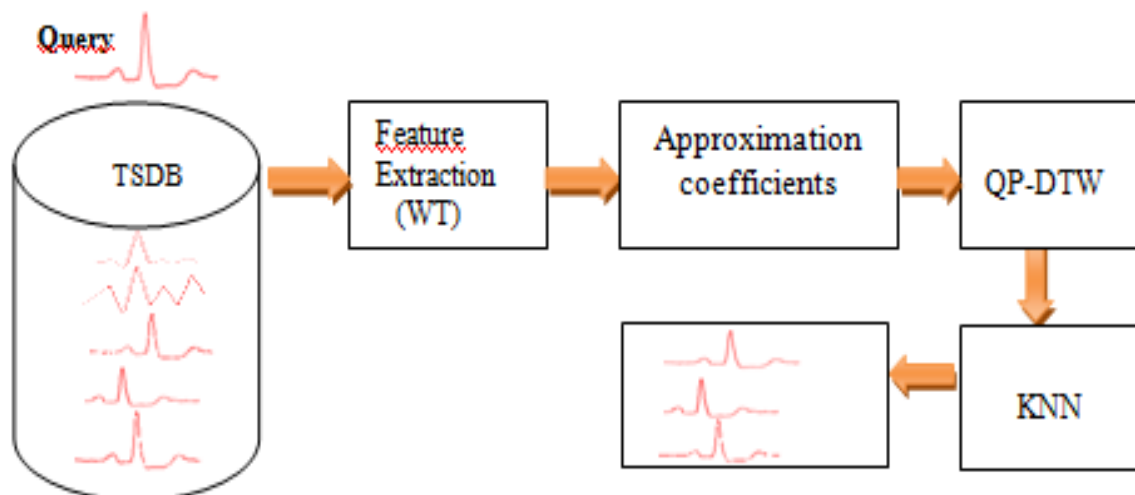


Fig.66. Diagramme illustratif de la méthode AQP-DTW proposée pour la recherche de similarité dans les ST.

VIII.3 Expérimentation

Dans cette section, nous évaluons l'efficacité et l'efficacite de notre algorithme sur le probleme de la recherche de similarite dans les series temporelles d'ECG. Les donnees, sont selectionnees dans les archives UCR [Chen, 2015]. L'UCR contient 7 series temporelles d'ECG et chacune d'entre elles possede une partition standard de l'ensemble d'entraınement et de test. Une description de l'ensemble de donnees utilise dans notre etude est presentee dans la Table 6.

Table 6. Description de la base de donnees ECG utilisee

| Nom de la base de donnees | Nombre de classes | Taille de l'ensemble d'apprentissage | Taille de l'ensemble de test | Longueur des ST |
|-----------------------------|-------------------|--------------------------------------|------------------------------|-----------------|
| ECG200 | 02 | 100 | 100 | 96 |
| ECGFiveDays | 02 | 23 | 861 | 136 |
| ECG5000 | 05 | 500 | 4500 | 140 |
| CinC_ECG_torso | 04 | 40 | 1380 | 1639 |
| TwoLeadECG | 02 | 23 | 1139 | 82 |
| NonInvasiveFatalECG Thorax1 | 42 | 1800 | 1965 | 750 |
| NonInvasiveFatalECG Thorax2 | 42 | 1800 | 1965 | 750 |

Nous commencons par evaluer la methode proposee AQP-DTW, en termes d'efficacite (duree d'execution) par rapport a la QP-DTW, sur le probleme de la recherche de similarite. Le temps d'execution est obtenu en calculant le temps d'execution moyen de toutes les requetes dans l'ensemble de test. La Figure 67, montre le temps d'execution en secondes pour A-QPDTW et QP-DTW, sur toutes les series temporelles d'ECG de l'UCR.

Nous remarquons que la methode AQP-DTW est plus efficace que la methode QP-DTW. En fait, elle est environ trois fois plus rapide que la methode QP-DTW dans tous les ensembles de donnees. En outre, dans le cas des longues series temporelles, nous montrons que notre methode est encore plus rapide. Par exemple, pour l'ensemble de donnees CinC_ECG_torso, le temps d'execution de la methode AQP-DTW est de 1046, 20 secondes alors qu'il est de 4228,75 secondes pour la methode QP-DTW (quatre fois plus rapide). Un

autre exemple, pour le petit ensemble de données ECG200, le temps d'exécution de l'AQP-DTW est de 0,37 secondes et de 1,12 secondes pour le QP-DTW.

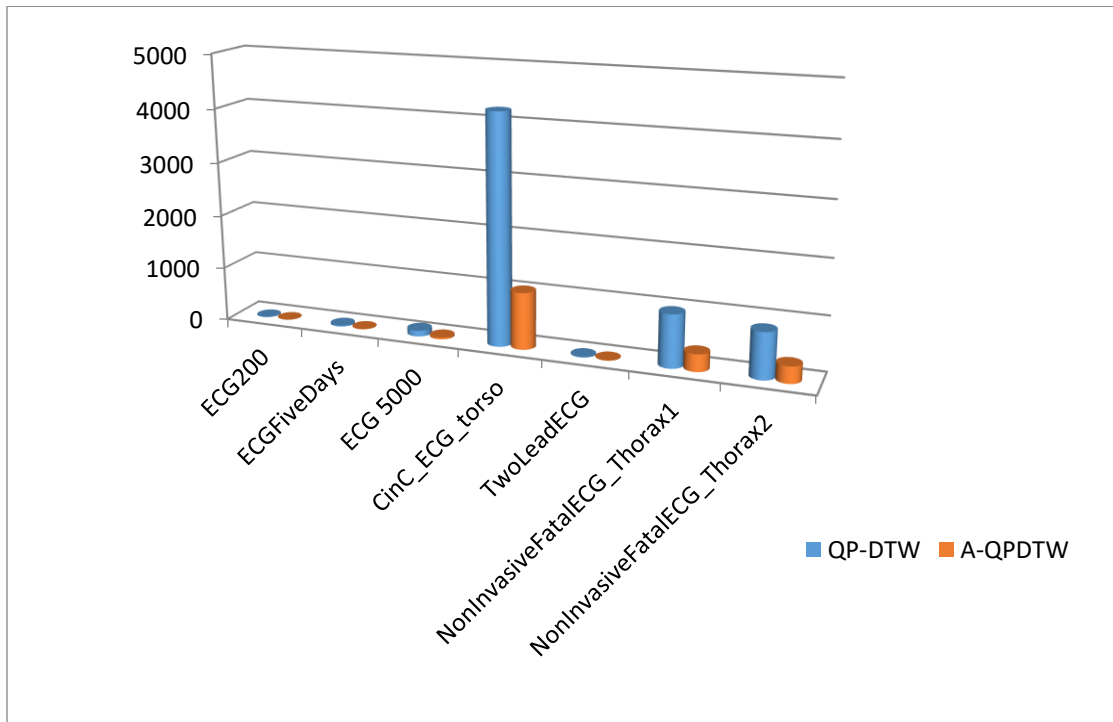


Fig.67. Temps d'exécution de A-QPDTW par rapport à QP-DTW.

D'autre part, nous comparons la précision de notre méthode AQP-DTW par rapport à la méthode QP-DTW dans le cadre du problème de la recherche de similarité. À cet égard, nous mentionnons que le critère de précision mesure la fraction des objets pertinents retournés pour une requête donnée par rapport au total des objets retournés [Garcia, 2014] (voir la figure 68).

Le tableau 7 montre les résultats expérimentaux de la comparaison des précisions de A-QPDTW et QP-DTW sur toutes les séries d'ECG de l'UCR. Les résultats montrent que les précisions des deux méthodes sont comparables entre elles. Pour l'ensemble de données "CinC_ECG_torso", la précision de la méthode QP-DTW est de 0,275 et celle de la méthode AQP-DTW de 0,273. Pour l'ensemble de données "NonInvasiveFatalECG_Thorax1", la précision de la méthode QP-DTW est de 0,698 et celle de la méthode AQP-DTW de 0,693. En effet, AQP-DTW gagne dans deux ensembles de données et perd dans cinq ensembles de données par rapport au QP-DTW.

Table 7. Précision de A-QPDTW V.S QP-DTW

| Ensemble de données | QP-DTW | A-QPDTW |
|-----------------------------|--------|---------|
| ECG200 | 0.693 | 0.632 |
| ECGFiveDays | 0.576 | 0.552 |
| ECG 5000 | 0.720 | 0.728 |
| CinC_ECG_torso | 0.275 | 0.273 |
| TwoLeadECG | 0.532 | 0.578 |
| NonInvasiveFatalECG_Thorax1 | 0.698 | 0.693 |
| NonInvasiveFatalECG_Thorax2 | 0.730 | 0.712 |

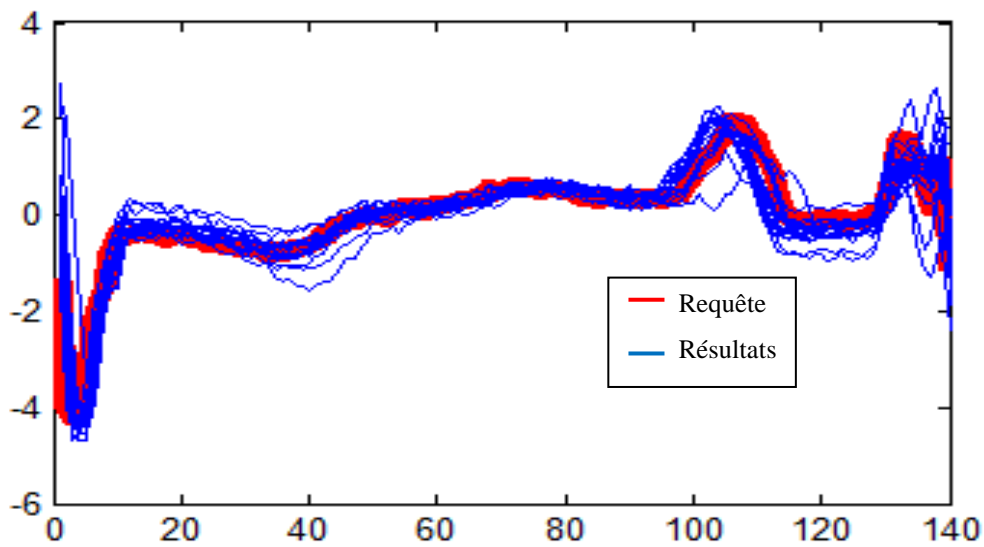


Figure 68. Précision de la méthode AQP-DTW. La ST en rouge représente l'objet requête et la ST bleue représente les 20 objets retournées qui sont similaires à la requête. ($k=20$).

VIII.4 Conclusion

Dans ce chapitre, nous nous sommes concentrés sur le problème de la recherche de similarité dans la base de données ECG. À cet égard, nous avons mis au point une mesure de similarité efficace. La méthode A-QP-DTW obtenue est basée sur la transformation de Haar en ondelettes et la mesure de similarité QP-DTW. Les résultats expérimentaux montrent que la méthode proposée est environ trois fois plus rapide que QP-DTW, sans pratiquement aucune perte de précision. Dans le cadre de travaux futurs, nous souhaitons appliquer la méthode proposée dans le domaine de la biométrie.

Chapitre IX. Méthode proposée basée 1NN-QPDTW pour la classification des séries temporelles

IX.1 Introduction

Plusieurs approches ont été proposées dans la littérature pour résoudre le problème de la classification des séries temporelles. Toutefois, les études montrent qu'un simple classifieur 1NN utilisé avec une mesure de similarité appropriée est très efficace et surpasse largement de nombreuses approches de classification complexes [Ding, 2008]. En particulier, il est rapporté que 1-NN avec DTW est difficile à battre [Xi, 2006], [Guisti, 2013], [Bagnall, 2014], [Petitjean, 2016].

Dans la première partie de ce chapitre, nous proposerons notre méthode de classification des séries temporelles à base d'une combinaison entre la méthode QP-DTW et le classifieur 1-NN. Dans la seconde partie de ce chapitre, nous ferons des tests expérimentaux sur le benchmark universel de l'UCR (Université de Californie à Riverside) [Chen, 2015].

IX.2 Méthode proposée

Le choix des mesures de distance (similarité) joue un rôle important dans la qualité de l'algorithme de classification de l'algorithme KNN [Hassanat, 2014]. La distance euclidienne (ED) est une mesure largement utilisée pour le classifieur KNN [Xing, 2010]. Cependant, elle exige que les deux séries soient de longueurs égales en comparaison. En plus de cette limitation, Ratanamahatana et Keogh ont souligné sa sensibilité à la distorsion dans le temps [Ratanamahatana, 2004-a]. D'autre part, la combinaison 1NN-DTW semble être la meilleure méthode pour la classification des séries temporelles [Batista, 2011], [Petitjean, 2016], mieux que les réseaux de neurones, les arbres de décision, les histogrammes multi-échelles et les modèles de Markov cachés, car ces classifieurs ont besoin de données d'entraînement et donc d'un point de vue d'experts [Wei, 2006].

Dans ce qui suit, nous présentons la méthode proposée 1NN-QPDTW (Figure 69), [Boulnemour, 2021].

Algorithme de classification 1NN-QPDTW :

```
Entrées:  $P = \{P_1, P_2, \dots, P_N\}$  ST pour l'apprentissage.  
   $Q$ : TS(Série temporelle à classifier)  
Sorties: Classe de  $Q$   
Début  
  Best  $\leftarrow -\infty$  // initialiser la meilleure distance jusqu'à présent  
  Classe  $\leftarrow \text{NaN}$   
  Pour  $i \leftarrow 1$  jusqu'à  $N$  faire  
    Distance( $Q, P_i$ )  $\leftarrow$  QPDTW ( $Q, P_i$ )  
    Si (Distance  $<$  Best)  
      Best  $\leftarrow$  Distance  
      Classe  $\leftarrow$  Classe de  $P_i$   
    Finsi  
  FinPour  
Retour (Classe)  
End.
```

Figure 69. Algorithme de classification des séries temporelles “One Nearest Neighbor Quasi Periodic Dynamic Time Warping, (1NN-QPDTW)”.

IX.3 Expérimentation

Dans cette section, nous évaluons les performances de notre méthode 1NN-QP-DTW pour la classification des séries temporelles. Nous avons ainsi choisi un seul voisin $K=1$ pour nous assurer que les performances des mesures de similarités ne soient pas liées au nombre de voisins choisis. Aussi, afin de vérifier le contraste entre les mesures de similarités, nous avons mené une vaste série d'expérimentations sur les 85 ensembles de données de classification des séries temporelles contenus dans la base de l'UCR. Ces archives de données comprennent 85 bases de ST réelles, synthétiques et génériques provenant de différents domaines d'application. Ils fournissent des jeux de séries temporelles comme référence pour tester et évaluer les algorithmes de classification des séries temporelles. Chaque série temporelle a une partition standard d'ensemble d'apprentissage et de test. Une description de l'ensemble des données utilisé est présentée dans la Table 8.

Table 8. a. Ensembles des séries temporelles

| Ensemble de données | Nombre de classes | Ensemble d'apprentissage | Ensemble de test | Longueur des séries temporelles |
|------------------------------|-------------------|--------------------------|------------------|---------------------------------|
| 50words | 50 | 450 | 455 | 270 |
| Adiac | 37 | 390 | 391 | 176 |
| ArrowHead | 3 | 36 | 175 | 251 |
| Beef | 5 | 30 | 30 | 470 |
| BeetleFly | 2 | 20 | 20 | 512 |
| BirdChicken | 2 | 20 | 20 | 512 |
| Car | 4 | 60 | 60 | 577 |
| CBF | 3 | 30 | 900 | 128 |
| ChlorineConcentration | 3 | 467 | 3840 | 166 |
| CinC_ECG_torso | 4 | 40 | 1380 | 1639 |
| Coffee | 2 | 28 | 28 | 286 |
| Computers | 2 | 250 | 250 | 720 |
| Cricket_X | 12 | 390 | 390 | 300 |
| Cricket_Y | 12 | 390 | 390 | 300 |
| Cricket_Z | 12 | 390 | 390 | 300 |
| DiatomSizeReduction | 4 | 16 | 306 | 345 |
| DistalPhalanxOutlineAgeGroup | 3 | 139 | 400 | 80 |
| DistalPhalanxOutlineCorrect | 2 | 276 | 600 | 80 |
| DistalPhalanxTW | 6 | 139 | 400 | 80 |
| Earthquakes | 2 | 139 | 322 | 512 |
| ECG200 | 2 | 100 | 100 | 96 |
| ECG5000 | 5 | 500 | 4500 | 140 |
| ECGFiveDays | 2 | 23 | 861 | 136 |
| ElectricDevices | 7 | 8926 | 7711 | 96 |
| FaceAll | 14 | 560 | 1690 | 131 |
| FaceFour | 4 | 24 | 88 | 350 |
| FacesUCR | 14 | 200 | 2050 | 131 |
| FISH | 7 | 175 | 175 | 463 |
| FordA | 2 | 1320 | 3601 | 500 |
| FordB | 2 | 810 | 3636 | 500 |
| Gun_Point | 2 | 50 | 150 | 150 |
| Ham | 2 | 109 | 105 | 431 |
| HandOutlines | 2 | 370 | 1000 | 2709 |
| Haptics | 5 | 155 | 308 | 1092 |
| Herring | 2 | 64 | 64 | 512 |
| InlineSkate | 7 | 100 | 550 | 1882 |
| InsectWingbeatSound | 11 | 220 | 1980 | 256 |
| ItalyPowerDemand | 2 | 67 | 1029 | 24 |
| LargeKitchenAppliances | 3 | 375 | 375 | 720 |
| Lighting2 | 2 | 60 | 61 | 637 |
| Lighting7 | 7 | 70 | 73 | 319 |

Table 8 b. Suite des ensembles de séries temporelles

| | | | | |
|--------------------------------|----|------|------|------|
| MALLAT | 8 | 55 | 2345 | 1024 |
| Meat | 3 | 60 | 60 | 448 |
| MedicalImages | 10 | 381 | 760 | 99 |
| MiddlePhalanxOutlineAgeGroup | 3 | 154 | 400 | 80 |
| MiddlePhalanxOutlineCorrect | 2 | 291 | 600 | 80 |
| MoteStrain | 2 | 20 | 1252 | 84 |
| NonInvasiveFatalECG_Thorax1 | 42 | 1800 | 1965 | 750 |
| NonInvasiveFatalECG_Thorax2 | 42 | 1800 | 1965 | 750 |
| OliveOil | 4 | 30 | 30 | 570 |
| OSULeaf | 6 | 200 | 242 | 427 |
| PhalangesOutlinesCorrect | 2 | 1800 | 858 | 80 |
| Phoneme | 39 | 214 | 1896 | 1024 |
| Plane | 7 | 105 | 105 | 144 |
| ProximalPhalanxOutlineAgeGroup | 3 | 400 | 205 | 80 |
| ProximalPhalanxOutlineCorrect | 2 | 600 | 291 | 80 |
| ProximalPhalanxTW | 6 | 205 | 400 | 80 |
| RefrigerationDevices | 3 | 375 | 375 | 720 |
| ScreenType | 3 | 375 | 375 | 720 |
| ShapeletSim | 2 | 20 | 180 | 500 |
| ShapesAll | 60 | 600 | 600 | 512 |
| SmallKitchenAppliances | 3 | 375 | 375 | 720 |
| SonyAIBORobotSurface | 2 | 20 | 601 | 70 |
| SonyAIBORobotSurfaceII | 2 | 27 | 953 | 65 |
| StarLightCurves | 3 | 1000 | 8236 | 1024 |
| Strawberry | 2 | 370 | 613 | 235 |
| SwedishLeaf | 15 | 500 | 625 | 128 |
| Symbols | 6 | 25 | 995 | 398 |
| synthetic_control | 6 | 300 | 300 | 60 |
| ToeSegmentation1 | 2 | 40 | 228 | 277 |
| ToeSegmentation2 | 2 | 36 | 130 | 343 |
| Trace | 4 | 100 | 100 | 275 |
| Two_Patterns | 4 | 1000 | 4000 | 128 |
| TwoLeadECG | 2 | 23 | 1139 | 82 |
| uWaveGestureLibrary_X | 8 | 896 | 3582 | 315 |
| uWaveGestureLibrary_Y | 8 | 896 | 3582 | 315 |
| uWaveGestureLibrary_Z | 8 | 896 | 3582 | 315 |
| UWaveGestureLibraryAll | 8 | 896 | 3582 | 945 |
| Wafer | 2 | 1000 | 6174 | 152 |
| Wine | 2 | 57 | 54 | 234 |
| WordsSynonyms | 25 | 267 | 638 | 270 |
| Worms | 5 | 77 | 181 | 900 |
| WormsTwoClass | 2 | 77 | 181 | 900 |
| Yoga | 2 | 300 | 3000 | 426 |

IX.4 Résultats et discussions

IX.4.1 Comparaison des techniques d'alignement DTW, QP-DTW et SEA

Nous avons comparé notre technique d'alignement QP-DTW avec les techniques d'alignement SEA et DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN.

La Table 9 illustre la moyenne des taux d'erreurs des méthodes DTW, SEA et QPDTW. Comme nous pouvons le voir à partir de la table9, la méthode QP-DTW a le taux d'erreur le plus bas suivie de la méthode DTW et de la méthode SEA. D'un point de vu précision, QP-DTW gagne dans 35 ensemble de données suivie de DTW qui gagne dans 34 ensemble de données et de SEA qui gagne dans 17 ensembles de données.

La Figure70 présente le diagramme de dispersion des précisions relatives aux méthodes DTW et QP-DTW. Chaque point représente un ensemble de données des archives UCR. Nous pouvons voir à travers cette figure que QP-DTW est aussi précise que DTW.

La Figure71 présente le diagramme de dispersion des précisions relatives aux méthodes SEA et QP-DTW. Nous pouvons voir à travers cette figure que QP-DTW est plus précise que SEA.

Table 9. Comparaison de QP-DTW avec DTW et SEA pour la classification 1NN

| | DTW | SEA | QP-DTW | | | | |
|------------------------------|--------------|--------------|--------------|--------------------------------|--------------|--------------|--------------|
| | | | | MALLAT | 0,066 | 0,212 | 0,016 |
| 50words | 0,31 | 0,729 | 0,413 | Meat | 0,067 | 0,066 | 0,2 |
| Adiac | 0,396 | 0,306 | 0,319 | MedicalImages | 0,263 | 0,43 | 0,348 |
| ArrowHead | 0,297 | 0,308 | 0,245 | MiddlePhalanxOutlineAgeGroup | 0,25 | 0,257 | 0,27 |
| Beef | 0,367 | 0,433 | 0,2 | MiddlePhalanxOutlineCorrect | 0,352 | 0,333 | 0,273 |
| BeetleFly | 0,3 | 0,3 | 0,1 | MiddlePhalanxTW | 0,416 | 0,426 | 0,443 |
| BirdChicken | 0,25 | 0,15 | 0,2 | MoteStrain | 0,165 | 0,218 | 0,183 |
| Car | 0,267 | 0,416 | 0,266 | NonInvasiveFatalECG_Thorax1 | 0,209 | 0,202 | 0,143 |
| CBF | 0,003 | 0,348 | 0,044 | NonInvasiveFatalECG_Thorax2 | 0,135 | 0,132 | 0,082 |
| ChlorineConcentration | 0,352 | 0,409 | 0,196 | OliveOil | 0,167 | 0,166 | 0,233 |
| CinC_ECG_torso | 0,349 | 0,365 | 0,125 | OSULeaf | 0,409 | 0,561 | 0,413 |
| Coffee | 0 | 0,214 | 0,071 | PhalangesOutlinesCorrect | 0,272 | 0,284 | 0,142 |
| Computers | 0,38 | 0,328 | 0,4 | Phoneme | 0,772 | 0,869 | 0,503 |
| Cricket_X | 0,246 | 0,482 | 0,273 | Plane | 0 | 0 | 0,009 |
| Cricket_Y | 0,256 | 0,525 | 0,302 | ProximalPhalanxOutlineAgeGroup | 0,195 | 0,234 | 0,214 |
| Cricket_Z | 0,246 | 0,451 | 0,369 | ProximalPhalanxOutlineCorrect | 0,216 | 0,233 | 0,223 |
| DiatomSizeReduction | 0,033 | 0,032 | 0,039 | ProximalPhalanxTW | 0,263 | 0,315 | 0,327 |
| DistalPhalanxOutlineAgeGroup | 0,208 | 0,247 | 0,247 | RefrigerationDevices | 0,536 | 0,506 | 0,518 |
| DistalPhalanxOutlineCorrect | 0,232 | 0,253 | 0,266 | ScreenType | 0,603 | 0,595 | 0,422 |
| DistalPhalanxTW | 0,29 | 0,275 | 0,32 | ShapeletSim | 0,35 | 0,561 | 0,377 |
| Earthquakes | 0,258 | 0,267 | 0,219 | ShapesAll | 0,232 | 0,318 | 0,117 |
| ECG200 | 0,23 | 0,15 | 0,16 | SmallKitchenAppliances | 0,357 | 0,411 | 0,134 |
| ECG5000 | 0,076 | 0,081 | 0,086 | SonyAIBORobotSurface | 0,275 | 0,286 | 0,169 |
| ECGFiveDays | 0,232 | 0,066 | 0,27 | SonyAIBORobotSurfaceII | 0,169 | 0,259 | 0,196 |
| ElectricDevices | 0,399 | 0,461 | 0,125 | StarLightCurves | 0,093 | 0,31 | 0,045 |
| FaceAll | 0,192 | 0,501 | 0,35 | Strawberry | 0,06 | 0,06 | 0,092 |
| FaceFour | 0,17 | 0,318 | 0,204 | SwedishLeaf | 0,208 | 0,276 | 0,192 |
| FacesUCR | 0,095 | 0,395 | 0,125 | Symbols | 0,05 | 0,393 | 0,129 |
| FISH | 0,177 | 0,285 | 0,228 | synthetic_control | 0,007 | 0,616 | 0,14 |
| FordA | 0,438 | 0,443 | 0,351 | ToeSegmentation1 | 0,228 | 0,21 | 0,32 |
| FordB | 0,406 | 0,446 | 0,327 | ToeSegmentation2 | 0,162 | 0,284 | 0,146 |
| Gun_Point | 0,093 | 0,08 | 0,18 | Trace | 0 | 0,08 | 0,16 |
| Ham | 0,533 | 0,485 | 0,543 | Two_Patterns | 0 | 0,59 | 0,007 |
| HandOutlines | 0,202 | 0,251 | 0,22 | TwoLeadECG | 0,096 | 0,077 | 0,239 |
| Haptics | 0,623 | 0,652 | 0,503 | uWaveGestureLibrary_X | 0,273 | 0,642 | 0,152 |
| Herring | 0,469 | 0,453 | 0,422 | uWaveGestureLibrary_Y | 0,366 | 0,697 | 0,241 |
| InlineSkate | 0,616 | 0,596 | 0,423 | uWaveGestureLibrary_Z | 0,342 | 0,638 | 0,224 |
| InsectWingbeatSound | 0,645 | 0,858 | 0,542 | UWaveGestureLibraryAll | 0,108 | 0,717 | 0,072 |
| ItalyPowerDemand | 0,05 | 0,21 | 0,072 | wafer | 0,02 | 0,019 | 0,048 |
| LargeKitchenAppliances | 0,205 | 0,282 | 0,195 | Wine | 0,426 | 0,444 | 0,5 |
| Lighting2 | 0,131 | 0,229 | 0,278 | WordsSynonyms | 0,252 | 0,753 | 0,326 |
| Lighting7 | 0,274 | 0,452 | 0,41 | Worms | 0,536 | 0,629 | 0,392 |
| | | | | WormsTwoClass | 0,337 | 0,464 | 0,215 |
| | | | | yoga | 0,164 | 0,214 | 0,16 |
| | | | | MOYENNE | 0,256 | 0,355 | 0,239 |
| | | | | WINS | 34 | 17 | 35 |

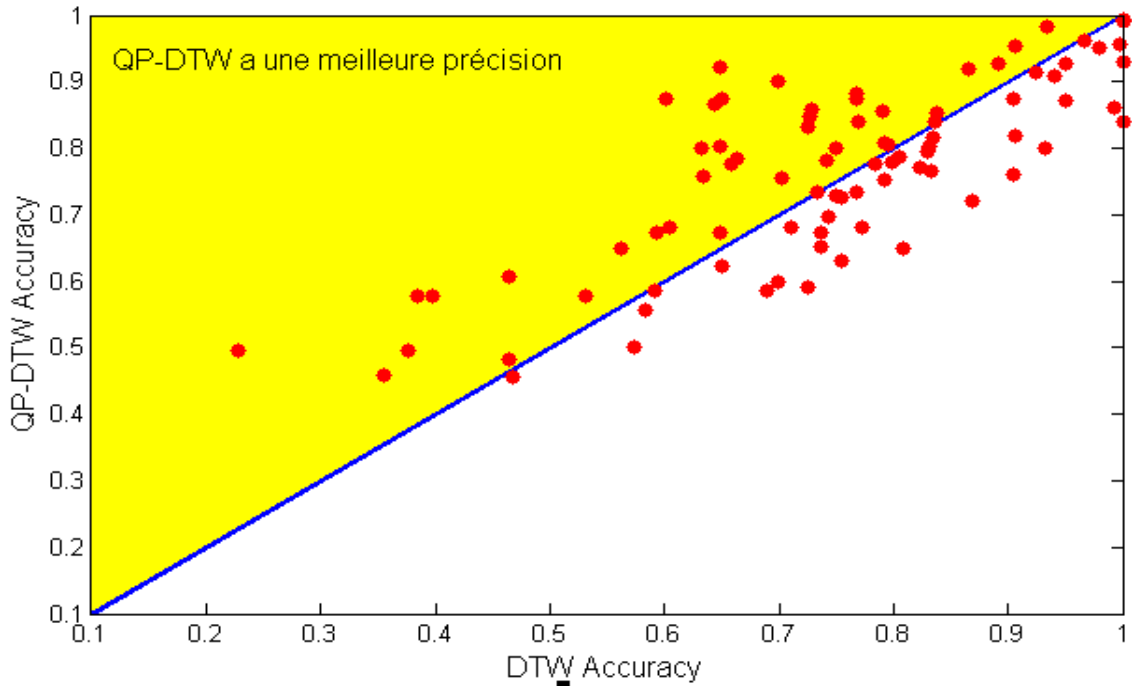


Fig.70. Comparaison de la précision de DTW et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN

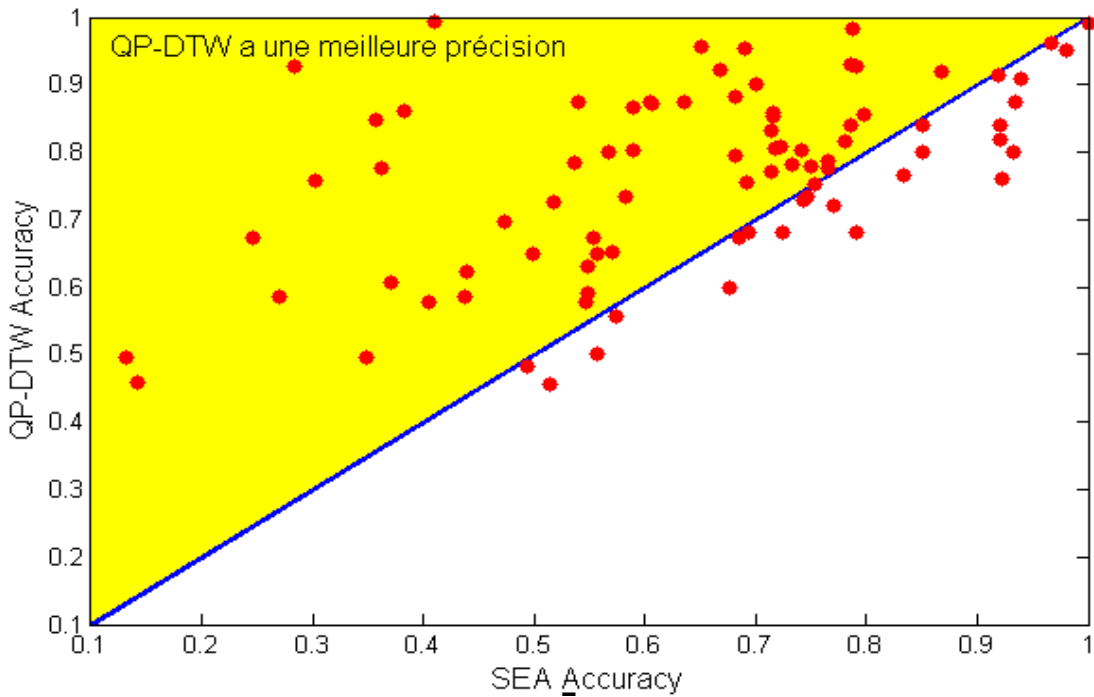


Fig.71. Comparaison de la précision de SEA et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN.

La Figure 65 illustre la comparaison des taux d'erreurs des méthodes DTW et QP-DTW. On peut voir à travers cette figure que la courbe du taux d'erreur de la méthode QP-DTW est plus basse que la courbe du taux d'erreur de la méthode DTW ce qui prouve que la méthode QP-DTW effectue une meilleure classification par rapport à la méthode DTW.

La Figure 66 illustre la comparaison des taux d'erreurs des méthodes SEA et QP-DTW. On peut voir à travers cette figure que la courbe du taux d'erreur de la méthode QP-DTW est plus basse que la courbe du taux d'erreur de la méthode SEA ce qui prouve que la méthode QP-DTW effectue une meilleure classification par rapport à la méthode SEA

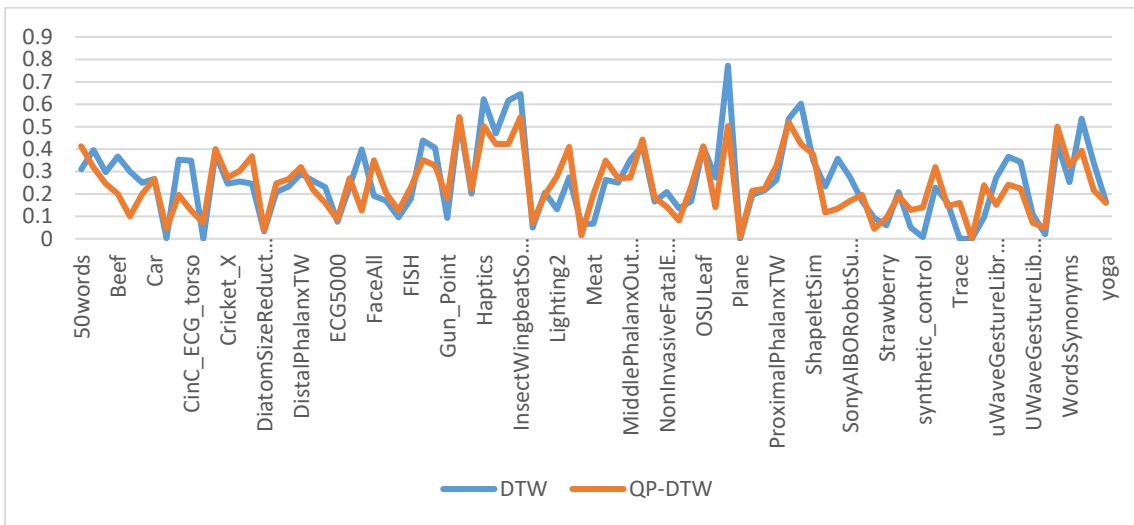


Figure 72. Comparaison du taux d'erreurs des méthodes DTW et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN.

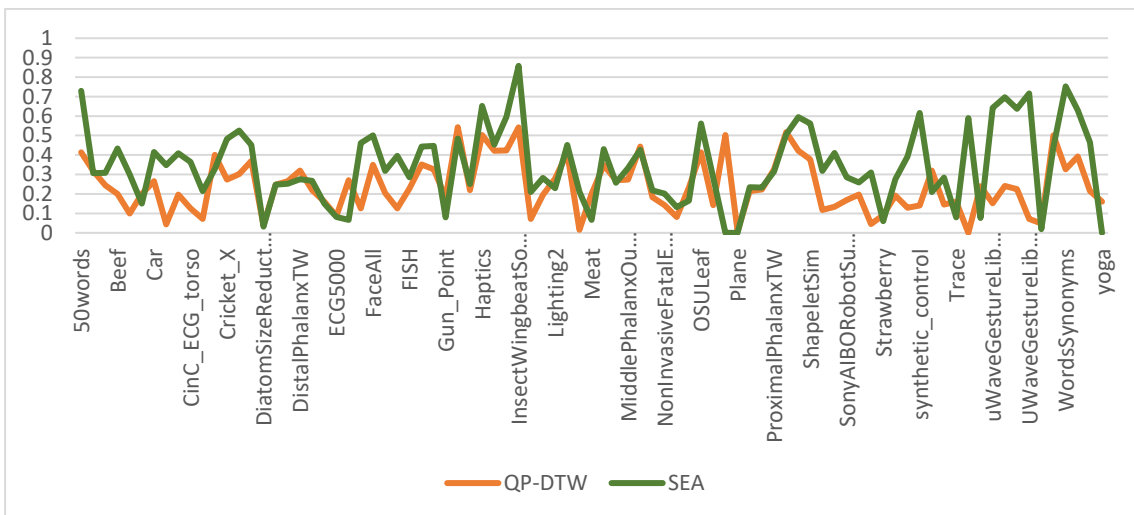


Figure 73. Comparaison du taux d'erreurs des méthodes SEA et QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN.

IX.4.2 Comparaison de 1NN-QP-DTW à d'autres classifieurs basés 1NN

Dans cette section, nous comparons le classifieur proposé 1NN-QPDTW avec 11 classifieurs de type 1NN utilisant différents types de distances/similarité pour la classification de 85 ensembles de données des archives UCR. La Table 8 illustre la moyenne des taux d'erreur des distances/mesures Euclidienne (ED), Manhattan (Manh), Chebyshev (Cheb), Squared Euclidian (SED), Divergence (DivD), $Average(L_1, L_\infty)$ (Avg), Hassanat (Hass), Lorentzian (Lortz), Canberra (Canbr), Squared-Chi-Squared(SCSD), la distance Invariable en fonction de la Complexité (CID) et de la méthode Quasi Periodic Dynamic Time Warping (QP-DTW). Les mesures utilisées sont présentées en détail dans le chapitre III de cette thèse.

Comme nous pouvons le voir dans Table 8, la technique QP-DTW donne de bons résultats pour la classification des séries temporelles, son taux d'erreur (0,239) est le plus bas par rapport aux 11 classifieurs basés mesures de distances/similarité.

La Figure 67 illustre la comparaison des 11 classifieurs basés mesures de distance/similarité avec le classifieur 1NN-QP-DTW. On peut voir que la courbe du taux d'erreur de la méthode QP-DTW est la plus basse suivie des courbes de CID, Lortz, Manh, $Average(L_1, L_\infty)$, Hass, ED, SED, Cambr, SCSD, DivD et Cheb.

Notons que les séries temporelles du même ensemble de données ne sont pas décalées les unes par rapport aux autres et ont la même longueur, ce qui nous a permis de comparer notre technique d'alignement « QP-DTW » avec les techniques de comparaison basées mesures (ED, Canberra, etc). Pour rappel, les techniques de comparaison basées mesures de distance/similarité comparent uniquement les séries temporelles de même longueur et ne prennent pas en compte les problèmes de décalages temporelles entre elles (Figure 19, chapitre III).

D'un autre côté, nous comparons la distance Euclidienne avec les autres mesures de distance puisque la distance Euclidienne est une mesure largement utilisée pour la classification des séries temporelles avec le classifieur 1-NN en raison de sa simplicité, et de ses bons résultats empiriques [Ding, 2008]. Étonnamment, les mesures de similarité qui ne sont pas souvent utilisées pour la classification des séries temporelles telles que Manhattan, $Average(L_1, L_\infty)$, Lorentzian, CID et Hassanat ont une meilleure précision par rapport à la distance Euclidienne comme le démontrent les résultats de la Table 8 et la Figure 67.

Table10. a. Comparaison des classifieur 1NN basés mesures de comparaison avec 1NN-QP-DTW sur la base du taux d'erreur de classification par classe et globalement.

| | ED | Manh | Cheb | SED | DivD | Avg | Hass | Lortz | Canbr | SCSD | CID | QP-DTW |
|------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|--------|
| 50words | 0,369 | 0,332 | 0,497 | 0,369 | 0,512 | 0,332 | 0,301 | 0,308 | 0,385 | 0,431 | 0,336 | 0,413 |
| Adiac | 0,389 | 0,402 | 0,425 | 0,389 | 0,34 | 0,402 | 0,435 | 0,404 | 0,366 | 0,35 | 0,373 | 0,319 |
| ArrowHead | 0,2 | 0,206 | 0,326 | 0,2 | 0,211 | 0,211 | 0,211 | 0,2 | 0,211 | 0,2 | 0,171 | 0,245 |
| Beef | 0,333 | 0,367 | 0,3 | 0,333 | 0,433 | 0,4 | 0,4 | 0,4 | 0,367 | 0,4 | 0,367 | 0,2 |
| BeetleFly | 0,25 | 0,35 | 0,3 | 0,25 | 0,15 | 0,35 | 0,35 | 0,35 | 0,25 | 0,15 | 0,3 | 0,1 |
| BirdChicken | 0,45 | 0,4 | 0,35 | 0,45 | 0,35 | 0,4 | 0,4 | 0,35 | 0,4 | 0,35 | 0,35 | 0,2 |
| Car | 0,267 | 0,283 | 0,283 | 0,267 | 0,283 | 0,283 | 0,283 | 0,283 | 0,217 | 0,25 | 0,267 | 0,266 |
| CBF | 0,148 | 0,111 | 0,133 | 0,148 | 0,027 | 0,114 | 0,133 | 0,108 | 0,108 | 0,15 | 0,016 | 0,044 |
| ChlorineConcentration | 0,35 | 0,373 | 0,337 | 0,35 | 0,41 | 0,369 | 0,377 | 0,375 | 0,379 | 0,37 | 0,351 | 0,196 |
| CinC_ECG_torso | 0,103 | 0,056 | 0,177 | 0,103 | 0,123 | 0,058 | 0,094 | 0,067 | 0,099 | 0,117 | 0,084 | 0,125 |
| Coffee | 0 | 0,036 | 0,036 | 0 | 0 | 0,036 | 0,107 | 0,071 | 0,036 | 0 | 0 | 0,071 |
| Computers | 0,424 | 0,412 | 0,528 | 0,424 | 0,404 | 0,416 | 0,448 | 0,424 | 0,432 | 0,46 | 0,404 | 0,4 |
| Cricket_X | 0,423 | 0,369 | 0,733 | 0,423 | 0,51 | 0,369 | 0,354 | 0,367 | 0,405 | 0,487 | 0,372 | 0,273 |
| Cricket_Y | 0,433 | 0,367 | 0,708 | 0,433 | 0,51 | 0,374 | 0,354 | 0,344 | 0,405 | 0,467 | 0,421 | 0,302 |
| Cricket_Z | 0,413 | 0,341 | 0,756 | 0,413 | 0,51 | 0,338 | 0,354 | 0,356 | 0,397 | 0,5 | 0,405 | 0,369 |
| DiatomSizeReduction | 0,065 | 0,072 | 0,042 | 0,065 | 0,046 | 0,072 | 0,069 | 0,075 | 0,059 | 0,046 | 0,065 | 0,039 |
| DistalPhalanxOutlineAgeGroup | 0,218 | 0,22 | 0,233 | 0,218 | 0,225 | 0,21 | 0,23 | 0,22 | 0,225 | 0,223 | 0,223 | 0,247 |
| DistalPhalanxOutlineCorrect | 0,248 | 0,238 | 0,272 | 0,248 | 0,252 | 0,243 | 0,245 | 0,248 | 0,255 | 0,253 | 0,255 | 0,266 |
| DistalPhalanxTW | 0,273 | 0,283 | 0,26 | 0,273 | 0,278 | 0,283 | 0,273 | 0,283 | 0,288 | 0,275 | 0,265 | 0,32 |
| Earthquakes | 0,326 | 0,286 | 0,193 | 0,326 | 0,351 | 0,286 | 0,298 | 0,311 | 0,317 | 0,27 | 0,286 | 0,219 |
| ECG200 | 0,12 | 0,11 | 0,18 | 0,12 | 0,09 | 0,11 | 0,11 | 0,11 | 0,09 | 0,1 | 0,11 | 0,16 |
| ECG5000 | 0,075 | 0,074 | 0,09 | 0,075 | 0,083 | 0,075 | 0,072 | 0,073 | 0,075 | 0,08 | 0,072 | 0,086 |
| ECGFiveDays | 0,203 | 0,215 | 0,319 | 0,203 | 0,319 | 0,213 | 0,226 | 0,206 | 0,217 | 0,24 | 0,218 | 0,27 |
| ElectricDevices | 0,45 | 0,429 | 0,726 | 0,449 | 0,452 | 0,431 | 0,427 | 0,429 | 0,433 | 0,463 | 0,44 | 0,125 |
| FaceAll | 0,286 | 0,279 | 0,485 | 0,286 | 0,263 | 0,282 | 0,273 | 0,27 | 0,282 | 0,227 | 0,269 | 0,35 |
| FaceFour | 0,216 | 0,159 | 0,489 | 0,216 | 0,25 | 0,159 | 0,125 | 0,148 | 0,216 | 0,25 | 0,193 | 0,204 |
| FacesUCR | 0,231 | 0,199 | 0,459 | 0,231 | 0,287 | 0,195 | 0,189 | 0,2 | 0,223 | 0,234 | 0,235 | 0,125 |
| FISH | 0,217 | 0,206 | 0,263 | 0,217 | 0,217 | 0,2 | 0,211 | 0,229 | 0,211 | 0,211 | 0,217 | 0,228 |
| FordA | 0,341 | 0,34 | 0,382 | 0,341 | 0,364 | 0,341 | 0,347 | 0,351 | 0,337 | 0,338 | 0,241 | 0,351 |
| FordB | 0,442 | 0,421 | 0,476 | 0,442 | 0,424 | 0,424 | 0,433 | 0,428 | 0,435 | 0,427 | 0,326 | 0,327 |
| Gun_Point | 0,087 | 0,047 | 0,207 | 0,087 | 0,153 | 0,04 | 0,047 | 0,047 | 0,08 | 0,1 | 0,073 | 0,18 |
| Ham | 0,4 | 0,486 | 0,467 | 0,4 | 0,448 | 0,495 | 0,486 | 0,495 | 0,486 | 0,486 | 0,486 | 0,549 |
| HandOutlines | 0,199 | 0,215 | 0,193 | 0,199 | 0,22 | 0,215 | 0,224 | 0,217 | 0,218 | 0,206 | 0,215 | 0,22 |
| Haptics | 0,63 | 0,64 | 0,62 | 0,63 | 0,636 | 0,63 | 0,627 | 0,643 | 0,63 | 0,61 | 0,584 | 0,503 |
| Herring | 0,484 | 0,406 | 0,406 | 0,484 | 0,391 | 0,406 | 0,5 | 0,406 | 0,375 | 0,391 | 0,453 | 0,421 |
| InlineSkate | 0,658 | 0,647 | 0,713 | 0,658 | 0,72 | 0,647 | 0,649 | 0,64 | 0,693 | 0,702 | 0,629 | 0,423 |
| InsectWingbeatSound | 0,438 | 0,431 | 0,501 | 0,438 | 0,501 | 0,433 | 0,422 | 0,432 | 0,446 | 0,455 | 0,437 | 0,542 |
| ItalyPowerDemand | 0,045 | 0,038 | 0,067 | 0,045 | 0,034 | 0,04 | 0,041 | 0,038 | 0,036 | 0,033 | 0,044 | 0,071 |
| LargeKitchenAppliances | 0,507 | 0,443 | 0,675 | 0,507 | 0,496 | 0,437 | 0,435 | 0,419 | 0,475 | 0,523 | 0,477 | 0,195 |
| Lighting2 | 0,246 | 0,18 | 0,393 | 0,246 | 0,41 | 0,197 | 0,164 | 0,18 | 0,262 | 0,344 | 0,246 | 0,278 |
| Lighting7 | 0,425 | 0,288 | 0,63 | 0,425 | 0,575 | 0,329 | 0,329 | 0,26 | 0,452 | 0,507 | 0,397 | 0,41 |
| MALLAT | 0,086 | 0,075 | 0,248 | 0,086 | 0,136 | 0,075 | 0,09 | 0,079 | 0,07 | 0,083 | 0,075 | 0,016 |
| Meat | 0,067 | 0,067 | 0,217 | 0,067 | 0,1 | 0,067 | 0,067 | 0,067 | 0,1 | 0,083 | 0,067 | 0,2 |
| MedicalImages | 0,316 | 0,293 | 0,396 | 0,316 | 0,37 | 0,291 | 0,287 | 0,295 | 0,309 | 0,332 | 0,309 | 0,348 |

Table10 b. Comparaison des classifieurs INN-basés mesures de comparaison avec 1-NN-QP-DTW sur la base du taux d'erreur de classification par classe et globalement.

| | | | | | | | | | | | | |
|--------------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| MiddlePhalanxOutlineAgeGroup | 0,26 | 0,268 | 0,258 | 0,26 | 0,255 | 0,263 | 0,268 | 0,268 | 0,258 | 0,268 | 0,265 | 0,27 |
| MiddlePhalanxOutlineCorrect | 0,247 | 0,26 | 0,285 | 0,247 | 0,225 | 0,255 | 0,267 | 0,255 | 0,267 | 0,23 | 0,253 | 0,273 |
| MiddlePhalanxTW | 0,439 | 0,424 | 0,451 | 0,439 | 0,471 | 0,426 | 0,414 | 0,421 | 0,441 | 0,459 | 0,426 | 0,443 |
| MoteStrain | 0,121 | 0,134 | 0,295 | 0,121 | 0,177 | 0,131 | 0,135 | 0,136 | 0,117 | 0,141 | 0,212 | 0,183 |
| NonInvasiveFatalECG_Thorax1 | 0,171 | 0,191 | 0,275 | 0,171 | 0,227 | 0,189 | 0,198 | 0,191 | 0,198 | 0,209 | 0,163 | 0,143 |
| NonInvasiveFatalECG_Thorax2 | 0,12 | 0,125 | 0,204 | 0,12 | 0,156 | 0,124 | 0,136 | 0,127 | 0,122 | 0,126 | 0,121 | 0,082 |
| OliveOil | 0,133 | 0,167 | 0,2 | 0,133 | 0,2 | 0,167 | 0,133 | 0,167 | 0,167 | 0,167 | 0,133 | 0,233 |
| OSULeaf | 0,479 | 0,455 | 0,455 | 0,479 | 0,512 | 0,455 | 0,43 | 0,446 | 0,475 | 0,5 | 0,438 | 0,413 |
| PhalangesOutlinesCorrect | 0,239 | 0,248 | 0,253 | 0,239 | 0,224 | 0,244 | 0,262 | 0,248 | 0,249 | 0,219 | 0,238 | 0,142 |
| Phoneme | 0,891 | 0,894 | 0,922 | 0,891 | 0,902 | 0,892 | 0,905 | 0,909 | 0,892 | 0,905 | 0,883 | 0,503 |
| Plane | 0,038 | 0,038 | 0,029 | 0,038 | 0,029 | 0,038 | 0,038 | 0,038 | 0,029 | 0,038 | 0,038 | 0,009 |
| ProximalPhalanxOutlineAgeGroup | 0,215 | 0,215 | 0,224 | 0,215 | 0,195 | 0,21 | 0,239 | 0,22 | 0,22 | 0,21 | 0,21 | 0,214 |
| ProximalPhalanxOutlineCorrect | 0,192 | 0,216 | 0,175 | 0,192 | 0,162 | 0,21 | 0,227 | 0,216 | 0,199 | 0,151 | 0,192 | 0,223 |
| ProximalPhalanxTW | 0,292 | 0,295 | 0,27 | 0,293 | 0,273 | 0,29 | 0,293 | 0,295 | 0,265 | 0,283 | 0,288 | 0,327 |
| RefrigerationDevices | 0,605 | 0,541 | 0,683 | 0,605 | 0,589 | 0,539 | 0,541 | 0,523 | 0,568 | 0,619 | 0,589 | 0,518 |
| ScreenType | 0,64 | 0,64 | 0,629 | 0,64 | 0,635 | 0,64 | 0,656 | 0,624 | 0,637 | 0,667 | 0,587 | 0,422 |
| ShapeletSim | 0,461 | 0,522 | 0,45 | 0,461 | 0,5 | 0,522 | 0,489 | 0,489 | 0,461 | 0,511 | 0,239 | 0,377 |
| ShapesAll | 0,248 | 0,24 | 0,287 | 0,248 | 0,26 | 0,24 | 0,243 | 0,242 | 0,238 | 0,255 | 0,227 | 0,117 |
| SmallKitchenAppliances | 0,659 | 0,603 | 0,664 | 0,656 | 0,635 | 0,603 | 0,499 | 0,512 | 0,64 | 0,664 | 0,608 | 0,134 |
| SonyAIBORobotSurface | 0,305 | 0,313 | 0,288 | 0,304 | 0,25 | 0,308 | 0,316 | 0,313 | 0,291 | 0,286 | 0,185 | 0,169 |
| SonyAIBORobotSurfaceII | 0,141 | 0,132 | 0,194 | 0,141 | 0,165 | 0,128 | 0,127 | 0,124 | 0,155 | 0,164 | 0,123 | 0,196 |
| StarLightCurves | 0,151 | 0,147 | 0,145 | 0,151 | 0,136 | 0,147 | 0,154 | 0,147 | 0,145 | 0,144 | 0,057 | 0,045 |
| Strawberry | 0,062 | 0,057 | 0,072 | 0,062 | 0,06 | 0,059 | 0,06 | 0,055 | 0,059 | 0,062 | 0,06 | 0,092 |
| SwedishLeaf | 0,211 | 0,211 | 0,285 | 0,211 | 0,214 | 0,213 | 0,227 | 0,216 | 0,206 | 0,206 | 0,123 | 0,192 |
| Symbols | 0,1 | 0,096 | 0,156 | 0,101 | 0,129 | 0,096 | 0,105 | 0,097 | 0,103 | 0,118 | 0,084 | 0,129 |
| synthetic_control | 0,12 | 0,12 | 0,19 | 0,12 | 0,15 | 0,12 | 0,137 | 0,133 | 0,123 | 0,12 | 0,05 | 0,14 |
| ToeSegmentation1 | 0,32 | 0,281 | 0,368 | 0,32 | 0,382 | 0,285 | 0,281 | 0,268 | 0,329 | 0,351 | 0,307 | 0,32 |
| ToeSegmentation2 | 0,192 | 0,146 | 0,531 | 0,192 | 0,292 | 0,146 | 0,115 | 0,123 | 0,177 | 0,208 | 0,185 | 0,146 |
| Trace | 0,24 | 0,24 | 0,27 | 0,24 | 0,23 | 0,25 | 0,25 | 0,21 | 0,25 | 0,24 | 0,14 | 0,16 |
| Two_Patterns | 0,09 | 0,039 | 0,504 | 0,093 | 0,239 | 0,04 | 0,034 | 0,034 | 0,064 | 0,103 | 0,121 | 0,007 |
| TwoLeadECG | 0,253 | 0,265 | 0,221 | 0,253 | 0,25 | 0,265 | 0,254 | 0,263 | 0,261 | 0,25 | 0,232 | 0,239 |
| uWaveGestureLibrary_X | 0,261 | 0,259 | 0,303 | 0,261 | 0,319 | 0,259 | 0,259 | 0,259 | 0,264 | 0,27 | 0,238 | 0,152 |
| uWaveGestureLibrary_Y | 0,338 | 0,334 | 0,405 | 0,338 | 0,387 | 0,334 | 0,331 | 0,332 | 0,338 | 0,355 | 0,29 | 0,241 |
| uWaveGestureLibrary_Z | 0,35 | 0,35 | 0,4 | 0,35 | 0,401 | 0,35 | 0,351 | 0,351 | 0,365 | 0,377 | 0,345 | 0,224 |
| UWaveGestureLibraryAll | 0,052 | 0,046 | 0,159 | 0,052 | 0,097 | 0,046 | 0,047 | 0,045 | 0,051 | 0,061 | 0,06 | 0,072 |
| Wafer | 0,005 | 0,005 | 0,02 | 0,005 | 0,017 | 0,005 | 0,005 | 0,005 | 0,005 | 0,008 | 0,006 | 0,048 |
| Ne | 0,389 | 0,352 | 0,574 | 0,389 | 0,556 | 0,352 | 0,315 | 0,352 | 0,519 | 0,519 | 0,389 | 0,5 |
| WordsSynonyms | 0,382 | 0,365 | 0,525 | 0,382 | 0,553 | 0,364 | 0,343 | 0,342 | 0,408 | 0,461 | 0,357 | 0,326 |
| Worms | 0,635 | 0,613 | 0,702 | 0,635 | 0,685 | 0,597 | 0,663 | 0,652 | 0,63 | 0,652 | 0,547 | 0,392 |
| WormsTwoClass | 0,414 | 0,398 | 0,409 | 0,414 | 0,425 | 0,392 | 0,431 | 0,425 | 0,392 | 0,414 | 0,348 | 0,215 |
| Yoga | 0,17 | 0,171 | 0,186 | 0,17 | 0,192 | 0,171 | 0,176 | 0,173 | 0,187 | 0,19 | 0,164 | 0,16 |
| Moyenne | 0,288 | 0,278 | 0,358 | 0,288 | 0,313 | 0,278 | 0,281 | 0,276 | 0,288 | 0,298 | 0,266 | 0,239 |

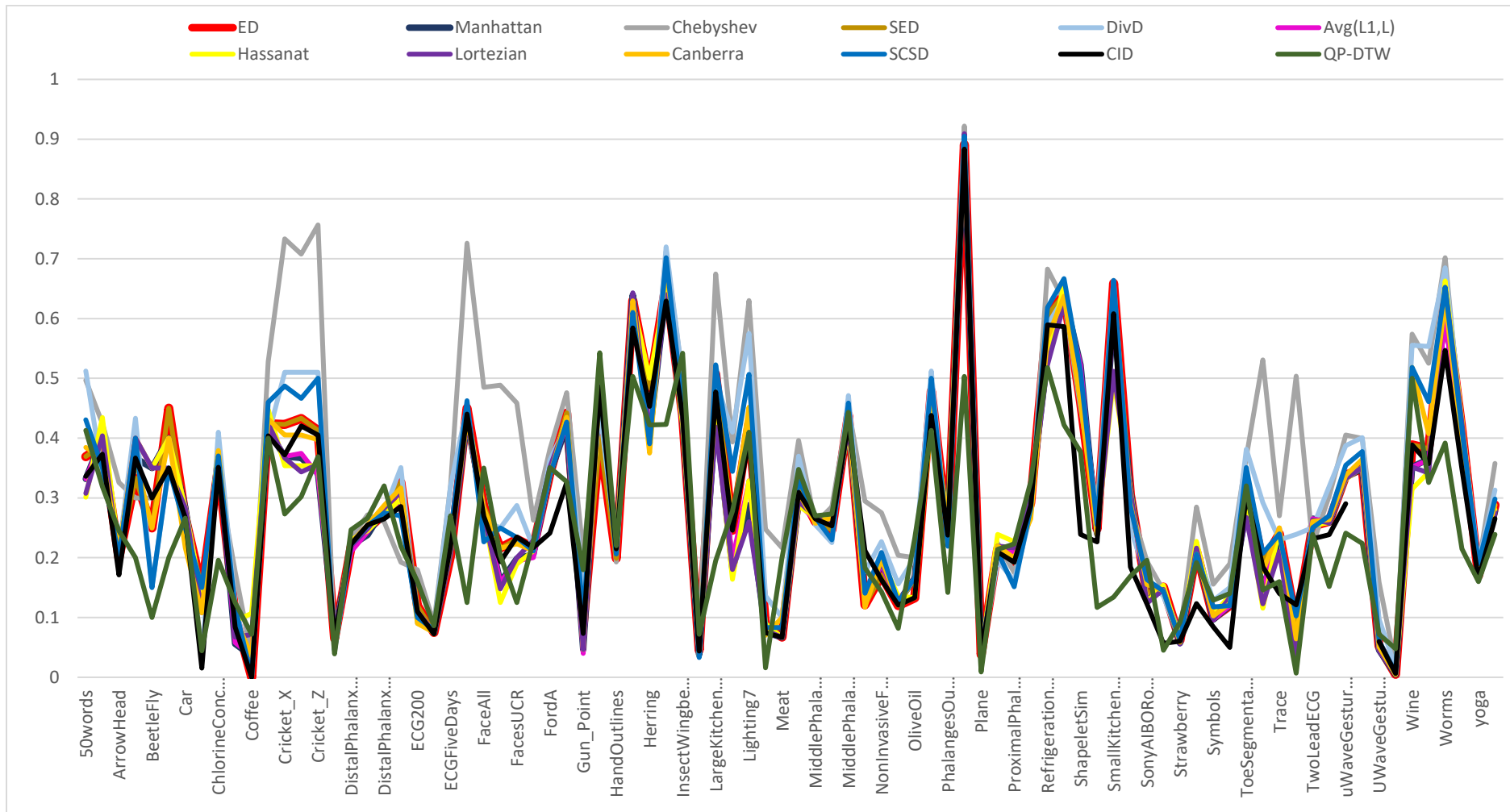


Fig.74. Comparaison des mesures de distances/similarité avec la QP-DTW pour la classification de 85 ensembles de données des archives UCR en utilisant le classifieur 1-NN.

IX.5 Conclusion

Dans ce chapitre, nous avons proposé notre méthode de classification des séries temporelles 1NN-QPDTW et nous l'avons comparé avec les méthodes SEA et DTW et avec plusieurs classifieurs basés mesures de distance/similarité.

Les résultats basés sur les expériences effectuées sur le benchmark universel UCR comprenant des séries temporelles réelles provenant de différents domaines d'application montrent que la méthode proposée « 1NN-QPDTW » est une méthode prometteuse pour la classification des séries temporelle.

Conclusion Générale

Nous nous sommes principalement intéressés dans ce travail de thèse au duo quasiment inséparable en matière de fouille de données ST et autres domaines : La Caractérisation des séries temporelles et les mesures de Distances/Similarités et Alignement qui vont avec pour différentes tâches de fouille. Le point de la caractérisation permet en même temps de réduire la dimensionnalité et d'améliorer l'efficacité et la performance des différentes tâches associées aux séries temporelles. Nous nous sommes intéressés spécialement dans cette thèse aux tâches de recherche et de classification des séries temporelles. Nous nous sommes intéressés aussi dans le cadre de cette thèse à l'épineux et complexe problème de l'appariement des séries temporelles. La démarche adoptée a consisté à chaque point étudié, puis, dans une deuxième phase, nous avons présenté nos contributions personnelles.

Dans la partie théorique, nous sommes forcés pour déterminer, critiquer et catégoriser les méthodes et approches les plus significatives dans chaque point abordé. Nous sommes ainsi sorties avec une vision très claire de la situation actuelle en matière de comparaison/alignement des ST, classification des ST et recherche par similarité dans les ST.

Dans la deuxième partie de cette thèse, nous avons pu contribuer avec des méthodes dans tous les axes et objectifs tracés pour cette thèse : Une méthode d'alignement/comparaison des ST (QPDTW), une technique de détection d'anomalies dans les ST de l'ECG basée QPDTW, une méthode de recherche par similarités dans les ST ECG basée aussi QPDTW et enfin, une méthode classification des ST générales de type 1NN basée elle aussi QPDTW. Les différentes études et contributions ont fait l'objet de plusieurs communications et de deux publications internationales.

Comme perspectives de travaux futurs, la considération d'autres signaux, notamment le Capnogramme est en vue. Nous comptons aussi rendre nos méthodes plus pratiques sous formes d'applications autonomes et, pourquoi pas, sous formes d'applications destinées aux smartphones.

Bibliographie

- [Adamek, 2003] T. Adamek, and N.E. O'Connor, "Efficient contour-based shape representation and matching", In *Multimedia Information Retrieval*, pp. 138-143, 2003.
- [Adamek, 2004] T. Adamek, and N.E. O'Connor, "A multiscale representation method for non-rigid shapes with a single closed contour", *IEEE Circuits and Systems for Video Technology*, vol.14, n°5, pp. 742-753, 2004.
- [Aßfalg, 2007] J. Aßfalg, H. Kriegel, P. Kroger, P. Kunath, A. Pryakhin, M. Renz, "Interval-focused Similarity Search in Time Series", In *Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA '07)*, Bangkok, Thailand, 2007.
- [Aghabozorgi, 2014] S. Aghabozorgi and T.Y. Wah, "Effective Clustering of Time-Series Data Using FCM", In *International Journal of Machine Learning and Computing*, vol.4, n°2, pp. 170-176, 2014.
- [Agrawal, 1993-a] R. Agrawal, C. Faloutsos, A. Swami. "Efficient Similarity Search in Sequence Databases", In *Proceedings of the International Conference on Foundations of Data Organization and Algorithms*, 1993.
- [Agrawal, 1993-b] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases", In *SIGMOD*, vol. 22, n°2, pp. 207-216, 1993.
- [Agrawal, 1995-a] R. Agrawal, K. Lin, H.S. Sawhne and K. Shim, "Fast similarity search in the presence of noise, scaling, and translation in time-series databases", In *Proceedings of the 21th International Conference on Very Large databases (VLDB'95)*, pp 490–501, 1995.
- [Agrawal, 1995-b] R. Agrawal, R. Srikant, "Mining sequential patterns", In *ICDE'95*, pp. 3–14, 1995.
- [Albert, 2002] R. Albert, A. L. Barabasi, "Statistical mechanics of complex networks *Reviews of Modern Physics*", vol. 74, n°47, 2002.
- [Alkasassbeh, 2015] M. Alkasassbeh, G. Altarawneh, A.B Hassanat, "On enhancing the performance of nearest neighbour classifiers using Hassanat distance metric", In *Canadian Journal of Pure and Applied Sciences*, vol.9, n°1, pp. 3291-3298, 2015.
- [Amr, 2012] T. Amr, "Survey on time-series data classification," in *TSDM journal*, 2012.
- [Annam, 2011] J.R. Annam, S.S. Mittapalli, R.S. Bapi, "Time series clustering and analysis of ECG heartbeats using dynamic time warping". In *Proceedings of Annual IEEE India Conference (INDICON)*, 2011.

- [Antonucci, 2015] A. Antonucci, M. Scanagatta, D.D. Mauà, C.P. Campos, “Early Classification of Time Series by Hidden Markov Models with Set-Valued Parameters”, In *Proceedings of the NIPS time series Workshop*, 2015.
- [Apostolico, 2002] A. Apostolico, M. E. Bock, and S. Lonardi, “Monotony of Surprise and Large-Scale Quest for Unusual Words”, In “*Proceedings of the 6th International Conference on Research in Computational Molecular Biology*”, pp. 22-31, Washington, 2002.
- [Argyros, 2003] T. Argyros, C. Ermopoulos “Efficient subsequence matching in time series databases under time and amplitude transformations”, Third IEEE International Conference on Data Mining, 2003.
- [Asuero, 2006] A. G. Asuero, A. Sayago, and A. G. Gonzalez, “The Correlation Coefficient: An Overview”, *Critical Reviews in Analytical Chemistry*, vol. 36, pp. 41–59, Department of Analytical Chemistry, Faculty of Pharmacy, The University of Seville, Seville, Spain, 2006.
- [Attalla, 2005] E. Attalla, and P. Siy, “Robust shape similarity retrieval based on contour segmentation polygonal multiresolution and elastic matching”, *Pattern Recognition*, vol. 38, n°12, pp. 2229-2241, 2005.
- [Backer, 1981] E. Backer and A. K. Jain, “A Clustering Performance Measure Based on Fuzzy Set Decomposition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-3, n°1, pp. 66-75, January 1981.
- [Bagnall, 2014] A. Bagnall, and J. Lines, “An experimental evaluation of nearest neighbour time series classification”, Technical report# CMP-C14-01, Department of Computing Sciences, University of East Anglia, 2014.
- [Bagnall, 2015] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, “Time-series classification with cote: the collective of transformation-based ensembles”. *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, No.9, pp. 2522–2535, 2015.
- [Bahri, 2004] A. Bahri, Y. Naïja, G. Jomier et M. Manouvrier, “Recherche par similarité de séquences temporelles dans les bases de données : un état de l'art”, *Manifestation de Jeunes Chercheurs STIC (MAJESTIC)*, Lamsade-Université Paris Dauphine, 2004.
- [Bajramovic, 2006] F. Bajramovic, F. Mattern, N. Butko, and J. Denzler, “A comparison of nearest neighbor search algorithms for generic object recognition”, In *Advanced Concepts for Intelligent Vision Systems*, Springer, pp. 1186-1197, 2006.
- [Bakshi, 1994] B. R. Bakshi, G. Stephanopoulos, “Representation of process trends - IV. Introduction of real-time patterns from operating data for diagnosis and supervisory control”, *Computers & Chemical Engineering*, vol.18, n°4, pp. 303-332, 1994.

- [Bartos, 2012] T. Bartos, T. Skopal, “Revisiting Techniques for Lowerbounding the Dynamic Time Warping Distance”, LNCS 7404, pp. 192–208, 2012.
- [Batista, 2011] G. E. Batista, X. Wang, and E. J. Keogh, “A complexity-invariant distance measure for time series. In *SDM*, vol. 11, pp. 699–710, 2011.
- [Batista, 2014] G.E.A.P.A. Batista, E. Keogh, O.M. Tataw, et al, “CID: an efficient complexity-invariant distance for timeseries”, *Data Mining and Knowledge Discovery*, vol. 28, Issue 3, pp. 634–669, 2014.
- [Bettaiah, 2014] V. Bettaiah, H. S. Ranganath, “Hierarchical Piecewise Linear Approximation, a Novel Representation of Time Series Data” In *proceedings of the Sixth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA)*, 2014
- [Baydogan, 2013] M. G. Baydogan, G. Runger, and E. Tuv, “A bag-of-features framework to classify time series”, *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, n°11, pp. 2796–2802, 2013.
- [Bayer, 1972] R. Bayer, E. Mc Creight, “Organization and Maintenance of Large Ordered Indexes”, *Acta Informatica*, vol.1, n°3, pp. 173-189, 1972.
- [Bayes, 1999] A. Bayes de Luna, “Clinical Electrocardiography: A Textbook”, New York: Futura Publishing, 1999.
- [Benali, 2013] R. Benali, “Analyse du signal ECG par réseau adaptif d’ondelettes en vue de la reconnaissance de pathologies cardiaques”, Thèse de doctorat, Université de Tlemcen, 2013.
- [Berndt, 1994] D. J. Berndt and J. Clifford, “Using dynamic time warping to find patterns in time series”, In *Proceedings of AAAI Workshop on Knowledge Discovery in Databases (KDD’94)*, pp. 359-370, Seattle, Washington, 1994.
- [Betancourt, 2013] J. P. Betancourt, C. Fatichah, M. L. Tangel, F. Yan, J.A. G. Sanchez, F. Dong, K. Hirota, “Similarity-Based Fuzzy Classification of ECG and Capnogram Signals”, *JACIII*, vol. 17, n°2, pp. 302-310, 2013.
- [Bhanu, 2004] B. Bhanu, and X. Zhou, “Face recognition from face profile using dynamic time warping”, In *Proceedings of International Conference on Pattern Recognition (ICPN’04)*, pp. 499-502, 2004.
- [Birney, 2001] E. Birney, “Hidden markov models in biological sequence analysis”. *IBM Journal of Research and Development*, vol. 45, n°3.4, pp. 449-454, 2001.
- [Boucheham, 2006] B. Boucheham, Y. Ferdi, M. Batouche, “Recursive versus sequential multiple errors reduction: A curve simplification approach to ECG data compression”, *Comput. Methods Programs Biomed*, vol. 81, pp. 162–173, 2006.

- [Boucheham, 2007] B. Boucheham, “ShaLTeRR: A contribution to short and long-term redundancy reduction in digital signals”, *Signal Process*, vol.87 n°10, pp. 2336–2347, 2007.
- [Boucheham, 2008] B. Boucheham, “Matching of quasi-periodic time series patterns by exchange of block-sorting signature”, *Pattern Recognition Letters*, Elsevier, vol. 29, pp. 501-514, 2008.
- [Boucheham, 2010] B. Boucheham, “Reduced data similarity-based matching for time series patterns alignment”, *Patterns Recognition Letter*, vol. 31, n°7, pp. 629-638, 2010.
- [Boucheham, 2011] B. Boucheham, “Abnormality Detection in Electrocardiograms by Time Series Alignment”, *Communications in Information Science and Management Engineering (CISME)*, Vol.1 No.3, pp. 7-11, 2011.
- [Boucheham, 2013] B. Boucheham, “Efficient matching for very complex time series”, *International Journal of Machine Learning and Cybernetics*, Springer, vol. 4, issue. 5, pp 537-550, 2013.
- [Boulnemour, 2015] I. Boulnemour, B. Boucheham, “I-SEA: Improved shape exchange algorithm for quasi-periodic time series alignment”, In *Proceedings of the IEEE International Conference on Computer Vision and Image Analysis Applications (ICCVIA)*, Sousse, Tunisia, pp. 1-6, 2015.
- [Boulnemour, 2016] I. Boulnemour, B. Boucheham, and S. Benloucif, “Improved dynamic time warping for abnormality detection in ECG time series”, In F. Ortuño and I. Rojas (eds), *Bioinformatics and Biomedical Engineering, Lecture Notes in Computer Science*, vol. 9656, Springer, Cham, Granada, Spain, 2016.
- [Boulnemour, 2018] I. Boulnemour and B. Boucheham, “QP-DTW: upgrading dynamic time warping to handle quasi-periodic time series alignment”, *Journal of Information Processing Systems*, vol. 14, n°4, pp. 851-876, 2018.
- [Boulnemour, 2020-a] I. Boulnemour, B. Boucheham, and A. Lahreche, “Accelerating Quasi Periodic Dynamic Time Warping For Similarity Search in ECG Time Series”, accepté pour être présenté à la *1^{ère} édition de la conférence nationale sur l’IA et les technologies de l’information (CNIATI)*, Taref, 2020.
- [Boulnemour, 2021] I. Boulnemour, B. Boucheham, and A. Lahreche, “On Enhancing the Accuracy of Nearest Neighbour Classifier Using Improved Shape Exchange Algorithm”, In *International Journal of Informatics and Applied Mathematics*, Vol.4, n°1, pp. 15-27, 2021.

- [Burrows, 1994] M. Burrows, D. J. Wheeler, “A Block-sorting Lossless Data Compression Algorithm”, Research Report 124, Digital Systems Research Center, Palo Alto, CA, 1994.
- [Cardone, 2003] A. Cardone, S.K. Gupta, and M. Karnik, “A survey of shape similarity assessment algorithms for product design and manufacturing applications”, *Journal of Computing and Information Science in Engineering (ASME)*, vol. 3, n°2, pp. 109-118, 2003.
- [Cassisi, 2012] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata, A. Pulvirenti, “Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining”, Chapter 3 published in IntechOpen, 2012.
- [Cha, 2007] S. H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions”, *International Journal of Mathematical Models and Methods in Applied Sciences*, vol.1, n°4, pp. 300-307, 2007.
- [Chan, 2003] F.K.P.Chan, A.W.C.Fu, C.Yu, Haar wavelets for efficient similarity search of time-series: with and without time warping, *IEEE Transaction on Knowledge Data Engineering*, vol. 15, pp. 686–705, 2003.
- [Chang, 2012] P.C. Chang, J.J. Lin, J.C. Hsieh, and J. Weng, “Myocardial infarction classification with multi-lead ECG using hidden Markov models and Gaussian mixture models” *Applied Soft Computing Journal*, vol. 12, n°10, pp. 3165–3175, 2012.
- [Chaovalit, 2011] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen, Discrete wavelet transform-based timeseries analysis and mining. *ACM Comput. Surv.* vol. 43, n°2, 2011,
- [Chávez, 2006] E. Chávez, and G. Navarro, “Metric Databases, Encyclopedia of Database Technologies and Applications”, Idea Group, 2006.
- [Chen, 2003] L. Chen, and M. T. Özsu, “Similarity-based Retrieval of Time-Series Data Using Multi-Scale Histograms”, Technical Report CS-2003-31, 2003.
- [Chen, 2004] L. Chen, and R.T. Ng, “On the marriage of L_p -norms and edit distance”, In *Proceedings of the Thirtieth international conference on Very large data bases, VLDB Endowment*, vol. 30, pp. 792-803, 2004.
- [Chen, 2005-a] L. Chen, M. T. Özsu, and V. Oria, “Robust and fast similarity search for moving object trajectories”, In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, ACM, pp 491-502, 2005.
- [Chen, 2005-b] L. Chen, “Similarity Search Over Time Series and Trajectory Data”, Thèse de doctorat, Université de Waterloo, Canada, 2005

- [Chen, 2015] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, “The UCR time series classification archive”, 2015, www.cs.ucr.edu/~eamonn/time_series_data/.
- [Chi, 2004] Y. Chi, R. R. Muntz, S. Nijssen, J. N. Kok, “Frequent subtree mining - an overview”. *Fundamenta Informaticae*, vol. 66, pp. 161–198, 2004.
- [Coviello, 2011] E. Coviello, A. B. Chan, and G. Lanckriet, “Time Series Models for Semantic Music Annotation”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, n°5, 2011.
- [Crochemore, 1994] M. Crochemore and W. Rytter, “Text Algorithms”, Oxford Univ. Press, 1994.
- [Dau, 2018] H. A. Dau et al., "Optimizing dynamic time warping's window width for time series data mining applications", *Data Min. Knowl. Discov.*, vol.32, n°4, pp. 1074-1120, 2018.
- [Deng, 2013] H. Deng, G. Runger, E. Tuv, M. Vladimir, “A time series forest for classification and feature extraction”, *Information Sciences*, vol. 239, pp. 142–153, 2013.
- [Deza, 2009] E. Deza, and M. M. Deza, “Encyclopedia of distances”, Springer, Berlin, Heidelberg, 2009.
- [Dietrich, 2004] C. Dietricha, G. Palma, K. Riedeb, F. Schwenker, “Classification of bioacoustic”, vol. 37, pp. 2293 – 2305, 2004.
- [Ding, 2008] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, “Querying and mining of time series data: experimental comparison of representations and distance measure”, In *Proceedings of the VLDB Endowment*, vol.1, n°2, pp. 1542–1552, 2008.
- [Dohnal, 2003] V. Dohnal, C. Gennaro, P. Savino, and P. Zezula, “Similarity Join in Metric Spaces”, In *Advances in Information Retrieval, 25th European Conference on IR Research, ECIR, Pisa, Italy, 2003*.
- [Douglas, 1973] D.H. Douglas, T.K. Peucker, “Algorithms for the reduction of the number of points required to represent a digitized line or its caricature”, *Canadian Cartographer*, vol. 10, n°2, pp. 112–122, 1973.
- [Dubois, 2004] R. Dubois, “Application des nouvelles méthodes d’apprentissage à la détection précoce d’anomalies en électrocardiographie”, thèse de doctorat, université Paris 6, 2004.
- [Duda, 2001] R. O. Duda, P. E. Hart, and D. C. Stork. “*Pattern Classification*” In John Wiley & Sons, New York, 2001.

- [Eads, 2005] D. Eads, K. Glocer, S. Perkins, and J. Theiler, “Grammar-Guided Feature Extraction for Time Series Classification”, In *Proceeding of Conference on Neural Information Processing Systems*, 2005.
- [Egondi, 2012] T. Egondi, C. Kyobutungi, S. Kovats, K. Muindi, R. Ettarh and J. Rocklöv, “Time-series analysis of weather and mortality patterns in Nairobi’s informal settlements”, *Glob Health Action* vol.5, n°1, pp. 23-32, 2012.
- [Esling, 2012] P. Esling, and C. Agon, “Time-series data mining. ACM Computing Surveys”, vol. 45, n°1, pp. 1-34, 2012.
- [Faloutsos, 1994] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, “Fast subsequence matching in time-series databases”, In *ACM SIGMOD*, pp. 419–429, Minneapolis MN, 1994.
- [Faloutsos, 2002] C. Faloutsos, “Image Databases: Search and Retrieval of Digital Imagery”, chapter in *Multimedia Indexing*, Wiley Inter-Science editors, Castelli, Bergman, pp. 435–464, 2002.
- [Farhi, 2015] A. Farhi, “Arythmie complète par fibrillation auriculaire”, 2015.
- [Fawaz, 2019] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P. A. Muller, F. Petitjean “InceptionTime: Finding AlexNet for Time Series Classification”, In *Machine Learning journal*, 2019.
- [Fix, 1951] E. Fix, J.L. Hodges, “Discriminatory analysis, non-parametric discrimination: Consistency properties”, In *Technical Report 4*, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
- [Fu, 2005] A. W. Fu, E. Keogh, L.Y. H. Lau and C.A. Ratanamahatana, “Scaling and Time Warping in Time Series Querying”, In *VLDB*, 2005.
- [Fu, 2011] T. Fu, “A review on time series data mining”, *Engineering Applications of Artificial Intelligence*, Elsevier, vol. 24, pp. 164–181, 2011.
- [Garcia, 2014] C. Garcia de Andrade, M. X. Ribeiro, “Similarity Search in Multidimensional Time Series using the Coulomb’s Law”, *Journal of Information and Data Management*, Vol. 5, No. 1, pp. 74–83, 2014.
- [Gardenhire, 1964] L.W. Gardenhire, “Redundancy reduction the key to adaptive telemetry”, In *Proceedings of the national telemetry conference*, pp. 1–16, 1964.
- [Gaudin, 2005] R. Gaudin and N. Nicoloyannis, “Apprentissage non supervisé des séries temporelles à l'aide des k-means et d'une nouvelle méthode d'agrégation de séries”, In *EGC*, vol. RNTI-E-3, pp. 201-212, 2005.

- [Gaudoin, 2012] O. Gaudoin, “Principes et Méthodes Statistiques”, Cours de statistiques, 2012.
- [Graves, 2013] A. Graves, A. Mohamed, G. Hinton “Speech recognition with deep recurrent neural networks”, In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [Guttman, 1984] A. Guttman, R-Tree: “A Dynamic Index Structure for Spatial Searching”, In *Proceedings of the, ACM SIGMOD International Conference on Management of data*, pp. 47-57, 1984.
- [Hamming, 1958] R. W. Hamming, “Error detecting and error correcting codes”, Bell System technical journal, vol. 131, n°1, pp. 147-160. 1958.
- [Hanyf, 2018] Y. Hanyf, H. Silkan, “A queries-based structure for similarity searching in static and dynamic metric spaces”, *Journal of King Saud University –Computer and Information Sciences*, Article in Press, 2018.
- [Hassanat, 2014] A.B. Hassanat, “Dimensionality invariant similarity measure”, *Journal of American Science*, vol. 10, n°8, pp. 221-226, 2014.
- [He, 2014] X. He, C. Shao, Y. Xiong, “A new similarity measure based on shape information for invariant with multiple distortions”, *Neurocomputing*, vol. 129, pp. 556-569, 2013.
- [He, 2015] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition”, *Computer Vision and Pattern Recognition*, 2015.
- [Hedges, 1976] T. Hedges, “An empirical modification to linear wave theory”, In *Proceedings of the Institution of Civil Engineers*, vol. 61, n°3, pp. 575-579, 1976.
- [Hills, 2014] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall, “Classification of time series by shapelet transformation,” *DAMI*, vol. 28, pp. 851–881, 2014.
- [Hu, 2006] J. Hu, B. Ray, and L. Han, “An interweaved HMM/DTW approach to robust time series clustering”, In *proceedings of the 18th International Conference on Pattern Recognition*, IEEE, vol.3, pp. 145–148, 2006.
- [Huriot, 1990] J.M. Huriot, J. Perreux, “ Distances, espaces et représentations (une revue) ”, Rapport de recherche, Institut de mathématiques économiques (IME), hal-01541305.
- [Huerta, 2012] R. Huerta, S. Vembu, M. K. Muezzinoglu, A. Vergara, “Dynamical SVM for Time Series Classification”, *Pattern Recognition. DAGM/OAGM. Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, vol. 7476, pp. 216-225, 2012.
- [Inokuchi, 2003] A. Inokuchi, T. Washio, H. Motoda, “Complete Mining of Frequent Patterns from Graphs: Mining Graph Data”, *Machine Learning*, vol. 50, pp. 321-354, 2003.

- [Itakura, 1975] F. Itakura, “Minimum prediction residual principle applied to speech recognition, Acoustics, Speech and Signal Processing”, In *IEEE Transactions on Signal Processing*, vol. 23, n°1, pp. 67-72, 1975.
- [Itakura, 1975] F. Itakura, “Minimum prediction residual principle applied to speech recognition”, In *proceedings of the IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, n°1, pp. 67-72, 1975.
- [Jambhale, 2014] S. S. Jambhale, and A. Khaparde, “Reconnaissance des gestes à l'aide de DTW & DTW par morceaux”, *Conférence internationale IEEE 2014 sur l'électronique et le système de communication*, pp 1-5, 2014.
- [Janacek, 2005] G. J. Janacek, A. J. Bagnall, and M. Powell, “A likelihood ratio distance measure for the similarity between the Fourier transform of time series”, In *Advances in Knowledge Discovery and Data Mining*, Springer, pp 737-743, 2005.
- [Jang, 2011] M. Jang, M.S. Han, J.h. Kim, and H.S. Yang, “Dynamic time warping-based k-means clustering for accelerometer-based handwriting recognition”, *Developing Concepts in Applied Intelligence*, Springer Berlin Heidelberg, vol. 363, pp. 21–26, 2011.
- [Jun, 2009] L. Junkui, W. Yuanzhen, “Early Abandon to Accelerate Exact Dynamic Time Warping”, In *International Arab Journal of Information Technology*, vol. 6, n° 2 April 2009.
- [Jung, 2008] H.Y. Jung, S.H. Park, S.J. Park, “Detection abnormal pattern in activities of daily living using sequence alignment method”, In *proceedings of IEEE conference on engineering medicine biology society*. 2008.
- [Kadir, 2012] A. Kadir, L. E. Nugroho, A. Susanto, and P.S. Insap, “Experiments of distance measurements in a foliage plant retrieval system”, *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol.5, n°2, pp. 47-60, 2012.
- [Karadayi, 2020] Y. Karadayi, “Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Datasets Using Deep Learning”, In V. Lemaire et al. (Eds.): *AALTD 2019*, LNAI 11986, Springer, pp. 167–182, 2020.
- [Keogh, 1998] E. Keogh, M. Pazzani, “An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback”, In *Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining*, pp. 239-241, AAAI Press, 1998.
- [Keogh, 2000] E. Keogh and M. Pazzani, “Scaling up Dynamic Time Warping for Data mining Applications”, In *Proceeding of the 6th International Conference on Knowledge Discovery and Data Mining*, Boston, MA, 2000.

- [Keogh, 2001-a] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, “Dimensionality reduction for fast similarity search in large time series databases”, *Knowledge and Information Systems*, vol.3, n°3, pp. 263–286, 2001.
- [Keogh, 2001-b] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, “Locally Adaptive Dimensionality Reduction for Similarity Search in Large Time Series Databases”, In SIGMOD 2001.
- [Keogh, 2001-c] E. Keogh and M. Pazzani, “Derivative Dynamic Time Warping”, In *the Proceedings of the SIAM International Conference on Data Mining*, Chicago, 2001.
- [Keogh, 2002] E. Keogh, C. Ratanamahatana, “Exact indexing of dynamic time warping”, In *the proceedings of the 28th International Conference on Very Large databases. Hong Kong*, pp. 406-417, 2002.
- [Keogh, 2005] E. Keogh, “Exact indexing of dynamic time warping”, In *Proceedings of the 28th international conference on very large data bases*, pp. 358–380, 2005.
- [Keogh, 2006] E. Keogh, L. Wei, X. Xi, S. Lee and M. Vlachos, “LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures”, In *Proceedings of the 32nd international conference on Very large data bases VLDB*, 2006. Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09).
- [Kim, 2001] S.W. Kim, S.H. Park, W.W. Chu, “An index-based approach for similarity search supporting time warping in large sequence databases”, In *Proceedings of the 17th international conference on data engineering*, pp. 607–614, 2001.
- [Korn, 1996] F. Korn, N. Sidiropoulos, C. Faloutsos, E. Siegel, and Z. Protopapas, “Fast nearest neighbor search in medical image databases”, In *Proceedings of the International Conference on Very Large databases*, pp. 215–226, 1996.
- [Korn, 1997] F. Korn, H. V. Jagadish, and C. Faloutsos, “Efficiently supporting ad hoc queries in large datasets of time sequences”. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data (SIGMOD’97)*, pp. 289–300, New York, NY, USA, 1997.
- [Kumar, 2005] N. Kumar, N. Lolla, E. Keogh, S. Lonardi, C. Ratanamahatana, and L. Wei, “Time-series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases”, In *Proceedings of the International Conference on Data Mining (SIAM’05)*, 2005.

- [Lacasa, 2008] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuno, “From time series to complex networks: The visibility graph”, In *Proceedings of the National Academy of Sciences*, vol.105, n°13, pp. 4972–4975, 2008.
- [Lacasa, 2009] L. Lacasa, B. Luque, F. Ballesteros, J. Luque, and J. C. Nuno, “The visibility graph: A new method for estimating the Hurst exponent of fractional Brownian motion”, In *EPL (Europhysics Letters)*, vol. 86, n°3, 2009.
- [Lahreche, 2017] A. Lahrech, B. Boucheham, “FastSEA: A Very Fast and Very effective Matching Technique for Very Complex Time Series”, In *Proceeding of the International Conference on Mathematics and Information Technology (ICMIT)*, 2017.
- [Lahreche, 2018] A. Lahrech, B. Boucheham, “LMDS-SEA: Upgrading the Shape Exchange Algorithm (SEA) to handle general time series classification by local matching and distance selection”, In *proceedings of the 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2018.
- [Lauer, 2006] F. Lauer, G. Bloch, “Méthodes SVM pour l'identification”, *Journées Identification et Modélisation Expérimentale (JIME'2006)*, Poitiers, France, 2006.
- [Laxman, 2006] S. Laxman, P.S. Sastry “A Survey of Temporal Data Mining”, *Sadhana journal*, vol. 31, Part 2, pp. 173–198, 2006.
- [Li, 1997] C. Li, P. Yu, V. Casteli, "MALM: a framework for mining sequence database at multiple abstraction levels", in *Proceedings of the 9th International Conference on Information and Knowledge Management*, pp.267-272, 1997
- [Li, 2012] K. Li, N. Du, and A. Zhang, “Detecting ECG abnormalities via transductive transfer learning”, In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB '12)*, pp. 210–217, 2012.
- [Li, 2013] C. Li, Z. Ma, L. Yao, and D. Zhang, “Improvements on EMG-based handwriting recognition with DTW algorithm”, In *Engineering in Medicine and Biology Society (EMBC), 35th Annual International Conference of the IEEE*, 2013.
- [Li, 2018] D. Li, J. Lin, T. F. Bissyandé, J. Klein, Y. Le Traon, “Extracting Statistical Graph Features for Accurate and Efficient Time Series Classification”, In *Proceedings of the 21st International Conference on Extending Database Technology*, 2018.
- [Liao, 2005] T. W. Liao, “Clustering of time series data - a survey”, *Pattern Recognition*, vol. 38, n°11, pp. 1857-1874, 2005.

- [Lin 2007] J. Lin, E. Keogh, W. Li, and S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series”. *Data Mining and Knowledge Discovery*, vol.15, n°2, pp. 107–144, October2007.
- [Lin, 2003] J. Lin, E. Keogh, and W. Truppel, “Clustering of streaming time series is meaningless”, In *Proceedings of the 2003 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, ACM Press, pp. 56-65, 2003.
- [Lin, 2012] J. Lin, R. Khade, and Y. Li, “Rotation-invariant similarity in time series using bag-of-patterns representation”, *Journal of International Informatics System*, pp. 1–29, 2012.
- [Lines, 2015] J. Lines, A. Bagnall, “Time series classification with ensembles of elastic distance measures”, *Data Mining and Knowledge Discovery* vol.29, n°3, pp. 565, 2015.
- [Lines, 2016] J. Lines, S. Taylor, A. Bagnal, “HIVE-COTE; The Hierarchical Vote Collective of Transformation – based Ensembles for Time Series Classification”, In *proceedings of the IEEE 16th International Conference on Data Mining (ICDM)*, 2016.
- [Lindgren, 2003] C. Lindgren, M.T. Johnson and R.J. Povinelli, “Speech recognition using phase space features”, In *Proceedings of the IEEE International conference on Acoustics Speech and Signal Processing*, China, 2003.
- [Liu, 2019] C. L. Liu, W. H. Hsaio, Y. C. Tu, “Time Series Classification With Multivariate Convolutional Neural Network”, In *Proceedings of the IEEE Transactions on Industrial Electronics*, vol. 66, n° 6, 2019.
- [Lonardi, 2001] S. Lonardi, “Global Detectors of Unusual Words: Design, Implementation, and Applications to Pattern Discovery in Bio sequences”, PhD thesis, Department of Computer sciences, Purdue University, 2001.
- [Lucas, 2018] B. Lucas · A. Shifaz, C. Pelletier, L. O’Neill, N. Zaidi, B. Goethals, F. Petitjean, Geoffrey I.Webb, “Proximity Forest: An effective and scalable distance-based classifier for time series”, *Machine Learning*, 2018.
- [Manne, 2012] S. Manne, S. Kotha, and S.S. Fatima, “Text categorization with K-nearest neighbor Approach”, In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications (INDIA 2012)*, Springer, vol. 132, pp. 413-420, India, 2012.
- [Marteau, 2009] P.F. Marteau, “Time warp edit distance with stiffness adjustment for time series matching”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, n°2, pp. 306-318, 2009.

- [Mefteh, 2011] B. Mefteh, “Contribution des réseaux de neurones temporels dans le traitement des images animées”, Thèse de doctorat en science, Université des Sciences et de la Technologie d’Oran, 2011.
- [Mi, 2010] S. Mi, W. Liping, Z. Kanjie, and Z. Jiangchao, “Multi-lead ECG classification based on independent component analysis and support vector machine”, In *Proceedings of the 3rd International Conference on Biomedical Engineering and Informatics (BMEI '10)*, pp. 960–964, 2010.
- [MITBIH, 2000] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. Ivanov, R. Mark, J. Mietus, G. Moody, C. Peng, H. Stanley, “PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals”. Vol. 101, n°23, pp. 215-220, 2000. [ElectronicPages; <http://circ.ahajournals.org/content/101/23/e215.full>, <http://www.physionet.org/physiobank/database/mitdb/>].
- [Morchen, 2006] F. Morchen, “Time series knowledge mining”, PhD thesis, Department of Mathematics and Computer Science, Philipps-University Marburg, Germany, 2006.
- [Morse, 2007] M. Morse and J.M. Patel, “An Efficient and Accurate Method for Evaluating Time Series Similarity”, In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data (SIGMOD '07)*, pp. 569-580, Beijing, China, 2007.
- [Mosbah, 2017] M. Mosbah and B. Boucheham, “Distance selection based on relevance feedback in the context of CBIR using the SFS meta-heuristic with one round”, *Egyptian Informatics Journal*, no 18, pp. 1-19, 2017.
- [Movchan, 2015] A. Movchan, M. Zymbler, “Time Series Subsequence Similarity Search Under Dynamic Time Warping Distance on the Intel Many-core Accelerators”, Chapter from book *Similarity Search and Applications, In proceedings of the 8th International Conference, SISAP*, Glasgow, UK, 2015.
- [Movellan, 2009] J.R. Movellan, “Primer on the Discrete Fourier Transform”, Tutorial, Machine Perception Laboratory, 2009.
- [Muhammad Fuad, 2011] M.M. Muhammad Fuad, “Similarity Search in High-dimensional Spaces with Applications to Time Series Data Mining and Information Retrieval”, PhD thesis, University of Britain, 2011.
- [Nanopoulos, 2001] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, “Feature-Based Classification of Time-Series Data”, *International Journal of Computer Research*, vol. 10, pp. 49-61, 2001.

- [Nguyen, 2017] T.L, Nguyen, S.Gsponer, G.Ifrim, “Time series classification by sequence learning in all-subsequence space”, In *Proceedings of the 33th International Conference on Data Engineering*, pp. 947–958. IEEE Press, San Diego, 2017.
- [Niennattrakul, 2007] V. Niennattrakul, C.A. Ratanamahatana, “On Clustering Multimedia Time Series Data Using K-Means and Dynamic Time Warping”, In *Proceedings of the International Conference on Multimedia and Ubiquitous Engineering (MUE '07)*, pp. 733-738, 2007.
- [Noé, 2005] L. Noé, “Recherche de similarités dans les séquences d’ADN : modèles et algorithmes pour la conception de graines efficaces”, Thèse de doctorat, Université Henri Poincaré - Nancy, 2005.
- [Nurjahan, 2014] B. Nurjahan, E. Keogh, “Rare Time Series Motif Discovery from Unbounded Streams” In *Proceedings of the VLDB Endowment*, vol. 8, n° 2, 2014.
- [Nygaard, 1997] R. Nygaard, D. Haugland, “Multiple error measures in ECG data compression”, in *Proceedings of the Norwegian ECG signal (NORSIG)*, Tromsø, Norway, pp. 134–139, 1997.
- [Oates, 1999] T. Oates, L. Firoiu, et P. Chen, “Clustering time series with hidden markov models and dynamic time warping”, In *Proceedings of the workshop on neural, symbolic and reinforcement learning methods for sequence learning (IJCAI'99)*, pp. 17–21, 1999.
- [Oates, 2001] T. Oates, L. Firoiu and P. Cohen, “Using dynamic time warping to bootstrap hmm-based clustering of time series”, *Sequence Learning*, pp. 35–52, 2001.
- [Obin, 2016] N. Obin, A. Roebel, “Similarity Search of Acted Voices for Automatic Voice Casting”, In *IEEE/ACM Transactions on Audio, Speech, and Language Processing* vol.29, n°9, 2016.
- [Oppenheim, 1999] A.V. Oppenheim, and R.W. Schaffer, “Discrete-Time Signal Processing”, 2nd Edition, Prentice Hall: Upper Saddle River, NJ, 1999.
- [Owsley, 1997] L. Owsley, L. Atlas, and G. Bernard, “Self-organizing feature maps and hidden markov models for machine-tool monitoring”, In *IEEE Transactions on Signal Processing* vol. 45, n°11, pp. 2787–2798, 1997.
- [Papadopoulos, 2009] A. Papadopoulos, A. Corral, A. Nanopoulos, Y. Theodoridis, “R-Tree (and Family)”. In: LIU L., ÖZSU M.T. (eds) *Encyclopedia of Database Systems*. Springer, Boston, MA, 2009.
- [Paquet, 1997] P. Paquet, “L’utilisation des réseaux de neurones artificiels en finance”, document de recherche n° 1, Laboratoire Orléanais de Gestion, 1997.

- [Paredes, 2009] R. Paredes, N. Reyes, “Solving similarity joins and range queries in metric spaces with the list of twin clusters”, *Journal of Discrete Algorithms*, Elsevier, vol. 7, n°1, pp. 18–35, 2009.
- [Park, 2000] S. Park, W. Chu, J. Yoon, and C. Hsu, “Fast Retrieval of Similar Subsequences of Different Lengths in Sequence Databases”, In *proceedings of the 16th IEEE International Conference on Data Engineering (ICDE)*, pp. 23–32, San Diego, USA, 2000.
- [Park, 2000b] S. Park, W. Chu, J. Yoon, and C. Hsu, “Efficient Searches for Similar Subsequences of Different Lengths in Sequence Databases”, In *Proceedings of the International Conference on Data Engineering (ICDE)*, pp. 23-32, 2000.
- [Petitjean, 2016] F. Petitjean, G. Forestier, G.I. Webb, et al. “Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification”, *Knowledge and Information Systems*, vol. 47, Issue 1, pp. 1-26, 2016.
- [Polat, 2007] K. Polat and S. Güneş, “Detection of ECG arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine”, *Applied Mathematics and Computation*, vol. 186, n°1, pp. 898–906, 2007.
- [Prasatha, 2017] V. B. S. Prasatha, H.A. Abu Alfeilat, O. Lasassmeh, and A. B. A. Hassanat, “Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-A Review”, *Journal Elsevier*, 2017.
- [Qi, 2018] C. Qi, Z. Liu, F. Su, “Accurate and Efficient Similarity Search for Large Scale Face Recognition”, *Computer Vision and Pattern Recognition*, Article in Press, 2018.
- [Raghavendra, 2011] S. Raghavendra, D. Bera, A. S. Bopardikar, and R. Narayanan, “Cardiac arrhythmia detection using dynamic time warping of ECG beats in e-healthcare systems”, In *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM '11)*, pp. 1–6, Lucca, Italy, 2011.
- [Rakthanmanon, 2012] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, “Searching and mining trillions of time series subsequences under dynamic time warping”, In *Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining*, ACM, pp. 262-270, 2012.
- [Ramgopal, 2014] S. Ramgopal, S. Thome-Souza, M. Jackson, N. E. Kadish, I. S. Fernandez, J. Klehm, W. Bosl, C. Reinsberger, S. Schachter, and T. Loddenkemper, “Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy”, *Epilepsy & Behavior*, vol. 37, pp. 291–307, 2014.

- [Ratanamahatana, 2004a] C.A. Ratanamahatana, E. Keogh, E, “Everything you know about dynamic time warping is wrong”. In Third Workshop on Mining Temporal and Sequential Data, pp. 22-25, 2004.
- [Ratanamahatana, 2004-b] C. A. Ratanamahatana, E. Keogh, “Making time-series classification more accurate using learned constraints”, In M. W. Berry, U. Dayal, C. Kamath, and D. B. Skillicorn (eds), In *Proceedings of the 4th SIAM International Conference on Data Mining (SDM'04)*, pp. 11-22. SIAM, 2004.
- [Ratanamahatana, 2005] C.A. Ratanamahatana, and E. Keogh, “Three Myths about Dynamic Time Warping”. In *Proceedings of SIAM International Conference on Data Mining (SDM '05)*, pp. 506-510, Newport Beach, CA, 2005.
- [Ratanamahatana, 2005] C.A. Ratanamahatana, E. Keogh, A.J. Bagnall, S. Lonardi, “A Novel Bit Level Time Series Representation with Implications for Similarity Search and Clustering”, In T.B. Ho, D. Cheung, H. Liu (eds) *Advances in Knowledge Discovery and Data Mining, (PAKD'05D)*, Lecture Notes in Computer Science, vol. 3518, Springer, Berlin, Heidelberg, 2005.
- [Ravinder, 2010] K .Ravinder, “Comparison of hmm and dtw for isolated word recognition system of punjabi language”, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 244–252, 2010.
- [Rodríguez, 2004] J.J. Rodríguez, C.J. Alonso, “Interval and dynamic time warping-based decision trees”. In *Proceedings of the ACM symposium on Applied computing (SAC)*, pp. 548-552, 2004.
- [Rodríguez, 2005] J.J. Rodríguez, C. Alonso, and J. Maestro, “Support Vector Machines of Interval-Based Features for Time Series Classification,” *Knowledge-Based Systems*, vol. 18, pp. 171-178, 2005.
- [Rubner, 2013] Y. Rubner, and C. Tomasi, “Perceptual metrics for image database navigation”, Springer, 2013.
- [Rüping, 2001] S .Rüping, “SVM kernels for time series analysis”, In *LLWA*, Citeseer, pp. 43–50, 2001.
- [Rutishauser, 2004] W. Rutishauser et J. Sztajzel, “Cardiologie Clinique. 2e édition”, Masson, 2004.
- [Sakoe, 1978] H. Sakoe, and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, n°1, pp. 43-49, 1978.

- [Salperwyck, 2015] C. Salperwyck, S. Maby, J. Cubillé, M. Lagacherie, “CourboSpark: Decision Tree for Time series on Spark”, In Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data, 2015.
- [Salvador, 2007] S. Salvador, and P Chan, “Toward accurate dynamic time warping in linear time and space”, *Intelligent Data Analysis*, vol.11, n°5, pp 561-580, 2007.
- [Samsudin, 2010] R. Samsudin, A. Shabri and P. Saad, “A Comparison of Time Series Forecasting using Support Vector Machine and Artificial Neural Network Model”, in *Journal of Applied Sciences*, Vol. 10, n°11, pp. 950-958, 2010.
- [Santos, 2015] F. Santos, “Arbres de décision”, Cours, Université de Bordeaux, 2015.
- [Schäfer, 2012] P. Schäfer, M. Höggqvist, “SFA: a Symbolic Fourier Approximation and index for similarity search in high dimensional datasets”. In: *EDBT*. ACM, 2012.
- [Schäfer, 2015] P. Schäfer, “The boss is concerned with time series classification in the presence of noise”, *Data Mining and Knowledge Discovery*, vol. 29, n° 6, pp. 1505–1530, 2015.
- [Schimke, 2007] S. Schimke, C. Vielhauer, “Similarity searching for on-line handwritten documents”, *Journal on Multimodal User Interfaces*, vol. 1, n° 2, 2007.
- [Senin, 2008] P.V. Senin, “Dynamic Time Warping Algorithm Review”, In Technical Report CSDL-08-04, Department of Information and Computer Sciences, University of Hawaii, 2008.
- [Senin, 2009] P. V. Senin, “Literature Review on Time Series Indexing”, CSDL Technical Report 09-08, Department of Information and Computer Sciences, University of Hawaii, 2009.
- [Serra, 2012] J. Serra and J. L. Arcos, “A competitive measure to assess the similarity between two time series”, In *Case-Based Reasoning Research and Development*, pp. 414–427, 2012.
- [Serfass, 2017] D. Serfass, “Dynamic Biometric Recognition of Handwritten Digits Using Symbolic Aggregate Approximation”, In *Proceedings of the ACM SE '17, South East Conference*, pp 1-4 2017.
- [Shannon, 2001] C.E. Shannon, “A mathematical theory of communication”. *ACM SIGMOBILE, Mobile Computing and Communications Review*, pp. 3-55.2001.
- [Sharma, 2017] A. Sharma, S. Sundaram, “A Novel Online Signature Verification System Based on GMM Features in a DTW Framework”, *IEEE Transactions on Information Forensics and Security*, vol. 12, n°3, pp. 705-718, 2017.

- [Shatkay, 1995] H. Shatkay, “The Fourier Transform - a Primer”. Technical Report CS-95-37, Department of Computer Science, Brown University, 1995.
- [Shatkay, 1996] H. Shatkay, S. B. Zdonik, “Approximate Queries and Representations for Large Data Sequences”, In Proceeding of the 12th International Conference on Data Engineering, 1996.
- [Shirkhorshidi, 2015] S. Shirkhorshidi, S. Aghabozorgi, and T.Y. Wah, “A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data”, PLOS one journal, vol. 10, n°12, pp. 10-12, 2015.
- [Silva, 2018] D. F. Silva, R. Giusti, E. Keogh, G. Batista, “Speeding up similarity search under dynamic time warping by pruning unpromising alignments”, In Data Mining and Knowledge Discovery, vol. 32, n°2, 2018.
- [Singha, 2018] S. Singha, S. K. Pandey, U. Pawar, R. Ram. Janghel, “Classification of ECG Arrhythmia using Recurrent Neural Networks”, Procedia Computer Science, Elsevier, vol. 132, pp. 1290-1297, 2018.
- [Smith, 2009], A.A. Smith, “Classification and Alignment of Gene-Expression Time-Series Data”, Thèse de doctorat, Université de Wisconsin–Madison, 2009
- [Stefan, 2013] A. Stefan, V. Athitsos, G. Das, “The Move-Split-Merge metric for time series”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, n°6, pp. 1425-1438, 2013.
- [Surya, 2017] V. B. S. Prasatha, H.A. Abu Alfeilat, O. Lasassmeh, and A. B. A. Hassanat, “Distance and Similarity Measures Effect on the Performance of K-Nearest Neighbor Classifier-A Review”, Journal Elsevier, 2017.
- [Terano, 2000] T. Terano, H. Liu, A. L. P. Chen, “Knowledge Discovery and Data Mining”, In Proceedings of the 4th Pacific Asia Conference, PAKDD, Kyoto, Japan. Series: Lecture Notes in Artificial Intelligence, Vol. 1805, 2000.
- [Tavenard, 2005] R. Tavenard, “Indexation de séquences de descripteurs pour exploitation audio et vidéo”, Rapport de stage, Ecole Centrale de Lyon – Université de Lyon I, IRISA, 2005.
- [Torkamani, 2017] S. Torkamani, V. Lohweg, “Survey on time series motif discovery” WIREs Data Mining and Knowledge Discovery, vol.7, n°2, 2017.
- [Viano, 2004] M.C. Viano, A. Philipe, "Cours de Séries Temporelles", Cours, Université des Sciences et Technologies de Lille, 2004.
- [Vlacos, 2003] M. Vlacos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, “Indexing Multi-Dimensional Time-Series with Support for Multiple Distance Measures”, In *Proceedings of*

the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 216-225, Washington, DC, USA, 2003.

[Vlachos, 2005] M. Vlachos, Z. Vagenas, P. S. Yu, and V. Athitsos, “Rotation invariant indexing of shapes and line drawings”, In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pp. 131-138, 2005.

[Wagner, 1974] R. A. Wagner, M. J. Fischer, “The String-to-String Correction Problem”, *Journal of the Association for Computing Machinery*, vol. 21, n°1, pp. 168-173, 1974.

[Walas, 2013] M. Walas, A. de Fombelle, A. Schmid, P. Scotto, A. Stella Caumel, V. Thyrault “Outils statistiques pour le management”, Livre publié par Pearson France, 2013.

[Wang, 2000] Z. Wang, Z. Chi, D. Feng, and Q. Wang, “Leaf Image Retrieval with Shape Features”, In *Proceedings of the 4th International Conference on Advances in Visual Information Systems*, pp. 477- 487, 2000.

[Wang, 2013] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, E. Keogh, “Experimental comparison of representation methods and distance measures for time series data”. *Data Mining and Knowledge Discovery*, vol. 26, n°2, pp. 275-309, 2013.

[Wei, 2005] L. Wei, E. Keogh, H. Van Herle, and A. Mafra-Neto, “Atomic Wedgie: Efficient Query Filtering for Streaming Time Series”, In *Proceeding of the 5th IEEE International Conference on Data Mining (ICDM 2005)*, pp 490-497, Houston, Texas, 2005.

[Wei, 2005] L. Wei, N. Kumar, V.N. Lolla, E. Keogh, and S. Lonardi, “Assumption-Free Anomaly Detection in Time Series”, In *Proceeding of the 17th International Conference on Scientific and Statistical Database Management, (SSDBM '05)*, pp. 237-240, University of California, Santa Barbara, 2005.

[Wei, 2006] L. Wei, and E. Keogh, “Semi-supervised time series classification”, In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*”, pp. 748-753, 2006.

[Wenbing, 2017] H. Wenbing, P. Qiuhui, P. Qianying, and H. Mingfeng, “A new method to analyze protein sequence similarity using Dynamic Time Warping”, *Genomics-journal-Elsevier*, vol.109, n°2, pp. 123-130, 2017.

[Wettschereck, 1997] D. Wettschereck, D.W. Aha, and T. Mohri, “A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms”, *Artificial Intelligence Review*, Springer, vol. 11, issue 1-5, pp. 273-314, 1997.

[Wikipedia.Maladie de Bouveret] https://fr.wikipedia.org/wiki/Maladie_de_Bouveret.

- [Williams, 1966] W. T. Williams, and G. N. Lance, “Computer programs for hierarchical polythetic classification \ similarity analyses”, *The Computer*, vol.9, n°1, pp. 60-64, 1966.
- [Wu, 1996] D. Wu, R. Agrawal, D. El Abbadi, A. Singh, and T. R. Smith, “Efficient retrieval for browsing large image databases”, In *Proceedings of the 5th International Conference on Information and knowledge management*, pp.11-18, Rockville, MD, 1996.
- [Wu, 2000] Y. Wu, D. Agrawal, A. EL Abbadi, “A comparison of DFT and DWT based similarity search in time series databases”, In *Proceedings of the 9th ACM CIKM international Conference on Information and knowledge management*, pp. 488-495, 2000.
- [Xi, 2006] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, “Fast time series classification using numerosity reduction”, In *Proceedings of the International Conference on Machine Learning*, pp. 1033–1040, 2006.
- [Xing, 2010] Z. Xing, J. Pei, E. Keogh, “A brief survey on sequence classification”. *ACM SIGKDD Explorations Newsletter*, vol. 12, n°1, pp. 40-48.
- [Xu, 2008] S. Xu, and Y. Wu, “An algorithm for remote sensing image classification based on artificial immune B-cell network”, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol.37, pp. 107-112, 2008.
- [Yang, 2000] Y. Yang, T. Ault, T. Pierce, and C.W. Lattimer, “Improving text categorization methods for event tracking”, In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 65-72, 2000.
- [Yazdani, 1996] N. Yazdani, M.Z.Ozsoyoglu, “Sequence matching of images”, In *Proceedings of the Eighth International Conference on Scientific and Statistical Database Management, IEEE Computer Society (SSDBM '96)*, pp. 53–62, Washington, DC, USA, 1996.
- [Ye, 2009] L. Ye, E.Keogh, “Time Series Shapelets: A New Primitive for Data Mining”, In *Proceedings of the KDD-09 Workshop on Statistical and Relational Learning in Bioinformatics*, Paris, France, 2009.
- [Ye, 2012] T. K. J. Ye, M. Murakawa, T. Higuchi, and N. Otsu, “Anomaly detection using multi-channel FLAC for supporting diagnosis of ECG”, *IEEJ Transactions on Electronics, Information and Systems*, vol. 132, pp. 111–119, 2012.
- [Yeh, 2010] Y.C. Yeh, W.J. Wang, C.W.Chiou, “A novel fuzzy c-means method for classifying heartbeat cases from ECG signals measurement”, 43, pp.1542-1545,2010.
- [Yi, 1998] B.Yi, K. Jagadish, H. Faloutsos, “Efficient retrieval of similar time sequences

under time warping”, In *proceedings of the International Conference on Data Engineering (ICDE’98)*, pp. 23–27.1998.

[Zhang, 2004] D. Zhang, and G. Lu, “Review of shape representation and description techniques”. *Pattern Recognition*, vol. 37, n°1, pp. 1-19, 2004.

[Zhu, 2003] Y. Zhu, D. Shasha, “Warping indexes with envelope transforms for query by humming”. In *Proceedings of ACM SIGMOD Conference*, pp. 181-192, San Diego, CA, 2003.

[Zhu, 2012] Q. Zhu, G. Batista, T. Rakthanmanon, and E. Keogh, “A Novel Approximation to Dynamic Time Warping allows Anytime Clustering of Massive Time Series Datasets”, In *Proceedings of SDM*, pp. 999-1010, 2012.

[Ziat, 2017] A.Y. Ziat, “Apprentissage de représentation pour la prédiction et la classification de séries temporelles”, Thèse de Doctorat en Réseau de neurones, Université Pierre et Marie Curie - Paris VI, 2017.

Annexe (Bibliographie de la candidate)

Travaux Publiés :

[Boulnemour, 2016] I. Boulnemour, B. Boucheham, and S. Benloucif, “Improved dynamic time warping for abnormality detection in ECG time series”, In F. Ortuño and I. Rojas (eds), *Bioinformatics and Biomedical Engineering, Lecture Notes in Computer Science*, vol. 9656, Springer, Cham, Granada, Spain, 2016.

[Boulnemour, 2018] I. Boulnemour and B. Boucheham, “QP-DTW: upgrading dynamic time warping to handle quasi-periodic time series alignment”, *Journal of Information Processing Systems*, vol. 14, n°4, pp. 851-876, 2018.

[[Boulnemour, 2021] I. Boulnemour, B. Boucheham, and A. Lahreche, “On Enhancing the Accuracy of Nearest Neighbour Classifier Using Improved Shape Exchange Algorithm”, In *International Journal of Informatics and Applied Mathematics* , Vol.4, n°1, pp. 15-27, 2021.

Conférences:

[Boulnemour, 2015] I. Boulnemour, B. Boucheham, “I-SEA: Improved shape exchange algorithm for quasi-periodic time series alignment”, In *Proceedings of the IEEE International Conference on Computer Vision and Image Analysis Applications (ICCVIA)*, Sousse, Tunisia, pp. 1-6, 2015.

[Boulnemour, 2020-a] I. Boulnemour, B. Boucheham, and A. Lahreche, “Accelerating Quasi Periodic Dynamic Time Warping For Similarity Search in ECG Time Series”, accepté pour être présenté à la *1^{ère} édition de la conférence nationale sur l’IA et les technologies de l’information (CNIATI)*, Tref, 2020.

Visibilité internationale de la candidate

| <i>Base d'index</i> | <i>Indice/Lectures</i> |
|------------------------------|--|
| Google Scholar/ ReaserchGate | <i>H index :2 /RG Score : 2,28/ Citation :13</i> Lectures : 4718 |
| DBLP | 2 |
| Scopus | 1 |
| Springer | 1 |
| ESCI Thomson Reuters | 1 |
| ACM | 1 |
| IEEE | 1 |
| Compendex | 1 |