Découverte des u-shapelets basée sur la corrélation pour le clustering de séries temporelles incertaines.

Vanel Steve Siyou Fotso[1] Engelbert Mephu Nguifo[1] Philippe Vaslin [1]

[1] University Clermont Auvergne, CNRS, LIMOS, F-63000 Clermont-Ferrand, France

{siyou, mephu, vaslin}@isima.fr

Abstract

An u-shapelet is a sub-sequence of a time series used for clustering a time series dataset. The purpose of this paper is to discover u-shapelets on uncertain time series. To achieve this goal, we propose a dissimilarity score called FOTS whose computation is based on the eigenvector decomposition and the comparison of the autocorrelation matrices of the time series. This score is robust to the presence of uncertainty ; it is not very sensitive to transient changes ; it allows capturing complex relationships between time series such as oscillations and trends, and it is also well adapted to the comparison of short time series. The FOTS score is used with the Scalable Unsupervised Shapelet Discovery algorithm for the clustering of 17 datasets, and it has shown a substantial improvement in the quality of the clustering with respect to the Rand Index. This work defines a novel framework for the clustering of uncertain time series.

Mots Clef

Clustering, UShapelet, Correlation, time series.

Résumé

Un u-shapelet est une sous-séquence d'une série temporelle utilisée comme propriété pour séparer un groupe de séries temporelles en deux sous-groupes. Un sous-groupe de séries temporelles contenant le shapelet et un sous-groupe de série temporelles ne contenant pas le shapelet. Le but de cet article est de découvrir des u-shapelets sur des séries temporelles incertaines. Pour atteindre cet objectif, nous proposons un score de dissimilarité appelé FOTS dont le calcul est basé sur la comparaison des vecteurs propres des matrices d'autocorrélation des séries temporelles. Ce score est robuste à la présence d'incertitude ; il n'est pas très sensible aux changements transitoires ; il permet de capturer des relations complexes entre des séries temporelles telles que les oscillations et les tendances, et il est également bien adapté à la comparaison de séries temporelles courtes. Le score FOTS est utilisé avec l'algorithme Scalable Unsupervised Shapelet Discovery pour la classification non supervisée de 17 ensembles de données, et il a montré une amélioration substantielle de la qualité de la classification non supervisée par rapport au Rand Index. Ce travail définit un nouveau cadre pour la classification non supervisée de séries temporelles incertaines.

Mots Clef

Classification non supervisée, UShapelet, Correlation, Séries temporelles.

1 Introduction

Toutes les mesures effectuées par un système mécanique ont une incertitude. En effet, le principe d'incertitude met en évidence les limites de la capacité des systèmes mécaniques à effectuer des mesures sur un système sans les perturber [1]. Ainsi, les séries temporelles des instruments de mesure sont incertaines. Ces séries temporelles produites par des capteurs constituent une vaste proportion des séries temporelles utilisées en science, que ce soit en médecine avec des Électrocardiogramme (enregistrement de l'activité électrique du cœur), en physique avec des mesures enregistrées par des télescopes, en informatique avec l'Internet des objets, etc. Ignorer l'incertitude des données au cours de leur analyse peut conduire à des conclusions approximatives ou inexactes, d'où la nécessité de mettre en œuvre des techniques de gestion des données incertaines.
Plusieurs études récentes ont porté sur le traitement de l'incertitude dans l'exploration de données. Deux approches principales permettent de prendre en compte l'incertitude dans les tâches de data mining : soit elle est prise en compte lors de la comparaison en utilisant les fonctions de distance appropriées [2–7], soit son impact est réduit par les transformations effectuées sur les données. Cette dernière stratégie est utilisée nativement par l'algorithme u-shapelet.

1.1 État de l'art sur les u-shapelets

Considérons un ensemble de données composé de 6 séries temporelles correspondant aux appels d'oiseaux : 3 séries temporelles correspondant à Moucherolle à côtés olive (séries temporelles vertes) et 3 séries temporelles correspondant aux appels de Moineau à couronne blanche (séries

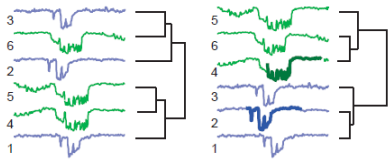

FIGURE 1 – Exemple de classification de séries temporelles utilisant d'une part la distance euclidienne (gauche), d'autre part des sous-séquences caractéristiques appelées u-shapelet (droite). [8].

temporelles bleues). Lorsque ces séries temporelles sont classées en utilisant la distance euclidienne comme mesure de dissimilarité (Fig. 1 gauche), les groupes obtenus ne sont pas homogènes ; en d'autres termes, l'algorithme n'identifie pas l'oiseau à partir de ses cris. Cependant, si nous recherchons des sous-séquences caractéristiques (u-shapelets) pour classer les séries temporelles, nous obtenons des groupes plus homogènes (Fig. 1 à droite).
Une fois cette observation faite, la question naturelle est de savoir comment trouver des sous-séquences qui caractérisent un groupe de séries temporelles, c'est-à-dire des sous-séquences qui ne sont observées que dans un sous-groupe de séries temporelles. L'algorithme de découverte d'u-shapelet répond à cette question et procède comme suit : l'algorithme prend la longueur du motif comme paramètre. Sur chaque série temporelle de la base, on fait glisser une fenêtre de la même longueur que le motif, chaque nouvelle sous-séquence obtenue par ce processus est un u-shapelet candidat.
Parmi les u-shapelets candidats, nous considérons comme un u-shapelet la sous-séquence capable de diviser l'ensemble de séries temporelles en deux sous-ensembles D_A et D_B de sorte que D_A contient toutes les séries temporelles qui possèdent le u-shapelet et D_B toutes celles qui ne contiennent pas l'u-shapelet. Deux autres contraintes sont prises en compte dans la découverte de motifs :
La première est la capacité de l'u-shapelet à construire des sous-ensembles bien séparés. La deuxième est la capacité de l'u-shapelet à construire des sous-ensembles qui ne sont pas déséquilibrés. C'est-à-dire que la taille de D_A doit être au plus k fois plus grande que celle de D_B et vice versa.

Definition 1 *Deux jeux de données D_A et D_B sont dit **r-équilibré** si et seulement si $\frac{1}{r} < \frac{|D_A|}{|D_B|} < (1 - \frac{1}{r}), r > 1$*

Definition 2 *Un **u-shapelet** est une sous-séquence qui a une longueur inférieur ou égale à la longueur de la plus petite série temporelle du jeu de données, et qui permet de diviser le jeux de données en deux sous-groupes **r-équilibrées** D_A et D_B ; où D_A est le groupe de séries temporelles qui contiennent un motif **similaire** au u-shapelet et D_B est le groupe de séries temporelles qui ne contiennent pas l'u-shapelet.*

La similarité entre une série temporelle et un u-shapelet est évaluée à l'aide d'une fonction de distance.

Definition 3 *La distance de sous-séquence **sdist(S, T)** entre une série temporelle T et une sous-séquence S est le minimum des distances entre la sous-séquence S et toutes les sous-séquences de T de longueur égale à celle de S.*

Cette définition ouvre la question de la mesure de distance à utiliser pour $sdist$. En général, la distance euclidienne omniprésente (ED) est utilisée, mais elle ne l'est pas appropriée pour les séries temporelles incertaines [5]. Dans la section suivante, nous présentons une fonction de dissimilarité plus adaptée à l'incertitude.
Le calcul de la $sdist$ entre un u-shapelet candidat et toutes les séries temporelles d'un jeux de données est appelé orderline.

Definition 4 *Un orderline est un vecteur de distance entre un u-shapelet et toutes les séries temporelles d'un jeu de données.*

Le calcul d'un orderline est coûteux en temps. Un orderline pour un seul u-shapelet candidat a une complexité en temps égale à $O(NMlog(M))$ où N est le nombre de séries temporelles du jeux de données, M est la longueur moyenne des séries temporelles. L'algorithme force brute pour la découverte de u-shapelet requiert K calculs d'orderline, où K est le nombre de sous-séquences candidate. La stratégie utilisée par [8] à filtrer les K sous-séquences candidates en considérant seulement celles permettant de construire deux groupes r-équilibrés. Cette sélection est faite efficacement grâce a un algorithme de hachage.
L'évaluation de la qualité des u-shapelets est basée sur leur pouvoir de séparation qui est calculé comme suit :

$$gap = \mu_B - \sigma_B - (\mu_A - \sigma_A), \quad (1)$$

Où μ_A (resp. μ_B) représente la moyenne(sdist(S, D_A)) (resp. moyenne(sdist(S, D_B))), et σ_A (resp. σ_B) représente l'écart-type de $sdist(S, D_A)$ (resp. écart-type de $sdist(S, D_B)$).
Si D_A ou D_B est constitué d'un seul élément (ou d'un nombre insuffisant de séries temporelles pour constituer un groupe), le gap prend la valeur zéro. Ceci assure qu'un gap élevé pour un u-shapelet correspond à une séparation réelle.

1.2 U-shapelets pour la classification non supervisée de séries temporelles incertaines

La classification non supervisée basée sur des u-shapelets est une approche introduite par [9] qui a suggéré de regrouper des séries temporelles à partir des propriétés locales de leurs sous-séquences plutôt qu'utiliser les caractéristiques globales de la série temporelle [10]. Dans ce but, le clustering basé sur les u-shapelets cherche d'abord un ensemble de sous-séquences caractéristiques des différentes catégories de séries temporelles et classe une série temporelle en fonction de la présence ou de l'absence de ces sous-séquences caractérisitiques.

La classification non supervisée de séries temporelles avec des u-shapelets présente plusieurs avantages. Premièrement, la classification non supervisée basée sur les u-shapelets est définie pour les ensembles de données dans lesquels les séries temporelles ont des longueurs différentes, ce qui n'est pas le cas pour la plupart des techniques décrites dans la littérature. En effet, dans de nombreux cas, l'hypothèse de longueur égale est implicite, et le découpage à longueur égale est effectué en exploitant des compétences humaines coûteuses [8]. Deuxièmement, la classification non supervisée basée sur les u-shapelets est beaucoup plus expressive en ce qui concerne le pouvoir de représentation. En effet, une série temporelle n'est associé à un groupe que si elle contient l'u-shapelet caractéristique de ce groupe. Ainsi, un une série temporelle pourrait n'être associé à aucun groupe.
De plus, il est très approprié d'utiliser la classification non supervisée basée sur les u-shapelets avec des séries temporelles incertaines parce que la stratégie de comparaison d'un u-shapelet et d'une série temporelle ignore les données non pertinentes de la série temporelle et ainsi réduire les effets négatifs de la présence d'incertitudes dans celle-ci. Malgré cet avantage, il est hautement souhaitable de prendre en compte l'impact négatif de l'incertitude lors de la découverte des u-shapelets.

1.3 Incertitude et découverte des u-shapelets

Les mesures traditionnelles de similarité comme la distance euclidienne (ED) ou la distorsion temporelle dynamique (DTW) ne fonctionnent pas toujours bien pour les séries temporelles incertaines. En effet, ces distances agrègent l'incertitude de chaque point de la série temporelle et amplifient ainsi l'impact négatif de l'incertitude. Cependant, ED joue un rôle fondamental dans la découverte des u-shapelets, car elle est utilisée pour calculer l'écart entre deux groupes formé par l'u-shapelet candidat. La découverte de u-shapelet sur des séries temporelles incertaines pourrait donc conduire à la sélection d'un mauvais candidat u-shapelet ou à l'assignation d'une série temporelle au mauvais groupe.
Dans cette étude, notre but n'est pas de définir un algorithme pour la découverte d'u-shapelets incertains, mais plutôt d'utiliser une fonction de dissimilarité robuste à l'incertitude pour améliorer la qualité des u-shapelets découverts et donc la qualité de clustering des séries temporelles incertaines.

1.4 Contributions

- Nous faisons un état de l'art sur les mesures de dissimilarité incertaines et nous les évaluation leur pertinence pour la comparaison de séries temporelles incertaines de petite taille.
- Nous introduisons une fonction de dissimilarité nommée correlation frobenius pour découverte d'u-shapelets sur les séries temporelles incertaines (FOTS) ;qui posséde des propriétés intéressantes pour la comparaison de séries temporelles incertaines de petite taille et ne fait aucune hypothèse sur la distribution de probabilité de l'incertitude.
- Nous mettons le code source à la disposition de la communauté scientifique pour permettre une extension de ce travail.

2 Définitions et travaux connexes

2.1 Définitions

Une série temporelle incertaine (UTS) $X =< X_1, \ldots, X_n >$ est une séquence de variable aléatoire où X_i est une variable aléatoire modélisant une valeur réelle inconnue à l'instant i. Deux modèles sont principalement utilisés pour représenter les séries temporelles incertaines : le modèle ensembliste, et le modèle basé sur la fonction de densité de probabilité de l'incertitude.

Dans le modèle ensembliste, chaque élément $X_i (1 \leq i \leq n)$ de l'UTS $X =< X_1, \ldots, X_n >$ est représenté par un ensemble $\{X_{i,1}, \ldots, X_{i,N_i}\}$ de valeurs observées et N_i représente le nombre d'observation à l'instant i.

Dans le modèle basé sur la distribution de probabilité, chaque élément $X_i, (1 \leq i \leq n)$ de l'UTS $X =< X_1, \ldots, X_n >$ est représenté par une variable aléatoire $X_i = x_i + X_{e_i}$, où x_i est la valeur exacte qui est inconnue, et X_{e_i} est une variable aléatoire représentant l'erreur. C'est ce modèle que nous considérons tout au long de notre travail.
Plusieurs mesures de similarité ont été proposées pour les séries temporelles incertaines. Elles peuvent être regroupées en deux catégories principales : les mesures de similarité traditionnelles et les mesures de similarité incertaines.

- les mesures de similarité traditionnelles telle que la distance euclidienne sont celles conventionnellement utilisées avec les séries temporelles. Elles utilisent une seul valeur incertaine à chaque instant comme approximation de la valeur réelle inconnue.
- les mesures de similarité incertaines utilisent des informations statistiques additionnelles qui mesurent l'incertitude associée à chaque approximation de la valeur réelle. C'est le cas notamment de DUST, PROUD, MUNICH [11]. [12] démontre que les performances des mesures de similarité incertaines associées au pré-traitement sont meilleures que les performances des mesures de similarité traditionnelles sur des jeux de données contenant de l'incertitude.

2.2 État de l'art sur les mesures de similarité incertaines

Les mesures de similarité incertaines peuvent être regroupées en deux grandes catégories : les mesures de similarité déterministes et les mesures de similarité probabilistes.

Mesure de similarité déterministe. Tout comme les mesures de similarité traditionnelles, les mesures de similarité déterministes renvoient un nombre réel représentant la dis-

tance entre deux séries temporelles incertaines. **DUST** est un exemple de mesure déterministe de similarité.

DUST [13] Etant donné deux séries temporelles incertaines $X =< X_1, \ldots, X_n >$ et $Y =< Y_1, \ldots, Y_n >$, la distance entre deux valeurs X_i, Y_i est définie comme étant la distance entre leurs valeurs réelles inconnues r(X_i), r(Y_i) : $dist(X_i, Y_i) = |r(X_i) - r(Y_i)|$. Cette distance est utilisée pour mesurer la similarité de deux valeurs incertaines.
$\varphi(|X_i - Y_i|)$ est la probabilité que les valeurs réelles à l'instant i soient égales, connaissant les valeurs réelles à l'instant i.

$$\varphi(|X_i - Y_i|) = Pr(dist(0, |X_i - Y_i|) = 0). \quad (2)$$

cette fonction de similarité est par la suite utilisée par la fonction de dissimilarité $dust$:

$$dust(X_i, Y_i) = \sqrt{-log(\varphi(|X_i - Y_i|)) + log(\varphi(0))}. \quad (3)$$

La distance entre les séries temporelles incertaines $X =< X_1, \ldots, X_n >$ et $Y =< Y_1, \ldots, Y_n >$ calculée à partir de $DUST$ est alors définie comme suit :

$$DUST(X, Y) = \sqrt{\sum_{i=1}^{n} dust(X_i, Y_i)^2}. \quad (4)$$

Le problème avec les distances déterministes incertaines comme $DUST$ est que leur expression varie en fonction de la distribution de probabilité de l'incertitude, et la distribution de probabilité de l'incertitude n'est pas toujours disponible pour les jeux de données de séries temporelles.

Mesure de similarité probabiliste. Les mesures des similarités probabilistes n'exigent pas la connaissance de la distribution des probabilités d'incertitude. De plus, elles fournissent aux utilisateurs plus d'informations sur la fiabilité du résultat. Il existe plusieurs fonctions de similarité probabiliste, comme MUNICH, PROUD, PROUDS ou Corrélation Locale.

MUNICH [14]. Cette fonction de distance convient aux séries temporelles incertaines représentées par le modèle ensembliste. La probabilité que la distance entre deux séries temporelles incertaines X et Y soit inférieure à un seuil ε est égale au nombre de distances entre X et Y, qui sont inférieures à ε, sur le nombre possible de distances :

$$Pr(distance(X, Y)) \leq \varepsilon = \frac{|\{d \in dists(X, Y) | d \leq \varepsilon\}|}{|dists(X, Y)|} \quad (5)$$

Le calcul de cette fonction de distance est très couteuse en temps.

PROUD [15] Soient $X =< X_1, ..., X_n >$ et $Y =< Y_1, ..., Y_n >$ deux UTS modélisées par des séquences de variables aléatoires, la distance PROUD entre X et Y est $d(X, Y) = \sum_{i=1}^{n} (X_i - Y_i)^2$. D'après le théorème central limite [16], la distribution cumulée approche asymptotiquement une loi normale

$$d(X, Y) \propto N(\sum_i E[(X_i - Y_i)^2], \sum_i Var[(X_i - Y_i)^2]) \quad (6)$$

Une conséquence de cette caractéristique de la distance PROUD est que le tableau de la loi réduite centrée normale peut être utilisé pour calculer la probabilité que la distance PROUD normalisée soit inférieure à un seuil :

$$Pr(d(X, Y)_{norm} \leq \epsilon). \quad (7)$$

Un inconvénient majeur de PROUD est son inadéquation pour la comparaison de séries temporelles de petites longueurs comme les u-shapelets. En effet, le calcul de la probabilité que la distance PROUD soit inférieure à une valeur est basé sur l'hypothèse qu'elle suit **asymptotiquement** une distribution normale. Ainsi, cette probabilité sera d'autant plus précise que les séries temporelles comparées sont longues (plus de 30 points de données).

PROUDS [12] est une version amélioré de PROUD qui suppose que les variables aléatoires qui constituent la série temporelle sont indépendantes et identiquement distribuées.

Definition 5 *La forme normalisée d'une série temporelle* $X =< X_1, \ldots, X_n >$ *est définie par* $\hat{X} =< \hat{X_1}, \ldots, \hat{X_n} >$ *tel que à chaque instant i* $(1 \leq i \leq n)$*, nous avons :*

$$\hat{X_i} = \frac{X_i - \bar{X}}{S_X}, \; \bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}, \; S_X = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{(n-1)}}. \quad (8)$$

PROUDS définit la distance entre deux séries temporelles normalisées $\hat{X} =< \hat{X_1} ... \hat{X_n} >$ and $\hat{Y} =< \hat{Y_1} ... \hat{Y_n} >$ (Définition 5) comme suit :

$$Eucl(\hat{X}, \hat{Y}) = 2(n-1) + 2 \sum_{i=1}^{n} \hat{X_i} \hat{Y_i} \quad (9)$$

Pour les mêmes raisons que PROUD, PROUDS ne conviennent pas à la comparaison de séries temporelles courtes. Un autre inconvénient de PROUDS est qu'il suppose que les variables aléatoires sont indépendantes : cette hypothèse est forte et particulièrement inappropriée pour des séries temporelles courtes comme les u-shapelets. Une

hypothèse plus réaliste avec les séries temporelles serait de considérer que les variables aléatoires constituant les séries temporelles sont M-dépendantes. Les variables aléatoires d'une série temporelle sont dites M-dépendantes si $X_i, X_{i+1}, ..., X_{i+M}$ sont dépendantes (corrélées) et les variables X_i et X_{i+M+1} sont indépendantes. Cependant, supposer que les variables aléatoires sont M-dépendantes complexifie l'écriture de PROUDS et rend son utilisation plus difficile car elle requiert dès lors d'affecter une valeur au paramètre M.

Corrélation incertaine [7] : Les techniques d'analyse de corrélation sont utiles pour la sélection de caractéristiques dans des séries temporelles incertaines. Ces informations permettent d'identifier les éléments redondants. La même stratégie peut être utile pour la découverte de u-shapelet. La corrélation incertaine est définie comme suit :

Definition 6 *(Correlation sur les séries temporelles incertaines) Etant données deux UTS $X =< X_1, \ldots, X_n >$ et $Y =< Y_1, \ldots, Y_n >$, leur correlation est définie par*

$$Corr(X, Y) = \sum_{i=1}^{n} \hat{X}_i \hat{Y}_i / (n-1), \tag{10}$$

Où $\hat{X}_i$ et $\hat{Y}_i$ sont les formes normales de X_i et Y_i (Définition 5) respectivement. X_i et Y_i sont des variables aléatoires indépendantes et continues.

Si nous connaissons la distribution de probabilité des variables aléatoires, il est possible de déterminer la fonction de densité de probabilité associée à la corrélation, qui servira par la suite à calculer la probabilité que la corrélation entre deux séries temporelles soit supérieure à un seuil donné. La corrélation incertaine a cependant quelques inconvénients :

- Il est trop sensible aux changements transitoires, ce qui conduit souvent à des scores très fluctuants ;
- Il ne peut pas capturer les relations complexes dans les séries temporelles ;
- Il faut connaître la fonction de distribution de probabilité de l'incertitude ou faire une hypothèse sur l'indépendance des variables aléatoires contenues dans les séries temporelles.

En raison de tous ces inconvénients, la corrélation incertaine ne peut pas être utilisée en l'état pour la découverte d'u-shapelet. Le paragraphe suivant présente une généralisation du coefficient de corrélation qui n'est pas une fonction de similarité incertaine mais qui reste intéressante pour la découverte des u-shapelets.

Corrélation locale [17] : La corrélation locale est une généralisation de la corrélation. Elle calcule un score de corrélation évolutif dans le temps qui suit une similarité locale sur des séries temporelles multivariées basées sur une matrice d'auto-corrélation locale. La matrice d'auto-corrélation **permet de capturer des relations complexes** dans des séries temporelles comme l'oscillation clé (par exemple, sinusoïdale) ainsi que les tendances apériodiques (par exemple, à la hausse ou à la baisse) qui sont présentes. L'utilisation de matrices d'auto-corrélation qui sont calculées sur la base de fenêtres se chevauchant permet de **réduire la sensibilité aux changements transitoires** dans les séries temporelles.

Definition 7 *(Auto-covariance, fenêtre glissante) Étant donnée une série temporelle X, un ensemble de fenêtre glissante w, l'estimateur de la matrice d'autocorrelation locale $\hat{\Gamma}_t$ utilisant une fenêtre glissante est définie à l'instant $t \in \mathbb{N}$ tel que (Eq.11) :*

$$\hat{\Gamma}_t(X, w, m) = \sum_{\tau=t-m+1}^{t} x_{\tau,w} \otimes x_{\tau,w}. \tag{11}$$

Où $\boldsymbol{x}_{\tau,\omega}$ est une sous-séquence de la série temporelle de longueur w et commençant à τ, $x \otimes y = xy^T$ est le produit extérieur de x et y. L'ensemble d'échantillons de m fenêtres est centré autour du temps t. Nous fixons généralement le nombre de fenêtres à m = w.

Étant donnée les estimations $\hat{\Gamma}_t(X)$ et $\hat{\Gamma}_t(Y)$ pour les deux séries temporelles, la prochaine étape est de savoir comment les comparer et extraire un score de corrélation. Cet objectif est atteint en utilisant la décomposition spectrale ; les vecteurs propres des matrices d'auto-corrélations capturent les principales tendances apériodiques et oscillatoires, même sur des **séries temporelles courtes**. Ainsi, les sous-espaces couverts par les premiers (k) vecteurs propres sont utilisés pour caractériser localement le comportement de chaque série. La définition 8 formalise cette notion :

Definition 8 *(LoCo score) Étant donnée deux séries temporelles X et Y leur score LoCo est défini par*

$$\ell_t(X, Y) = \frac{1}{2}(\|\boldsymbol{U}_X^T \boldsymbol{u}_Y\| + \|\boldsymbol{U}_Y^T \boldsymbol{u}_X\|) \tag{12}$$

Où $\boldsymbol{U}_X$ et $\boldsymbol{U}_Y$ sont les k premiers vecteurs propres des matrices d'auto-corrélation locales $\hat{\Gamma}_t(X)$ et $\hat{\Gamma}_t(Y)$ respectivement, et u_X et u_Y sont les vecteurs propres ayant la plus large valeur propre.
Intuitivement, deux séries temporelles X et Y seront considérées comme étant proches lorsque l'angle formé par l'espace portant les informations de la série temporelle X et de la série temporelle Y est nul. En d'autres termes, X et Y seront proches lorsque la valeur de $cos(\alpha)$ sera 1. La seule l'hypothèse faite pour le calcul de la similitude LoCo est que la moyenne des points de la série temporelle est nulle. Cette hypothèse peut facilement être vérifiée, il suffit pour cela de normaliser les séries temporelles en cours de comparaison. La fonction de similarité LoCo a de nombreuses propriétés intéressantes et ne nécessite pas :

- de connaître la distribution de probabilité de l'incertitude,

— de supposer l'indépendance des variables aléatoires ou de faire une hypothèse sur la longueur de l'u-shapelet.

Elle est donc intéressante pour la découverte de motifs, mais nous avons encore besoin d'une dissimilarité pour pouvoir découvrir des u-shapelets. Dans le paragraphe suivant, nous allons définir une fonction de dissimilarité qui a les mêmes propriétés que LoCo et c'est-à-dire, qui est robuste à la présence d'incertitude.

3 Notre approche

3.1 Fonction de dissimilarité

La fonction de similarité LoCo définie sur deux séries temporelles multivariées X et Y correspond approximativement à la valeur absolue du cosinus de l'angle formé par les espaces propres de X et Y ($|cos(\alpha)|$). Une idée simple serait d'utiliser la valeur $sin(\alpha)$ ou α comme fonction de dissimilarité mais cette approche ne fonctionne pas si bien ; le sinus et l'angle ne sont pas assez discriminants pour la comparaison de vecteurs propres à des fins de clustering. Nous proposons donc la mesure de dissimilarité suivante (Définition. 9).

Definition 9 *(FOTS : Frobenius cOrrelation for uncertain Time series u-Shapelet discovery)*
Étant données deux séries X et Y, leur score FOTS est défini par

$$FOTS(X,Y) = \|U_X - U_Y\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{k}(U_X - U_Y)_{ij}^2} \quad (13)$$

où $\|\|_F$ est la norme de Frobenius, m est la longueur de la série temporelle et k est le nombre de vecteurs propres.

Comme le calcul FOTS est basé sur la comparaison des k-premiers vecteurs propres des matrices d'autocorrélation de la série temporelle, il a les mêmes propriétés souhaitables de la fonction de similarité LoCo, c'est-à-dire :

— Il **permet de capturer des relations complexes** dans des séries temporelles comme les tendances oscillatoires clés (par exemple, sinusoïdales) ainsi que les tendances apériodiques (par exemple, à la hausse ou à la baisse) qui sont présentes ;
— Il permet de **réduire la sensibilité aux changements transitoires** dans les séries temporelles ;
— Il est approprié pour le **comparison de séries temporelles courtes**.

De plus, la fonction de dissimilarité FOTS est **robust à la présence d'incertitude** due à la décomposition spectrale des matrices d'autocorrélation des séries temporelles. La robustesse du FOTS à l'incertitude est confirmée par le théorème suivant :

Théorème 1 ***(HoffmanWielandt)*** *[18] Si X et $X+E$ sont matrices $n \times n$ symétriques, alors :*

$$\sum_{i=1}^{n}(\lambda_i(X+E) - \lambda_i(X))^2 \leq ||E||_F^2. \quad (14)$$

où $\lambda_i(X)$ est la plus grande valeur propre de X, et $|||E|E||_F^F 2$. est le carré de la norme Frobenius de E.

La section suivante explique comment le FOTS est intégré dans l'algorithme de découverte d'u-shapelets.

3.2 Algorithme des u-shapelets avec score FOTS

Dans cette section, nous ne définissons pas un nouvel algorithme SUShapelet, mais nous expliquons comment nous utilisons l'algorithme SUShapelet avec le score FOTS (FOTS-SUSh) pour faire face à l'incertitude.

Le gap est un critère essentiel pour la sélection des u-shapelets. Il est sujet à l'incertitude parce que son calcul est basé sur la distance euclidienne. Pour y remédier, nous proposons d'utiliser le score de FOTS au lieu d'une simple distance euclidienne lors du calcul du gap. L'algorithme 1 explique comment calculer l'orderline en utilisant le score de FOTS. L'algorithme 2 calcul l'orderline et trie les série temporelles en fonction de leur proximité au u-shapelet candidat (ligne 2 et 3). Un u-shapelet est considéré comme étant présent dans une série temporelle si sa distance à la série temporelle est inférieure ou égale à un seuil. Ainsi, l'algorithme de sélection de seuil construit un cluster D_A dont la taille varie entre lb et ub (ligne 5). L'algorithme cherche alors parmi les seuils sélectionnés celui ayant un gap maximal (ligne 6 et 11).

Definition 10 *La fonction de dissimilarité $\boldsymbol{sd_f(S,T)}$ entre une série temporelle T et une sous-séquence S est le minimum des valeurs du score de FOTS entre la sous-séquence S et toutes les sous-séquences possible de T de longueur égales à S.*

Algorithme 1 : ComputeOrderline

Input : u-shapeletCandidate : s,
Jeu de données : D
Output : Distance entre l'u-shapelet candidat et toutes les séries temporelles du jeu de données

1 **function** ComputeOrderline(s, D)
2 $dis \leftarrow \{\}$ $s \leftarrow zNorm(s)$
3 **forall** $i \in \{1, 2, \ldots, |D|\}$ **do**
4 $ts \leftarrow D(i,:)$
5 $dis(i) \leftarrow sd_f(s, ts)$
6 **return** $dis/_{|s|}$

Algorithme 2 : ComputeGap

Input : u-shapeletCandidate : s,
Jeu de données : D,
lb, ub : lower/upper bound : nombre minimum et maximum de séries temporelles par groupe
Output : gap : distance entre deux groupes

function ComputeGap(s, D, lb, ub)
  $dis \leftarrow ComputeOrderline(s, D)$
  $dis \leftarrow sort(dis)$ $gap \leftarrow 0$
  for $i \leftarrow lb$ **to** ub **do**
    $D_A \leftarrow dis \leq dis(i)$, $D_B \leftarrow dis > dis(i)$
    $m_A \leftarrow mean(D_A)$, $m_B \leftarrow mean(D_B)$
    $s_A \leftarrow std(D_A)$, $s_B \leftarrow std(D_B)$
    $currGap \leftarrow m_B - s_B - (m_A + s_A)$
    if $currGap > gap$ **then**
      $gap \leftarrow currGap$
  return gap

4 Evaluation expérimentale

4.1 Classification non supervisée avec les u-shapelets

Il existe de nombreuses façons de regrouper les séries temporelles décrites par des u-shapelets. Dans cette expérience, l'algorithme divise itérativement le jeu de données à partir de chaque u-shapelet découvert : chaque u-shapelet divise l'ensemble de données en deux groupes D_A et D_B. Les séries temporelles qui appartiennent à D_A sont celles contenant le u-shapelet et sont ensuite supprimées du jeu de données. Une nouvelle recherche de u-shapelet se poursuit avec le reste des données jusqu'à ce qu'il n'y ait plus de séries temporelles dans le jeu de données ou jusqu'à ce que l'algorithme ne soit plus capable de trouver d'u-shapelet. En guise de critère d'arrêt, nous considérons la baisse du gap. L'algorithme s'arrête lorsque le gap de l'u-shapelet nouvellement trouvé devient la moitié du gap du premier u-shapelet découvert. Cette approche est une mise en oeuvre directe de la définition d'u-shapelet.

Choisir la longueur N des u-shapelet : Le choix de la longueur de l'u-shapelet est guidé par la connaissance du domaine d'application duquel provient les séries temporelles. Dans le cadre de ces expériences, nous avons testé tous les nombres entre 4 et la moitié de la longueur des séries temporelles. Nous considérons comme longueur de l'u-shapelet celle permettant de mieux regrouper les séries temporelles.

Choisir la longueur w de la fenêtre : L'utilisation de fenêtres qui se chevauchent pour le calcul de la matrice d'auto-corrélation permet de capturer les oscillations présentes dans la série temporelle. Au cours de ces expériences, nous considérons que la taille de la fenêtre est égale à la moitié de la longueur de la forme en U.

Choisir le nombre k de vecteurs propres : Un choix pratique est de fixer k à une petite valeur ; nous utilisons $k = 4$ tout au long de ces expériences. En effet, les tendances apériodiques clés sont capturées par un seul vecteurs propres, tandis que les principales tendances oscillatoires se manifestent dans une paire de vecteurs propres.

5 Métriques d'évalutation

Différentes mesures de la qualité de classification non supervisée de séries temporelles ont été proposées, notamment le score Jaccard, l'indice Rand, l'indice Folkes et l'indice Mallow, etc. Cependant, dans notre cas, nous avons des étiquettes de classe pour les jeux de données, nous pouvons donc utiliser cette information externe pour évaluer la véritable qualité de la classification non supervisée en utilisant l'indice Rand. De plus, l'indice Rand semble être la mesure de la qualité des groupes couramment utilisée [8–10].

5.1 Comparaison avec u-shapelet

De même que [11], nous avons testé notre méthode sur 17 jeux de données du monde réel provenant des archives UCR [19] représentant un large éventail de domaines d'application. Les ensembles de d'apprentissage et de test ont été réunis pour obtenir des jeux de données plus importants. Le tableau 1 présente des informations détaillées sur les ensembles de données testés.

Data-set	Size of dataset	Length	No. of Classes	Type
50words	905	270	50	IMAGE
Adiac	781	176	37	IMAGE
Beef	60	470	5	SPECTRO
Car	120	577	4	SENSOR
CBF	930	128	3	SIMULATED
Coffee	56	286	2	SPECTRO
ECG200	200	96	2	ECG
FaceFour	112	350	4	IMAGE
FISH	350	463	7	IMAGE
Gun_Point	200	150	2	MOTION
Lighting2	121	637	2	SENSOR
Lighting7	143	319	7	SENSOR
OliveOil	60	570	4	SPECTRO
OSULeaf	442	427	6	IMAGE
SwedishLeaf	1125	128	15	IMAGE
synthetic_control	600	60	6	SIMULATED
FaceAll	2250	131	14	IMAGE

TABLE 1 – Jeux de données

Le tableau 2 présente une comparaison entre les deux algorithmes.

5.2 Comparaison avec k-shape et USLM

k-Shape et USLM sont deux algorithmes de clustering basés sur les u-shapelets pour les séries temporelles présentées dans [10]. Dans cette section, nous comparons l'indice Rand index obtenu par FOTS-UShapelet et celui obtenu par

Datasets	RI_SUSh	RI_FOTS
50words	0.811	**0.877**
Adiac	0.796	**0.905**
Beef	0.897	**0.910**
Car	0.708	**0.723**
CBF	0.578	**0.909**
Coffee	0.782	**0.896**
ECG200	0.717	**0.866**
FaceFour	0.859	**0.910**
FISH	0.775	**0.899**
Gun_Point	0.710	**0.894**
Lighting2	0.794	**0.911**
Lighting7	0.757	**0.910**
OliveOil	0.714	**0.910**
OSULeaf	0.847	**0.905**
SwedishLeaf	0.305	**0.909**
synthetic_control	0.723	**0.899**
FaceAll	0.907	**0.908**

TABLE 2 – Comparison du Rand Index de SUSH (RI_SUSh) et de FOTS-SUSh (RI_FOTS). Le meilleur Rand Index est en gras

k-Shape et USLM sur 5 jeux de données [1] (Tableau 3). Les résultats de k-Shape et USLM ont été précédemment rapportés dans [10]. Cette comparaison montre qu'en général, FOTS-UShapelet donne de meilleurs résultats que k-Shape et USLM.

TABLE 3 – Comparison entre k-Shape, USLM et FOTS-UShapelet

Rand Index	k-Shape	USLM	FOTS-UShapelet
CBF	0.74	**1**	0.909
ECG200	0.70	0.76	**0.866**
Fac.F.	0.64	0.79	**0.910**
Lig2	0.65	0.80	**0.911**
Lig.7	0.74	0.79	**0.910**
OSU L.	0.66	0.82	**0.905**

5.3 Discussion

L'utilisation du score FOTS associé à l'algorithme de SU-Shapelet fait qu'il est possible de découvrir d'autres u-shapelets que ceux trouvés par la distance Euclidienne. Le FOTS-SUSh améliore les résultats de la classification des séries temporelles parce que le score FOTS prend en compte les propriétés intrinsèques de la série temporelle et est robuste à la présence d'incertitude. Cette amélioration est particulièrement significative lorsque le score FOTS est utilisé pour la classification non supervisée de séries temporelles contenant plusieurs petites oscillations. En effet, ces oscillations ne sont pas capturées par la distance euclidienne mais par le score FOTS dont le calcul est basé sur la matrice d'autocorrélation. Cette observation est illustrée par le résultat obtenu sur jeu de données SwedishLeaf.

1. Nous considérons 5 jeux de données car se sont les jeux de données pour lesquels nous avons les résultats des algorithmes k-shape et USLM.

Analyse de la complexité en temps. ED peut être calculé en $\mathcal{O}(n)$ et le score FOTS est calculé en $\mathcal{O}(n^{\omega}), \leq \omega \leq \omega \leq\leq 3$ en raison de la complexité temporelle des décompositions des vecteurs propres [20]. Le calcul du score FOTS est alors plus coûteux que celui de ED. Cependant, son utilisation reste pertinente pour la recherche d' u-shapelets, car ils sont souvent de petite taille.

6 Conclusion et perspective

Le but de ce travail était de découvrir des u-shapelets sur des séries temporelles incertaines. Pour répondre à cette question, nous avons proposé un score de dissimilarité (FOTS) adapté à la comparaison de séries temporelles courtes, dont le calcul est basé sur la comparaison du vecteurs propres des matrices d'autocorrélation des séries temporelles. Ce score est robuste à la présence d'incertitude, il n'est pas très sensible aux changements transitoires, et il permet de capturer des relations complexes entre des séries temporelles telles que les oscillations et les tendances. Le score FOTS a été utilisé avec l'algorithme Scalable Unsupervised Shapelet Discovery pour la classification non supervisée de 17 jeux de données de la littérature et a montré une amélioration de la qualité du regroupement évalué à l'aide de l'indice Rand. En combinant les avantages de l'algorithme des u-shapelets, qui réduit les effets néfastes de l'incertitude, et les avantages du score FOTS, qui est robuste à la présence de l'incertitude, ce travail définit un cadre original pour la classification non supervisée de séries temporelles incertaines. Dans la perspective de ce travail, nous prévoyons d'utiliser le score FOTS pour la classification non supervisée floue de séries temporelles incertaines.

Remerciements

Nous remercions cordialement le Ministère Français de l'enseignement supérieur et de la recherche qui a financé ce travail et nous remercions les reviewers pour leurs commentaires et leurs suggestions qui ont aidé à améliorer la qualité de ce travail.

Références

[1] G. B. Folland and A. Sitaram, "The uncertainty principle : a mathematical survey," *Journal of Fourier analysis and applications*, vol. 3, no. 3, pp. 207–238, 1997.

[2] N. B. Rizvandi, J. Taheri, R. Moraveji, and A. Y. Zomaya, "A study on using uncertain time series matching algorithms for MapReduce applications," *Concurrency and Computation : Practice and Experience*, vol. 25, no. 12, pp. 1699–1718, aug 2013.

[3] J. Hwang, Y. Kozawa, T. Amagasa, and H. Kitagawa, "GPU Acceleration of Similarity Search for Uncertain Time Series," in *2014 17th International*

Conference on Network-Based Information Systems. IEEE, sep 2014, pp. 627–632.

[4] K. Rehfeld and J. Kurths, "Similarity estimators for irregular and age-uncertain time series," *Climate of the Past*, vol. 10, no. 1, pp. 107–122, 2014.

[5] M. Orang and N. Shiri, "An experimental evaluation of similarity measures for uncertain time series," in *Proceedings of the 18th International Database Engineering & Applications Symposium on - IDEAS '14*. New York, New York, USA : ACM Press, 2014, pp. 261–264.

[6] W. Wang, G. Liu, and D. Liu, "Chebyshev Similarity Match between Uncertain Time Series," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–13, 2015.

[7] M. Orang and N. Shiri, "Correlation analysis techniques for uncertain time series," *Knowledge and Information Systems*, vol. 50, no. 1, pp. 79–116, jan 2017.

[8] L. Ulanova, N. Begum, and E. Keogh, "Scalable clustering of time series with u-shapelets," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. SIAM, 2015, pp. 900–908.

[9] J. Zakaria, A. Mueen, and E. Keogh, "Clustering time series using unsupervised-shapelets," in *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE, 2012, pp. 785–794.

[10] Q. Zhang, J. Wu, H. Yang, Y. Tian, and C. Zhang, "Unsupervised feature learning from time series." in *IJCAI*, 2016, pp. 2322–2328.

[11] M. Dallachiesa, B. Nushi, K. Mirylenka, and T. Palpanas, "Uncertain time-series similarity : return to the basics," *Proceedings of the VLDB Endowment*, vol. 5, no. 11, pp. 1662–1673, 2012.

[12] M. Orang and N. Shiri, "Improving performance of similarity measures for uncertain time series using preprocessing techniques," in *Proceedings of the 27th International Conference on Scientific and Statistical Database Management - SSDBM '15*. New York, New York, USA : ACM Press, 2015, pp. 1–12.

[13] K. Murthy and S. R. Sarangi, "Generalized notion of similarities between uncertain time series," Mar. 26 2013, uS Patent 8,407,221.

[14] J. Aßfalg, H.-P. Kriegel, P. Kröger, and M. Renz, "Probabilistic similarity search for uncertain time series." in *SSDBM*. Springer, 2009, pp. 435–443.

[15] M.-Y. Yeh, K.-L. Wu, P. S. Yu, and M.-S. Chen, "Proud : a probabilistic approach to processing similarity queries over uncertain data streams," in *Proceedings of the 12th International Conference on Extending Database Technology : Advances in Database Technology*. ACM, 2009, pp. 684–695.

[16] J. Hoffmann-Jørgensen and G. Pisier, "The law of large numbers and the central limit theorem in banach spaces," *The Annals of Probability*, pp. 587–599, 1976.

[17] S. Papadimitriou, J. Sun, and S. Y. Philip, "Local correlation tracking in time series," in *Data Mining, 2006. ICDM'06. Sixth International Conference on*. IEEE, 2006, pp. 456–465.

[18] R. Bhatia and T. Bhattacharyya, "A generalization of the Hoffman-Wielandt theorem," *Linear Algebra and its Applications*, vol. 179, pp. 11–17, jan 1993.

[19] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, "The ucr time series classification archive," July 2015.

[20] V. Y. Pan and Z. Q. Chen, "The complexity of the matrix eigenproblem," in *Proceedings of the thirty-first annual ACM symposium on Theory of computing*. ACM, 1999, pp. 507–516.