TD Data Mining Avancé M2

Université d'Orléans 2021-2022

14 mars 2022

1 Exercice 1 : Bag of SFA Symbols (BOSS)

Soit un ensemble de 9 sous séquences issues d'une série temporelle univariée $X = \{S_0, ..., S_8\}$ sur lesquelles on a appliqué la transformation de Fourier discrète (DFT). Si on considère uniquement le premier coefficient de Fourier, on obtient le tableau 1, où les colonnes C_0 et C_1 désignent la partie réelle et imaginaire du premier coefficient.

1. À partir du tableau 1, décrire une méthode qui permet de découper chaque colonne en c partitions dont les valeurs triées par ordre croissant (ie. $max(c_i) < min(c_{i+1})$, $\forall i \in [0, c[$), contenant approximativement le même nombre d'éléments. La sortie de cette méthode donnera les bornes inférieures et supérieures associées à chaque partition c_i.
2. Calculer avec votre méthode les bornes inférieures et supérieures pour C_0 et C_1 et $c = 3$.
3. À l'aide de la méthode du Multiple Coefficient Binning (MCB) et des bornes calculées précédemment, transformer chaque ligne de la table 1 en mot SFA.
4. En considérant l'ordre d'occurrence décrit par l'index de la table 1, appliquer sur l'ensemble des mots SFA la technique de réduction vue en cours. Quelles lignes devrions-nous conserver ?
5. Quel serait l'histogramme des mots SFA après la réduction ? Compléter la table 2 avec votre résultat.

index X_i	C_0	C_1	mot SFA
0	1.45	0.96	
1	1.30	0.87	
2	1.12	1.23	
3	1.02	0.55	
4	0.85	0.97	
5	0.96	1.09	
6	0.85	1.34	
7	0.51	1.22	
8	0.89	0.58	

TABLE 1 – Coefficients de Fourier obtenus par DFT($\mathcal{X}$).

AA	AB	AC	BA	BB	BC	CA	CB	CC

TABLE 2 – Histogramme des mots SFA résultants de la réduction.

2 Exercice 2 : Shapelet et distance normalisée

On rappelle que la fonction de distance entre une shapelet $S = \{s_0, ..., s_{l-1}\}$ et une série $X = \{x_0, ..., x_{m-1}\}$ est égale à :

$$d(S, X) = v_0, ..., v_{m-l} \ , \ v_i = \sum_{j=0}^{l-1} |X_{i+j} - s_j| \tag{1}$$

Sois une série $X_1 = \{1, 0, 1, 1, 0, 2, 1, 2\}$ et une shapelet $S = \{0, 1, 0.5\}$.

1. Calculer le vecteur de distance $d(S, X_1)$.
2. Calculer le vecteur de distance $d(S, X_1)$ en utilisant la distance z-normalisée [1] vue en cours. On arrondira tous les résultats (déviation standard, moyenne ...) à 2 décimales.
3. Pour chaque vecteur de distance, extraire $min(d(S, X_1))$ et $argmin(d(S, X_1))$.
4. Dans une tâche de classification, dans quel cas choisirions-nous d'utiliser cette distance z-normalisée par rapport à la distance classique ?

Imaginons que vous avez mis en application votre compréhension de l'algorithme de transformation par shapelets et réussi à résoudre une tâche de classification grâce à une seule shapelet utilisant une distance z-normalisée.

Vous souhaitez maintenant valider vos résultats avec une visualisation en traçant la shapelet sur l'emplacement de son minimum dans une série de vos données de test (i.e sur $argmin(d(S, X))$, mais vous n'avez stocké que la valeur z-normalisée de cette shapelet $S = \{-1.25, 1.25, 0\}$! Cela vous donne une visualisation qui ne correspond pas vraiment à la logique de l'algorithme, tel que celle de la figure 1(a).

Comment faire pour obtenir la visualisation "correcte" donnée par la figure 1(b) ?

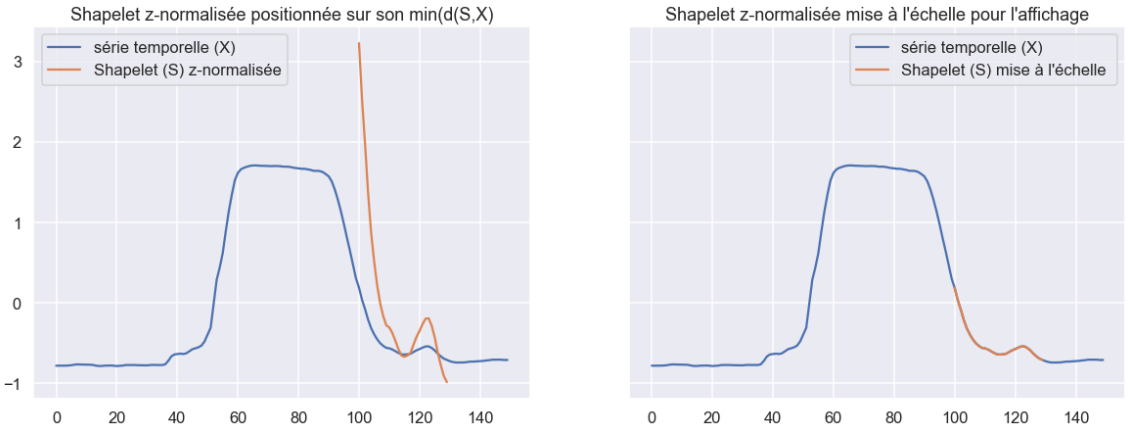

FIGURE 1 – (a) Exemple de shapelet S dont les valeurs sont z-normalisée, positionnés sur $argmin(d(S, X))$ avec d la distance z-normalisée. (b) Shapelets dont les valeurs z-normalisée ont été mises à l'échelle par rapport à la sous séquence associée dans X.

3 Exercice 3 : Le cas des séries multivariées

On définit une série multivariée comme une série qui suit l'évolution de plusieurs variables au fil du temps. Plus formellement, nous avions jusqu'à présent des séries univariées tel que $X = \{x_0, ..., x_m - 1\}$, dans un contexte multivarié avec n variable, cette définition devient $X = \{(x_{0,0}, ...x_{n-1,0}), (x_{0,1}, ..., x_{n-1,1}), ..., (x_{0,m-1}, ..., x_{n-1,m-1})\}$ avec $x_{i,j}$ la valeur de la i^{eme} variable au j^{eme} pas de temps.

1. Proposer une ou plusieurs approches pour faire en sorte d'adapter l'algorithme des shapelets au cas multivariée.
2. Formaliser la nouvelle fonction de distance que vous utilisez dans ces approches.

1. z-normalisation d'un vecteur S : $(S - mean(S))/std(S)$