Bref etat de l'art sur la classification de séries temporelles

Sem EGLOH LOKOH

March 4, 2023

1 Approches basées l'extaction de features

1.1 Shapelet-based algorithms

Une shapelet est définie comme une sous-séquence d'observations consécutives d'une série temporelle. L'idée ici est d'extraire toutes les sous-séries possibles (de même logneur l) de chaque série temporelle puis de selectionner un nombre k des meilleurs sous-séries. Les meilleurs shapelet sont selectionnées en focntion de leur capacité à discriminer les différentes classes de séries temporelles en se basant sur un critère comme distance de Mahalanobis ou la distance de la corrélation. L'objectif étant de caractériser chaque classe avec un set de shapelet (paternes fréquents dans une classe et rare dans une autre classe).

On compare chaque série temporelle à chaque shapelet extrait précédemment, et on attribue une "distance" entre la série temporelle et chaque shapelet. Ensuite, un classificateur (par exemple un classificateur basé sur les k plus proches voisins) est utilisé pour classer la série temporelle en fonction de ces distances.

Articles :

- Michael Franklin Mbouopda, Engelbert Mephu Nguifo. Classification des Séries Temporelles Incertaines Par Transformation "Shapelet". Conférence Nationale en Intelligence Artificielle (CNIA), Jun 2020, Angers, France. pp.14-21. ffhal-03099395f
- A Shapelet Transform for Multivariate Time Series Classification

Implémentation sur github :

- https://github.com/DataPop/RandomShapeletClassifier
- https://github.com/rong-hash/AQOURSNet
- https://github.com/jorwatkapola/shapelet_classification

1.2 Tree-based algorithms

1.2.1 Time series forest

L'idée toute simple est d'extraire trois (03) features (moyenne, écartype et pente) de chaque sous-séries aléatoirment selectionnées des séries temporelles et d'en constituer un ensemble de features (3* le nombre de sous-séries sélectionnées sur le même intervalle). L'ensemble de feature est alonrs utiliseé dans les arbres pour de la classification.

Articles :

- A time series forest for classification and feature extraction, Houtao Deng , George Runger, Eugene Tuv, Martyanov Vladimir

1.2.2 Time series bag-of-features

Un peu plus avancée que le Time series forest, cette approche subdivise chaque sous-série en sous intervalle non chevauchant et en extrait les mêmes caractéritiques. On ontbient alors pour chaque sous série 4*3 features et ainsi pour tout le dataset 12* le nombre de sous-séries sélectionnées.

Articles :

- Baydogan MG, Runger G, Tuv E (2013) A Bag-ofFeatures Framework to Classify Time Series. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11):2796–2802

1.2.3 Proximity forest

Il s'agit d'une approche qui prend directement en entrée les séries temporelles et utilise une métrique de distance (adaptée pour les séries temporelles) comme critère de séparation lors de la construction des arbres.

1.2.4 Entropie de l'ensemble

L'entropie de l'ensemble est une mesure de la complexité d'une série temporelle qui prend en compte la répartition des valeurs dans la série temporelle. L'approche consiste à calculer l'entropie pour chaque série temporelle, puis à utiliser ces valeurs pour entraîner un arbre de décision qui est utilisé pour classer les nouvelles séries temporelles. On peut éventuellememnt aussi utiliser des statistiques dérivées de l'entropie.

1.3 Ensemble de caractéristiques qui peuvent être extraites des séries temporelles pour les utiliser dans un modèle de classification

- **Moyenne :** La valeur moyenne de la série temporelle peut fournir des informations sur la tendance générale de la série.
- **Écart type :** L'écart type de la série temporelle peut indiquer la variabilité ou la volatilité de la série.
- **Moyenne mobile :** La moyenne mobile d'une série temporelle est calculée en prenant la moyenne des n dernières observations. Elle peut aider à lisser les données et à identifier les tendances à court terme.
- **Variance :** La variance de la série temporelle peut fournir des informations sur la dispersion des données autour de la moyenne.
- **Coefficient de corrélation :** Le coefficient de corrélation entre deux séries temporelles peut indiquer si elles évoluent de manière similaire ou opposée.
- **Transformée de Fourier :** La transformée de Fourier peut être utilisée pour extraire les fréquences principales d'une série temporelle. Elle peut aider à identifier les cycles à différentes échelles de temps.
- **Coefficients d'autocorrélation :** Les coefficients d'autocorrélation peuvent être utilisés pour mesurer la corrélation entre des observations éloignées dans la série temporelle. Ils peuvent aider à identifier les tendances à long terme ou les saisons dans les données.
- **Les caractéristiques basées sur l'entropie :** Ces caractéristiques peuvent fournir des informations sur la complexité et la régularité de la série temporelle.
- **Séries chronologiques :** Les séries chronologiques sont les séries temporelles avec une unité de temps liées à la date, comme les heures, les jours, les mois, les années, etc.
- **Les caractéristiques basées sur les sous-séries :** Ces caractéristiques incluent les motifs fréquents, les motifs rares et les shapelets. Ces caractéristiques peuvent fournir des informations sur les caractéristiques distinctives et les structures de la série temporelle.
- **Les features extraits à l'aide des réseaux de neurones**
- **Trend strength** La force de la tendance est une mesure statistique qui quantifie le degré de tendance ou de mouvement directionnel dans une série chronologique. Une mesure de la force de la tendance est le coefficient de détermination (R-carré), qui est une mesure de la proportion de la variabilité de la série chronologique qui est expliquée par le modèle de régression linéaire. Des valeurs de R-carré plus élevées indiquent une relation linéaire plus forte et donc une tendance plus forte dans la série chronologique.
- **seasonality strength** La force de la saisonnalité est une mesure statistique qui quantifie le degré de périodicité d'une série chronologique. Une mesure couramment utilisée est l'indice saisonnier, qui représente l'écart moyen par rapport à la tendance générale pour une période donnée. L'intensité de la saisonnalité est ensuite calculée comme la variance des indices saisonniers. Des variances plus importantes indiquent une saisonnalité plus forte, tandis que des variances plus faibles indiquent une saisonnalité plus faible.

- **spikiness** La spikiness d'une série temporelle mesure la variabilité ou l'irrégularité des résidus de la série après suppression de toute tendance et saisonnalité. Elle est définie comme la variance des variances de la composante résiduelle. Pour la calculer, on divise la série en sous séquences non chevauchantes de longueur égale, et la composante résiduelle est calculée pour chaque morceau. La variance de la composante résiduelle est ensuite calculée en supprimant un morceau à la fois et en recalculant la composante résiduelle pour les données restantes. La variance de ces variances de la composante résiduelle est le caractère piquant de la série temporelle.
- **peak** Il s'agit de l'indice du point de pic de la série
- **trough** Il s'agit de l'indice du point de creux de la série
- **level shift idx** Indice de la différence maximale de la valeur moyenne, entre deux fenêtres coulissantes consécutives
- **level shift size** Taille de la différence maximale de la valeur moyenne, entre deux fenêtres coulissantes consécutives
- **y acf1** Première valeur de l'autocorrelation ACF de la série
- **y acf5** Somme des carrés des 5 premières valeurs l'autocorrelation ACF de la série
- **diff1y acf1** Première valeur de l'autocorrelation ACF de la série différentiée d'ordre 1
- **diff1y acf5** Somme des carrés des 5 premières valeurs l'autocorrelation ACF de la série différentiée d'ordre 1
- **diff2y acf1** Première valeur de l'autocorrelation ACF de la série différentiée d'ordre 2
- **diff2y acf5** Somme des carrés des 5 premières valeurs l'autocorrelation ACF de la série différentiée d'ordre 2
- **y pacf5** Somme des carrés des 5 premières valeurs l'autocorrelation partielle PACF de la série
- **diff1y pacf5** Somme des carrés des 5 premières valeurs l'autocorrelation partielle PACF de la série différentiée d'ordre 1
- **diff2y pacf5** Somme des carrés des 5 premières valeurs l'autocorrelation partielle PACF de la série différentiée d'ordre 2
- **seas acf1** Coefficient d'autocorrélation au premier retard saisonnier
- **seas pacf1** Coefficient d'autocorrélation partiel au premier retard saisonnier
- **firstmin ac** Indice du premier minimum des valeurs d'autocorelation. Cet indice est utilisé pour déterminer l'ordre optimal du modèle MA en identifiant le nombre de termes d'erreur nécessaires pour modéliser la série chronologique.
- **firstzero ac** Indice du premier passage à zéro des valeurs d'autocorelation. Cette valeur indique le nombre de décalages à partir desquels la série chronologique n'a plus de corrélation significative.
- **holt alpha** Le paramètre de niveau du modèle de lissage exponentiel linéaire de Holt sans saisonalité.
- **holt beta** Le paramètre tendance du modèle de lissage exponentiel linéaire de Holt sans saisonalité.
- **hw alpha** Le paramètre de niveau du modèle de lissage exponentiel linéaire de Holt winters avec saisonalité.
- **hw beta** Le paramètre tendance du modèle de lissage exponentiel linéaire de Holt winters avec saisonalité.
- **hw gamma** Le paramètre saisonnier du modèle de lissage exponentiel linéaire de Holt winters avec saisonalité.
- **length** Taille de la série

- **entropy** L'entropie mesure le degré d'aléatoire ou de désordre. L'entropie d'une série temporelle mesure la quantité d'aléatoire dans les données. Il s'agit d'une valeur scalaire qui fournit une mesure du caractère prévisible ou imprévisible des données. La valeur de l'entropie de Shannon normalisée est comprise entre 0 et 1, les valeurs les plus élevées indiquant une plus grande complexité ou imprévisibilité de la série chronologique. Une valeur de 0 indique que la série chronologique est complètement prévisible ou ordonnée, tandis qu'une valeur de 1 indique une imprévisibilité ou un caractère aléatoire maximal.

- **lumpiness** L'homogénéité est une propriété statistique d'une série temporelle qui mesure à quel point la variance de la série varie dans le temps. Plus précisément, l'homogénéité est définie comme la variance des variances par sous séquences non chevauchant d'une série temporelle.

- **stability** La stabilité est une propriété statistique d'une série temporelle qui mesure à quel point la moyenne de la série varie dans le temps. Plus précisément, la stabilité est définie comme la variance des moyennes par morceaux d'une série temporelle.

- **flat spots** Les Flat spots font référence à des segments d'une série chronologique dans lesquels les valeurs restent approximativement constantes.

- **hurst** L'exposant de Hurst est utilisé comme une mesure de la mémoire à long terme des séries temporelles.

- **std1st der** Ecart-type de la dérivée première de la série temporelle en utilisant la fonction gadient de numpy. Le calcul de l'écart type de la dérivée première d'une série temporelle peut donner un aperçu de la variabilité du taux de variation de la série temporelle dans le temps. La dérivée première est une mesure du taux de changement de la série temporelle à chaque point de temps, et peut être utilisée pour analyser la rapidité ou la lenteur avec laquelle la série temporelle évolue dans le temps.

- **crossing points** Le nombre de points de croisement. Les points de croisement se produisent lorsqu'une série temporelle croise la ligne médiane. Les corring pointpeuvent idiquer un changement dans la tendance sous-jacente ou un changement saisonnier.

- **binarize mean** Valeur moyenne de la série binarisée (1 si la valeurs de la série temporelle supérieure à sa moyenne, et 0 si elle est inférieure).

- **unitroot kpss** La statistique du test de racine unitaire de KPSS. Test résalisé avec l'hypothèse nulle selon laquelle la série chronologique observable est stationnaire autour d'une tendance déterministe.

- **heterogeneity** Statistique de test du multiplicateur de Lagrange, basé sur le test d'Engle pour l'hétéroscédasticité conditionnelle autogressive (ARCH).

- **histogram mode** Mesure la valeur modale de la série à l'aide d'histogramme. Elle peut être utile pour identifier la centralité de la série.

- **linearity** Caractéristique de linéarité : R^2 d'une régression linéaire ajustée de la série

- **CUSUM Features :** CUSUM est l'abréviation de cumulative sum, c'est un algorithme de détection des points de changement dans kats.

  Dans l'implémentation de Kats, il a deux composantes principales :

  1. Localiser le point de changement : L'algorithme estime itérativement les moyennes avant et après le point de changement et trouve le point de changemen en maximisant/minimisant la valeur du cusum jusqu'à ce que le point de changement ait convergé.

  2. Test d'hypothèse : Réalisation d'un test du rapport de vraisemblance logarithmique où l'hypothèse nulle n'a pas de point de changement avec une moyenne et l'hypothèse alternative a un point de changement avec deux moyennes.

  Et voici quelques éléments qui méritent d'être mentionnés : * Il n'y a qu'un seul point de changement d'augmentation/diminution ; * La distribution gaussienne comme modèle sous-jacent pour calculer la valeur du cusum et effectuer le test d'hypothèse.

  - **cusum num** Nombre de points de changement

  - **cusum conf** Confiance du point de changement détecté, 0 si pas point de changement
  - **cusum cp index** Position du point de changement détecté dans la série temporelle
  - **cusum delta** Delta des niveaux moyens avant et après le point de changement
  - **cusum llr** La log-vraisemblance du point de changement
  - **cusum regression detected** Vrai ou faux - si la régression est détectée par l'algorithme CUSUM
  - **cusum stable changepoint** Vrai ou faux - si le point de changement est stable
  - **cusum p value** P-value du point de changement

- **robust num** Nombre de points de changement détectés par le Robust Stat de Kats
- **robust metric mean** Moyenne des valeurs métriques du détecteur robuste de statistiques
- **bocp num** Nombre de points de changement détectés par le BOCP detector de Kats
- **bocp conf max** Valeur maximale de l'intervalle de confiance des points de changement détectés
- **bocp conf mean** Valeur moyenne de l'intervalle de confiance des points de changement détectés
- **outlier num** Nombre de valeurs aberrantes identifiées
- **trend num** Nombre de tendances détectées par le détecteur de tendances de Kats
- **trend num increasing** Nombre de tendances haussière
- **trend avg abs tau**
- **nowcast roc** Taux de changement moyen de la série
- **nowcast ma**
- **nowcast mom**
- **nowcast lag**
- **nowcast macd**
- **nowcast macdsign**
- **nowcast macddiff**
- **seasonal period** Période de saisonnalité détectée
- **trend mag** Magnitude de la tendance qui est équivaut à la pente du modèle de régression linéaire simple sur la composante tendance
- **residual std** Écart-type de la composante résiduelle de la série
- **seasonality mag** Magnitude de la saisonnalité calculée par différence entre le percentile 95 et le percentile 5 de la composante saisonnière.

2 Approches basées la similarité

2.1 KNN classification avec la DTW (distortion temporelle dynamique)

Proposée par Sakoe et Chiba en 1978 le DTW est une méthode de déformation non linéaire de séries temporelles pour déterminer leur similarité. La DTW consiste à trouver le chemin optimal qui minimise la somme des distances entre les éléments correspondants de deux séries. La DTW est la somme des distances du chemin optimal entre deux séries. NB:

Implémentation sur github :

- https://github.com/markdregan/K-Nearest-Neighbors-with-Dynamic-Time-Warping
- https://github.com/alexminnaar/time-series-classification-and-clustering
- Time series classification with ensembles of elastic distance measures, Jason Lines, Anthony Bagnall

2.2 Ensemble de mesure de similarite et de distance de séries temporelles

- **Plus longue sous-séquence :** La métrique de la plus longue sous-séquence commune (LCS) est une mesure de similarité entre deux séries temporelles. Elle est définie comme la longueur de la plus longue sous-séquence commune de deux séries temporelles données. Une sous-séquence est une séquence qui peut être obtenue à partir d'une autre séquence en supprimant certains éléments sans changer l'ordre des éléments restants.

- **DTW :** Dynamic Time Warping est une mesure de similarité entre deux séries temporelles qui permet de comparer des séries de longueurs différentes, même lorsqu'elles ont des motifs (shapes) différents. DTW prend en compte les différences de vitesse ou de phase entre les séries, ce qui la rend plus souple que les mesures de similarité classiques telles que la distance euclidienne.

- **Edit distance :** La distance d'édition est une mesure de la similitude entre deux séries temporelles qui est utilisée pour mesurer la distance entre deux chaînes de caractères. Elle mesure le nombre minimum d'opérations d'édition nécessaires pour transformer une série temporelle en une autre. Ces opérations d'édition peuvent être l'insertion, la suppression ou la substitution d'éléments.

- **Distance euclidenne :** La distance euclidienne est une mesure de similarité entre deux séries temporelles. Elle mesure la distance géométrique entre deux points dans un espace à n dimensions. Pour deux séries temporelles, cela signifie qu'elle mesure la distance entre deux points dans un espace à deux dimensions, où chaque dimension représente la valeur de la série temporelle à un moment donné.

- **Correlation distance :** La distance de corrélation est calculée comme la différence entre 1 et le coefficient de corrélation de Pearson entre les deux séries temporelles. Elle mesure la similarité entre deux séries temporelles basée sur leur corrélation. Cette métrique est utilisée pour évaluer à quel point deux séries temporelles sont corrélées.

- **Cumulative differences distance :** La métrique de distance cumulée des différences est une mesure de similarité entre deux séries temporelles. Elle consiste à calculer la somme des différences absolues entre les valeurs correspondantes de chaque série à chaque instant, et à cumuler ces différences sur l'ensemble de la série.

- **Dynamique des fractale :** Cette métrique est basée sur l'analyse de la complexité et de la structure auto-similaire des séries temporelles. Elle peut être utile pour étudier les propriétés fractales des signaux sociaux.

- **Densité spectracle :** Elle permet d'analyser les variations de fréquence des signaux temporels. Elle peut être utile pour identifier les composantes périodiques des signaux sociaux, par exemple pour étudier les modèles de saisonnalité.

3 Approche Réseaux de neurones

3.1 LSTM avec attention)

Articles et Implémentation sur github :

- https://github.com/anaramirli/human-activity-recognition
- LSTMs for Human Activity Recognition Time Series Classification
- LSTM-MFCN: A time series classifier based on multi-scale spatial–temporal features
- Early Time-Series Classification Algorithms: An Empirical Comparison, Charilaos Akasiadis, Evgenios Kladis, Evangelos Michelioudakis, Elias Alevizos, and Alexander Artikis
- Deep learning for time series classification
- https://github.com/hfawaz/dl-4-tsc

3.2 Réseaux convolutifs (CNN)

InceptionTime (2020) est probablement la principale architecture d'apprentissage profond pour la classification des séries temporelles. InceptionTime est un ensemble de réseaux neuronaux composé de cinq réseaux Inception.

4 Usefull links

- Time Series Classification: A review of Algorithms and Implementations, Johann Faouzi
- Deep learning for time series classification: a review., Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, Pierre-Alain Muller. Data Mining and Knowledge Discovery, 2019, 33 (4), pp.917-963. ff10.1007/s10618-019-00619-1ff. ffhal-02365025v2
- `http://www.timeseriesclassification.com`
- `https://github.com/facebookresearch/Kats`

5 Ensemble de paramètres qui peuvent caractériser un changement temporel dans une série temporelle

- **La fréquence :** La fréquence d'une série temporelle est la mesure de la rapidité des variations dans la série temporelle. Les séries temporelles avec des fréquences élevées ont des variations plus rapides que les séries temporelles avec des fréquences plus faibles.
- **L'amplitude :**L'amplitude d'une série temporelle est la mesure de l'écart entre les valeurs maximales et minimales dans la série temporelle. Les séries temporelles avec des amplitudes élevées ont des variations plus importantes que les séries temporelles avec des amplitudes plus faibles.
- **La moyenne :** La moyenne d'une série temporelle est la valeur moyenne des valeurs de la série temporelle. Les changements de la moyenne d'une série temporelle peuvent indiquer un changement dans la tendance globale de la série temporelle.
- **La variance :** La variance d'une série temporelle est une mesure de la dispersion des valeurs de la série temporelle autour de la moyenne. Les changements de variance d'une série temporelle peuvent indiquer un changement dans la régularité de la série temporelle.
- **Les raies spectrales :** Les raies spectrales d'une série temporelle sont les fréquences dominantes dans la série temporelle. Les changements dans les raies spectrales peuvent indiquer un changement dans les fréquences dominantes de la série temporelle.

Sources:

- A review and experimental evaluation of recent advances in time series classification