Transcription de séries temporelles en séquences temporelles via conservation des caractéristiques de variation

Guillaume Savarit, Karell Bertet, Christophe Demko

Transcription de séries temporelles en séquences temporelles via conservation des caractéristiques de variation

Guillaume Savarit, Karell Bertet, and Christophe Demko

La Rochelle Université, La Rochelle, France

Abstract. Dans ce papier nous nous intéressons à la possibilité de transcrire les séries temporelles en séquences temporelles en conservant des caractéristiques issue de la série, de manière à réduire la volumétrie des données et profiter de l'analyse multiséquentielle et hétérogène apportée par des outils déjà existant de l'analyse de séquence. Nous utilisons en exemple d'application les données issues des capteurs de niveau de la marée du Port Atlantique de La Rochelle.

Keywords: Série temporelle · Séquence temporelle · Analyse de données

1 Introduction

Toutes données collectées avec un capteur au cours du temps vont former une série temporelle, sous la forme d'une association entre une valeur x_i et une unité de temps t_i.

$$y =< (x_i, t_i) > \tag{1}$$

Bien que de fait extrêmement commune, l'analyse des séries temporelles brutes soulève de nombreuses problématiques lors de l'analyse de celle-ci. Leurs tailles dépendant de la fréquence et de la durée de captation, il s'agit de données pouvant très vite atteindre un volume conséquent.

Pour réduire ce volume, il est possible de résumer ces séries en identifiant leurs caractéristiques. Selon l'objet de l'analyse, il est intéressant de détecter les outliers, les valeurs aberrantes que prendraient la série ponctuellement, ou de détecter les périodes de celle-ci. Pour cela, des algorithmes de changements de points ou de détection d'erreurs sont utilisés. Mais dans ce cas, le temps devient une indication des instants où les événements analysés ont eu lieu.

Un autre objet d'analyse des séries consiste à faire de la prédiction de nouvelles valeurs en fonction de celles captées. Dans ce cas, les méthodes de régression, linéaire ou statistique, sont utilisées. Il est ainsi possible d'approcher la série « idéale », un modèle représentant nos valeurs. Mais dans ce cas, les informations portées par les valeurs ne sont pas expliquées.

Pour tenter malgré tout d'expliquer les séries tout en conservant la notion temporelle, et en réduisant le volume des données, nous allons proposer ici une approche des séries temporelles sous forme de séquences conservant les caractéristiques de la série et pouvant permettre une analyse multiséquentielle hétérogène via la plateforme Galactic [3].

Nous allons tout d'abord définir les séries temporelles et les séquences, puis expliquer la chaine de traitement généralisé permettant de construire une séquence à partir d'une série, et enfin mettre en perspective les futurs traitements.

2 État de l'art

2.1 Séries temporelles

La notion de série temporelle est assez ancienne, réellement apparue dans la littérature dans les années 20. L'une des premières définitions est celle de Warren M. Persons en statistique [9] qui définit une série temporelle comme « un nombre agrégé ou moyenné ou relatif appliqué à un interval défini ou un temps défini », et donne trois conditions : même unité de temps (par exemple le jour, l'année, le mois), consécutifs dans le temps, construit par un critère fixe ou un standard. Cette définition a pour objectif de rapprocher une série temporelle d'une fonction en plus restreinte.

C'est en économie que Harold T. Davis [5] va poser une définition plus générale, qui est plus proche de celle actuellement utilisée : « une série de données observées successivement dans le temps ». Il exprime mathématiquement celles-ci comme une séquence tel que :

$$y = y_t \text{ avec } t = 1, 2, 3, \ldots, n \tag{2}$$

Contrairement à la définition précédente qui considérait les séries temporelles comme des fonctions, la définition de Davis implique la discontinuité des séries temporelles. Néanmoins il considère que pour des intervalle de temps suffisamment court nous pouvons considérer que la variable t est continue et écrire :

$$y = f(t) \tag{3}$$

Davis décrit aussi la différence entre les séries temporelles utilisées en astronomie et en économie. Il introduit ainsi différents types de séries temporelles, celles reposant sur des variables peu nombreuses et connues qu'il est possible de mettre en équation (astronomie) et celles reposant sur de multiples variables et des tendances (économie).

Dans les années 80 va émerger une précision sur la définition des séries temporelles. Considérant que toutes les séries reposent sur des fonctions π, il est possible d'écrire :

$$y_t = \sum_{k=1}^{\infty} \pi_k y_{t-k} + a_t \tag{4}$$

Où k est le lag et a le bruit.

Il existe plusieurs classification des séries temporelles. Comme de nombreux autres types de données on distingue les séries univariées (reposant sur une seule variable) et multivariées (reposant sur plusieurs variables). Il existe aussi une classification temporelle, entre les séries à haute fréquence (petit interval de temps) et basse fréquence (grand interval de temps). Cette distinction n'entre en compte que dans la comparaison de plusieurs séries temporelles.

L'analyse de séries temporelles se fait majoritairement en utilisant des modèles. Ainsi, dans leur article de 1972, qui sera ensuite utilisé pour poser les bases du modèle ARIMA (*AutoRegressive Integrated Moving Average*, un modèle statistique utilisant les valeurs précédentes pour prédire les valeurs suivantes), Amemiya et Wu [2] expliquent qu'un modèle d'agrégation fonctionnant à haute fréquence fonctionnera aussi à basse fréquence. Brewer en 1973 [4] va aussi soulever le problème de la fréquence, et le résoudre en généralisant le modèle.

En 1987, Lüktepohl [7] montre que les modèles ARIMA entres autres ne fonctionnent que sur des données univariées. Lorsqu'appliqué sur des données multivariées, la question de la fréquence redevient importante, notamment dû à l'hétéroscédasticité. En effet un modèle régressif ne fonctionne que si la variance de l'erreur des variables est faible (homoscédasticité). Dans le cas inverse, les erreurs de prédictions du modèle seront grandes.

2.2 Séquences temporelles

Les séquences sont un type de données très utilisé pour representer notamment les déplacements, les phrases et autres. Une séquence temporelle A est définie par un ensemble ordonné d'événements. Chaque événement associe un symbole s_i à un interval non nul $[\underline{t_i}, \overline{t_i}]$ où le symbole peut être une valeur symbolique issue d'un dictionnaire, ou une valeur numérique.

$$A =< (s_i, [\underline{t_i}, \overline{t_i}]) > \tag{5}$$

La fouille de motifs sequentiels repose originellement sur des algorithmes comme GSP d'Agrewal et Srikant [1], qui travaille sur les séquences sans notion d'intervalle. Guyet et Quiniou en 2008 [6] avec *QTempIntMiner* et Nakagaito en 2009 [8] avec *QTPSpan* s'intéressent à la fouille de séquences d'intervalles.

En 2020, Boukhetta [3] utilise l'analyse formelle de concept via le framework Galactic pour analyser les séquences temporelles. Cette approche permet aussi l'hétérogénéité des données analysées ainsi qu'une analyse multiséquence.

3 Chaîne de traitement

3.1 Général

Notre méthode de traitement consiste à transformer les séries temporelles en séquences temporelles en utilisant la composante sémantique pour décomposer les séries en différents épisodes de valeurs consecutives de même sémantique. Les séquences ainsi obtenues résument les séries temporelles tout en réduisant leur volumétrie, et il est possible d'utiliser des outils existants d'analyse de séquences.

L'analyste choisit une sémantique applicable à la série temporelle en construisant un dictionnaire représentant cette sémantique. Dans la suite nous nous intéresserons à un cas général, celui de l'analyse par la variation de la courbe, en étudiant le signe de la dérivée première de celle-ci. Le dictionnaire sera ainsi composé de 3 éléments au moins :

$$\Sigma = \{Croissant, Decroissant, Stable\} \tag{6}$$

Toujours dans le cas général, la stabilité peut être scindée en deux types. Soit le pic d'une courbe, entre deux intervalles croissant puis décroissant, soit la vallée, entre deux intervalles décroissant puis croissant.

Un algorithme dédié de changements de points, qui découvre les points d'inflexions de la série, sépare ensuite les valeurs de la série en différents épisodes consécutifs afin d'obtenir une séquence temporelle.

Enfin, les séquences sont annotées grâce au dictionnaire choisi. Nous obtenons ainsi à partir d'une série temporelle donnée la transformation suivante :

$$< (x_i, t_i) > \rightarrow < (s_j, [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{C, D, S\} \tag{7}$$

La chaîne de traitement est donc celle-ci ainsi :

Choix de la sémantique L'analyste choisit un dictionnaire qui représente la sémantique de ses données.
Découpage en séquence La série est transformée en séquence respectant le dictionnaire.
Annotation Le dictionnaire est appliqué à la séquence.

3.2 Application

Dans la suite nous allons montrer l'application sur un cas concret. Il s'agit d'une série temporelle issue des données de capteurs de niveaux de la marée dans le bassin à flot du Port Atlantique de La Rochelle. Ces capteurs sont placés autour du bassin à flot et retransmettent les niveaux d'eau de l'océan, du sas et du bassin à flot au cours du temps. Chaque mesure est prise toutes les 60 secondes en moyenne, et réprésente les données de mars 2021 à mars 2022. Cela représente quelques 981 070 mesures.

Nous nous intéressons en particulier aux mesures du coté de l'océan et à la question de prédiction des marées, nous allons utiliser une sémantique représentant ce que nous connaissons de la marée avec un dictionnaire à 4 éléments :

$$\Sigma = \{Montante, Descendante, Haute, Basse\} \tag{8}$$

Par rapport à notre dictionnaire général, les notions de marée *Montante* et *Descendante* sont équivalentes à la croissance et la décroissance. Les marées *Haute* et *Basse* quant à elles sont une séparation de la stabilité selon les symboles précédent et suivant de la séquence. Une marée stable après une marée montante et suivie par une marée descendante sera une marée haute.

Une fois la série traduite en séquence, nous obtenons :

$$< (x_i, t_i) > \rightarrow < (s_j, [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{M, D, H, B\} \tag{9}$$

Une fois le traitement terminé, nous obtenons 5000 séquences pour l'année. Ceci réduit drastiquement le volume de données pour l'analyse tout en résumant l'information portée par la série temporelle.

4 Perspective

Pour contrebalancer la perte d'information induite par la transformation en séquence, il est possible par la suite d'extraire d'autres caractéristiques internes aux séries, tel que les extremums, les moyennes, voire les valeurs elle-même. En tirant partie de la possibilité d'analyser des séquences avec des données hétérogènes offerte par GALACTIC, il est ainsi possible d'analyser les séquences avec plus d'information.

Il serait ainsi possible d'analyser la séquence suivante, prenant en compte les extremums :

$$< (x_i, t_i) > \rightarrow < (s_j, [x_{min}, x_{max}], [\underline{t_j}, \overline{t_j}]) > \text{avec} \quad s_j \in \{C, D, S\} \tag{10}$$

Ceci afin de conserver des informations des séries temporelles tout en réduisant le volume des données en tirant partie de l'analyse multiséquentielle.

En utilisant une telle approche, nous allons dans le futur mettre en place la possibilité d'expliquer les séries temporelles via les outils de l'analyse formel de concept, notamment GALACTIC, en conservant les caractéristiques de la série sous forme de séquence, réduisant le volume de donnée et permettant le traitement de celle-ci.

References

1. R. Agrawal and R. Srikant. Mining sequential patterns. In *Proceedings of the Eleventh International Conference on Data Engineering*, pages 3–14, March 1995.
2. T. Amemiya and R. Y. Wu. The Effect of Aggregation on Prediction in the Autoregressive Model. *Journal of the American Statistical Association*, 67(339):628–632, September 1972. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1972.10481264.
3. S. E. Boukhetta, C. Demko, K. Bertet, J. Richard, and C. Cayèré. Temporal Sequence Mining Using FCA and GALACTIC. In Tanya Braun, Marcel Gehrke, Tom Hanika, and Nathalie Hernandez, editors, *Graph-Based Representation and Reasoning*, Lecture Notes in Computer Science, pages 185–199, Cham, 2021. Springer International Publishing.
4. K. R. W. Brewer. Some consequences of temporal aggregation and systematic sampling for ARMA and ARMAX models. *Journal of Econometrics*, 1(2):133–154, June 1973.
5. H. T. Davis. *Analysis of economic time series. Principia Press, Bloomington, 1941.* 1941.
6. T. Guyet and R. Quiniou. Mining Temporal Patterns with Quantitative Intervals. In *2008 IEEE International Conference on Data Mining Workshops*, pages 218–227, December 2008. ISSN: 2375-9259.
7. H. Lütkepohl. Forecasting Aggregated Vector ARMA Processes, 1987.
8. F. Nakagaito, T. Ozaki, and T. Ohkawa. Discovery of Quantitative Sequential Patterns from Event Sequences. In *2009 IEEE International Conference on Data Mining Workshops*, pages 31–36, December 2009. ISSN: 2375-9259.
9. W. M. Persons. Correlation of Time Series. *Journal of the American Statistical Association*, 18(142):713–726, June 1923. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1923.10502103.