08 JAN

14h00 à 16h00
Espace Gilbert Simondon, 1B36,
ENS Paris-Saclay

THÈSES ET HDR

Sylvain COMBETTES : soutenance de thèse

Titre : Représentations symboliques de séries temporelles
Direction : L. Oudre, C. Truong
Soutenance le 08/01/2024 à 14h00 en 1B36

AJOUTER AU CALENDRIER

Sylvain COMBETTES

Représentations symboliques de séries temporelles

Résumé de la thèse

Ce travail porte sur la représentation et la comparaison de signaux physiologiques, qu'ils soient univariés ou multivariés. Dans de nombreuses applications, par exemple en neurologie comportementale, les chercheurs doivent interpréter et comparer de manière interactive une grande quantité de signaux physiologiques multivariés. Les objectifs de cette thèse sont de définir de nouvelles représentations et distances qui peuvent traiter des signaux physiologiques avec une structure complexe : multivariés, non-stationnaires et multimodaux. De plus, cette représentation doit conserver l'information temporelle tout en étant interprétable et rapide à calculer.
Après avoir passé en revue les différentes techniques de symbolisation (qui transforment une série de valeurs réelles en une série discrète plus courte) et avoir étudié les distances sur les séries temporelles, les chaînes de caractères et sur les séquences symboliques, nous introduisons de nouvelles représentations symboliques en proposant une distance entre les séquences symboliques obtenues.
La première contribution est une représentation symbolique pour un ensemble de séries temporelles univariées appelée ASTRIDE. Contrairement à la plupart des procédures de symbolisation, ASTRIDE est adaptative à la fois à l'étape de segmentation en effectuant une détection de ruptures et à l'étape de quantification en utilisant des quantiles. Au lieu de transformer signal après signal, ASTRIDE construit un dictionnaire de symboles commun à tous les signaux d'un ensemble de données. Nous introduisons également une nouvelle distance sur les représentations symboliques qui basée sur la distance d'édition générale avec des poids personnalisés. Nous montrons les performances d'ASTRIDE par rapport à 4 autres représentations symboliques en termes de reconstruction et, le cas échéant, sur des tâches de classification.
La deuxième contribution est une représentation symbolique pour un ensemble de séries temporelles multivariées pouvant être non stationnaires, appelée d_symb, ainsi qu'un outil d'exploration en ligne appelé d_symb playground. Contrairement à la plupart des distances sur les signaux multivariés, d_symb prend en compte de leur non-stationnarité grâce à une étape de symbolisation. Cette étape est basée sur une procédure de détection de ruptures qui divise un signal non stationnaire en plusieurs segments stationnaires, suivie d'une quantification à l'aide d'un partitionnement en K-moyennes. La distance proposée est basée sur la distance d'édition générale. Les avantages de d_symb sont montrés sur trois jeux de données de signaux physiologiques. Les expériences montrent à quel point la symbolisation est interprétable : un simple coup d'œil aux séquences symboliques fournit une compréhension immédiate et complète d'un ensemble de données. De plus, par rapport à 9 distances multivariées dites élastiques dans une tâche de regroupement, d_symb atteint des performances compétitives tout en étant plusieurs ordres de grandeur plus rapide que les autres méthodes. Avec ces caractéristiques souhaitables, nous avons développé d_symb playground, un outil en ligne qui permet aux chercheurs de téléverser leurs données et d'y appliquer d_symb.

Mots clés

Reconnaissance de formes, Approches symboliques, Détection de ruptures, Apprentissage de représentation

Symbolic representations of time series

Abstract of the thesis

This work addresses the problem of representing and comparing physiological signals that can be univariate or multivariate. In many applications, such as behavioral neurology, researchers have to interpret and compare large amounts of multivariate time series in an interactive and interpretable way. The objectives of this thesis are to define novel symbolic representations and distance measures that can handle physiological signals with a complex structure: multivariate, non-stationary, and multimodal. Moreover, the representation should preserve the time information and be interpretable and fast to compute.
After reviewing symbolization techniques (that transform a real↓alued series into a shorter discrete↓alued series) and conducting a survey of distance measures on time series, strings, and symbolic sequences, we introduce novel symbolic representations and define a distance measure between the resulting symbolic sequences.
The first contribution is a symbolic representation for a data set of univariate time series called ASTRIDE. Unlike most symbolization procedures, ASTRIDE is adaptive during both the segmentation step by performing change-point detection and the quantization step using quantiles. Instead of proceeding signal by signal, ASTRIDE builds a dictionary of symbols that is common to all signals in a data set. We also introduce a novel distance measure on symbolic representations that is based on the general edit distance with custom weights. We show the performance of ASTRIDE compared to 4 other symbolic representations on reconstruction and, when applicable, on classification tasks.
The second contribution is a symbolic representation for a data set of multivariate time series that can be non-stationary, called d_symb, along with an online exploration tool, called the d_symb playground. Unlike most distance measures on multivariate signals, d_symb takes into account their non-stationarity thanks to a symbolization step. This step is based on a change-point detection procedure that splits a non-stationary signal into several stationary segments, followed by quantization using K-means clustering. The proposed distance measure leverages the general edit distance.
The advantages of d_symb are shown on three data sets of physiological signals. Experiments show how interpretable the symbolization is: a single glance at the symbolic sequences provides an immediate and comprehensive understanding of a data set. Moreover, compared to nine multivariate elastic distances on a clustering task, d_symb achieves a competitive performance while being several orders of magnitude faster than the other methods. With these desirable characteristics, we developed the d_symb playground, an online tool, that allows researchers to apply d_symb to their uploaded data.

Key words

Change-point detection, Pattern recognition, Symbolic approaches, Representation learning

Jury

- Germain FORESTIER, Professeur des Universités, Université de Haute-Alsace (France), rapporteur et examinateur
- Romain TAVENARD, Professeur des Universités, Université de Rennes 2 (France), rapporteur et examinateur
- Mathilde MOUGEOT, Professeure des Universités, ENSIIE et affiliée à l'ENS Paris-Saclay (France), examinatrice
- Themis PALPANAS, Professeur des Universités, Université Paris-Cité (France), examinateur
- Patrick SCHÄFER, Chercheur, Humboldt-Universität zu Berlin (Allemagne), examinateur