Extracción de características en series de tiempo

ITESO, Universidad
Jesuita de Guadalajara

Dr. Gaddiel Desirena López

Verano 2025

Contenido

Series de tiempo

- La dinámica de un proceso o fenómeno se puede capturar como un conjunto de mediciones repetidas.
- Las series de tiempo son un tipo de datos para comprender la dinámica en los sistemas del mundo real.
  - se muestrean en un período de muestreo constante,
  - o periodo de muestreo variable.

La representación de muestreo constante permite que otros tipos de datos secuenciales se representen de la misma manera

- Espectros de frecuencia,
- Oraciones en un libro,
- Forma de objetos, etc.

Extracción de características en series de tiempo

- Variable ordenada
  - medidas de su tendencia,
  - distribución,
  - entropía
- Serie de tiempo
  - entropía espectral,
  - tendencia,
  - estacionalidad,
  - autocorrelación,
  - parámetro óptimo de transformación de Box-Cox,
  - componentes de Fourier.

Series de tiempo

- Medidas simples
  - la media es dispersa o con alta concentración de valores,
  - presencia de valores atípicos,
  - la distribución se aproxima a la Gaussiana, entre otros.
- Características temporales
  - qué tan relacionada es la serie consigo misma,
  - qué tan ruidosa es la serie de tiempo,
  - periodicidad,
  - cómo cambian las propiedades estadísticas a lo largo del tiempo (estacionariedad),
  - transformada discreta de Fourier,
  - entropía. Como medida de complejidad o predicibilidad de la serie.

Características globales

Métrica de estacionalidad:

$$StatAv(\tau) = \frac{std(\{\bar{x}_{1,w}, \bar{x}_{w+1,2w}, \ldots, \bar{x}_{(m-1)w+1,mw}\})}{std(x)}$$

donde la desviación estándar se toma a través del conjunto de medias calculadas en m ventanas no superpuestas de la serie de tiempo, cada una de longitud w.
Autocorrelación:

$$C(\tau) = \langle x_t x_{t+\tau} \rangle = \frac{1}{\sigma_x^2 (N-\tau)} \sum_{i=1}^{N-\tau} (x_t - \bar{x})(x_{t+\tau} - \bar{x}),$$

donde $\tau = t_2 - t_1$ es el intervalo de tiempo de interés.

Intervalos

- Motivación
  - clasificación de series de tiempo.
    - buscan aprender la ubicación de subsecuencias discriminatorias y las características que separan diferentes clases.
- Objetivo
  - Obtención de características
    - Media
    - Desviación estándar
    - Tendencia (pendiente)

Intervalos

- media:

$$\bar{x}(t_1, t_2) = \frac{1}{t_2 - t_1 + 1} \sum_{i=t_1}^{t_2} x_i,$$

- varianza muestreada:

$$\sigma_x^2(t_1, t_2) = \frac{1}{t_2 - t_1} \sum_{i=t_1}^{t_2} (x_i - \bar{x}(t_1, t_2))^2,$$

  donde $\bar{x}$ es la media de la serie de tiempo x en el intervalo $[t_1, t_2]$.

- pendiente: calculada a partir de una línea de regresión de mínimos cuadrados a través del intervalo

Intervalos

Esta información se utiliza para comprender qué propiedades de series de tiempo impulsan una clasificación exitosa en cada momento. El proceso es:

1. Muestreo aleatorio de intervalos.
2. Usar un clasificador para cada uno de ellos.
3. Evaluar la contribución de cada característica en el modelo.

Otro trabajo ha utilizado matrices de covarianza característica-característica para capturar propiedades de subsecuencia para la clasificación.

Shapelets

- Consiste en encontrar subsecuencias con alta predecibilidad.
- El método consiste en:
  - Definir una subsecuencia de intervalo aleatorio. Será la shapelet candidata.
  - Comparar las distancias entre la shapelet y las serie de tiempo discretizada.
  - La subsecuencia que genere menor error será la shaplet.

$$
\begin{aligned}
s &= \min d(s, x) \\
d(a, b) &= \sqrt{2(1 - C(a, b))}
\end{aligned}
$$

$$
C(a, b) = \frac{\sum_{i=1}^{m} ab - m\mu_a\mu_b}{m\sigma_a\sigma_b}
$$

Shapelets

- Consiste en encontrar subsecuencias con alta predecibilidad.
- El método consiste en:
  - Definir una subsecuencia de intervalo aleatorio. Será la shapelet candidata.
  - Comparar las distancias entre la shapelet y las serie de tiempo discretizada.
  - La subsecuencia que genere menor error será la shaplet.

$$
\begin{aligned}
s &= \min d(s, x) \\
d(a, b) &= \sqrt{2(1 - C(a, b))}
\end{aligned}
$$

$$
C(a, b) = \min_{0 \leq l \leq n-m} \frac{\sum_{i=1}^{m} a_i b_{i+l} - m\mu_a\mu_b}{m\sigma_a\sigma_b}
$$

Diccionario de patrones

- Resume el numero de repeticiones encontradas a lo largo de la serie de tiempo.
- Encontrar estos patrones discriminativos y luego caracterizar cada serie de tiempo por la frecuencia de cada patrón, proporciona información útil sobre el peso que representa de cada shapelet.