ITESO, Universidad Jesuita de Guadalajara

Departamento de Matemáticas y Física

Ingeniería de características

Extracción de características en series de tiempo

Dr. Gaddiel Desirena López

Primavera 2024

El paso del tiempo y la dinámica de los procesos es un componente fundamental de la experiencia humana. La dinámica se puede capturar como un conjunto de mediciones repetidas del sistema a lo largo del tiempo o como una serie de tiempo. Las series de tiempo son un tipo de datos fundamental para comprender la dinámica en los sistemas del mundo real.

Las series de tiempo suelen representarse con un período de muestreo constante, Δt, lo que facilita una representación como un vector ordenado $x = \{x_1,\ x_2,\ \ldots,\ x_N\}$, donde N mediciones han tomado en los tiempos $t = \{0,\ \Delta t,\ 2\Delta t,\ \ldots,\ (N-1)\Delta t\}$. Esto permite que otros tipos de datos secuenciales de valor real se representen de la misma manera, como espectros (orden por frecuencia), secuencias de longitud de palabras de oraciones en libros (orden a través del texto), la forma de los objetos (donde la distancia desde un punto central en una forma se puede medir y ordenar por el ángulo de rotación de la forma).

Cada variable ordenada se convierte en un conjunto de características globales (como medidas de su tendencia, distribución o entropía), que se utilizan para definir la similitud entre pares de series de tiempo. Saber qué representaciones basadas en características proporcionan un buen rendimiento para una tarea determinada puede proporcionar una comprensión conceptual de las propiedades de los datos que impulsan la toma de decisiones precisa. En el caso de características globales de series de tiempo, se codifican conceptos teóricos más profundos (como entropía, estacionalidad o componentes de Fourier).

Cada serie de tiempo se representa como un vector que contiene seis características (entropía espectral, tendencia, estacionalidad, período estacional, autocorrelación de retardo-1 y el parámetro óptimo de transformación de Box-Cox), después de lo cual se obtiene el conjunto de datos de la serie de tiempo completo.

Los modelos que se ajustan bien a los datos observados se pueden simular en el tiempo para hacer predicciones sobre el estado futuro del sistema, una tarea conocida como *forecasting*.

1. Características globales

Distinguimos entre características de conjuntos desordenados: medidas simples, y características con dependencia temporal. Las primeras, parten de la distribución de datos

- la media es dispersa o con alta concentración de valores,
- presencia de valores atípicos,
- la distribución se aproxima a la Gaussiana, entre otros.

Las características temporales son, por ejemplo:

- 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 de Fourier,
- entropía. Como medida de complejidad o predicibilidad de la serie.

Una serie de tiempo estacionaria se produce a partir de un sistema con parámetros fijos y constantes o distribuciones de probabilidad sobre parámetros que no varían. Las medidas estacionaras señalan cómo varían las dependencias temporales. Por ejemplo, la métrica StatAv:

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

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. Las series de tiempo en las que la media en las ventanas de longitud w varían más que la serie de tiempo completa en su conjunto tienen valores más altos de StatAv en esta escala de tiempo en relación con las series de tiempo en las que las medias de las ventanas son menos variables.

La autocorrelación mide la correlación entre los valores de series de tiempo separados por un lapso de tiempo determinado. Lo siguiente proporciona es una estimación:

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

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

2. Intervalos

Algunos problemas de clasificación de series de tiempo pueden involucrar diferencias de clase en las propiedades de las series de tiempo que están restringidas a intervalos de tiempo discriminativos específicos.

Los clasificadores de intervalo buscan aprender la ubicación de subsecuencias discriminatorias y las características que separan diferentes clases, que se pueden aprender calculando características simples en muchas subsecuencias, y luego construyendo clasificadores buscando tanto características como intervalos de tiempo.

Tres características simples para capturar las propiedades de las subsecuencias de la serie temporal (para un intervalo $t_1 \leq x \leq t_2$) para ayudar a la interpretabilidad y la eficiencia computacional [1]:

- 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

De esta manera, cada subsecuencia de la serie de tiempo está representada por los valores de estas tres características. 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 [2].

3. Shapelets

Otra representación de las series de tiempo es en términos de las propias subsecuencias individuales.

En el contexto de la clasificación de series de tiempo, las subsecuencias que son altamente predictivas de las diferencias de clase se conocen como shapelets y brindan información interpretable sobre los tipos de patrones secuenciales (o formas).

El problema del descubrimiento del shapelet se puede enmarcar en la determinación de subsecuencias, s, que distinguen mejor las diferentes clases de series de tiempo por su distancia al shapelet, $d(s, x)$. La distancia entre una serie de tiempo, x, y una subsecuencia, s, de longitud m, se define como la distancia euclidiana mínima entre la normalización de la subsecuencia a lo largo de la serie de tiempo:

$$s = \text{mín}(s, x_{k,\ldots,k+m}),$$

para una función de distancia euclidiana, d definida como sigue [3]:

$$d(s, x) = \sqrt{2(1 - C(s, x))},$$

donde $C(s, x)$ representa la correlación entre estas series de tiempo

$$C(s, x) = \min_{0 \le l \le n-m} \frac{\sum_{i=1}^{m} s_i x_{i+l} - m\mu_s\mu_x}{m\sigma_s\sigma_x},$$

donde μ_s y μ_x son respectivamente la media de la shapelet candidata y la serie de tiempo. Así mismo, las desviaciones estándar σ_s y σ_x correspondientes a la shapelet candidata y el conjunto de datos de la serie de tiempo.

De esta manera, los clasificadores basados en shapelets pueden aprender subsecuencias discriminatorias interpretables para problemas de clasificación de series de tiempo.

4. Diccionario de patrones

Si bien los shapelets pueden capturar qué tan bien coincide una subsecuencia con una serie de tiempo (y son adecuados para series de tiempo cortas basadas en patrones), no pueden capturar cuántas veces se representa una subsecuencia dada en una grabación de serie de tiempo extendida.

Aprender estos patrones discriminativos y luego caracterizar cada serie de tiempo por la frecuencia de cada patrón a lo largo de la grabación, proporciona información útil sobre el peso que representa de cada shapelet.

Referencias

[1] `https://arxiv.org/pdf/1302.2277.pdf`

[2] Ergezer, H., & Leblebicioğlu, K. (2018). Time series classification with feature covariance matrices. Knowledge and Information Systems, 55(3), 695-718.

[3] `http://alumni.cs.ucr.edu/~mueen/pdf/Logical-Shapelet.pdf`