Universitat de València

Cajamar UniversityHack 2022

Informe

Fase local

The Backpropagation Boyz

Miguel Hortelano Busto
José Manuel Sánchez Aquilué
Arturo Sirvent Fresneda

marzo, 2022

Escuela Técnica Superior de Ingeniería (ETSE-UV)

1. Introducción

La predicción de series temporales juega un papel importante en multitud de campos como la industria, finanzas, medicina, etc. La expresión y descripción de los diferentes procesos y comportamientos de un sistema a lo largo del tiempo suele ser altamente no lineal, plagado de fenómenos aleatorios, fluctuaciones y estacionalidades. Por ello suponen una cuestión recurrente en el mundo del machine learning, donde se han llegado a proponer multitud de métodos para resolver estas predicciones.

Entre los más populares encontramos las redes neuronales. La cantidad de modelos propuestos es ingente, basándose en diferentes arquitecturas, y su viabilidad depende muchas veces del problema propuesto. Sin embargo podemos destacar las redes de tipo LSTM (Long Short Term Memory), cuya ventaja principal es la capacidad de mantener dependencias de la serie temporal a largo plazo sin caer en el desvanecimiento (o explosión) del gradiente. Como ejemplo de su uso, una red neuronal de esta familia aparecía junto con un suavizado exponencial en el algoritmo ganador del concurso de predicción de series temporales M4 Makridakis.

Sobre los datos presentados, sabemos que corresponden a contadores de agua de los cuales tenemos el identificador único del contador (id), la fecha en la que se ha tomado cada muestra y tanto la lectura del contador como su consumo extrapolado de las diferentes lecturas. Las partes entera y decimal de estas dos últimas variables se dan en dos columnas diferentes, es decir, tenemos 4 columnas numéricas.

2. Análisis exploratorio

Contamos con series temporales cuyas muestras se han tomado a lo largo de un año entero empezando en febrero. En total contamos con 2747 series con diferente número de datos, desde apenas cuatro muestras hasta 32000 con una mayoría de ids con (aproximadamente el 70 %). Con el fin de disponer de un dataset más reducido sin perder información del conjunto, optamos por agregarlos datos por día y declarar una única variable con la parte entera y decimal, tanto para delta como para la lectura.

Una vez cargado el conjunto de datos, definimos un array de series temporales con la evolución del valor delta. Con tan solo visualizar las diez primeras series caemos en la cuenta de que la evolución del consumo de cada depósito es de lo más diversa. Algunos tienen un consumo constante todos los días del año, otros solamente gastan agua en una determinada época, alguno presenta un consumo medianamente constante y un día en concreto se dispara, etc. Por tanto es una buena idea agrupar

las series en distintos clústers. Debido al enorme volumen de datos que estamos manejando no es posible entrenar el modelo con todas las muestras, hace falta dividir el conjunto en dos partes. Con una quinta parte del conjunto construiremos el modelo a partir del cual conseguiremos las etiquetas de cada clúster.

Tras el proceso de clústering se comprueba que estén bien balanceados y no hayan quedado grupos muy específicos junto a otros que actúen de cajón de sastre. A continuación se muestra por pantalla la representación de cada grupo, tanto las series normalizadas como sin normalizar. Para cada clúster mostramos por pantalla la distribución por día de la semana, día del mes y año. Aquí ya se ven diferencias entre grupos, como, por ejemplo, algunos muestran un mayor consumo los meses de verano mientras que otros gastan más en otoño.

En principio sí que se aprecia cierta homogeneidad en elementos de un mismo grupo. Sin embargo la mejor manera de comprobarlo es mediante el coeficiente de silhouette lo haremos de manera numérica como gráfica (Figura 1). A la vista de los resultados, podemos comprobar que quizás este tipo de técnicas de clústering no son las más apropiadas para estos problemas. En la sección del modelo utilizaremos en su lugar un clústering jerárquico.

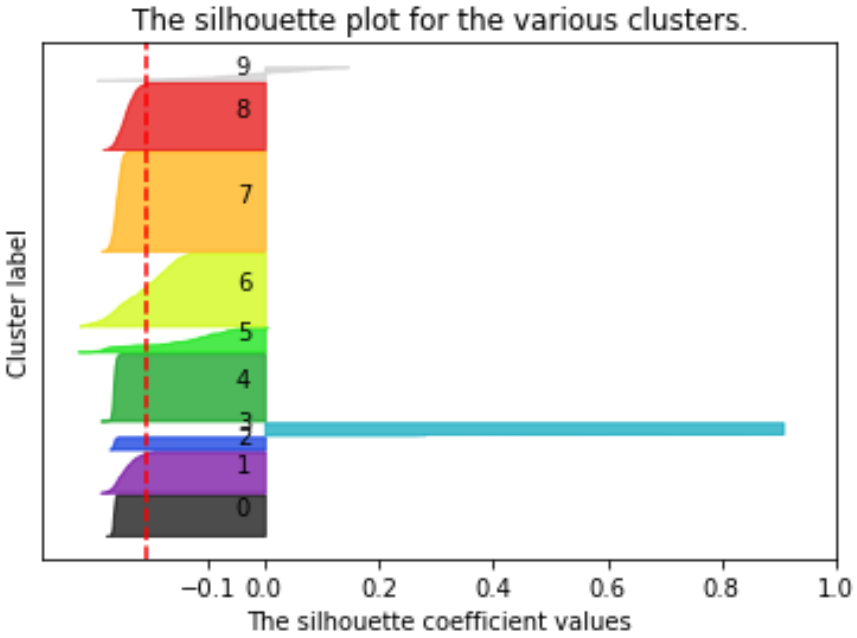

Figura 1: Diagrama de silhouette del agrupamiento.

3. Preprocesado de datos

En primer lugar, ordenamos las lineas del fichero Modelar_UH2022.txt, de manera queden las muestras agrupadas por su id de menor a mayor y dentro de la agrupación aparezcan en orden cronológico. No es práctico trabajar con la gran can-

tidad de muestras que tenemos en el fichero original, necesitamos agregar los datos de alguna manera. Así que repetimos el proceso de reducción de datos aplicado también en el EDA.

Una vez tengamos el fichero definitivo es de interés conseguir que todas las series tengan el mismo número de muestras para entrenar el modelo con más comodidad. Por tanto recurrimos al zero padding, asignando valor 0 en los días sin datos de aquellas series que lo requieran. Como vamos a utilizar los datos del consumo diario en vez de la lectura del contador, nos interesa aplicar un filtro gaussiano que reduzca el comportamiento a priori caótico y ruidoso de esta variable. Aunque pueda parecer que al estar modificando los datos originales, a la hora de validar el modelo tendremos peores resultados, el suavizado ayuda al algoritmo a aprender mejor y acabamos obteniendo mejores predicciones comparando con el conjunto original sin filtrar. Naturalmente, para hacer funcionar el modelo de la mejor forma posible es conveniente escalar los datos. En este tipo de problemas se suele utilizar un escalado minmax.

Con los datos ya escalados calculamos las distancias euclidianas entro los diferentes vectores de datos. Con la matriz de distancias resultante realizamos un clústering jerárquico. Con la información del clúster dividiremos el conjunto de series en cuatro partes. Esta división pretende facilitar el aprendizaje del modelo agrupando aquellas series que tengan características y patrones parecidos.

Como complemento a los propios datos de la serie temporal, también contaremos con datos externos que agregaremos a los otorgados en el concurso. En concreto datos climáticos extraídos de la plataforma web de AEMET (Agencia Estatal de Meteorología) correspondientes a estaciones de la costa valenciana. Utilizaremos la media de temperatura diaria, la insolación y las precipitacio. Dos factores que, consideramos, afectan al consumo de agua. Dos ejemplos rápidos podrían ser la necesidad de refrescarse (o calentarse) como los cambios actividades de ocio o el consumo de regadío en los días de lluvia. También se ha valorado la inclusión de más variables climáticas, pero estas no mejoraban los resultados del modelo final.

4. Modelo

Como se ha comentado en este documento las redes neuronales de tipo LSTM gozan de cierta popularidad en la predicción de series temporales. Por ello escogemos un modelo de este tipo al funcionar bien con series largas y de nuestras características.

La red predice los valores de consumo pertinentes a una semana desde el úl-

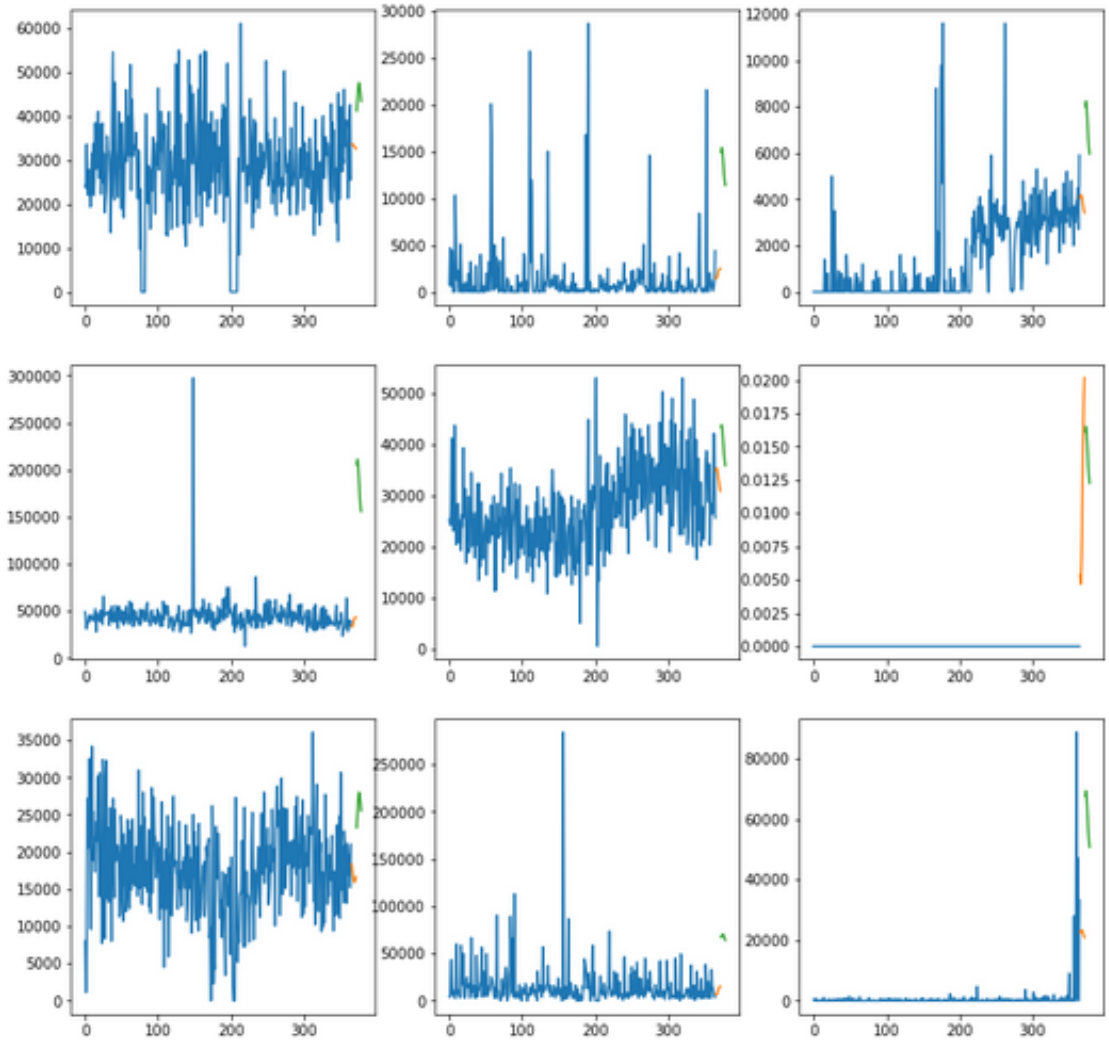

Figura 2: Predicciones de los 9 primeros depósitos.

timo día. Para hacer la predicción el modelo ha sido entrenado para recibir una ventana temporal de 40 días como input (además del), y sacar los 7 días siguientes como output. El modelo completo se compone de un ensemble de 4 modelos, cada uno aplicado a un cluster de series temporales, permitiendo así que este se especialice en las características de las diferentes series (pues unas se pueden deber a industria, otras a hogares etc...). La arquitectura de todos los modelos es la misma: LSTM(30) + LSTM(20) + DENSE(100,RELU") + DENSE(50,RELU") + DENSE(7,"SIGMOID").