ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA
Universidad de Córdoba

Petición de tema de trabajo de fin de grado

Grado en Ingeniería Informática

Recuperación de información en series temporales: aplicación a problemas de altura de ola.

Autor
Victoriano Pedrajas Fernández

Directores
Juan Carlos Fernández Caballero
David Guijo Rubio

Marzo, 2020

UNIVERSIDAD D CORDOBA

Autor:

Fdo: Victoriano Pedrajas Fernández

Directores:

Fdo: Dr. Juan Carlos Fernández Caballero

Fdo: D. David Guijo Rubio

Índice general

Capítulo 1

Datos del Proyecto.

Titulo: Recuperación de información en series temporales: aplicación a problemas de altura de ola.

Autor:

- **Nombre**: Victoriano Pedrajas Fernández
- **e-mail**: i62pefev@uco.es
- **DNI**: 31023427S
- **Titulación**: Grado en Ingeniería Informática
- **Mención**: Computación

Directores del proyecto:

- **Nombre**: Dr. Juan Carlos Fernández Caballero
- Profesor Titular de Universidad
- **Departamento**: Informática y Análisis Numérico
- Escuela Politécnica Superior de Córdoba
- Universidad de Córdoba
- **e-mail**: jfcaballero@uco.es

- **Nombre**: D. David Guijo Rubio
- Contratado FPU
- **Departamento**: Informática y Análisis Numérico
- Escuela Politécnica Superior de Córdoba
- Universidad de Córdoba
- **e-mail**: dguijo@uco.es

Capítulo 2

Introducción.

Las boyas oceanográficas son instrumentos de control empleados para caracterizar las propiedades de olas generadas por acción del viento. Las boyas se encuentran instaladas en aguas costeras y mar adentro y disponen de sensores que miden diferentes observaciones meteorológicas. Entre ellas se encuentran la altura de ola y el periodo de energía de la ola, a partir de las cuales se puede obtener el flujo de energía, el cual es empleado para estimar la cantidad de energía mareomotriz que puede ser explotada. Además, la altura de ola también resulta interesante en otro tipo de aplicaciones: diseño de estructuras, navegación, etc. El problema es que la precisión y disponibilidad de dichas boyas es muy inestable, resultando en completas temporadas donde falta información. Los eventos que pueden dar lugar a esta clase de fallos pueden ser accidentes de navegación, periodos de mantenimiento, tormentas [1], etc.). Esto repercute negativamente en los sistemas de energía renovable que dependen de la medición de dichas boyas que, tras los fallos producidos en las boyas, repercuten en aun más errores de medición y predicción. Por ello, cuando la información de un cierto punto de la serie temporal no se encuentra disponible, es de gran interés el uso de un algoritmo que reconstruya los valores perdidos.

A lo largo del tiempo, se han ideado muchas soluciones para intentar recuperar los datos perdidos. Entre ellos se destacan los trabajos que hacen uso de algoritmos autorregresivos con el fin de reconstruir series de altura de olas a partir de las propias series [2, 3], y el uso de métodos de fractales e interpolaciones cúbicas [4], entre otros.

En los últimos años, debido a la explosión de la ciencia de datos se han comenzado a desarrollar muchos trabajos que aplican técnicas basadas en *machine learning*. Se destaca sobretodo el uso de redes neuronales como método de predicción principal [5]. En términos de la recuperación de datos en series de olas, las redes neuronales realizan un trabajo especialmente preciso a la hora de llevar a cabo estimaciones sobre los datos de olas perdidos [6].

El objetivo principal de este proyecto fin de grado se basa en la creación de un algoritmo que haga uso de redes neuronales para la recuperación de información en series temporales de altura de olas con el fin de sobrellevar los problemas citados anteriormente.

A continuación, se desarrollan los principales conceptos relacionados con este Trabajo Fin de Grado (TFG).

2.1. *Machine Learning*

El aprendizaje automático, o *machine learning*, es una técnica del campo de la inteligencia artificial que permite obtener conocimiento sin la necesidad de programarlo

explícitamente. Este aprendizaje esta basado en una serie de datos que se le proporcionan al programa y, en base a ellos, el programa transmitirá conocimiento con el fin de realizar una determinada tarea. El objetivo principal de este proceso es que los sistemas adquieran conocimiento sin la intervención humana.

De forma fundamental, existen varios tipos de aprendizaje automático:

- **Aprendizaje supervisado**: Extrae información de patrones que ya han sido etiquetados por un experto.
- **Aprendizaje no supervisado**: Recibe datos que no han sido etiquetados y trata de realizar agrupaciones de los mismos.
- **Aprendizaje semisupervisado**: Usa tanto datos etiquetados como no etiquetados para realizar el aprendizaje.
- **Aprendizaje por refuerzo**: Interactúa con un entorno el cual proporciona errores y refuerzos positivos a los modelos.

Este proyecto se basa en el aprendizaje supervisado, ya que está enfocado a la recuperación de información de series temporales compuestas por valores numéricos, concretamente, altura de ola en metros.

Capítulo 3

Objetivos.

El objetivo principal de este proyecto es la recuperación de información en series temporales de altura de ola. Los principales objetivos a destacar de este trabajo de fin de grado son los siguientes:

1. Realizar un estudio teórico de la literatura en la recuperación de información de series temporales.
2. Obtener los datos y formatear (de acuerdo a lo especificado en el artículo [10]) los conjuntos de datos a partir de los cuales se realiza la recuperación de información.
3. Implementar los métodos de recuperación de información, entre los cuales se encuentran los métodos de funciones de transferencia y métodos de los vecinos.
4. Aplicar redes neuronales para mejorar la recuperación de información obtenida a partir de los métodos especificados en el objetivo anterior.
5. Realizar estudio comparativo con diferentes metodologías de aprendizaje automático a partir de los datos recuperados.

Capítulo 4

Antecedentes.

A día de hoy existe una gran variedad de herramientas en el estado del arte del aprendizaje automático. Debido al ámbito del problema considerado en este Trabajo Fin de Grado (TFG), recuperación de información en series temporales, nos enfocaremos en métodos de aprendizaje automático supervisado. La recuperación de información en series temporales se puede afrontar desde dos perspectivas diferentes: 1) de forma longitudinal, más relacionada con el área de predicción de series temporales, y 2) desde un punto de vista transversal, más relacionado con los métodos de regresión. Concretamente, en este TFG nos vamos a centrar en la segunda forma.

Más concretamente, nuestro problema se puede afrontar desde las dos siguientes perspectivas: aplicando técnicas de regresión para estimar, en metros, la altura de ola, o mediante técnicas de clasificación, para las cuales habría que discretizar la variable en tantos valores diferentes como se quiera, un ejemplo es: "de 0 a 0.25m, de 0.25m a 0.5m...". Aún así, aunque este problema podría ser afrontado desde ambas perspectivas si se discretizase, en este proyecto el enfoque se encuentra en los métodos de regresión.

4.1. Metodologías generales de recuperación de información para series temporales.

De forma fundamental se aplicaran funciones de transferencia y el método de los vecinos para la recuperación de información en series temporales. Posteriormente, se aplicarán métodos de aprendizaje supervisados para mejorar la recuperación de información obtenida mediante las técnicas previamente mencionadas.

4.1.1. Funciones de transferencia

Este método genera una regresión lineal a partir de dos series temporales: Una de ellas contiene valores perdidos mientras que la segunda tiene toda la información. El proceso como tal se compone de los siguientes pasos:

1. Se calculan las regresiones con los datos comunes de la serie completa y la serie incompleta.

2. Se aplica dicha regresión a los datos faltantes con el fin de recuperar así todos los valores de la serie.

4.1.2. Método de los vecinos.

El método de los vecinos aplica la siguiente ecuación para cada valor perdido de la serie incompleta.

$$\hat{Y_t} = \left(\frac{1}{N} \sum_{i=1}^{N} \frac{\bar{Y} * X_{it}}{\bar{X_i}} \right)$$

Siendo cada una de las variables empleadas:

- $\hat{Y_t}$: Valor predicho de altura de ola en el instante t.
- N : Número de boyas con dicho valor real.
- $\bar{Y}$: Media de los valores existentes de la serie temporal de altura de ola.
- X_{it} : Valor de la boya i en el instante t.
- $\bar{X_i}$: Media de la altura de ola de la boya i.

4.1.3. Redes Neuronales Artificiales.

Se llaman Redes Neuronales Artificiales (RNA), o *artificial neural networks*, a aquellos modelos no lineales que tratan de emular el comportamiento de las neuronas de nuestro cerebro. Para ello, las neuronas (o unidades más básicas del modelo neuronal) se representan como una determinada función matemática. Con el fin de lograr comportamientos más complejos a partir de estas unidades más sencillas, se unen multitud de ellas en lo que se denominan redes neuronales. Dependiendo de como se realicen estas conexiones, se puede dar lugar a un comportamiento u otro de dicho conjunto.

Con el fin de ampliar los conceptos relacionados a las RNAs, se puede apreciar en la figura 4.1 la existencia de todas sus diferentes partes que la componen.

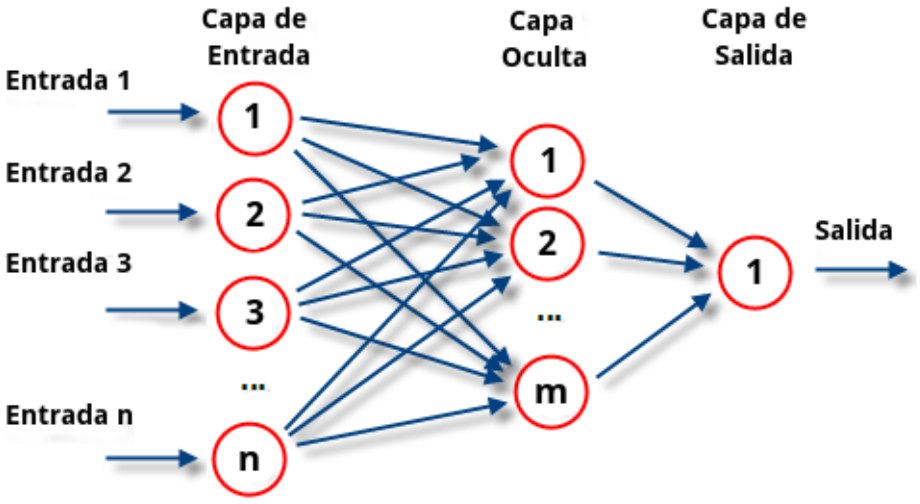

Figura 4.1: Arquitectura de una RNA.

De esta manera, se puede observar la existencia de varios tipos de capas.

1. **Capa de entrada**: Son las entradas de la red neuronal, donde cada neurona recibe un atributo del conjunto de datos.

2. **Capa oculta**: La capa de entrada envía datos a través de la sinapsis a esta capa de neuronas. Llamamos sinapsis al proceso por el cual, unos parámetros (a los que se llamaran de pesos a partir de ahora) manipularán los datos de las neuronas mediante cálculos al pasar de una capa oculta a otra. La última de las capas ocultas envía datos a través de la sinapsis a esta capa de salida. Cabe destacar como, en este caso, se puede ver que sólo existe una capa oculta aunque pueden existir más dependiendo de la complejidad del problema.

3. **Capa de salida**: Los resultados de los valores de esta capa, son las salidas de la RNA.

4.2. Grupo de Investigación AYRNA.

El Grupo de Investigación Aprendizaje y Redes Neuronales Artificiales (AYRNA) [15] tiene una amplia experiencia en el campo de las redes neuronales. Asimismo, la obtención de los datos de altura de ola se realizará haciendo uso de la herramienta SPAMDA [7], de la cual, son los autores y desarrolladores. La herramienta SPAMDA se trata de una herramienta software para el preprocesamiento y análisis de datos de estaciones del *National Data Buoy Centre* NDBC [12] y del proyecto *NCEP/NCAR Reanalysis Project* (NNRP), siendo de esta manera de gran ayuda a la hora de obtener análisis de datos preprocesados, entre otros.

Capítulo 5

Fases de Desarrollo.

Durante el desarrollo del proyecto, se pueden distinguir una serie de fases. Dichas fases serán identificadas a continuación.

5.1. Obtención de conocimientos.

En esta fase inicial se busca estudiar y analizar el problema desde una perspectiva teórica, con el fin de identificar todas las necesidades que se produzcan en respuesta a la consecución de los objetivos indicados en la sección 3.

Asimismo, se estudiará el lenguaje de programación *Python* [8] y la herramienta SPAMDA.

5.2. Estudio y análisis del problema.

El problema que se va a abordar puede ser dividido a su vez en otros dos:

- Problema real: Recuperación de información en series temporales de altura de ola. Los datos obtenidos pertenecerán principalmente a la *National Data Buoy Centre* (NDBC) de los EE.UU.

- Problema técnico: Creación de un algoritmo de recuperación de información en series temporales haciendo uso de funciones de transferencia, regresiones a partir del método de los vecinos y mediante el uso de RNAs.

5.3. Diseño

Una vez analizados y estudiados los requisitos y objetivos de nuestro modelo, se procederá a diseñarlo para posteriormente codificarlo. Se analizará la funcionalidad del sistema, el diseño de los datos, el flujo de control que estos seguirán, etc. Se creará un *framework* sencillo que facilite el desarrollo de nuevos métodos de recuperación de información en series temporales.

5.4. Implementación

Se usará el lenguaje de programación *Python*, debido a los siguientes motivos:

- *Python* es un lenguaje de programación multiparadigma. Es decir, que obliga al usuario a adoptar un estilo particular de programación que permite a su vez varios estilos, como pueden ser la programación orientada a objetos, la programación imperativa y la programación funcional.
- Hace uso de memoria dinámica.
- Posee una característica denominada resolución dinámica de nombres, que hace que se enlace un método con un nombre de variable.
- Al ser fácil el diseño del lenguaje, hace que sea fácil la extensión de nuevos módulos.
- *Python* permite la inclusión en cualquier tipo de aplicación, librería, etc. que necesite una interfaz programable.
- Permite el uso de un modo interactivo.
- Por último, *Python* dispone de una gran cantidad de librerías matemáticas y científicas.

Se empleará la herramienta SPAMDA con el fin de poseer una herramienta software que pueda crear conjuntos de datos con las siguientes ventajas:

- Permite el almacenamiento y la gestión de la información proveniente del NDBC y NNRP que habilitaría al usuario el alta, consulta y actualización de la misma en la propia aplicación.
- Permite la creación de nuevos conjuntos de datos a partir de otras ya existentes y la conversión de los mismos a los formatos ARFF (*Attribute-Relation File Format*) [13] y CSV (*Comma-separated values*) [14].
- Visualización del contenido de los conjuntos de datos y muestra valores estadísticos de los atributos.

5.5. Pruebas y experimentación

El modelo será sometido a una serie de pruebas que nos permitirán certificar su correcto funcionamiento y la corrección de todos los posibles errores que se produzcan, además de que se comprobará si los resultados finales cumplen los objetivos planteados, es decir, si el software construido cumple con unos determinados criterios de precisión a la hora de recuperar información en series temporales de forma precisa.

Capítulo 6

Recursos.

6.1. Recursos humanos.

Autor: Victoriano Pedrajas Fernández
Alumno de cuarto curso del Grado de Ingeniería Informática.

- **e-mail**: i62pefev@uco.es
- **DNI**: 31023427S
- **Titulación**: Grado en Ingeniería Informática
- **Mención**: Computación

Director: Dr. Juan Carlos Fernández Caballero

Profesor Titular de Universidad e investigador en el Grupo de Investigación AYRNA. Será encargado de dar soporte técnico y teórico para las tareas de estudio, análisis, diseño e implementación del trabajo.

- **Departamento**: Informática y Análisis Numérico
- Escuela Politécnica Superior de Córdoba
- Universidad de Córdoba
- **e-mail**: jfcaballero@uco.es

Director: D. David Guijo Rubio

Contratado FPU e investigador en el Grupo de Investigación AYRNA. Será encargado de dar soporte técnico y teórico para las tareas de análisis, diseño y programación del trabajo.

- **Departamento**: Informática y Análisis Numérico
- Escuela Politécnica Superior de Córdoba
- Universidad de Córdoba
- **e-mail**: dguijo@uco.es

6.2. Recursos software.

- **Ubuntu 19.10**: Sistema operativo para la programación, prueba de la herramientas y generación de la documentación.
- **Python**: Lenguaje de programación.
- **Google Collab**: Entorno de desarrollo.
- **Git**: Programa que aportará la gestión de versiones.
- **Github**: Plataforma para acceder al código y a sus distintas versiones en la nube.
- **LaTeX**: Sistema de composición de textos orientado a la creación de documentos escritos que presenten una alta calidad tipográfica.
- **Overleaf**: Editor en linea de LaTeX.

6.3. Recursos hardware.

Con el fin de realizar la programación en este proyecto, se hará uso del ordenador personal del proyectista, el cual, posee las siguientes características.

- Procesador Intel(R) Core(TM) i5-6200U @ 2.80 GHz x 4 (64 bits)
- Memoria RAM de 8 GB
- Memoria SSD de 50 GB
- Gráficos Intel HD Graphics 520.

Capítulo 7

Planificación Temporal.

Con el fin de hacer un correcto uso del tiempo empleado en este proyecto, a continuación se desarrolla la siguiente tabla con una estimación de la planificación horaria para cada fase del desarrollo del proyecto.

Fase de Desarrollo	**Mes 1**	**Mes 2**	**Mes 3**	**Mes 4**	**Horas Est.**
Estudio y análisis del problema	40	0	0	0	40
Diseño	25	25	0	0	50
Implementación	0	50	40	0	90
Pruebas y experimentación	0	0	20	40	60
Documentación	10	10	10	30	60
Total	75	85	70	70	300

Cuadro 7.1: Estimación de la organización temporal.

Bibliografía

[1] Rao, S., Mandal, S., 2005. Hindcasting of storm waves using neural networks. Ocean Enf. 32, 667-684

[2] Soares, C.G., Cunha, C., 2000. Bivariate autoregressive models for the time series of significant wave height and mean period. Coast. Eng. 40 (4), 297-311.

[3] Agrawal, J., Deo, M., 2002. On-line wave prediction. Marine Struct. 15 (1), 57-74. AAltunkaynak, A., 2013. Prediction of significant wave height using geno-multilayer perceptron. Ocean Eng. 58 (0), 144-153.

[4] Liu, X., Xia, J., Gunson, J., Wright, G., Arnold, L., 2014. Comparssion of wave height interpolation with wavelet refined cubic spline and fractal methods. Ocean Eng. 87 (0), 136-150.

[5] Haykin, S., 1998. Neural Networks: A Comprehensive Foundation, 2nd ed. Prentice Hall, PTR, Upper Saddle River, NJ, USA.

[6] Balas, C., Koç, L., Balas, L., 2004. Predictions of missing wave data by recurrent neuronets. J. Waterw. Port Coast. Ocean Eng. 130 (5), 256-265.

[7] A. M. Gómez Orellana, J. C. Fernández Caballero, M. Dorado Moreno, 2018. Herramienta software para el preprocesamiento y análisis de datos de estaciones del *National Data Buoy Center* (NDBC) y del proyecto *NCEP/NCAR Reanalysis Project* (NNRP).

[8] The Python Software Foundation. Python documentation, 2015. `https://docs.python.org/3/`.

[9] A.M. Durán-Rosal, Juan Carlos Fernández, P.A. Gutiérrez, and C.Hervás-Martínez. Detection and prediction of segments containing extreme significant wave heights. *Ocean Engineering*, 142:268–279, 09 2017.

[10] A.M. Durán-Rosal, C. Hervás-Martínez, A.J. Tallón-Ballesteros, A.C. Martínez-Estudillo, S. Salcedo-Sanz. Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks. *Ocean Engineering*, 117:292-301, 2016.

[11] F. J. Jiménez-Romero, D. Guijo-Rubio, F. R. Lara-Raya, A. Ruíz-González, C. Hervás-Martínez. Validation of artificial neurla networks to model the acoustic behaviour of induction motors. 10-11, 2020.

[12] National Data Buoy Center. Home. `http://www.ndbc.noaa.gov/`, 2020. Último acceso: 05/03/2020.

[13] Attribute-Relation File Format (ARFF). Home.
`https://www.cs.waikato.ac.nz/ml/weka/arff.html`, 2008.
Último acceso: 09/03/2020.

[14] Wikipedia. Comma-separated values.
`https://en.wikipedia.org/wiki/Comma-separated_values`.
Último acceso: 05/03/2020.

[15] Aprendizaje y Redes Neuronales Artificiales AYRNA. Inicio.
`http://www.uco.es/ayrna/`, 2020. Último acceso: 05/03/2020.