TRABAJO FIN DE MÁSTER

Máster Oficial en Ciencia de Datos e Ingeniería de Computadores

Detección de Anomalías en Series Temporales basada en técnicas Deep Learning

Biblioteca de algoritmos

Autor
Ignacio Aguilera Martos

Director
Francisco Herrera Triguero

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Facultad de Ciencias

—

Granada, 10 de Septiembre de 2020

Detección de Anomalías en Series Temporales basada en técnicas Deep Learning

Biblioteca de algoritmos

Autor
Ignacio Aguilera Martos

Directores
Francisco Herrera Triguero

Detección de Anomalías en Series Temporales basada en técnicas Deep Learning: Biblioteca de algoritmos

Ignacio Aguilera Martos

Palabras clave:

Resumen

Outlier Detection in Time Series using Deep Learning: Library implementation

Ignacio Aguilera Martos

Keywords:

Abstract

Agradecimientos

Índice general

Capítulo 1

Introducción

Antes de comenzar con el desarrollo en sí del estudio acometido en este trabajo, vamos a contextualizar el mismo y vamos a establecer un marco de trabajo teórico previo a la experimentación, que nos otorgará de rigurosidad para la parte práctica del mismo.

El estudio realizado en este trabajo versa sobre la aplicación de estructuras de Aprendizaje Profundo (Deep Learning) para obtención y detección de anomalías, en concreto, en series temporales. Dentro de este trabajo se van a desarrollar las técnicas conocidas como Autoencoders y Redes Neuronales para predicción de series temporales.

Lo primero que atacaremos en este estudio es la definición de anomalía, para luego pasar a una introducción teórica de Estadística Multivariante y Machine Learning en general. Estas dos secciones nos van a aportar el rigor que necesitamos para adentrarnos teóricamente dentro del Deep Learning y entender los fundamentos de las arquitecturas de redes neuronales que aplicaremos en la práctica.

Tras esto se realizará una descripción de la experimentación realizada, los datos que se emplearán en dicha experimentación y los resultados de la misma.

1.1. Contextualización

Lo primero que debemos de hacer antes de empezar, es establecer el problema u objetivo a resolver de este estudio. Para ello vamos a hacer una breve introducción a los datos (o al problema propuesto que es lo mismo

en este ámbito) y a explicar por qué precisamos de un trabajo arduo y prolongado, es decir, por qué no es un problema trivial.

1.1.1. Definición del problema

El ámbito de trabajo va a ser el de las series temporales, pues el conjunto de datos que nos define el problema es una serie temporal. Esta serie temporal mide la sensórica de una máquina de la empresa ArcelorMittal, que no podemos especificar por motivos de privacidad. En este sentido tenemos 106 variables de tipo numérico con las que vamos a trabajar y 468 días de datos con una granularidad de una medida por segundo. Esto hace que el volumen de datos del que disponemos sea inmenso, haciendo que tenga sentido el uso del Deep Learning por la enorme cantidad de datos de entrenamiento de los que vamos a disponer.

Como hemos comentado los datos son medidas de sensores de una cierta máquina. Esta máquina experimenta errores graves de vez en cuando, que hacen que se deba detener completamente para labores de mantenimiento. Nuestro objetivo es ser capaces de detectar estas labores de mantenimiento mediante técnicas de detección de anomalías. El principio subyacente es sencillo: esperamos un comportamiento normal de la máquina en la mayoría del tiempo salvo cuando haya necesidad de un mantenimiento, momento en el cual la sensórica arrojará medidas alteradas que nos den pie a pensar en un posible fallo.

Este tipo de problemas son conocidos como mantenimiento predictivo, pues lo que pretenden precisamente es anticipar la necesidad de dichas labores.

Con esto dicho nuestro objetivo será tomar los datos de entrada (la sensórica) para nuestros modelos Deep Learning y, de alguna manera, saber diferenciar lo que son datos normales y datos anómalos.

1.2. Contenido básico y fuentes

El trabajo contiene una primera sección en la que se incluye una introducción de Aprendizaje Automático orientado específicamente a nuestro problema. Para ello primero se hace una contextualización del concepto de aprendizaje así como los principios inductivos que guían el mismo hacia un buen resultado como por ejemplo el ERM o minimización del error empírico. Se aportan también algunas reflexiones y conceptos en cuanto a la aproxima-

ción de funciones, que no es más que el objetivo del aprendizaje automático.

Todos estos conocimientos están basados en la teoría estadística de Vapnik y Chervonenkis que es brevemente repasada y en la que se dan cotas sobre el aprendizaje y su rendimiento. Esta introducción ha sido escrita basándose en tres libros: Learning from Data de Yaser Abu-Mostafa [**?**], Learning from Data de Cherkassky y Mulier [**?**] y Outlier Ensembles de Aggarwal y Sathe [**?**].

Este marco nos dirige hacia la primera definición del concepto de anomalía que está basada en distancias y rangos intercuartil que se describen en el libro Outlier Analysis [**?**].

Para dar una definición alternativa y una buena introducción para los modelos debemos hacer una breve introducción estadística. En esta introducción se define un vector aleatorio así como su función de densidad, su función característica y su función de distribución. Se refieren los conceptos de independencia y probabilidad y esperanza condicionada. Por último y aprovechando este contexto se enuncian y demuestran algunas desigualdades y fórmulas famosas. Este contenido viene dado por los apuntes de la asignatura Estadística Multivariante del grado en Matemáticas, los apuntes de la asignatura Procesos Estocásticos del grado en Matemáticas y el libro Probability Theory de M. Loève [**?**].

Tras esto puede ser introducido el concepto probabilístico y basado en densidad de una anomalía. Este concepto viene apoyado en el paper [**?**] que describe el algoritmo HICS.

Con los dos conceptos de probabilidad y el marco teórico ya planteado se introducen los modelos implementados y el concepto de algoritmos de ensamblaje. Estos conceptos sobre los algoritmos vienen de los libros Outlier Analysis [**?**] y Outlier Ensembles [**?**]. Se aporta en esta sección la explicación teórica de cada modelo así como la implementación desarrollada por mí mismo en Python. Los artículos en los que se basa cada algoritmo son [**?**], [**?**], [**?**] y [**?**].

Finalmente se analiza el comportamiento de todos los modelos en la sección de resultados frente a los algoritmos considerados como clásicos. Se aportan conclusiones tras todo el trabajo y, al haber margen de mejora, se aportan algunas ideas que podrían aplicarse en un futuro para desarrollar un modelo propio.

1.3. Objetivos

Por todo lo descrito anteriormente el trabajo tiene los siguientes objetivos claros:

- Desarrollar un marco teórico sobre el Machine Learning.
- Desarrollar un marco teórico sobre el Deep Learning.
- Estudiar el estado del arte de los algoritmos de detección de anomalías que emplean Deep Learning.
- Estudiar la teoría estadística que rodea el Machine Learning y el Deep Learning.
- Entender los fundamentos teóricos y el funcionamiento de los modelos implementados.
- Desarrollar una implementación de los modelos.
- Obtener una comparativa entre los modelos clásicos y los Deep Learning.

Todos estos objetivos han sido alcanzados en el desarrollo de este estudio, obteniendo además algunas ideas nuevas que pudieran ser la base de un modelo propio.

Parte I

Machine Learning, Deep Learning y el concepto de anomalía

Capítulo 2

Concepto de Anomalía

2.1. jaja

jeje

Bibliografía