Universidad de Granada

TRABAJO FIN DE GRADO

Doble Grado Ingeniería Informática y Matemáticas

Detección de anomalías basada en técnicas de ensembles

Biblioteca de algoritmos

Autor
Ignacio Aguilera Martos

Directores
Francisco Herrera Triguero
Jacinto Carrasco Castillo

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

FACULTAD DE CIENCIAS

—

Granada, mes de 201

Detección de anomalías basada en técnicas de ensembles

Biblioteca de algoritmos

Autor
Ignacio Aguilera Martos

Directores
Francisco Herrera Triguero
Jacinto Carrasco Castillo

Detección de anomalías basada en técnicas de ensembles: Biblioteca de algoritmos

Ignacio Aguilera Martos

Palabras clave: outlier, anomalía, ensemble, python

Resumen

Poner aquí el resumen.

Outlier detection based in ensemble methods: Library implementation

Ignacio Aguilera Martos

Keywords: outlier, ensemble, anomaly, python

Abstract

Write here the abstract in English.

Yo, **Ignacio Aguilera Martos**, alumno de la titulación Doble Grado en Ingeniería Informática y Matemáticas de la **Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación de la Universidad de Granada** y la **Facultad de Ciencias**, con DNI 77448262-V, autorizo la ubicación de la siguiente copia de mi Trabajo Fin de Grado en la biblioteca del centro para que pueda ser consultada por las personas que lo deseen.

Fdo: Ignacio Aguilera Martos

Granada a X de mes de 201 .

D. **Francisco Herrera Triguero**, Profesor del Departamento Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

D. **Jacinto Carrasco Castillo**, Doctorando del Departamento Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada.

Informan:

Que el presente trabajo, titulado ***Detección de anomalías basada en técnicas de ensembles, Biblioteca de Algoritmos***, ha sido realizado bajo su supervisión por **Ignacio Aguilera Martos**, y autorizamos la defensa de dicho trabajo ante el tribunal que corresponda.

Y para que conste, expiden y firman el presente informe en Granada a X de mes de 201 .

Los directores:

Francisco Herrera Triguero **Jacinto Carrasco Castillo**

Agradecimientos

Poner aquí agradecimientos...

Índice general

Capítulo 1

Introducción

Antes de comenzar el objeto de estudio de este trabajo, lo primero que debemos hacer es contextualizar el mismo y establecer un marco de trabajo en cuanto a teoría que se empleará en la posterior experimentación y desarrollo del mismo.

El estudio realizado y plasmado en este trabajo se centra en la obtención de técnicas para la detección de anomalías en conjuntos de datos, concepto que introduciremos posteriormente. En concreto las técnicas que se van a desarrollar son las conocidas como técnicas de ensemble que se basan en el estudio del problema o bien combinando modelos existentes o bien haciendo un estudio pormenorizado aplicando algún criterio por subespacios del conjunto de datos.

En primer lugar el trabajo desarrollará una introducción a la teoría de aprendizaje y resolución de problemas mediante datos y no por diseño así como la teoría matemática que esto involucra. Esta primera sección nos dará suficiente estructura al trabajo para poder definir en términos de distancias lo que significa que una instancia de un conjunto de datos sea una anomalía.

Posteriormente se desarrollará brevemente algunos conocimientos estadísticos básicos de estadística multivariante para poder introducir el concepto de anomalía desde la perspectiva de las probabilidades condicionadas.

Tras esto podremos entrar en el terreno de la experimentación, desarrollo y explicación de técnicas y puesta en contraste con los algoritmos clásicos para comprobar el desempeño de las nuevas técnicas.

Por último se presentarán las conclusiones obtenidas tras todo este estudio.

1.1. Contextualización

En este contexto y sin haber comenzado la discusión del problema podríamos considerar el establecimiento de una solución al problema de detección de anomalías empleando un algoritmo preciso y óptimo, aunque costoso en tiempo, pero esta solución no es viable en este contexto. Pensemos por un momento en cómo detectar una anomalía en un conjunto de datos. Para poder dar un algoritmo concreto que resuelva siempre el problema de forma óptima tenemos que estar seguros de que somos capaces de definir de forma clara y para todo conjunto de datos cuándo estamos en presencia de un dato anómalo. Pero, ¿sabemos exactamente cuándo sobrepasamos la barrera de dato ligeramente desviado, ruido o una anomalía?

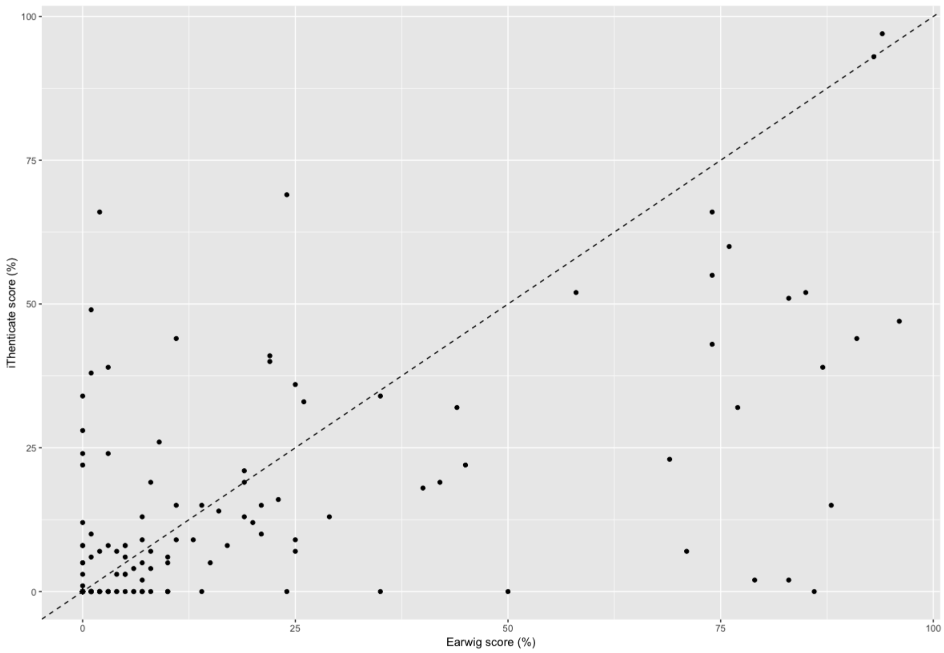

Figura 1.1: Ejemplo de conjunto de datos (Wikimedia)

Pensemos por ejemplo en este conjunto de datos. Tenemos un cluster muy bien definido en la esquina inferior izquierda y datos dispersos. ¿Podemos afirmar claramente cual es la línea que distingue los datos anómalos de los normales?

Esta discusión nos lleva a la primera barrera que tenemos que superar en este trabajo. El problema no es resoluble por diseño y por tanto debemos realizar una aproximación a través de los datos. Esto nos va a llevar a aprender de los mismos, valorar cada dato con una puntuación que nos acabará discriminando los datos normales y anómalos.

En el contexto de los problemas resueltos a través de los datos tenemos dos tipos: clasificación y regresión.

Podríamos pensar que estamos ante un problema de clasificación ordinario, es decir, aprender a través de los datos de entrenamiento cuáles son las anomalías y cuales los datos normales pero no es así. Pensemos que no sabemos ni siquiera nosotros clasificar claramente estas anomalías salvo en algunos ejemplos prácticos. Por ejemplo si le diéramos a una persona un conjunto de datos ya establecido como el que hemos visto anteriormente 1.1 no sería nada fácil e incluso no podemos afirmar que sepamos hacer esta tarea. Por contra si el conjunto de datos es generado en un proceso de trabajo, por ejemplo un conductor de camiones haciendo su ruta, podemos saber cuándo hay una anomalía porque el trabajador la está viviendo y puede señalarla.

Por tanto este problema no es de clasificación al uso pues no es un problema supervisado. Esto significa que, en general no vamos a tener un conjunto etiquetado en el que poder probar cómo aprende nuestro modelo y tendremos que abordar otras técnicas para comprobar el desempeño.

Ahora que comprendemos el alcance del problema podemos intentar profundizar un poco más en la base teórica del mismo analizando las herramientas que el machine learning nos provee para abordarlo.

Parte I

Machine Learning y el concepto de Anomalía

En esta sección vamos a centrarnos en dos aspectos: el Machine Learning para establecer las herramientas usadas en el estudio y el propio concepto de anomalía y algunas reflexiones acerca del mismo.

Capítulo 2

Machine Learning

En este capítulo vamos a hacer un repaso sobre los conceptos asociados al Machine Learning, el aprendizaje y la teoría matemática que involucra. Estas herramientas y conceptos los utilizaremos posteriormente para resolver el problema de detección de anomalías.

2.1. Contextualización del aprendizaje

Para comenzar tenemos que empezar definiendo en que consiste el proceso de aprender sobre unos datos. Supongamos que tenemos un problema en el que tenemos una entrada y una salida, por ejemplo una entrada válida podría ser un vector $x \in \mathbb{R}^d$ y una salida un valor real o un número natural. El problema de aprendizaje intenta estimar una estructura de tipo entrada-salida como la descrita usando únicamente un número finito de observaciones.

Podemos definirlo de forma más general empleando tres conceptos:

- Generador: El generador se encarga de obtener las entradas $x \in \mathbb{R}^d$ mediante una distribución de probabilidad $p(x)$ desconocida y fijada de antemano.
- Sistema: El sistema es el que produce la salida "y" (correcta) para cada entrada $x \in \mathbb{R}^d$ mediante la distribución de probabilidad $p(x|y)$ desconocida y fijada de antemano.
- Máquina de aprendizaje: esta es la que va a obtener información de las entradas y salidas conocidas para intentar predecir la salida correcta

para una entrada nueva que se nos de. De forma abstracta esta máquina lo que hace es tomar una serie de funciones de un conjunto general de forma que para una entrada dada x la función $f(x, \omega)$ con $\omega \in \Omega$ nos de la salida que corresponde para x donde ω es una forma de indexar las funciones tomadas para generalizar la salida del conjunto más general de funciones que hemos indicado.

El único cabo que hemos dejado sin atar en las definiciones que acabamos de ver es el conjunto de funciones del cual tomaremos algunas para adaptar la máquina de aprendizaje a los datos recibidos. Este conjunto de funciones, que notaremos como $\mathcal{H}$, es de momento la única forma que tenemos de aplicar un conocimiento a priori en la máquina de aprendizaje.

Para finalizar esta breve introducción y poder continuar profundizando vamos a exponer algunos ejemplos de clases de funciones para que podamos visualizar el contexto.

- Funciones lineales: En este caso la clase de funciones $\mathcal{H}$ está formada por funciones de la forma $h(x) = w_0 + \sum_{i=1}^{d} x_i w_i$ donde $w \in \mathbb{R}^{d+1}$. Este es el modelo de funciones más clásico.
- Funciones trigonométricas: Un ejemplo de una clase de funciones trigonométricas podría ser $f_m(x, v_m, w_m) = \sum_{j=1}^{m-1}(v_j \sin(jx) + w_j \cos(jx)) + w_0$ donde en este caso la entrada es un único valor real. Este tipo de clases de funciones serán útiles en problemas de regresión que luego explicaremos con algo más de detalle aunque sin centrarnos mucho en ello pues no es el objetivo del estudio.

2.1.1. Objetivo del aprendizaje

Cuando hablamos de aprendizaje queremos obtener algo de dicho aprendizaje a partir de los datos. Como ya se ha mencionado, intentamos obtener una función de una familia de funciones que aproxime o modele de buena manera la salida del sistema. Por tanto, ese es nuestro objetivo: obtener una función de la familia de funciones que minimice el error.

El problema que enfrentamos es que sólo disponemos de un número finito, por ejemplo n, de observaciones de datos y su correspondiente salida. Esto nos va a hacer que no podamos tener una garantía de optimalidad a no ser que tendamos n a infinito.

Sin embargo si que podemos cuantificar cómo de buena es una aproximación con respecto a otra mediante la función pérdida o error que denotaremos como $L(y, f(x, \omega))$. Esta función nos va a medir la diferencia entre la salida real del sistema y la salida dada por la función f para la entrada x siendo siempre $L(y, f(x, \omega)) \geq 0$.

Recordemos además que el Generador obtiene datos mediante una distribución desconocida pero fijada de antemano y que son independientes e idénticamente distribuidos con respecto a la distribución conjunta, es decir:

$$p(x, y) = p(x)p(y|x)$$

Una vez definido todo esto podemos obtener el valor esperado de pérdida o error mediante el funcional

$$R(\omega) = \int L(y, f(x, \omega))p(x, y)dxdy$$

Ahora podemos concretar un poco más lo que entendemos como objetivo del aprendizaje. El objetivo será encontrar una función $f \in \mathcal{H}$ que nos minimice el valor del funcional $R(\omega)$. Pero recordemos que $p(x, y)$ es desconocida para nosotros, por lo que no podemos saber cómo se distribuyen los datos y por tanto el valor del funcional no es calculable para nosotros y por tanto la solución puramente de cálculo no es accesible.

Por tanto, la única forma realmente potente y útil de encontrar una buena aproximación será incorporar el conocimiento a priori que tenemos del sistema. En la sección anterior hemos visto que una forma de incorporar dicho conocimiento es mediante la selección de la clase de funciones, pero además será muy relevante el hecho de cómo los datos son empleados en el proceso de aprendizaje. En este apartado de decisión tendremos que resolver primero la codificación de los datos, el algoritmo empleado y el uso de técnicas como la regularización que veremos después para incorporar nuestro conocimiento en el camino que nos lleve a la solución.

Capítulo 3

Concepto de anomalía