Facultat d'Informàtica de Barcelona

Quatrimestre de otoño, 2022/2023

Práctica de Aprendizaje Automático: Seismic Bumps

Guillem González
(guillem.gonzalez.valdivia@Estudiantat.upc.edu)
Carles Orriols
(carles.orriols.gonzalez@Estudiantat.upc.edu)

Contents

1 Introducción

"La actividad minera ha estado y está siempre relacionada con la aparición de peligros que se denominan comúnmente peligros mineros. Un caso especial de este tipo de amenaza es el peligro sísmico que se produce con frecuencia en muchas minas subterráneas. El peligro sísmico es el más difícil de detectar y predecir de los peligros naturales y, en este sentido, es comparable a un terremoto. Los sistemas de control sísmico y sismoacústico cada vez más avanzados, y permiten una mejor comprensión de los procesos del macizo rocoso y la definición de métodos de predicción del riesgo sísmico.

Sin embargo, la precisión de los métodos creados hasta ahora dista mucho de ser perfecta. La complejidad de los procesos sísmicos y la gran desproporción entre el número de eventos sísmicos de baja energía y el número de fenómenos de alta energía hace que las técnicas estadísticas sean insuficientes para predecir peligro sísmico. Por lo tanto, es esencial buscar nuevas oportunidades de mejorar la predicción de la peligrosidad utilizando también métodos de aprendizaje automático."[1]

El dataset elegido se corresponde con un conjunto de datos relacionados con el sector minero, concretamente fueron obtenidos en una mina de carbón de Polonia. El dataset describe, por cada muestra, una situación en la que se ha dado un evento sísmico, describiendo factores que determinan la fuerza del mismo, como los diferentes niveles de energía por cada evento, y otros parámetros.

En estos casos, y según se especifica en el enunciado, la estadística es inefectiva para la predicción de eventos, por lo que se requiere del uso de técnicas más avanzadas. A partir de estos datos, se busca determinar, a partir de técnicas de aprendizaje, predecir futuras situaciones, para discernir si son situaciones de peligro o no peligro.

2 Objetivos

El objetivo de este estudio es determinar, a partir de los atributos de la población, el atributo class de futuras muestras, partiendo del aprendizaje de los datos de la tabla.

La variable class determina, de manera dicotómica, si el evento registrado supone un evento de riesgo o no. Los valores a 0 determinan que la situación no es de riesgo, mientras que los valores a 1 determinan que sí hay riesgo.

3 Dataset

La tabla de descripción del dataset que se presenta en la documentación es la siguiente:

Data Set Characteristics	Multivariate
Number of Instances	2584
Area	N/A
Attribute Characteristics	Real
Number of Attributes	19
Date Donated	2013-04-03
Associated Tasks	Classification
Missing Values?	N/A
Number of Web Hits	85975

Veamos las descripciones de cada uno de los atributos. Recuperamos estas descripciones de la página de la UCI:

- seismic: Resultado de la evaluación de la peligrosidad sísmica por turnos en la explotación minera obtenida por el método sísmico (a - ausencia de peligro, b - peligro bajo, c - peligro alto, d - estado de peligro)
- seismoacoustic: Resultado de la evaluación de la peligrosidad sísmica por turnos en la explotación minera obtenido por el método sismoacústico

- shift: Información sobre el tipo de turno (W - turno de obtención de carbón, N - turno de preparación)
- Genergy: Energía sísmica registrada en el turno anterior por el geófono más activo (GMax) de los geófonos que monitorizan el tajo largo
- gpuls: Número de pulsos registrados en el turno anterior por el GMax
- gdenergy: Desviación de la energía registrada en el turno anterior por el GMax con respecto a la energía media registrada durante ocho turnos anteriores
- gdpuls: Desviación del número de impulsos registrados en el turno anterior por GMax con respecto al número medio de impulsos registrado durante ocho turnos anteriores
- ghazard: Resultado de la evaluación de la peligrosidad sísmica en el turno de trabajo de la mina obtenido por el método sismoacústico basado en el registro procedente únicamente de GMax
- nbumps: Número de golpes sísmicos registrados en el turno anterior
- nbumps2: El número de golpes sísmicos (en el rango de energía $[10^{\mathbf{2}},10^{\mathbf{3}}]$) registrados en el turno anterior
- nbumps3: El número de golpes sísmicos (en el rango de energía $[10^{\mathbf{3}},10^{4})$) registrados dentro del turno anterior
- nbumps4: El número de golpes sísmicos (en el rango de energía $[10^{4},10^{5})$) registrados en el turno anterior
- nbumps5: El número de golpes sísmicos (en el rango de energía $[10^{5},10^{6})$) registrados en el último turno
- nbumps6: El número de golpes sísmicos (en el rango de energía $[10^{6},10^{7})$) registrados en el turno anterior
- nbumps7: El número de golpes sísmicos (en el rango de energía $[10^{7},10^{8})$) registrados en el turno anterior
- nbumps89: El número de golpes sísmicos (en el rango de energía $[10^{8},10^{10})$) registrados en el turno anterior
- energy: Energía total de los golpes sísmicos registrados en el turno anterior
- maxenergy: La energía máxima de las protuberancias sísmicas registradas en el turno anterior
- class: El atributo de decisión o atributo a predecir: "1" significa que se ha producido un golpe sísmico de alta energía en el turno siguiente ("estado peligroso"), "0" significa que no se produjeron golpes sísmicos de alta energía en el turno siguiente ("estado no peligroso")

El objetivo es predecir si la siguiente estimación tendrá valor class = 0 o class = 1 (no peligro/peligro).

Observamos que el dataset consta de 19 atributos y 2584 observaciones. Todos los atributos son numéricos menos seismic, seismoacustic y shift, que son categóricos. La variable objetivo es class, que es un atributo booleano que determina si existe peligro (=1) o no (=0).

Una vez estudiados los datos (véase las tablas estadísticas y gráficos en el archivo jupyter notebook) podemos mencionar lo siguiente:

Observamos que, a priori, ninguno de ellos consta de una distribución normal. Las variables categóricas (seismic, seismoacustic y shift) se corresponden con la peligrosidad de la actividad y el tipo de actividad (shift/coal-getting). Las variables numéricas se refieren a las mediciones registradas de energía, números de baches sísmicos registrados en rangos de energía, estadísticas de energía promedio y máximos de energía y, finalmente, resultado de peligrosidad / no peligrosidad.

Observamos también que los valores de los atributos no se reparten de forma homogenea, sino que se concentran en valores determinados, por eso encontramos picos tan altos en los histogramas.

En cuanto a la correlación de las variables, encontramos que la variable objetivo "class" no tiene una correlación buena con ninguna de las demás variables. Los valores de correlación más cercanos se encuentran entre el 0.2 y 0.3. Por otra parte, sí que observamos buenas correlaciones entre otras variables, siendo la correlación más fuerte entre maxenergy y energy, con un 1 de correlación; rbumps3 con rbumps, con un valor aproximado entre 0.8 y 0.9; y rbumps2 y rbumps, con un valor aproximado de 0.8.

4 Preprocesamiento de los datos

5 Referencias y estudio bibliográfico

Utilizamos citación en formato APA para esta práctica:

1. UCI Machine Learning Repository: seismic-bumps Data Set. (s. f.). Recuperado 24 de octubre de 2022, de http://archive.ics.uci.edu/ml/datasets/seismic-bumps
2. Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn: Machine learning without learning the machinery. GetMobile Mobile Computing and Communications, 19(1), 29–33.
3. Bilogur, A. (s/f). missingno: Missing data visualization module for Python.
4. Seaborn: Statistical data visualization — seaborn 0.12.1 documentation. (s/f). Pydata.org. Recuperado el 5 de diciembre de 2022, de https://seaborn.pydata.org/
5. Pandas. (s/f). Pydata.org. Recuperado el 5 de diciembre de 2022, de https://pandas.pydata.org/
6. Introduction — statsmodels. (s/f). Statsmodels.org. Recuperado el 5 de diciembre de 2022, de https://www.statsmodels.org/stable/index.html
7. Bisong, E. (2019). NumPy. En Building Machine Learning and Deep Learning Models on Google Cloud Platform (pp. 91–113). Apress.
8. Matplotlib — visualization with python. (s/f). Matplotlib.org. Recuperado el 5 de diciembre de 2022, de https://matplotlib.org/
9. Yellowbrick: Machine Learning Visualization — Yellowbrick v1.5 documentation. (s/f). Scikit-yb.org. Recuperado el 5 de diciembre de 2022, de https://www.scikit-yb.org/en/latest/
10. Surhone, L. M., Timpledon, M. T., Marseken, S. F. (Eds.). (2010). Scipy. Betascript Publishing.