UNIVERSIDAD DE CANTABRIA

Aplicación de técnicas de Machine Learning a la predicción de fallos de discos mediante el uso de Spark

Máster en Ciencia de Datos

Alejandro Villanueva Noriega

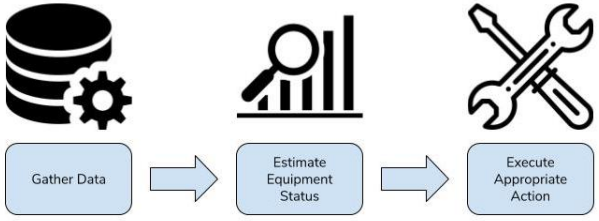
Facultad de Ciencias

27 de Septiembre de 2021

Mantenimiento Predictivo

¿Cómo podemos definir el mantenimiento predictivo? ¿A qué se debe este auge?

- "Run to Failure" R2F $\rightarrow$ Mantenimiento correctivo
- Mantenimiento preventivo
- Mantenimiento predictivo



Hard Disk Drives

- Los discos duros fueron inventados hace más de 50 años y se utilizan en los ordenadores personales desde mediados de la década de 1980.
- Memoria no volátil. No necesita de un aporte energético constante para conservar dicha información.
- Almacenamiento en la nube $\rightarrow$ importante reto al que se enfrentan las empresas tecnológicas (HDDs 80 % de los fallas en Data Centers)
- SMART: Self Monitoring Analysis and Reporting Technology

Series temporales

Una serie temporal multivariante o vectorial se representa de la siguiente forma:

$$\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N \; ; \; (\boldsymbol{y}_t)_{t=1}^{N} \; ; \; (\boldsymbol{y}_t : t = 1, ..., N)$$

dónde $\boldsymbol{y}_t \equiv [y_{t1}, y_{t2}, ..., y_{tM}]'$ $(M \geq 2)$ es la observación número t $(1 \leq t \leq N)$ de la serie, N es el número de observaciones y M es la variable observada. Por consecuente, esta serie puede venir representada por una matriz $\boldsymbol{Y}$ de orden $N \times M$:

$$\boldsymbol{Y} \equiv \begin{bmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_N' \end{bmatrix} \equiv \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1M} \\ y_{21} & y_{22} & \cdots & y_{2M} \\ \vdots & \vdots & \ldots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NM} \end{bmatrix}$$

dónde y_{tj} es la observación número t $(1 \leq t \leq N)$ de la serie sobre la variable número j $(1 \leq j \leq M)$.

Estado del arte I

- Enfoque por clasificación binaria: clasificar cada punto de la TS como falla o "saludable". Más extendido y utilizado.
- Enfoque de detección de anomalías.
- Enfoque de predicción del tiempo hasta el fallo como variable continua.

Estado del arte II : Clasificación Binaria

- Enfoque por clasificación binaria: clasificar cada punto de la TS como falla o "saludable". Más extendido y utilizado.
  - Botezatu et al. [1]: Transfer-Learning y Regularized Greedy Forests. 10-15 días antelación
  - Xiao et al. [2]: Random Forest. Recopilación de mas datos. Online
  - Murray et al. [3]: Aprendizaje de múltiples instancias y el clasificador Naive-Bayes.
  - Sun et al.[4]: Redes neuronales convolucionales

Estado del arte III : Detección de anomalías & Predicción tiempo al fallo

- Enfoque de detección de anomalías
  - Wang et al. [5]: Distancia de Mahalanobis, transformación de Box-Cox y pruebas de razon de verosimilitud generalizada.
  - Hamerly y Elkan [6]: Mezcla de un clasificador NaiveBayes con maximización de expectativas y otro clasificador Naive-Bayes para detectar las anomalías.
  - Aussel et al. [7]: SVM, Random Forest y XGBoost.
  - Zhu et al. [8]: Red neuronal recurrente y un modelo basado en SVMs para la predicción de la salud del disco duro.
- Enfoque de predicción del tiempo hasta el fallo como variable continua
  - Chaves et al. [9]: Red bayesiana para intentar obtener este tiempo hasta el fallo.

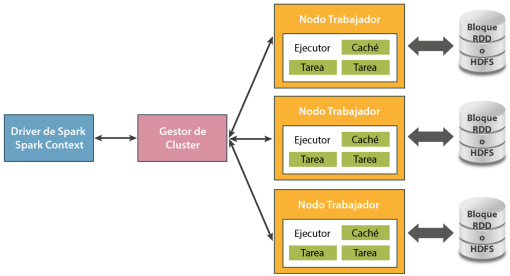
Algoritmos de clasificación

- Regresión logística binaria
- Linear Support Vector Machine
- Random Forest

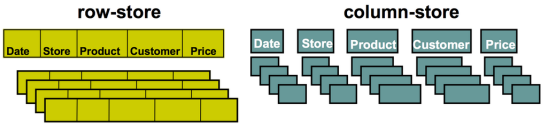
* Reducción de la dimensión: Análisis de componentes principales
* ¿Dataset? 30 Atributos SMART de 32 discos independientes cada 15 minutos desde el 01/01/2018 hasta el 01/06/2021.

Spark: Computación distribuida

AMPLab, Univ. Berkeley 2009. MLlib, Spark-SQL, Structured Streaming, GraphX

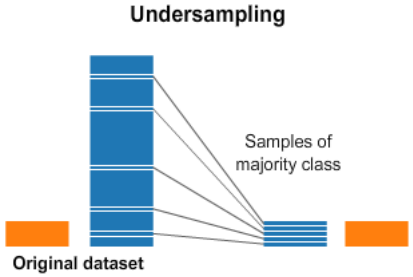


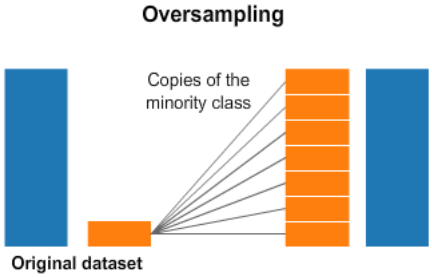
Parquet: orientado a columnas, autodescriptivo, mayor rendimiento (subset columnas memoria)



Imbalanced data classification

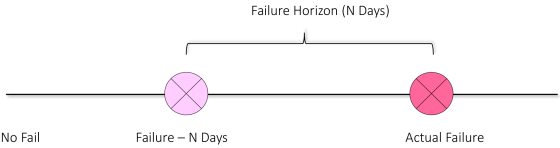
- Clase A $\ll$ Clase B (detección de anomalías, diagnósticos médicos, etc.)
- Sesgo hacia clase mayoritaria $\rightarrow$ difícil generalización
- Métricas de validación en clasificación desequilibrada: ROC, Matriz de confusión, G-Mean, F-Score
- Solución: Reequilibrar la distribución de clases mediante muestreo. SMOTE / Submuestreo





Metodología

- Problema: predicción a N días de antelación. Horizonte o lag $= 1, 2, 7$ días. Modelos independientes.
- Fallo $1 \rightarrow t \in [day_{fallo} - lag, day_{fallo}]$, "Saludable" $0 \rightarrow t < day_{fallo}$



- Distintas distribuciones de las clases para los diferentes lags.

Clase	$Lag = 7$ días	$Lag = 2$ días	$Lag = 1$ días
1 - Fallo	14,415	3,916	1,920
0 - No-Fallo	1,594,700	1,605,199	1,607,195
Proporción	1:110	1:410	1:840

Table: Distribución y proporción de las clases para los distintos horizontes analizados.

Procesado de datos

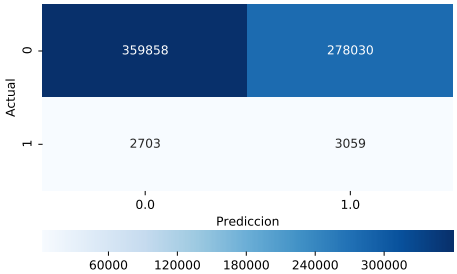
- SMOTE $\rightarrow$ Submuestreo aleatorio sin reemplazo "mejoraba" resultados

Clase	*Lag = 7* días	*Lag = 2* días	*Lag = 1* días
1 - Fallo	8,653	2,356	1,127
0 - No-Fallo	8.469	2,307	1,079

Table: Distribución y proporción de las clases para los distintos horizontes analizados.

- PCA: ciertas SMARTs que distorsionaban generalización del modelo $\rightarrow$ SMART más destacadas por Literatura.
- Entrenamiento 0.6 aleatorio sobre el Dataframe total.

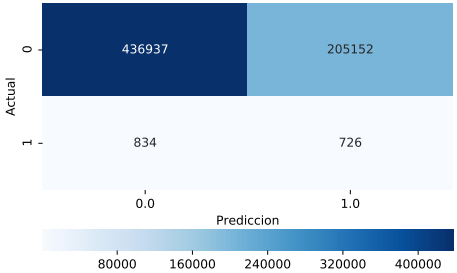
Análisis y resultados I : lag = 7 días → Random Forest



Clase	*Precision*	*F-Score*	*Recall*
1 - Fallo	0.011	0.021	0.531
0 - No-Fallo	0.993	0.719	0.564

Table: *Sensitivity* = 0.564; *Specificity* = 0.531; AUC = 0.548; *G-Mean* = 0.547 .

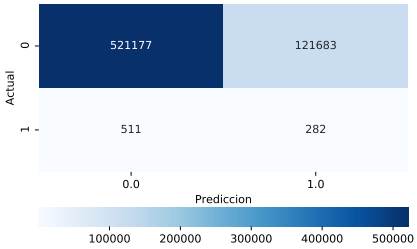
Análisis y resultados II: lag = 2 días → Random Forest

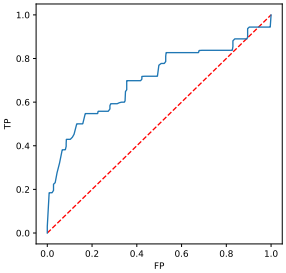


Clase	*Precision*	*F-Score*	*Recall*
1 - Fallo	0.004	0.007	0.465
0 - No-Fallo	0.998	0.810	0.681

Table: *Sensitivity* = 0.465; *Specificity* = 0.681; AUC = 0.611; *G-Mean* = 0.563 .

Análisis y resultados III: lag = 1 día → Regresión Logística Binaria





Clase	*Precision*	*F-Score*	*Recall*
1 - Fallo	0.002	0.004	0.356
0 - No-Fallo	0.999	0.895	0.811

Figure: *Sensitivity* = 0.356; *Specificity* = 0.811; *G-Mean* = 0.537 ; AUC = 0.637

Conclusiones

- Los resultados de los diferentes modelos evidencian la dificultad de los problemas de clasificación imbalanceada, obteniendo modelos que no tienen una capacidad óptima de predicción, con un gran número de falsas alarmas.
- Se ha evidenciado la potencia de procesamiento de PySpark para el proceso de la ETL, reduciendo los tiempos de ejecución.
- Se ha comprendido el principal problema del mantenimiento predictivo: predicción a N días de antelación.
- Dificultad de predicción de las fallas en discos dada la poca estandarización de las distintas medidas SMART.
- Una posible mejora: modelos con memoria como las LSTM, o añadir agregaciones $\rightarrow$ con el fin de dar cuenta del desgaste acumulado en los discos.

Referencias

Mirela Madalina Botezatu, Ioana Giurgiu, Jasmina Bogojeska, and Dorothea Wies-mann. Predicting disk replacement towards reliable data centers. InProceedings ofthe 22nd ACM SIGKDD International Conference on Knowledge Discovery and DataMining, pages 39–48, 2016.

Jiang Xiao, Zhuang Xiong, Song Wu, Yusheng Yi, Hai Jin, and Kan Hu. Disk failureprediction in data centers via online learning. InProceedings of the 47th InternationalConference on Parallel Processing, pages 1–10, 2018.

Joseph F Murray, Gordon F Hughes, Kenneth Kreutz-Delgado, and Dale Schuurmans.Machine learning methods for predicting failures in hard drives: A multiple-instanceapplication.Journal of Machine Learning Research, 6(5), 2005.

Xiaoyi Sun, Krishnendu Chakrabarty, Ruirui Huang, Yiquan Chen, Bing Zhao, HaiCao, Yinhe Han, Xiaoyao Liang, and Li Jiang. System-level hardware failure predictionusing deep learning. 2019.

Yu Wang, Eden WM Ma, Tommy WS Chow, and Kwok-Leung Tsui. A two-stepparametric method for failure prediction in hard disk drives.IEEE Transactions onindustrial informatics, 10(1):419–430, 2013.

Greg Hamerly, Charles Elkan, et al. Bayesian approaches to failure prediction for diskdrives. InICML, Citeseer, 2001.

Nicolas Aussel, Samuel Jaulin, Guillaume Gandon, Yohan Petetin, Eriza Fazli, andSophie Chabridon. Predictive models of hard drive failures based on operational data, 2017.

Bingpeng Zhu, Gang Wang, Xiaoguang Liu, Dianming Hu, Sheng Lin, and JingweiMa. Proactive drive failure prediction for large scale storage systems. In2013 IEEE29th symposium on mass storage systems and technologies (MSST), pages 1–5. IEEE,2013.

Iago C Chaves, Manoel Rui P de Paula, Lucas GM Leite, Joao Paulo P Gomes, andJavam C Machado. Hard disk drive failure prediction method based on a bayesiannetwork. 2018

Gracias por su atención.