Laredo: democratización de análisis de flujos de datos para el mantenimiento predictivo*

Ricardo Dintén[0000−0002−3163−0473] and Marta Zorrilla[0000−0002−0475−8834]

Grupo ISTR, Universidad de Cantabria, Santander 39005, ESPAÑA
{ricardo.dinten,marta.zorrilla}@unican.es

Resumen La llegada del IoT y la IA a las fábricas permitirá optimizar sus procesos productivos. Pero la complejidad del proceso de minería de datos, la escasez de profesionales cualificados y la falta de herramientas dirigidas a no expertos, frena su despliegue en el sector industrial. Por ello, en este trabajo se analiza la literatura científica relacionada con herramientas para el análisis de flujos de datos y se plantea una propuesta de servicio, dirigido a usuarios no expertos (no científicos de datos), que permita acercar la construcción de *workflows* científicos escalables y distribuidos enfocados al mantenimiento predictivo y prescriptivo para su despliegue sobre la arquitectura industrial RAI4.0.

Keywords: Data stream mining · Non-expert data scientists · Predictive Maintenance

1. Introducción y Objetivos

Durante los últimos años, las tecnologías IoT y la inteligencia artificial, así como, la proliferación de servicios en la nube han supuesto una revolución en la Industria. Tanto es así, que según un informe elaborado por Microsoft [7] en 2022, el 72 % de las empresas de fabricación a nivel global está desarrollando estrategias para adoptar estas tecnologías. Uno de los aspectos más relevantes y en los que se ha centrado gran parte de la investigación es en el mantenimiento predictivo y prescriptivo como evolución del mantenimiento preventivo. Ambos emplean técnicas avanzandas de minería de datos: el primero, para precedir futuros fallos o estimar la vida útil de la máquinaria, y el segundo, para seleccionar la acción de mantenimiento más apropiada para hacer frente a esa avería. El uso de estas técnicas permiten planificar el mantenimiento, evitar paradas del sistema no previstas y reducir el coste global. Sin embargo, la definición, especificación, diseño e implementación de este tipo de sistemas no es trivial, y requiere de expertos para poder aprovechar todo su potencial. Pero este perfil profesional es escaso. Según [7], ocho de cada diez fabricantes tienen dificultades para encontrar perfiles relacionados con ciencia de datos, IA y ciberseguridad. Esto justifica el

* Financiado por MCIN/ AEI/10.13039/501100011033/ FEDER bajo la subvención PID2021-124502OB-C42 (PRESECREL) y por la Ayuda Concepción Arenal (BOC 18-10-2021).

desarrollo de heramientas que democraticen el uso de estas técnicas, no solo en las grandes corporaciones sino en todo el tejido empresarial incluidas las PYMEs, que en España representan el 99 % del total, contando el 93 % con menos de 10 empleados.[1]

Este trabajo presenta una linea de investigación de un proyecto de tesis en desarrollo, que busca democratizar el análisis de flujos de datos mediante el desarrollo de un servicio que asista a usuarios no expertos en las fases de diseño, construción, configuración e implantación de *data pipelines* enfocados al mantenimiento predictivo y prescriptivo.

2. Trabajos relacionados

En este apartado se presenta un breve análisis de algunas características de los sistemas de asistencia al análisis de datos encontrados en la literatura después de realizar una búsqueda sistemática para el periodo comprendido entre enero de 2010 y julio de 2022 en tres bases de datos: SCOPUS, Web of Science e INSPEC. Una vez filtrados los resultados de la búsqueda, nos quedamos con solo 8 trabajos.

En su análisis se ha prestado atención a los siguientes aspectos: uso de tecnologías big data y capacidad de distribución; cobertura completa del estándar de facto para el desarrollo de proyectos de análisis de datos, CRISP-DM; tipo de interfaz empleada para el desarrollo del análisis de datos y el nivel de experiencia requerido para utilizarlo; si son de propósito general o específico; y, si incluyen soporte para el análisis de flujos de datos (*data stream mining*).

De los artículos analizados, [3,8,9] soportan uso de tecnologías Big Data en concreto en [3,9] hacen uso de Apache Spark para distribuir la ejecución de los modelos predictivos. El resto [1,2,4,5,6] no ofrecen la posibilidad de distribuir el entrenamiento o las prediciones sino que lo ejecutan en un único nodo.

Por otro lado, ninguno de los sistemas aborda todas las etapas de CRISP-DM, en particular, la mayoría deja la etapa de evaluación fuera del alcance de la herramienta. Esta etapa es importante, ya que en ella se pueden detectar modelos con bajo rendimiento que deben ser rediseñados y reentrenados.

En cuanto a la interfaz empleada se distinguen 4 alternativas: (i) biblioteca de código, la cual no es apta para usuarios no expertos, ya que requiere que los operadores se utilicen mediante algún lenguaje de programación; (ii) interfaz basada en bloques, que evita al usuario la necesidad de escribir código; sin embargo, requiere que conozca la función de cada uno de los bloques y en qué orden deben conectarse para hacer funcionar el *pipeline*; (iii) interfaz de tipo asistente, ésta no requiere codificación y guía al usuario a través de todas las etapas necesarias durante la construción de su sistema predictivo; (iv) interfaz que pide al usuario el conjunto de datos y realiza todo el proceso de forma transparente, sin interacción del usuario. Los asistentes y los sistemas automáticos se consideran las opciones más apropiadas para usuarios no expertos, siendo los útimos

[1] `https://plataformapyme.es/Publicaciones/Marco%20Estrat%C3%A9gico%20de%20la%20PYME/Informe_Seguimiento_Anual_2021.pdf`

adecuados solamente para problemas muy concretos cuya solución se tiene muy probada. En los trabajos analizados, 5 de los 8 han optado por una interfaz de tipo asistente y solo uno es totalmente automático.

Ninguna de las herramientas anteriores tiene como propósito el mantenimiento predicitivo, por tanto, se ha analizado la versatilidad que ofrecen para ser adaptadas a otras tareas. De las analizadas, [1,2,6,8] tienen un propósito muy específico, limitando tanto el formato de los datos soportado como el tipo de operaciones que se pueden realizar sobre ellos. [3,4,5,9] son herramientas genéricas de análisis de datos, por lo que aunque podrían emplearse para mantenimiento predictivo, cuentan con algoritmos y operadores no aplicables en el contexto del mantenimiento predictivo y podrían generar confusión en usuarios no expertos.

Otro aspecto detectado es que solo una [3] proporciona soporte para aprendizaje en linea sobre los flujos de datos. El resto [1,2,4,5,6,8,9], se limitan a analizar conjuntos de datos estáticos o entrenar modelos predictivos que luego pueden desplegarse en algún servicio web o en un motor de procesamiento de eventos para realizar prediciones sobre los datos nuevos. Sin embargo, no ofrecen la posibilidad de extraer patrones o reentrenar los modelos con esos nuevos datos.

Por último, se analizaron también tres de las herramientas comerciales más populares como son: Rapidminer, Weka o Knime. Estas son herramientas mucho más completas que permiten el análisis de conjuntos de datos estáticos así como de flujos de datos. Optan por una estrategia de diseño mediante bloques y cubren gran parte de las etapas de CRISP-DM. Sin embargo, se descartan como opción apropiada para no expertos, ya que el gran número de algoritmos disponibles (más de 1500 en el caso de Rapidminer) abrumaría a un usuario con pocos conocimientos de minería de datos, provocando que no sepa qué operaciones debe emplear o que haga un uso incorrecto de las mismas.

3. Propuesta y trabajo en curso

A continuación, se muestran los requisitos que debe cumplir el servicio propuesto:

- El servicio deberá guiar al usuario por cada una de las fases del proceso de creación de *pipelines* de análisis de flujos de datos mediante un asistente o *wizard*. Estas fases serán: analisis exploratorio de datos, preparación de los datos, modelado, evaluación y despliegue.
- Basándose en los datos y el problema a resolver, el servicio hará recomendaciones sobre las transformaciones y modelos predictivos más apropiados para lograr su propósito.
- La plataforma dará soporte a las problemáticas más comunes en el ámbito del mantenimiento predictivo que son: estimación de la vida útil restante, detección de anomalías, y diagnóstico de fallos.
- La plataforma empleará tecnologías Big Data como Spark, Storm o Flink que permitan distribuir las tareas de entrenamiento y predicción de los modelos predictivos.

Actualmente, estamos trabajando en la parte experimental de creación de *pipelines* para los diferentes tipos de problema, con el objetivo de desarrollar bloques de código reutilizables. En paralelo, se están valorando los beneficios de incluir aprendizaje federado y *ensemble* como técnicas para distribuir el aprendizaje y mejorar los resultados, y se están recopilando conjuntos de datos con los que poder validar los desarrollos realizados.

Referencias

1. Cachucho, R., Liu, K., Nijssen, S., Knobbe, A.: Bipeline: A web-based visualization tool for biclustering of multivariate time series (2016). https://doi.org/10.1007/978-3-319-46131-1_3
2. Frías, M., Iturbide, M., Manzanas, R., Bedia, J., Fernández, J., Herrera, S., Cofiño, A., Gutiérrez, J.: An R package to visualize and communicate uncertainty in seasonal climate prediction. Environmental Modelling & Software **99**, 101–110 (jan 2018). https://doi.org/10.1016/j.envsoft.2017.09.008
3. Giatrakos, N., Arnu, D., Bitsakis, T., Deligiannakis, A., Garofalakis, M., Klinkenberg, R., Konidaris, A., Kontaxakis, A., Kotidis, Y., Samoladas, V., Simitsis, A., Stamatakis, G., Temme, F., Torok, M., Yaqub, E., Montagud, A., Ponce De León, M., Arndt, H., Burkard, S.: INforE: Interactive Cross-platform Analytics for Everyone. In: International Conference on Information and Knowledge Management, Proceedings. pp. 3389–3392. Athena Research Center and Technical University of Crete, Maroussi. Athens, Greece (2020). https://doi.org/10.1145/3340531.3417435
4. Han, J., Park, K.S., Lee, K.M.: An Automated Machine Learning Platform for Non-experts. In: ACM International Conference Proceeding Series. pp. 84–86. Department of Computer Science, Chungbuk National University, Cheongju,Chungbuk, South Korea (2020). https://doi.org/10.1145/3400286.3418276
5. Karatzoglidi, M., Kerasiotis, P., Kantere, V.: Automated energy consumption forecasting with enforce. Proceedings of the VLDB Endowment **14**(12), 2771–2774 (2021). https://doi.org/10.14778/3476311.3476341
6. La Ferlita, A., Alaimo, S., Di Bella, S., Martorana, E., Laliotis, G.I., Bertoni, F., Cascione, L., Tsichlis, P.N., Ferro, A., Bosotti, R., Pulvirenti, A.: RNAdetector: a free user-friendly stand-alone and cloud-based system for RNA-Seq data analysis. BMC Bioinformatics **22**(1), 298 (dec 2021). https://doi.org/10.1186/s12859-021-04211-7
7. Ladha, P.: Top 6 findings from iot signals: Manufacturing spotlight (Aug 2022), https://www.microsoft.com/en-us/industry/blog/manufacturing/2022/08/11/top-6-findings-from-iot-signals-manufacturing-spotlight/
8. Mousheimish, R., Taher, Y., Zeitouni, K.: Demo: Complex event processing for the non-expert with auto CEP. In: DEBS 2016 - Proceedings of the 10th ACM International Conference on Distributed and Event-Based Systems. pp. 340–343. DAVID Laboratory, University of Versailles UVSQ, Versailles, 78000, France (2016). https://doi.org/10.1145/2933267.2933296
9. Shahoud, S., Winter, M., Khalloof, H., Duepmeier, C., Hagenmeyer, V.: An extended Meta Learning Approach for Automating Model Selection in Big Data Environments using Microservice and Container Virtualizationz Technologies. Internet of Things **16**, 100432 (dec 2021). https://doi.org/10.1016/j.iot.2021.100432