Minería de flujos de datos en entornos heterogéneos y distribuidos: aplicación en la Industria 4.0

Ricardo Dintén[0000-0002-3163-0473], Patricia López[0000-0002-9562-6342], Juan Yebenes[0000-0002-2906-5065] and Marta Zorrilla[0000-0002-0475-8834]

Grupo de Ingeniería de Software y Tiempo Real, Universidad de Cantabria, Santander 39005, ESPAÑA
{ricardo.dinten, lopezpa, marta.zorrilla}@unican.es,
juan-rafael.yebenes@alumnos.unican.es

Resumen. Uno de los principales objetivos de la Industria 4.0 es lograr la necesaria integración horizontal y vertical del sistema de producción. Para ello es necesario desplegar una plataforma digital que integre y procese la ingente cantidad de datos generados en el entorno. Mucha de esta información procede del IoT, y, en concreto, corresponde a sensores que emiten flujos continuos de datos cuyo análisis mediante técnicas de minería de datos permitiría mejorar los procesos industriales, como por ejemplo construyendo modelos dirigidos al mantenimiento preventivo y predictivo de los sistemas físicos, donde aún hay retos abiertos. El objeto de este artículo es describir el punto de partida de esta investigación que es el resultado de un proyecto del plan nacional y discutir su extensión señalando las líneas de trabajo que se pretenden abordar y los resultados que se persigue conseguir para contribuir al avance de la I4.0.

Palabras clave: Big data, Inteligencia artificial, Cloud computing, Arquitecturas intensivas en datos, IoT

1 Antecedentes

La «Industria 4.0 (I4.0)», o también conocida como la «Cuarta Revolución Industrial», persigue no solo la digitalización de los procesos sino su optimización, control y automatización a partir del análisis de la ingente cantidad de datos disponibles en el entorno.

En un escenario I4.0, las máquinas son capaces de comunicarse entre sí para recibir o transmitir información y ejecutar acciones que conlleven a procesos productivos más depurados y a la fabricación de productos personalizados gracias al despliegue de sensores (IoT). Estos sensores, que están embebidos en dispositivos y máquinas, proporcionan enormes cantidades de datos que, analizados mediante la aplicación de técnicas inteligentes, permitirán obtener resultados significativos en la mejora de la eficiencia operativa y el desempeño organizacional, su sostenibilidad y resiliencia.

Para poder integrar y procesar toda la información generada en el entorno se requiere una plataforma tecnológica adaptada y adecuada para abordar la velocidad, variedad y el volumen ingente de datos, así como tecnologías y herramientas que puedan explotar el paralelismo, la computación distribuida y la escalabilidad dinámica con objeto de poder acotar los tiempos de respuesta (tiempo real no estricto). Al respecto, nuestro

grupo de investigación ha diseñado RAI4.0 [1], una arquitectura de referencia centrada en el dato que satisface los requisitos de la I4.0 dentro del proyecto nacional (TIN2017-86520-C3-3 R). Sobre ella se han instanciado varios casos de uso con los que se ha demostrado su viabilidad y adecuación técnica. En [2] se realizaron benchmarks sobre la escalabilidad y rendimiento de los procesos desplegados en la arquitectura; en [3] se caracterizaron tecnologías de procesamiento de datos en streaming; en [4] se estudiaron y extrajeron criterios de diseño para el gestor de datos distribuido y escalable Cassandra y, finalmente, en [5] se abordó la construcción y despliegue de un sistema de mantenimiento predictivo mediante la aplicación de técnicas de minería de datos. Este último trabajo ha permitido conocer y experimentar las limitaciones actuales en el campo conocido como data stream mining [6] y la necesidad de ahondar en su investigación.

Por último, y persiguiendo que los datos sean una ventaja competitiva para la industria, en paralelo se ha definido un marco para la construcción de sistemas de gobernanza de datos en la I4.0 [7][8], que aún es necesario desarrollar para que permita su efectiva instanciación en las organizaciones, ofreciendo modelos y herramientas que ayuden a su adopción.

2 Objetivos

Teniendo en cuenta los antecedentes mencionados en la sección anterior, se plantean los siguientes objetivos:

- Extender la arquitectura RAI 4.0 incluyendo soporte para desplegar aplicaciones virtualizadas en entornos mixtos (fog/edge/cloud). Con esta extensión se espera reducir latencias y adecuarse a requisitos temporales (tiempo real no estricto).
- Aplicar técnicas de minería de flujos de datos a casos de uso reales, haciendo uso de la plataforma desarrollada, así como, desarrollar herramientas que permitan aplicar estas técnicas a usuarios no expertos. Se espera que esto permita sacar provecho de sus datos a un mayor número de empresas, especialmente a las que no pueden permitirse la contratación de personal específico para estas tareas.
- Desarrollar el framework de gobernanza de datos para poder instanciarlo en organizaciones reales, de manera que les aporte una serie de modelos y herramientas con los que conseguir una ventaja competitiva.

3 Líneas de trabajo y enfoque para su desarrollo

Para la consecución de los objetivos que se pretenden conseguir se requiere avanzar en diferentes líneas de trabajo.

- Línea 1. Arquitectura para aplicaciones de uso intensivo de datos para I4.0

Actualmente se está trabajando en la extensión de la arquitectura RAI4.0 para soportar la virtualización y el despliegue en entornos mixtos (fog/edge/cloud). Esto conlleva, por una parte, la ampliación del metamodelo actual [1] que da soporte al desarrollo de modelos de plataformas digitales conformes a la arquitectura y la actualización

de la herramienta que facilita su configuración y despliegue en base a dichos modelos; y, por otra, el desarrollo de diferentes casos de uso para su validación. Al respecto, ya se han realizado avances en los dos primeros elementos y se están refactorizando los casos de uso. En particular, se está adaptando el caso de uso de análisis de contaminación presentado en [1] para su despliegue en un entorno mixto basado en tecnologías Docker o Kubernetes. Con este despliegue se pretende validar la capacidad del metamodelo para diseñar y configurar aplicaciones intensivas de datos en entornos mixtos basados en contenedores y probar el desempeño de la herramienta, así como medir el grado de reutilización de los elementos del modelo de la aplicación (caso de uso) para diferentes configuraciones de despliegue.

- Línea 2. Minería de flujos de datos continuos

El proceso de análisis y extracción de conocimiento no es una tarea trivial y se enfrenta a diferentes retos. En concreto, para el análisis de flujos de datos se pueden mencionar [9]: i) la detección del concept drift, esto es, cambios en las propiedades estadísticas de los modelos que hacen que el predictor comience a fallar; ii) la paralelización del proceso de entrenamiento para construir modelos en tiempos acotados, ya que el volumen de datos ingente hace inviable el entrenamiento en una sola máquina; y iii) los problemas inherentes al preprocesado e integración de fuentes de datos dinámicas como son la preservación del orden de llegada de los datos, tratar con datos espúreos, faltantes o desincronizados, etc.

Para abordar el primer reto, inicialmente se trabajará con técnicas existentes, de forma independiente, así como combinada (*ensemble*), y proponiendo modificaciones de comportamiento de los algoritmos para obtener una precisión satisfactoria. El segundo reto afecta a la parte de entrenamiento de los modelos. Inicialmente se abordará mediante la construcción paralela de modelos para su posterior ensamblado y la detección de necesidad de reentrenamientos y su actualización como una combinación entre modelos. Por último, el tercer reto se abordará mediante la correcta parametrización de los *stream processor* (tamaño de ventana, gestión del estado, tolerancia a fallos y orden de eventos).

Posteriormente, se empaquetarán y diseñarán *workflows* científicos que permitan automatizar y democratizar el uso de estas técnicas a usuario no expertos.

- Línea 3. Gobierno del dato

Los casos de uso sobre los que se trabajará estarán principalmente enmarcados en el mantenimiento preventivo y predictivo del sector industrial y logístico fruto de nuestros convenios de colaboración. En esta línea se trabajará en el diagnóstico e identificación de propuestas de mejoras para el gobierno de las fuentes de datos de la organización, dándose así los pasos hacia la instanciación de su marco de gobernanza. Se perseguirá que este sea mantenible y extensible y que permita a las organizaciones definir objetivos, políticas, estándares, roles y responsabilidades sobre los activos de datos para realizar su gestión y monitorización de forma inteligente y semiautomatizada, apoyándose en técnicas de IA y aplicando metodologías ágiles bajo el enfoque "Continuous Governance", "DataGovOps" y "Governance as code".

4 Conclusiones

En este trabajo se han expuesto las líneas de investigación en desarrollo en el grupo de investigación en el campo de la I4.0 con objeto de conseguir los siguientes resultados:

- Un metamodelo para el diseño de una arquitectura que de soporte al ecosistema tecnológico necesario para el IoT y su despliegue en entornos mixtos.
- La construcción de herramientas basadas en modelos para el diseño de aplicaciones intensivas en datos conformes al metamodelo RAI4.0 creado.
- La ejecución de *benchmarks* de técnicas de minería de datos e IA para casos de uso industriales. Disposición en abierto de conjuntos de datos etiquetados.
- La construcción de paquetes científicos que permitan a usuarios no expertos utilizar tecnologías IA para construir modelos.
- La formulación de modelos para gestionar los metadatos, el linaje, la calidad y la seguridad en los activos de datos, así como modelos de seguimiento y evaluación de la implantación del sistema de gobernanza basados en nuestro marco.

Reconocimientos. Este trabajo ha sido parcialmente apoyado por MCIN/ AEI /10.13039/501100011033/ FEDER "Una manera de hacer Europa" bajo la subvención TIN2017-86520-C3-3-R (PRECON-I4) y por la Ayuda Concepción Arenal del Programa de Personal Investigador en formación Predoctoral de la Universidad de Cantabria y el Gobierno de Cantabria (BOC 18-10-2021).

Referencias

1. López Martínez, P., Dintén, R., Drake, J.M. and Zorrilla, M.: A big data-centric architecture metamodel for Industry 4.0, Future Generation Computer Systems, Volume 125, pp. 263-284 (2021).
2. de la Rubia, L. M., Algorri, M., Zorrilla, M., Drake, J.M.: Descripción de pruebas de benchmark para plataformas de tercera generación. Jornadas de Ingeniería del Software y Bases de Datos (JISBD), Sevilla, Spain (2018).
3. Algorri M., Drake J. M., Zorrilla M.: Actualización reactiva de bases de datos usando cadenas de procesadores de flujo de datos. Jornadas de Ingeniería del Software y Bases de Datos (JISBD), La Laguna (Tenerife), Spain (2017).
4. Dintén R., Zorrilla M. E., López Martínez P. and Drake J.M.: Apache Cassandra como Gestor de Persistencia de la plataforma P3forI4 diseñada para la Industria 4.0. XXI Jornadas de Tiempo Real (JTR2019). Ferrol (2019).
5. Dintén, R.: Minería de flujos de datos para mantenimiento predictivo en entornos industriales. (Trabajo Fin de Máster). Universidad de Cantabria, Máster Universitario en Ingeniería Informática. Santander (2019).
6. Gama, J.: Knowledge discovery from data streams. CRC Press, Boca Raton (2010).
7. Yebenes, J.: Marco para la construcción de sistemas de Gobernanza de Datos en entornos de la Industria 4.0. Tesis Doctoral. Universidad de Cantabria (2022).
8. Zorrilla, M. Yebenes, J.: A reference framework for the implementation of data governance systems for industry 4.0, Computer Standards & Interfaces, 81 (2022).
9. Kolajo, T., Daramola, O. and Adebiyi, A.: Big data stream analysis: a systematic literature review. Journal of Big Data. 6. (2019).