Verificación Formal de Software en Sistemas de Big DATA

Fernando Asteasuain –Rafael Aragón – Luciana Rodriguez Caldeira- Nicolás Granata - Nahuel Patera- Pablo Gamboa – Hang Shao Feng
Contacto Principal: fernando.asteasuain@uai.edu.ar
Centro de Altos Estudio CAETI – Universidad Abierta Interamericana
Emails: <nombre.apellido>@uai.edu.ar

Resumen

La Ingeniería de Software debe evolucionar para poder enfrentar los desafíos de un mundo moderno híper conectado y con grandes volúmenes de información y datos disponibles para ser analizados. En este contexto, nuevas disciplinas como la denominada Ciencia de Datos [6] han surgido recientemente. Para llevar adelante estos desafíos se deben contar con herramientas para la verificación formal de sistemas basados en BIG DATA que cuentan con una fuerte interacción con áreas de la Inteligencia Artificial como Aprendizaje Automático [1, 2, 3, 4] para poder mantener los estándares esperados de rigurosidad y calidad.

Esta evolución requiere de novedosas técnicas para componer un sistema a través de sus múltiples aristas, junto con herramientas que sean eficaces pero también lo suficientemente flexibles y expresivas [5].

La presente investigación pretende dar un paso en pos de enfrentar este desafío, combinando técnicas de verificación formal con la Inteligencia Artificial, en especial con la teoría de juegos, la síntesis de comportamiento y el aprendizaje automático.

Palabras claves: *Big Data, Verificación Formal, Síntesis de Comportamiento.*

Contexto

La presente investigación se encuentra enmarcado dentro del proyecto "Framework para el Desarrollo de Software mediante Modularización Avanzada" desarrollado en el Centro de altos Estudios CAETI de la Universidad Abierta Interamericana.

El proyecto se va renovando anualmente, siendo este el sexto año consecutivo del desarrollo de la investigación.

El grupo de investigación está constituido por un investigador formado, dos en formación, un estudiante de maestría, dos estudiantes de grado haciendo sus tesis de licenciatura y un estudiante de grado haciendo la práctica profesional supervisada dentro del marco del proyecto. El proyecto es financiado 100% por la Universidad Abierta Interamericana.

La composición del grupo de investigación ha sido estable en estos años, lo cual es fundamental para poder profundizar en las líneas de investigación. Si existe una dinámica más fluida en los estudiantes de grado, ya que la mayoría al terminar sus tesis de grado o práctica profesional opta por trabajos en la industria. Sin embargo, la posibilidad de continuar con tesis de Maestría/Doctorado es una opción que se ha logrado capitalizar en algunas ocasiones. En particular, para el presente año contamos con un estudiante de posgrado.

Existe también la posibilidad de transferencia y servicios al sector industrial y productivo. Es importante mencionar que las líneas de Investigación del proyecto tienen impacto en áreas prioritarias del *Plan Nacional de Ciencia y Técnica 2020* como AgroIndustria, Biodiversidad e Innovación Productiva.

1. Introducción

La creciente demanda de sistemas basados en datos implica un gran esfuerzo para la comunidad de la Ingeniería de Software para poder validar y verificar el comportamiento esperado de sistemas en este dominio. Las características de los sistemas de BIG DATA se pueden resumir en lo que se conoce como las cinco "V": Variedad, Velocidad, Volumen, Valor y Veracidad [17, 18]:

1) Volumen: la cantidad creciente de información obtenida.

2) Variedad: Diversas y heterogéneas fuentes y tipos de dato.

3) Velocidad: para la manipulación y adquisición de datos, streaming en tiempo real, y datos de tiempo variable.

4) Valor: El peso de cada dato dentro de toda la información disponible.

5) Veracidad: confianza en los datos obtenidos y su procesamiento

Sin embargo, en la aplicación de técnicas y herramientas tradicionales de la Ingeniería de Software en este tipo de sistemas se han encontrado algunas debilidades como la falta de modelado y diseño y dificultades a la hora de validar y verificar el comportamiento esperado [9,10,11,12,13]. Incluso algunos autores han denominado al big data como "no testeable", proponiendo técnicas como testing metamórfico [20,21,22]. La verificación formal es el área que más necesita crecer dentro de la Ingeniería de Software [13]. Un enfoque en este sentido son los estándares de calidad establecidos para cuantificar la calidad de los dato [19]. Sin embargo, según lo reportado en [13] solo 2 de los casi 200 trabajos analizados vinculando Ingeniería de Software y BIG DATA se enfocan en temas de verificación formal. La mayoría de estos enfoques se concentran en sólo dos de las cinco "V": Velocidad y Volumen, buscando mejorar la performance y el tamaño de exploración de los sistemas a analizar. Ejemplos exitosos han sido versiones paralelas o distribuidas de herramientas como model checkers. Sin embargo, las otras tres "V", Variedad, Valor y Veracidad no han recibido la atención necesaria. Dichas características necesitan formalismos y notaciones que sean expresivos y flexibles, para poder lidiar con un volumen gigantesco y poco estructurado información de información y datos.

Un enfoque novedoso en la búsqueda de formalismos más poderosos puede darse desde el área de Síntesis de comportamiento [7,8]. Se llama síntesis de comportamiento al proceso de obtener automáticamente un controlador de un sistema a partir de su especificación de manera que por construcción se garantice que las propiedades que describen el comportamiento del sistema son satisfechas.

Un controlador es en esencia un autómata que recibe información de sensores, la procesa, y envía instrucciones a actuadores. La construcción de un controlador se realiza a través de técnicas de la Inteligencia Artificial y la teoría de juegos: se trata de obtener una estrategia ganadora para nuestro sistema que no importe que acciones se disparen desde el ambiente, se podrán cumplir los objetivos de comportamiento propuestos. Este esquema es clásico en entornos de sistemas abiertos como Robótica o Internet de las cosas.

La inmensidad de este desafío requiere entonces la necesidad combinar técnicas como la síntesis de comportamiento y el aprendizaje automático con herramientas tradicionales de verificación formal como model checking [14,15,16].

Una solución a los mencionados problemas tendrá impacto en las mencionadas áreas consolidando métodos, técnicas y herramientas de la Ingeniería de Software atacando cada una de las cinco "V" que caracterizan a Big Data.

2. Líneas de Investigación y Desarrollo

En el presente trabajo se explorarán las siguientes líneas de investigación:

- Aplicación de métodos formales de Ingeniería de Software al proceso de aprendizaje automático.
- Aplicación de métodos formales de Ingeniería de Software en redes neuronales.
- Aplicación de métodos formales de Ingeniería de Software en sistemas de BIG DATA.
- Aplicación de métodos formales de Ingeniería de Software en sistemas basados en Internet de las Cosas.
- Explorar la Síntesis de Comportamiento como un mecanismo poderoso para especificar comportamiento en sistemas orientados a datos.
- Analizar posibilidad de mejoras en los algoritmos para obtener controladores y en la expresividad de los lenguajes de especificación.
- Continuar y profundizar el desarrollo de herramientas de software que den aplicabilidad a los conocimientos adquiridos.
- Reforzar la interacción con model checkers distribuidos y/o paralelos. Las soluciones de software distribuidas y/o paralelas son claves para la optimización en tiempo que requieren los sistemas de BIG DATA.
- Contribuir en facilitar el proceso de verificar sistemas de BIG DATA.
- Potenciar y fusionar la cada vez más fuerte interacción entre la Ingeniería de Software y la Inteligencia Artificial.
- Expandir las nociones de estándares de calidad sobre manejo y manipulación de datos.
- Obtener demostraciones formales de la correctitud y completitud de los procesos involucrados.
- Expandir las nociones de testing metamórfico a la especificación de comportamiento. Esto implica buscar propiedades "metamórficas" para la verificación de sistemas en este dominio.

3. Resultados Obtenidos/Esperados

El objetivo principal del presente proyecto de Investigación es aplicar nuevos métodos y técnicas rigurosas y formales de la Ingeniería de Software al procesamiento de grandes volúmenes de datos conocido como sistemas Big Data.

El desafío involucra la interacción de diversas áreas como Inteligencia Artificial, Redes Neuronales, Aprendizaje Automático, Síntesis de Comportamiento o el Procesamiento Dinámico de Información. Se combinarán técnicas de aprendizaje automático y síntesis de comportamiento para aplicar todo el potencial de la Ingeniería de Software a sistemas orientados a BIG DATA y ciencia de datos.

Como objetivos específicos podemos mencionar:

a) Analizar los desafíos que requieren las técnicas, procesos y herramientas de Ingeniería de Software para lidiar con los sistemas de Big DATA.

b) Razonar sobre la manera de poder validar el comportamiento esperado en sistemas de BIG DATA.

c) Explorar sistemas paralelos de verificación formal para atacar los problemas de performance en sistemas con grandes volúmenes de información.

d) Modelar, abstraer y razonar para la descripción de eventos de alto nivel de interés en áreas de Inteligencia Artificial.

e) Aplicar herramientas y técnicas formales de la Ingeniería de Software para el desarrollo de sistemas de Big Data.

f) Divulgar los resultados de la investigación en congresos y revistas científicas de interés para los temas de la investigación.

g) Consolidar los recursos humanos en inicios de las tareas de investigación como estudiantes avanzados de la carrera, mediante la participación activa en el proyecto o mediante la realización de tesis de final de carrera de grado.

Los resultados esperados incluyen:

- Consolidar y profundizar técnicas formales de la Ingeniería de Software en un área de aplicación de vanguardia como Big Data e Internet de las Cosas y Ciencia de Datos.

- Dirección de tesis de licenciatura, de maestría y supervisión de prácticas profesionales (PPS).

4. Formación de Recursos Humanos

El presente proyecto buscará potenciar a los investigadores en formación y de posgrado. En cuanto a los estudiantes de grado, se buscará en primer lugar que finalicen su tesis de grado o su práctica profesional. Sin embargo, se los pretende motivar como para que puedan continuar dentro del proyecto mediante la realización de algún posgrado.

En este sentido, una táctica que ha resultado exitosa fue la presencia activa de los estudiantes de grado en los congresos científicos del área, ya sean virtuales o presenciales.

5. Bibliografía

1. ALPAYDIN, Ethem. Machine learning. MIT Press, 2021.

2. WANG, Ping; LI, Yan; REDDY, Chandan K. Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR), 2019, vol. 51, no 6, p. 1-36.

3. HAN, Rui; JOHN, Lizy Kurian; ZHAN, Jianfeng. Benchmarking big data systems: A review. IEEE Transactions on Services Computing, 2017, vol. 11, no 3, p. 580-597.

4. ELSHAWI, Radwa, et al. Big data systems meet machine learning challenges: towards big data science as a service. Big data research, 2018, vol. 14, p. 1-11.

5. VARDI, Moshe Y. Branching vs. linear time: Final showdown. International conference on tools and algorithms for the construction and analysis of systems. Springer, Berlin, Heidelberg, 2001. p. 1-22.

6. PROVOST, Foster; FAWCETT, Tom. Data science and its relationship to big data and data-driven decision making. Big data, 2013, vol. 1, no 1, p. 51-59

7. S. Maoz and Y. Sa'ar. Aspectltl: an aspect language for ltl specifications. In AOSD, pages 19-30. ACM, 2011.

8. N. DIppolito, V. Braberman, N. Piterman, and S. Uchitel. Synthesising nonanomalous event-based controllers for liveness goals. ACM Tran, 22(9), 2013.

9. Embley, David W., and Stephen W. Liddle. "Big data—conceptual modeling to the rescue." International Conference on Conceptual Modeling. Springer, Berlin, Heidelberg, 2013.

10. Arndt, Timothy. "Big Data and software engineering: prospects for mutual enrichment." Iran Journal of Computer Science 1.1 (2018): 3-10.

11. Camilli, Matteo. "Coping with the State Explosion Problem in Formal Methods: Advanced Abstraction Techniques and Big Data Approaches." (2015).

12. Hummel, Oliver, et al. "A collection of software engineering challenges for big data system development." 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE, 2018.

13. Kumar, Vijay Dipti, and Paulo Alencar. "Software engineering for big data projects:

Domains, methodologies and gaps." 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016.
14. SOURI, Alireza, et al. Formal verification of a hybrid machine learning-based fault prediction model in Internet of Things applications. IEEE Access, 2020, vol. 8, p. 23863-23874.

15. URBAN, Caterina; MINÉ, Antoine. A review of formal methods applied to machine learning. arXiv preprint arXiv:2104.02466, 2021.

16. W. Nam and H. Kil, "Formal Verification of Blockchain Smart Contracts via ATL Model Checking," in IEEE Access, vol. 10, pp. 8151-8162, 2022, doi: 10.1109/ACCESS.2022.3143145.

17. CAPPA, Francesco, et al. Big data for creating and capturing value in the digitalized environment: Unpacking the effects of volume, variety, and veracity on firm performance. Journal of Product Innovation Management, 2021, vol. 38, no 1, p. 49-67.

18. NAEEM, Muhammad, et al. Trends and future perspective challenges in big data. En Advances in Intelligent Data Analysis and Applications. Springer, Singapore, 2022. p. 309-325.

19. ISO 25012 (2008), "Ingeniería de software - Requisitos de calidad y evaluación de productos de software (SQuaRE) - Modelo de calidad de datos", Disponible en: https://www.iso.org/obp/ui/es/#iso:std:iso-iec:25012:ed-1:v1:en, consultado en febrero 2022.

20. CHEN, Tsong Yueh, et al. Metamorphic testing: A review of challenges and opportunities. ACM Computing Surveys (CSUR), 2018, vol. 51, no 1, p. 1-27.

21. SEGURA, Sergio, et al. A survey on metamorphic testing. IEEE Transactions on software engineering, 2016, vol. 42, no 9, p. 805-824.

22. DING, Junhua; ZHANG, Dongmei; HU, Xin-Hua. A framework for ensuring the quality of a big data service. En 2016 IEEE International Conference on Services Computing (SCC). IEEE, 2016. p. 82-89.