CONSORCIO DE I+D+I
CLOUD COMPUTING,
BIG DATA &
EMERGING TOPICS

Encuentro de Cooperación en Postgrado

Libro de Resúmenes

Noviembre 2021

Encuentro de Cooperación en Postgrado

Libro de Resúmenes

La Plata, Buenos Aires, Argentina.
10 de noviembre de 2021

Encuentro de Cooperación en Postgrado del Consorcio de Cloud Computing, Big Data & Emerging Topics : Libro de Resúmenes / compilación de Armando De Giusti ... [et al.]. - 1a ed. - La Plata : Universidad Nacional de La Plata. Facultad de Informática, 2021.
Libro digital, PDF

Archivo Digital: descarga y online
ISBN 978-950-34-2075-1

1. Tesis. 2. Tesis Doctorales. 3. Tesis de Maestría. I. De Giusti, Armando,comp.
CDD 004.0711

Prefacio

Este libro compila las presentaciones realizadas durante el Encuentro de Cooperación en Postgrado del Consorcio CCBD&ET que tuvo lugar en Noviembre de 2021. Los objetivos del Encuentro fueron:

- Fomentar la exposición de propuestas de Tesis Doctorales y/o de Maestría desarrolladas en el ámbito del Consorcio y/o grupos e investigadores relacionados con los temas del Consorcio.
- Reunir a Tesis y sus Directores, para potenciar el intercambio de ideas relacionadas con los temas de las Propuestas de Tesis en curso.
- Impulsar el desarrollo de Tesis co-dirigidas por investigadores de diferentes grupos del Consorcio.

En esta edición, se presentaron propuestas de 16 de tesis doctorales, 1 de tesis de maestría y 1 de trabajo posdoctoral.

Índice

Sistema de Gestión de Almacenamiento para Tolerancia a Fallos en Computación de Altas Prestaciones

Doctorado en Informática, Universitat Autònoma de Barcelona, España.

Tesista: Betzabeth León [1]
Directores: Dolores Rexachs[1], Daniel Franco[1]
Directores: Emilio Luque[1]
[1] Departamento de Arquitectura de Computadores y Sistemas Operativos, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, España
{betzabeth.leon, dolores.rexachs, daniel.franco, emilio.luque)@uab.es
Palabras claves: HPC, Tolerancia a fallos, Checkpoint, I/O.

Motivación

La solución de grandes problemas científicos reales puede necesitar el uso de grandes recursos computacionales, tanto en términos de esfuerzo de CPU como de requisitos de memoria. Muchas aplicaciones científicas se desarrollan para ejecutarse en una gran cantidad de procesadores. En entornos HPC es importante que las aplicaciones mantengan su disponibilidad, para ello, existen una serie de estrategias que protegen las aplicaciones en caso de que ocurran fallos de algún componente del sistema físico.
Entre las estrategias de redundancia y recuperación del Rollback Recovery se encuentra el Checkpoint/Restart, que almacena la información, de manera periódica, en un sistema de almacenamiento estable, suspendiendo la ejecución y consumiendo recursos de E/S. Con los checkpoint coordinados [1], todos los procesos deben sincronizarse para tener una línea de recuperación consistente y para crear un estado global coherente, esto simplifica la recuperación. En [2] indicaron que las aplicaciones que se ejecutan a gran escala pueden gastar más del 50% de su tiempo total almacenando checkpoints, reiniciando y rehaciendo el trabajo perdido, siendo esto una de las principales causas de overhead. Así como también, el costo en términos de tiempo de cómputo, uso de la red o recursos de almacenamiento puede ser una limitación para su uso práctico [3].
Conociendo en profundidad la estructura de la imagen que guarda el checkpoint, se puede saber qué elementos lo componen y como influyen en su tamaño, su escalabilidad, a fin de predecir el espacio de almacenamiento requerido por el mismo. Por lo tanto, el diseño de metodologías para predecir su tamaño cuando varía el número de procesos y establecer con pocos recursos opciones adecuadas de configuración, puede ayudar con un número limitado de recursos a predecir la cantidad de espacio de almacenamiento que necesitamos a mayor escala y elegir los parámetros del sistema que cumplan con los requisitos de sobrecarga especificados (capacidad de almacenamiento, configuración del sistema de almacenamiento, configuración del checkpoint (compresión, intervalo)). Por consiguiente, con el conocimiento de la forma más adecuada en que se gestionan estos elementos, podemos tener protección contra fallas que mantenga

la disponibilidad de nuestras aplicaciones y que afecten su comportamiento de menor manera.

En la presente investigación pretendemos analizar el impacto del almacenamiento estable de los mecanismos de protección y recuperación frente a fallos. Para esto, es necesario caracterizar los patrones de E/S en diferentes tipos de aplicaciones, generados por las estrategias de tolerancia a fallos. Así como, proponer una metodología para configurar la entrada y salida paralela, mediante la cual se pueda establecer la configuración adecuada para cada aplicación. Además, pretendemos replicar el comportamiento de los patrones de E/S de los protocolos de rollback recovery con pocos recursos, para predecir el comportamiento a gran escala y por último utilizar aplicaciones científicas reales para validar las soluciones obtenidas de los experimentos realizados en las fases de experimentación con benchmark conocidos.

Objetivos y Aportes

Nuestra propuesta se centra en la caracterización de los patrones de E/S generados por las estrategias de tolerancia a fallos, debido a que esto contribuirá a conocer su comportamiento en términos de su interacción con el sistema de archivos y la forma más adecuada de administrarlos. Por ello, conocer los archivos generados y los patrones de acceso nos permitirá obtener información para la selección y configuración del subsistema de E/S, así como los elementos que influyen en la escalabilidad.

Para poder llevar a cabo los experimentos necesarios con la finalidad de caracterizar los patrones de E/S generados por la tolerancia a fallos, se propuso sistematizar en una metodología el procedimiento, para identificar y analizar de manera adecuada todos los elementos involucrados en la ejecución de una aplicación con tolerancia a fallos. En la figura 1 se muestran los pasos necesarios:

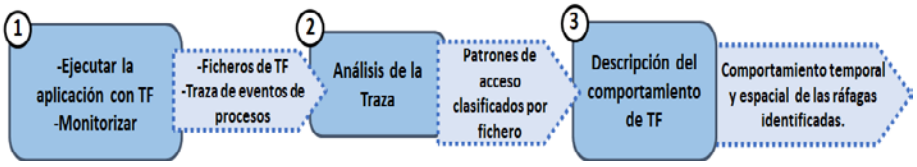


Figura 1 Metodología para el análisis de los patrones de la Tolerancia a Fallos

Esta metodología ha permitido seguir un procedimiento para realizar el análisis de los patrones del checkpoint coordinado. El primer paso consiste en ejecutar la aplicación con tolerancia a fallos y monitorizarla para así poder obtener los ficheros generados por la tolerancia a fallos y la traza de los eventos por proceso. Luego con esta información se realiza un análisis de la traza, con la finalidad de obtener los patrones de acceso clasificados por fichero. A continuación, se lleva a cabo una descripción del comportamiento de la tolerancia a fallos y es posible describir el comportamiento temporal y espacial de las ráfagas identificadas.

Con el checkpoint coordinado, la aplicación comienza ejecutarse y a un mismo intervalo de tiempo todos los procesos se detienen y deben coordinarse de manera síncrona para hacer el checkpoint. A partir de aquí se genera un fichero de checkpoint por proceso, estos son guardados en un sistema de almacenamiento estable. Cuando se produce una avería en cualquier proceso, todo el sistema retrocede a la última imagen del

checkpoint para continuar con el cálculo [4]. Cada uno de estos ficheros posee la imagen del estado global, la cual está conformada por los datos de la aplicación, las librerías y los buffers de comunicación y si la aplicación hace E/S también tendrá la información relativa a los buffers de E/S.

Así mismo, se observó que existen una serie de elementos que pueden influir en el comportamiento del tamaño del checkpoint e impactar en la información que almacena, algunos de estos elementos son: la compresión de ficheros, la implementación MPI utilizada, el mapping y el sistema de ficheros. Luego de analizar estos elementos se diseñó una metodología para predecir el comportamiento del tamaño del checkpoint, cuando se modifica el número de procesos y el mapping.

Como se mencionó anteriormente el checkpoint está conformado por tres zonas, las cuales son: la zona de librerias (LB), la zona de datos (DTAPP) y la zona de memoria compartida (SHMEM). La zona de DTAPP depende del workload de la aplicación, la de LB depende del sistema y lo que la aplicación necesita para ejecutarse y la zona SHMEM depende del número de procesos dentro de un nodo. Esta metodología indica cómo se puede predecir el tamaño del checkpoint, para ello se caracteriza primero el entorno y se ejecuta la aplicación para poder identificar el tamaño de la zona LB y en el caso de la zona DTAPP obtener la ecuación de regresión para poder estimar su tamaño con cualquier número de procesos. En el caso de la memoria compartida (SHMEM), además de obtener el tamaño de esta zona a través de ecuaciones de regresión, se diseñó un modelo de la memoria compartida utilizada por MPICH. Este modelo tiene una aproximación muy cercana del tamaño de la memoria compartida con múltiples cores. De esta manera, con este enfoque, tenemos una idea del funcionamiento lógico de la memoria compartida dentro del mismo nodo. Este modelo puede servir como herramienta para representar predicciones o simulaciones que requieren el uso y la representación de este elemento.

En [5][6] indican la importancia sobre la consideración de los intervalos del checkpoint y la necesidad de proponer estrategias para determinar y predecir el tiempo de los mismos. En nuestra investigación luego de conocer y poder predecir el tamaño del contenido que almacena el checkpoint, se diseñó un modelo para estimar el número de checkpoints a ejecutar según un porcentaje de tiempo dado sobre el tiempo de ejecución de la aplicación. Este modelo nos permite comprender qué sucede cuando se crea un checkpoint en un sistema HPC, para tomar decisiones que se adapten a los requisitos del usuario.

Estado Actual y Trabajo Futuro

Identificar el impacto en la E/S de la tolerancia a fallos, puede ayudar a establecer metodologías y configuraciones para reducir el overhead generado por el almacenamiento de estas estrategias. El estado global de las aplicaciones sin E/S que almacena el checkpoint está integrado por tres zonas, las cuales son: zona de datos, zona de librerías y zona de memoria compartida. En el caso de aquellas aplicaciones que hacen E/S se agrega una nueva zona correspondiente a los buffers de E/S. El número de procesos y el mapping utilizado influye de manera directa en el tamaño de cada una de estas zonas, por lo cual también influye en el tamaño de los ficheros de checkpoint. El patrón

de acceso (número de escrituras y tamaños de las escrituras) observado en los checkpoint coordinados ha mostrado una gran irregularidad, debido a que las ráfagas de escrituras generadas están compuestas de muchas escrituras de tamaños muy pequeños y pocas muy grandes, por lo tanto, no hay una regularidad en el patrón de acceso secuencial. Estas escrituras grandes son las que consumen la mayor parte del tiempo del checkpoint. La metodología para predecir el tamaño del checkpoint y el modelo para predecir el tamaño de la memoria compartida, constituye una herramienta útil para saber de antemano el espacio de almacenamiento necesario cuando realizamos cambios en el número de procesos y para conseguir el mapping más apropiado. A partir de un overhead máximo las restricciones de sobrecarga (espacio de almacenamiento y tiempo del overhead) aceptado, se ha diseñado un modelo para establecer el número de checkpoints que teniendo en cuenta la restricción se deben realizar durante la ejecución de la aplicación. Este modelo es útil porque se adapta a los requerimientos del usuario, según los recursos con que dispone.

Además del checkpoint coordinado, existen otros tipos de estrategias de tolerancia a fallos como el checkpoint no coordinado, semi coordinado y los logs de mensajes, los cuales presentan patrones distintos de comportamiento de almacenamiento y ejecución. Como líneas futuras se puede analizar el comportamiento del checkpoint en este tipo de estrategias y así extender el modelo propuesto en esta investigación. Además de esto se pueden abordar otro tipo de aplicaciones como las de machine learning, muchas de las cuales también utilizan checkpoint. Así como desarrollar utilidades que se centren en el comportamiento de la E/S para evaluar su impacto y reducirlo.

Referencias

[1] Kumar M, Choudhary A, Kumar V (2014) A comparison between diferent checkpoint schemes with advantages and disadvantages. Int J Comput Appl Nat Semin Recent Adv Wireless Netw Commun 3:36

[2] J. Elliott, K. Kharbas, D. Fiala, F. Mueller, K. Ferreira, and C. Engelmann (2012). Combining partial redundancy and checkpointing for hpc. IEEE 32nd International Conference on Distributed Computing Systems, pages 615–626, 2012.

[3] N. Losada, M. J. Martín, G. Rodríguez, and P. González (2015). I/O optimization in the checkpointing of openmp parallel applications. In 23rd Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, pages 222–229.

[4] Lemarinier Bouteiller, Capello Krawezik (2003) Coordinated checkpoint versus message log for fault tolerant MPI. Proceedings IEEE International Conference on Cluster Computing, pp. 242–250. https://doi.org/10.1109/CLUSTR.2003.1253321.

[5] D. Dauwe, S. Pasricha, A. A. Maciejewski and H. J. Siegel (2018). "An Analysis of Multilevel Checkpoint Performance Models". *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 783-792, doi: 10.1109/IPDPSW.2018.00125.

[6] Muhammad Abrar Akber S, Chen H, Wang Y, Jin H (2018) Minimizing Overheads of Checkpoints in Distributed Stream Processing Systems, In: 2018 IEEE 7th International Conference on Cloud Networking (CloudNet), pp. 1–4. https://doi.org/10.1109/CloudNet.2018.854954.

Arquitecturas de crowdsourcing para procesamiento de imágenes en el contexto de desastres naturales

Doctorado en Ciencias de la Computación, Universidad Nacional de San Luis, Argentina

Tesista: Fernando Loor [1]
Directora: Verónica Gil-Costa[1]
Co-Director: Mauricio Marín[2]

[1] Universidad Nacional de San Luis, Argentina
{floor,gvcosta}@unsl.edu.ar
[2] DIINF, CITIAPS, Universidad de Santiago, Chile
mauricio.marin@usach.cl

Palabras claves: Corwdsourcing, redes P2P, desastres naturales, evaluación de rendimiento.

1 Motivación

Los desastres naturales como terremotos, huracanes, erupciones volcánicas e incendios afectan a millones de personas en el mundo cada año. Estos eventos producen daños físicos y económicos arrasando viviendas e incluso pueblos enteros. Más aún, estos eventos dramáticos pueden cobrarse la vida de seres que viven en las áreas afectadas. En estos escenarios, los equipos de respuesta de emergencia tienen un rol fundamental en ayudar a reducir el impacto de los desastres y típicamente son coordinados por una organización nacional responsable de enviar los recursos apropiados a las áreas más afectadas.

Cuando un desastre ocurre, la gente que se encuentra en el lugar afectado captura y comparte imágenes con información georreferenciada que puede ser de gran importancia para producir mapas de situación. Sin embargo, algunas fotos o videos pueden ser difíciles de clasificar automáticamente. Algunas fotos pueden estar fuera de foco o no ser nítidas. Incluso algunas fotos pueden no pertenecer al desastre natural actual. En ese contexto, el crowdsourcing emerge como un mecanismo poderoso a través del cual voluntarios pueden ayudar a llevar a cabo diferentes tareas como el procesamiento de imágenes complejas usando técnicas de etiquetado y clasificación.

2 Objetivos y Aportes

El objetivo de este trabajo es reducir el tiempo de ejecución requerido para etiquetar y procesar fotos que no pueden ser procesadas o clasificadas automáticamente, a través de la utilización eficiente de recursos y realizando un procesamiento distribuido de las

tareas. El enfoque de nuestro trabajo es en campañas de crowdsourcing de corto plazo. Esto implica reducir los costos causados por la transferencia de datos, así como tomar ventaja de las características de las redes alternativas como las redes P2P. Producir una solución práctica a este problema es desafiante ya que queremos evitar saturar la carga de trabajo de voluntarios, del servidor y de las redes de comunicación para prevenir que se comprometa la eficiencia del tiempo de procesamiento de cada foto.

3 Estado Actual y Trabajo Futuro

La plataforma propuesta en [1] y [2] acelera el etiquetado de las imágenes con voluntarios digitales compartiendo de manera más eficiente la información, haciendo uso de redes P2P construidas con STB. Los STB, incorporados en los smartTVs actualmente, tienen determinadas capacidades de procesamiento y de almacenamiento que sirven a tal fin. Los usuarios pueden colaborar compartiendo la capacidad de cómputo y almacenamiento de los equipos o trabajando activamente como voluntarios en la clasificación o etiquetado de las imágenes.

La plataforma se compone de 3 capas: (1) la capa de arquitectura, que comprende la conexión del servidor con el ISP y con la red P2P formada por los STB, (2) la capa con el algoritmo para el flujo de tareas, que controla cómo se transmiten las tareas, las imágenes y los resultados a cada entidad de la red, y (3) la capa de módulos, en la que se han incorporado propuestas de procesamiento o escenarios presentes en la literatura, de los cuáles se evaluó el impacto en el rendimiento de la plataforma.

La metodología de trabajo utilizada es de desarrollo de software, en espiral e incremental. Se desarrolló un simulador en la librería de eventos discretos en C++ LibCppSim, incorporando las entidades de la plataforma, parámetros y métricas de rendimiento; el simulador representa una de las contribuciones de [2]. Es importante destacar que no se encontró una propuesta similar en la literatura. La librería elegida fue LibCppSim, pero también se ensayaron versiones del modelo de la aplicación sobre la red P2P en los simuladores PeerSim [1] y Cadmium [11].

Las entidades principales de la plataforma se modelan utilizando corrutinas. En el caso de la capa de arquitectura, el servidor cuenta con una base de datos donde almacena los resultados, el ISP distribuye el trabajo hacia los peer, envía los objetos y recopila los resultados, y la red P2P permite compartir la información de las tareas entre los usuarios conectados a cada peer. La red P2P implementa el protocolo Pastry para la división del espacio de objetos, que en este caso son las imágenes correspondientes a cada tarea.

Para distribuir las tareas entre los voluntarios, el servidor envía solicitudes de tareas a los usuarios, y al recibir la confirmación de trabajo, envía las imágenes correspondientes. Los usuarios realizan el etiquetado y envían los resultados al servidor. Cuando el servidor recibe todas las respuestas de una tarea, realiza la agregación de los resultados y decide si hay consenso sobre la tarea, o si la misma debe ser reenviada a usuarios expertos para lograr el consenso. En el primer enfoque propuesto, el servidor envía las solicitudes de trabajo a los voluntarios, pero los mismos solicitan

las imágenes a etiquetar a otros peers de la red. El peer responsable de cada imagen recibe la imagen desde el servidor y la almacena en memoria caché, de manera de satisfacer las solicitudes de esa imagen que reciba desde otros usuarios en adelante; la agregación de las respuestas sigue siendo realizada por el servidor. En el segundo enfoque propuesto, la agregación es realizada por determinados peers de la red, que envían el consenso calculado al servidor; de esta manera, se logra reducir la carga de trabajo del servidor y la comunicación a través del ISP.

Para evaluar el rendimiento de la plataforma bajo altas tasas de arribo de tareas se generaron diversas distribuciones de arribo con configuraciones diferentes para los parámetros de la plataforma, se corrieron las simulaciones correspondientes, y se utilizó análisis de sensibilidad global [5][6] para determinar qué parámetros tienen una mayor influencia sobre la salida del simulador.

De estos experimentos, en primer lugar, se obtuvieron mapas de calor que relacionaban las variables de entrada y de salida del sistema, y a partir de allí, se fijaron los valores menos influyentes y se estudió el comportamiento de las métricas de salida más influyentes en relación al resto de los parámetros. Los resultados de los experimentos muestran que para tasas de arribo elevadas, las propuestas que hacen uso de la red P2P mejoran las métricas de throughput, reducen la latencia en el ISP, y completan de manera más rápida la cantidad de respuestas necesarias para que las tareas tengan consenso.

Debido a la naturaleza dinámica del arribo de los datos a ser procesados, actualmente estamos desarrollando un módulo de control que ajuste algunos de los parámetros en tiempo real, en orden a poder adaptar el funcionamiento del servidor según las características del trabajo que llega al mismo. En trabajos como [9][10] se muestra que la información puede llegar en forma de ráfagas o con picos de arribos de miles de tareas, lo que puede saturar los recursos de la plataforma, sobre todo al servidor y a las redes, ocasionando que las tareas caduquen sin lograr reunir la cantidad de respuestas necesarias para declarar el consenso. Una manera de lidiar con este problema es mediante un control dinámico de los parámetros más importantes de la plataforma.

El módulo de control se basa en índices métricos [7][8], y el proceso de control tiene 2 etapas: offline y online. En la etapa offline, se realizan múltiples ejecuciones del simulador de manera de muestrear el espacio de parámetros, y los datos de las simulaciones son indexados a través de un índice métrico. En la etapa online, el estado de la plataforma compuesto por los parámetros y las métricas de rendimiento se utiliza para construir un vector de consulta del índice métrico. El resultado de la búsqueda es idealmente una configuración de parámetros de entrada similar a la del estado actual del simulador, pero con mejores valores en las métricas de rendimiento. Actualmente estamos evaluando el desempeño de esta metodología y planeamos a futuro compararla con enfoques basados en Reinforcement Learning.

Referencias

1. Manriquez, M., Loor, F., Gil-Costa V., Marin, M.: A digital TV-based distributed image processing platform for natural disasters. In Proceedings of the Winter Simulation Conference (WSC), pp. 2689-2700. IEEE, (2019).

2. Loor, F., Manriquez, M., Gil-Costa, V., Marin, M.: Feasibility of P2P-STB based crowdsourcing to speed-up photo classification for natural disasters. Cluster Computing, 1-24. (2021)
3. Alam, F., Ofli, F., Imran, M.: Processing social media images by combining human and machine computing during crises. International Journal of Human–Computer Interaction 34(4), 311-327. (2018).
4. Moradi, M., Moradi, M., Bayat, F., Nadjaran Toosi, A.: Collective hybrid intelligence: towards a conceptual framework. International Journal of Crowd Science (2019).
5. Iooss, B., Lemaître, P.: A review on global sensitivity analysis methods. In Uncertainty management in simulation-optimization of complex systems. Springer, Boston, MA, pp. 101-122 (2015).
6. Herman, J., Usher, W.: SALib: an open-source Python library for sensitivity analysis. Journal of Open Source Software 2(9) (2017).
7. Bozkaya, T., Ozsoyoglu, M.: Indexing large metric spaces for similarity search queries. ACM Transactions on Database Systems (TODS) 24 (3), 361-404. (1999).
8. Chen, L., Gao, Y., Song, X., Li, Z., Miao, X., Jensen, C. S.: Indexing metric spaces for exact similarity search. arXiv preprint arXiv:2005.03468 (2020).
9. Kurkcu, A., Zuo, F., Gao, J., Morgul, E. F., Ozbay, K.: Crowdsourcing incident information for disaster response using twitter. In Proceedings of the 65th Annual Meeting of Transportation Research Board. (2017).
10. Bhavaraju, S. K. T., Beyney, C., Nicholson, C.: Quantitative analysis of social media sensitivity to natural disasters. International journal of disaster risk reduction 39. (2019).
11. Loor, F., Gil-Costa, V., Wainer, G., A Comparative Study Between Cadmium DEVS and LibCppSim. In Proceedings of the HPCS Conference. (2020).

Modelización y Gestión del Consumo Energético en un Sistema de Altas Prestaciones con Tolerancia a Fallos

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina.

Tesista: Marina Morán [1]
Director: Javier Balladini [1]
Directora: Dolores Rexachs del Rosario [2]
Co-Director: Enzo Rucci [3]

[1] Ing. de Computadoras, Facultad de Informática, Universidad Nacional del Comahue
Neuquén, Argentina
{marina,javier.balladini}@fi.uncoma.edu.ar
[2] Arq. de Computadores y SO, Universitat Autónoma de Barcelona,
Bellaterra, España
dolores.rexachs@uab.es
[2] III-LIDI, Facultad de Informática, Universidad Nacional de la Plata - CIC
La Plata, Argentina
erucci@lidi.info.unlp.edu.ar

Palabras claves: consume energético, ahorro energético, resiliencia, tolerancia a fallos memoria distribuida, HPC, MPI, DVFS, ACPI, checkpoint.

1 Motivación

Los sistemas de computación de altas prestaciones actuales utilizan desde cientos hasta miles de millones de unidades de procesamiento y la tendencia es a continuar aumentando su poder de cómputo incrementando el número de componentes. Sin embargo, este crecimiento en el poder de cómputo conlleva un aumento en el consumo energético, y los costos para abastecer de energía a grandes sistemas de cómputo de altas prestaciones (HPC) se pueden volver inviables, por lo que la eficiencia energética se ha convertido en un nuevo reto. Dadas las limitaciones que existen para abastecer de energía a este tipo de computadoras, se hace necesario conocer el comportamiento de su consumo energético para encontrar formas de limitarlo y disminuirlo.

El método de tolerancia a fallos más usado actualmente en HPC es rollback recovery. Con este método, los procesos involucrados en la aplicación se detienen para poder resguardar el estado actual de ejecución. Estos métodos agregan un consumo energético adicional al propio de la ejecución de la aplicación. Nos preguntamos entonces, ¿qué oportunidades de ahorro energético presenta la tolerancia a fallos? ¿Es posible disminuir el consumo energético de la ejecución de una aplicación al hacer más eficiente energéticamente su método de tolerancia a fallos? Conocer el comportamiento energético de los sistemas actuales y futuros de HPC se ha vuelto indispen-

sable. Los métodos de tolerancia a fallos son una parte fundamental de estos sistemas, por lo que conocer cómo es su consumo de energía es una línea importante de investigación actual.

2 Objetivos y Aportes

El objetivo general de esta investigación es conocer y gestionar las posibilidades de ahorro energético que se presentan ante un fallo al usar rollback recovery en sistemas HPC.

3 Estado Actual y Trabajo Futuro

3.1 Predicción del Consumo Energético del Checkpoint/Restart en HPC

En este trabajo [1] proponemos un modelo y un método que nos permita predecir el consumo energético de las operaciones de checkpoint y restart (CR), considerando diferentes parámetros del sistema y de la aplicación. Nos enfocamos en CR coordinado a nivel de sistema, en aplicaciones Simple Programa Múltiple Data (SPMD), sobre clusters homogéneos.

Se caracterizó el sistema para conocer cómo se comporta con respecto a la potencia disipada y al tiempo al realizar CR. Se toman mediciones de potencia y tiempo, para diferentes frecuencias de reloj y tamaños del problema. Estas mediciones conforman una nube de puntos que deberá analizarse para encontrar la función que mejor la aproxime, y luego obtener los coeficientes de dicha función mediante el método de mínimos cuadrados. Se validan las funciones obtenidas utilizando otros valores de frecuencia de reloj y tamaños del problema, diferentes a los utilizados para obtener las ecuaciones. Tanto para Checkpoint como para Restar, la ecuación para la potencia disipada depende de la frecuencia de reloj, mientras que la ecuación para el tiempo depende de la frecuencia de reloj y del tiempo que demanda la operación. Estas ecuaciones pueden usarse para realizar la predicción del consumo de energía del CR para otras frecuencias de reloj y tamaños del problema.

Se analizan algunos factores que afectan el consumo y/o la calidad de la predicción del consumo de energía: Los estados C del procesador, la configuración del montado del NFS, la compresión de los archivos de checkpoint.

3.2 Modelización y Gestión del consumo energético en un Sistema de Altas Prestaciones con Tolerancia a Fallos

Una aplicación paralela de paso de mensajes pude verse afectada por fallos de los componentes del sistema de cómputo. Cuando un nodo falla, es posible emplear un esquema de checkpoints no coordinados donde los procesos de los nodos que no han fallado continúan ejecutando. Estos procesos eventualmente se detendrán cuando requieran comunicarse con algún proceso que está en recuperación. Como allí ocurri-

rá una espera, pensamos que este escenario presenta oportunidades para el ahorro energético.

En este trabajo [2] evaluamos una serie de estrategias que pueden aplicarse buscando mejorar la eficiencia energética a partir de la ocurrencia de un fallo. Las estrategias utilizan la Advanced Configuration and Power Interface (ACPI), en particular, consideramos el uso de técnicas de Dynamic Voltage and Frequency Scaling (DVFS) e hibernación del sistema a nivel de nodo.

Se definió un modelo que estima el ahorro energético logrado a partir de la aplicación de las estrategias seleccionadas cuando ocurre un fallo.

Hemos desarrollado un simulador basado en eventos que nos permite evaluar las estrategias bajo diferentes configuraciones del sistema, diferentes características de la aplicación y diferentes del momento del fallo. Se simula el fallo de un nodo en una aplicación paralela de paso de mensajes, con checkpoints no coordinados a nivel de sistema. Al momento del fallo, el simulador evalúa cada uno de los procesos vivos con cada una de las frecuencias de reloj provistas, y determina la mejor estrategia para aplicar. La salida del simulador incluye el ahorro energético estimado al aplicar la estrategia seleccionada, y una traza para visualizar el comportamiento de la aplicación con la herramienta Paraver.

La experimentación con el simulador nos permite observar que las estrategias pueden lograr importantes ahorros de energía, especialmente cuando los tiempos de reejecución son largos. Adicionalmente, encontramos que las esperas activas presentan oportunidades de ahorro energético.

Referencias

1. M. Morán, J. Balladini, D. Rexachs and E. Luque, "Prediction of Energy Consumption by Checkpoint/Restart in HPC," in IEEE Access, vol. 7, pp. 71791-71803, 2019, doi: 10.1109/ACCESS.2019.2919970.
2. M. Morán, J. Balladini, D. Rexachs and E. Rucci, "Towards Management of Energy Consumption in HPC Systems with Fault Tolerance," 2020 IEEE Congreso Bienal de Argentina (ARGENCON), 2020, pp. 1-8, doi: 10.1109/ARGENCON49523.2020.9505498.

Modelización y Simulación basada en Agentes aplicada a la Arquitectura de Entrada/Salida de los Computadores Paralelos

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina.

Tesista: Diego Encinas [1]
Director externo: Emilio Luque [2]
Co-Director externo: Dolores Rexachs [2]
Director local: Marcelo Naiouf [1]
Co-Director local: Armando De Giusti [1]

[1] Instituto de Investigación en Informática LIDI – UNLP – CIC. Argentina
{dencinas, mnaiouf, degiusti}@lidi.info.unlp.edu.ar
[2] High Performance Computing for Efficient Applications and Simulation – UAB, España
{dolores.rexachs,emilio.luque}@uab.es

Palabras claves: Modelado y Simulación, ABMS, Sistema de E/S en HPC.

1 Motivación

Mejorar el procesamiento y el almacenamiento de grandes cantidades de datos se ha convertido en un reto en los sistemas paralelos y distribuidos. El paralelismo en E/S o almacenamiento paralelo, es una técnica utilizada para acceder a los datos en un sistema de almacenamiento con varios discos simultáneamente, desde distintos procesos de aplicaciones, con el objetivo de maximizar el ancho de banda y acelerar las operaciones. Para implementarlo es muy importante contar con un sistema de archivos paralelo o distribuido, en caso contrario el sistema de archivos probablemente gestionará las peticiones de E/S que reciba en forma secuencial, sin que el paralelismo E/S ofrezca ninguna ventaja concreta.
Generalmente evaluar las prestaciones de un sistema de E/S con diferentes configuraciones y la misma aplicación permite seleccionar la disposición óptima. Pero para hacer cambios en la configuración puede ser una gran ventaja analizar la performance que van a obtener las aplicaciones antes de configurar el sistema (hardware y software). Una manera de conseguir predecir el comportamiento de las aplicaciones en el sistema de cómputo ante distintas configuraciones, es utilizando técnicas de modelado y simulación.
Analizar y diseñar modelos de simulación basados en la arquitectura de E/S paralela, permite disminuir la complejidad y cubrir las exigencias de las aplicaciones en HPC, al poder identificar y evaluar los factores que influyen en las prestaciones.
En el área de E/S para HPC se han desarrollado algunas plataformas de simulación que se mencionan a continuación. Simulator Framework for Computer Architectures

and Storage Networks (SIMCAN) [1] orientada a optimizar las comunicaciones y algoritmos de E/S. Paralel I/O Simulator of Hierarchical Data (PIOSimHD) [2] desarrollado para analizar la performance de Message Passing Interface-Input/Output (MPI-I/O). Co-design of Exascale Storage System (CODES) [3] es un framework desarrollado para evaluar el diseño del sistema de almacenamiento exaescalar. High-Performance Simulator for Hybrid Parallel I/O and Storage System (HPIS3) [4] que modela la carga de trabajo de las aplicaciones. CODES y HPIS3 están basados en ROSS (Rensselaer's Optimistic Simulation System) [5] que es una plataforma de simulación paralela.

Todas las herramientas mencionadas se basan en el paradigma de simulación basada en eventos (Discrete Event Simulation, DES). En este trabajo se propone utilizar el modelado y simulación basado en agentes (Agent-Based Modelling and Simulation, ABMS) para desarrollar un simulador que permita evaluar las prestaciones de la pila de software de E/S. Este tipo de simulación se caracteriza por la existencia de muchos agentes que interactúan usualmente en ausencia de un controlador central, en el que los comportamientos emergentes que se presentan son difíciles de predecir y/o anticipar debido a la capacidad adaptativa entre las partes. ABMS permite un fácil cambio en el nivel de análisis: se puede focalizar tanto en el nivel macro como en el micro (cómo nace el comportamiento agregado de los agentes del sistema y, además, analizar el comportamiento individual de cada uno).

El paradigma de agentes tiene aplicaciones en varias áreas de la ciencia y es de gran interés dentro del campo de la Inteligencia Artificial (IA- Intelligence Artificial), permite la resolución de problemas con complejidad de manera satisfactoria en comparación con otras técnicas clásicas. Es una técnica de simulación que recrea la funcionalidad de diferentes componentes de un sistema real mediante el modelado de entidades denominadas agentes. Básicamente, un agente es una entidad capaz de percibir y actuar dependiendo los cambios en su entorno. Asimismo, posee la habilidad de interactuar con otros agentes, ejecutando y coordinando sus acciones con el fin de alcanzar los objetivos.

Generalmente, los dos paradigmas funcionan con un tiempo discreto, pero DES es usado para un nivel de abstracción de bajo a medio. En ABMS el comportamiento del sistema es definido a nivel individual y el comportamiento global emerge al iniciarse la comunicación e interacción de los agentes en un mismo entorno. De hecho, ABMS es más sencillo de modificar ya que la depuración del modelo habitualmente es a nivel local y no global. Requiere validación del modelo: verificación (comprobar que el código computacional funcione según las especificaciones del modelo), calibración y validación (que el modelo represente adecuadamente al sistema real).

2 Objetivos y Aportes

El objetivo general de esta investigación es proponer un modelo de la Entrada/Salida en Computadoras de Altas Prestaciones que permita predecir cómo cambios realizados en los diferentes componentes del mismo afectan a la funcionalidad y el rendimiento del sistema.

Los objetivos específicos son:

- Analizar las propuestas de la comunidad científica para el modelado y simulación del sistema de E/S paralelo.
- Identificar los componentes del sistema real que permiten modelar la funcionalidad del sistema de E/S.
- Caracterizar los agentes que permiten modelar el software y hardware del sistema de E/S en computadoras de altas prestaciones.
- Analizar las características de los componentes del sistema de E/S paralelo que inciden en el rendimiento.
- Proponer y diseñar un modelo de simulación para el sistema de E/S de computadoras de altas prestaciones en base al análisis anterior.
- Evaluar la eficacia del modelo propuesto para simular la funcionalidad y predecir el rendimiento a partir de cambios en los componentes del sistema de E/S.

El aporte de este trabajo de tesis será un modelo del sistema de E/S paralela, que permitirá evaluar su rendimiento a partir de cambios en sus componentes principales (red de almacenamiento, dispositivos de E/S, componentes de la pila de software, entre otros). Para ello se utilizarán técnicas de simulación basadas en modelado y simulación basado en agentes (ABMS, Agent-Based Modeling&Simulation).

3 Estado Actual y Trabajo Futuro

En este trabajo de tesis doctoral se propuso un modelado utilizando el concepto de caja blanca con el propósito de observar el comportamiento específico en cada uno de los módulos o capas del sistema. De esta manera se analizaron las capas de la pila de software de E/S utilizando instrumentación de código y monitorización no invasiva en las funciones correspondientes a las operaciones de E/S.

Se inició con un análisis del sistema de E/S en HPC y se prosiguió con el desarrollo de un modelo estructural, definiendo a los agentes y sus interacciones. De esta manera, se modelizó el funcionamiento de diferentes módulos de las distintas capas de la pila de software de E/S como así también de componentes hardware por medio de agentes y entornos.

Se definió una plataforma de prueba para llevar a cabo las distintas monitorizaciones del sistema con el fin de obtener conjuntos de datos que se emplearon en las etapas de verificación, calibración y validación. Se utilizaron clústeres físicos como virtuales en clouds públicos. También se implementó instrumentación de código como también herramientas de monitorización para la obtención de métricas.

Por otro lado, se optó por el lenguaje y entorno de desarrollo NetLogo para el desarrollo del simulador. Con esta herramienta se realizó un primer simulador sólo funcional (funcionamiento de los agentes y las comunicaciones entre los mismos) y así generar las primeras pruebas de concepto. Luego, se prosiguió con la implementación de un simulador funcional junto con las características temporales del sistema para llegar a validar los primeros escenarios. Actualmente se ha conseguido un modelo y simulador funcional, temporal y espacial que permite obtener predicción de tiempos y bytes

de operaciones básicas (datos, control y comunicaciones). Los trabajos futuros a conseguir se relacionan con el detalle de modelización y simulación de operaciones de E/S, partición de datos y bufferes.

Los resultados de las distintas etapas de desarrollo del modelado y simulación del sistema se han presentado en diversos eventos académicos. A continuación, se listan las publicaciones en revistas, congresos y jornadas relacionadas directamente a este trabajo de tesis doctoral.

- International Conference on Advances in System Simulation (Congreso): "Modeling I/O System in HPC: An ABMS Approach"
- Journal of computer science & technology (Revista): "Using AWS EC2 as Test-Bed infrastructure in the I/O system configuration for HPC applications"
- Jornadas de Cloud Computing & Big Data: "Análisis funcional de la pila de software de E/S paralela utilizando IaaS".
- International Conference on Advances in System Simulation (Congreso): "On the Calibration, Verification and Validation of an Agent-Based Model of the HPC Input/Output System" (Best Paper Award)
- International journal on advances in systems and measurements (Revista) "An Agent-Based Model for Analyzing the HPC Input/Output System".

Referencias

1. A. Núñez, J. Fernández, J. García, F. García, and J. Carretero, "New Techniques for Simulating High Performance MPI Applications on Large Storage Networks," The Journal of Supercomputing, 2010, 51:40–57.
2. J. Kunkel, "Using Simulation to Validate Performance of MPI(-IO) Implementations," in Supercomputing, ser. Lecture Notes in Computer Science, J. M. Kunkel, T. Ludwig, and H. W. Meuer, Eds., no. 7905. Berlin, Heidelberg: Springer, 06 2013, pp. 181–195.
3. N. Liu et al., "Modeling a leadership-scale storage system." in PPAM (1), ser. Lecture Notes in Computer Science, R. Wyrzykowski, J. Dongarra, K. Karczewski, and J. Wasniewski, Eds., vol. 7203. Springer, 2011, pp. 10–19.
4. B. Feng, N. Liu, S. He, and X.-H. Sun, "HPIS3: Towards a Highperformance Simulator for Hybrid Parallel I/O and Storage Systems," in Proceedings of the 9th Parallel Data Storage Workshop, ser. PDSW'14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 37–42.
5. C. Carothers, D. Bauer, and S. Pearce, "ROSS: a high-performance, low memory, modular time warp system," in Parallel and Distributed Simulation, 2000. PADS 2000. Proceedings. Fourteenth Workshop on, 2000, pp. 53–60.

Modelo de Proceso para el Análisis Inteligente de Datos guiado por Ingeniería del Conocimiento

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina

Tesista: Cinthia Vegega [1]
Director: José Ángel Olivas Varela [2]
Co-Director: María Florencia Pollo Cattaneo[1]

[1] Grupo de Estudio de Metodologías para Ingeniería en Software (GEMIS)
Universidad Tecnológica Nacional, Facultad Regional Buenos Aires, Argentina
[2] Soft Management of Internet and Learning (SMILe)
Universidad de Castilla-La Mancha, España
cinthia.vegega@gmail.com, flo.pollo@gmail.com,
joseangel.olivas@uclm.es

Palabras claves: Ingeniería del Conocimiento, Inteligencia de Datos, Inteligencia Artificial, Aprendizaje Automático.

1 Motivación

En esta sección se presenta el contexto académico (sección 1.1) y el contexto científico (sección 1.2) que justifica el desarrollo del trabajo de tesis.

1.1 Contexto Académico

Dentro del ámbito de la Universidad de Castilla-La Mancha (España), el grupo de investigación *Soft Management of Internet and Learning* (SMILe) [1] bajo la guía del Dr. José Ángel Olivas Varela desarrolla su trabajo dentro del campo de la aplicación de la Computación Suave, los Sistemas de Apoyo a la Decisiones, el Análisis de Sentimientos y Opiniones y el Análisis Inteligente de Datos, entre otros tópicos de intéres. Asimismo, dentro de la Facultad Regional Buenos Aires de la Universidad Tecnológica Nacional (Argentina), el *Grupo de Estudio de Metodologías para Ingeniería en Software* (GEMIS) [2], a cargo de la coordinación de la Dra. María Florencia Pollo Cattaneo realiza abordajes desde el punto de vista metodológico en el contexto de la Ingeniería en Sistemas de Información, Ingeniería en Software e Ingeniería del Conocimiento. Esto lleva al interés y la cooperación entre ambos grupos de investigación a través de las actividades vinculadas que desarrollan en cuanto a la Ingeniería del Conocimiento y el Análisis Inteligente de Datos.

1.2 Contexto Científico

En la actualidad, existen foros internacionales que comienzan a abordar la importancia de la relación entre la Ingeniería del Conocimiento y el Aprendizaje Automático

(o Machine Learning, en inglés), tal como AAAI-MAKE [3], KEPS [4], IKE [5] y KEOD [6].

La combinación de la Ingeniería del Conocimiento y el Aprendizaje Automático permite complementar las fortalezas y debilidades de ambos, abriendo nuevas posibilidades para la organización del conocimiento, ya que el conocimiento se utiliza normalmente en una forma anárquica. Si bien existen metodologías, tales como CRISP-DM [7], SEMMA [8] o P3TQ [9] que definen un entendimiento del dominio, dichas metodologías, no sistematizan la forma de utilizarlo y no lo relacionan de una forma directa con el Análisis de Datos.

2 Objetivos y Aportes

La Ingeniería del Conocimiento [10,11] es una disciplina que se encuentra vinculada con la Inteligencia Artificial y se orienta a la construcción de Sistemas Inteligentes, los cuales son artefactos que presentan algún comportamiento inteligente en el sentido humano. Dentro de los Sistemas Inteligentes se encuentran los Sistemas Basados en Conocimientos cuya fuente de conocimiento puede provenir de los datos o del conocimiento del dominio.

Los métodos de análisis de datos normalmente utilizan como entradas datos numéricos normalizados y estructurados, sin considerar la naturaleza heterogénea de los repositorios de datos. Si se agregara a este análisis de datos, el conocimiento del dominio se enriquecería y complementaría el desarrollo de los Sistemas Inteligentes. En este sentido, la Ingeniería del Conocimiento puede ayudar con el trabajo de los datos que pueden ser imprecisos, inciertos, tener errores, contradicciones y sesgos. La Computación Suave (o Soft Computing, en inglés), por ejemplo, abarca técnicas tolerantes a la imprecisión y la incertidumbre dentro de la representación del conocimiento. Asimismo, la práctica habitual para encontrar relaciones entre los datos existentes implica “ir a ciegas” en su procesamiento utilizando la aplicación de algoritmos o herramientas de análisis, tal como herramientas estadísticas o de aprendizaje automático [12]. Esto se realiza en forma unidireccional con el propósito de poder interpretar la salida generada por estos algoritmos a fin de verificar si el conjunto de datos puede ser útil o no para la implementación de un Sistema Inteligente. Esta práctica no es útil para un Sistema Inteligente a raíz de su naturaleza compleja.

Teniendo en cuenta este contexto, se propone a partir de los datos existentes y el conocimiento del dominio, sistematizar el proceso de análisis inteligente de los datos utilizando técnicas y herramientas pertenecientes a la Ingeniería del Conocimiento. Este análisis se focalizará en diferentes elementos tales como el contraste y establecimiento de hipótesis entre datos y conocimiento del dominio, el equilibrio bidireccional entre la correlación y la causalidad buscando reglas de asociación en ambos sentidos y la construcción de modelos cognitivos formalizables como taxonomías y ontologías que enriquezcan el análisis.

3 Estado Actual y Trabajo Futuro

El presente trabajo de tesis se encuentra en su propuesta inicial en donde se está analizando el contexto y la temática a desarrollar a fin de elaborar el plan de tesis y presentarlo para su aprobación en el Doctorado en Ciencias Informáticas de la Universidad Nacional de La Plata (Argentina).

Referencias

1. SMILe, https://blog.uclm.es/gruposmile, último acceso 24/11/2021.
2. GEMIS, https://grupogemis.com.ar, último acceso 24/11/2021.
3. AAAI-MAKE, https://www.aaai-make.info, último acceso 24/11/2021.
4. KEPS, https://icaps21.icaps-conference.org/workshops/KEPS, último acceso 24/11/2021.
5. IKE, https://www.american-cse.org/csce2021/conferences-IKE, último acceso 24/11/2021.
6. KEOD, https://keod.scitevents.org, último acceso 24/11/2021.
7. Chapman, P., Clinton, J., Keber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R, CRISP-DM 1.0 Step by step BI guide. Edited by SPSS, https://www.the-modeling-agency.com/crisp-dm.pdf, último acceso 24/11/2021.
8. SAS, Introduction to SEMMA, https://documentation.sas.com, último acceso 24/11/2021.
9. Pyle, D., Business Modeling and Data Mining, Morgan Kaufmann Publishers (2003).
10. García Martínez, R., Britos, P.: Ingeniería de Sistemas Expertos, Editorial Nueva Librería (2004).
11. Palma, J. T., Paniagua-Arís, E., Martín, F., Marín, R.: Ingeniería del Conocimiento. De la Extracción al Modelado de Conocimiento, Inteligencia Artificial, vol. 11, pp. 46-72 (2000).
12. Olivas Varela, J. A.: Inteligencia Artificial, Inteligencia Computacional y Análisis Inteligente de Datos, OBS Business School (2021).

Aprendizaje automático para clasificación anticipada en datos secuenciales [1] [2]

Doctorado en Ciencias de la Computación, Universidad Nacional de San Luis, Argentina.

Tesista: Juan Martín Loyola [1,2]
Director: Marcelo Luis Errecalde [1]
Co-Director: Esteban Gabriel Jobbágy Gampel [2]

[1] Universidad Nacional de San Luis (UNSL), Ejército de Los Andes 950, San Luis, C.P. 5700, Argentina
[2] Instituto de Matemática Aplicada San Luis (IMASL), CONICET-UNSL, Av. Italia 1556, San Luis, C.P. 5700, Argentina

Palabras claves: Detección Anticipada de Riesgos, Clasificación Anticipada de Texto.

1 Motivación

En la formulación tradicional del aprendizaje automático (supervisado) el problema es construir un clasificador que pueda predecir correctamente las clases de nuevos objetos, dados ejemplos de entrenamiento de viejos objetos. El supuesto en este caso es que los ejemplos de entrenamiento corresponden a datos aislados, e independientes entre sí, con suficiente información relevante auto-contenida como para hacer un análisis individual (clasificación) aceptable.

Sin embargo, este esquema de trabajo no se adapta a muchas situaciones del mundo real donde la efectividad del sistema de clasificación depende directamente de considerar las observaciones/datos respetando la secuencia en que se fueron generando. Tomemos, por ejemplo, un modelo del lenguaje que predice la probabilidad de ocurrencia de la siguiente letra. Si el sistema leyó una "Q", la probabilidad de ocurrencia de una "u" será significativamente más alta que la de cualquier otra letra. De igual manera, la interpretación del significado de una palabra como "banco", no será el mismo si previamente dije que "para comprar esta casa debo retirar dinero del" <banco>, que si hubiera dicho "me sentía cansado, por lo que decidí sentarme en el" <banco>. En ambos casos, la palabra polisémica "banco", requiere de las secuencias previas de palabras emitidas, para eliminar cualquier ambigüedad sobre el significado que tiene en cada caso. Esta situación, que hemos ejemplificado con palabras, se repite en un sinnúmero de situaciones involucrando sonidos, imágenes y las más diversas señales sensoriales, en las cuales la correcta interpretación del dato actual de entrada sólo puede realizarse en forma realista,

[1] Video de exposición: https://www.youtube.com/watch?v=y1h6RYVXB2Q
[2] Diapositivas: https://jmloyola.github.io/files/talks/2021_encuentro_posgrado.pdf

considerando la secuencia de datos previos, e incluso en muchos casos, dependiendo de datos producidos muchos pasos hacia atrás en esa secuencia.

En este contexto, esta tesis se enmarca en el área del aprendizaje automático con datos secuenciales (AADS), es decir, asumiremos que el algoritmo de aprendizaje automático explícitamente considera que la entrada es una secuencia.

Varios autores han categorizado las aplicaciones de AADS de distintas formas, dependiendo de las características de la entrada y de la salida. En particular, Graves [1], utiliza como marco de referencia el etiquetado de secuencias (sequence labelling) cuyo objetivo es asignar secuencias de etiquetas (tomadas de un alfabeto fijo), a las secuencias de entrada. En este contexto, el tipo de tarea se vincula a las distintas restricciones que se imponen en ese proceso de etiquetado.

Cuando las secuencias de etiquetas son restringidas a tener longitud uno la tarea recibe el nombre de “clasificación de secuencia”. Si las secuencias de salida consisten en muchas etiquetas, pero los puntos de la secuencia de entrada donde estas etiquetas deben ser producidas son conocidas de antemano, las tareas son referenciadas como de “clasificación de segmentos”. Por último, el escenario que Graves llama “clasificación temporal”, no impone ningún tipo de alineamiento entre las secuencias de entrada y salida, e incluso la de salida puede ser vacía. El elemento crucial que se incorpora en este caso es que el sistema requiere de un algoritmo para decidir en qué lugar de la secuencia de entrada se debería generar la clasificación (etiqueta) correspondiente.

Esta última nomenclatura es de interés para nuestro trabajo, ya que incorpora el aspecto de la decisión de “cuándo” (en qué lugar de la secuencia de entrada) se debería tomar la decisión de generar la etiqueta (clasificación) correspondiente. Este es un aspecto fundamental en un tipo de clasificación temporal que suele ser referenciada como de “clasificación anticipada” (CA). La idea subyacente a la CA es que el clasificador debería ser capaz de poder clasificar la secuencia de entrada tan pronto tenga la información relevante necesaria para poder realizar esta clasificación de manera confiable. La clasificación anticipada suele ser un aspecto deseable, ya que puede en algunos casos evitar algún tipo de costo asociado con la lectura completa de la secuencia de entrada o bien producir una mayor utilidad/beneficio al clasificar anticipadamente el flujo de entrada.

Sin embargo, existen casos donde la CA no es sólo “deseable”, sino también “crítica” ya que existe un riesgo asociado con la demora en la clasificación de la secuencia. Estos escenarios, que serán uno de los ejes de esta propuesta de tesis, se han popularizado últimamente con el nombre de “detección anticipada de riesgos” (DAR) (en inglés “early risk detection”).

2 Objetivos y Aportes

El objetivo principal de esta tesis es el estudio, formulación y desarrollo de representaciones y métodos de aprendizaje automático para datos secuenciales. El interés principal en nuestro caso estará dado en aquellos dominios en los cuales existe una demanda concreta por clasificar las secuencias con la mayor antelación posible.

Particularmente, se busca:

- Relevar el estado del arte en enfoques de AADS y DAR.
- Construir o adaptar colecciones de datos existentes que sean adecuadas para el entrenamiento y evaluación de enfoques de AADS y DAR.
- Definir nuevas representaciones y algoritmos para AADS y DAR que sean representativas del estado del arte.

3 Estado Actual y Trabajo Futuro

La primera etapa del trabajo de tesis estuvo principalmente enfocada en el estudio del estado del arte tanto en enfoques de Aprendizaje Automático con Datos Secuenciales, como en los problemas de Clasificación Anticipada y Detección Anticipada de Riesgo. A partir de esto, se formalizó el marco de trabajo requerido para problemas de Clasificación Anticipada [2] haciendo hincapié en los dos componentes que se deben resolver: Clasificación con Información Parcial (CIP) y la Decisión del Momento de Clasificación (DMC).

Al ser un problema no abordado previamente, no existían, a mediados de 2017, conjuntos de datos con los que se pudieran evaluar los enfoques de clasificación anticipada y comparar con otros grupos de investigación. Afortunadamente, a finales de 2017 se creó el laboratorio de predicción temprana de riesgos en Internet, eRisk[3]. El objetivo del laboratorio es explorar las metodologías de evaluación, las métricas de efectividad y las aplicaciones prácticas de la detección temprana de riesgos en Internet. Todos los años, el laboratorio provee una serie de conjuntos de datos de entrenamiento donde los diferentes grupos entrenan sus modelos, y luego comparan el desempeño de todos en los respectivos conjuntos de datos de prueba.

Una particularidad de los problemas de detección anticipada de riesgo es que suelen tener un desbalance de clases considerable. Esto se debe a que, en general, los casos de riesgo son mucho menores en cantidad que los casos de no riesgo. Así, para mejorar el desempeño de los modelos propuestos, se amplió el conjunto de datos de entrenamiento utilizando información de Reddit.

En la edición del año 2021 del laboratorio eRisk nuestro grupo presentó tres tipos de modelos distintos para detección anticipada de riesgo [3]:

- EarlyModel: modelo simple basado en el marco de clasificación anticipada propuesto en [2]. El rol del CIP puede ser llevado a cabo por cualquier clasificador de texto que retorne la probabilidad de la clase predicha. Por otro lado, para la DMC se utilizó un árbol de decisión.
- SS3: modelo similar al anterior donde el rol del CIP es llevado a cabo por el modelo SS3 [4]. Por otro lado, para la DMC se utilizó una función que considera el contexto de todos los documentos siendo procesados en paralelo.

[3] https://early.irlab.org/

- EARLIEST: modelo de aprendizaje profundo *end-to-end* entrenado utilizando Aprendizaje por Refuerzo para aprender cuándo detener la lectura de la entrada y clasificar. La representación aprendida por el modelo es utilizada tanto para clasificar la entrada en riesgo o no-riesgo, como para determinar si se debe detener la lectura o no.

Como se puede ver en la tabla a continuación, los resultados obtenidos por estos modelos fueron muy alentadores, obteniendo los mejores resultados para la medida F_1 y las medidas que consideran el tiempo de clasificación [5].

team name	run id	P	R	$F1$	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	$speed$	$latency\text{-}weighted\ F1$
UNSL (EarlyModel)	0	.336	.914	.491	.125	**.034**	11	.961	.472
UNSL (EARLIEST)	1	.11	.987	.198	.093	.092	**1**	**1.0**	.198
UNSL (EARLIEST)	2	.129	.934	.226	.098	.085	**1**	**1.0**	.226
UNSL (SS3)	3	.464	.803	.588	.064	.038	3	.992	.583
UNSL (SS3)	4	.532	.763	**.627**	.064	.038	3	.992	**.622**
NLP-UNED	4	.453	.816	.582	.088	.04	9	.969	.564
Birmingham	0	.584	.526	.554	.068	.054	2	.996	.551
Birmingham	2	**.757**	.349	.477	.085	.07	4	.988	.472
EFE	2	.366	.796	.501	.12	.043	12	.957	.48
BLUE	2	.454	.849	.592	.079	.037	7	.977	.578
UPV-Symanto	1	.276	.638	.385	**.059**	.056	1	1.0	.385

Queda pendiente como trabajo futuro analizar por qué el modelo EARLIEST no tuvo el desempeño esperado y proponer mejoras al modelo. Además, nos interesa determinar si agregar más información del contexto puede beneficiar el desempeño de los modelos.

Referencias

1. Graves, A.: Supervised sequence labelling. In: Supervised sequence labelling with recurrent neural networks, vol. 385, pp. 5-13. Springer, Berlin, Heidelberg (2012).
2. Loyola, J.M., Errecalde, M.L., Escalante, H.J., Montes y Gomez, M.: Learning when to classify for early text classification. In Argentine Congress of Computer Science, pp. 24-34. Springer, Cham (2017, October).
3. Loyola, J.M., Burdisso, S.G., Thompson, H., Cagnina, L., Errecalde, M.L.: UNSL at eRisk 2021 A Comparison of Three Early Alert Policies for Early Risk Detection. In Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania (2021, September).
4. Burdisso, S.G., Errecalde, M.L., Montes-y-Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications, vol. 133, 182-197 (2019).
5. Parapar, J., Martín-Rodilla, P., Losada, D.E., Crestani, F.: Overview of erisk 2021 Early risk prediction on the internet. In Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum, Bucarest, Romania (2021, September).

Análisis y diseño de técnicas de preprocesamiento de instancias escalables para problemas no balanceados en Big Data. Aplicaciones en situaciones de emergencias humanitarias.

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata (UNLP), Argentina.

Doctorado en Tecnologías de la Información y la Comunicación, Universidad de Granada (UGR), España.

Tesista: María José Basgall [1,2,3]
Director (UNLP): Marcelo Naiouf[2]
Director (UGR): Alberto Fernández Hilario[3]

[1] UNLP, CONICET, III-LIDI, La Plata, Argentina
[2] Instituto de Investigación en Informática (III-LIDI), Facultad de Informática - Universidad Nacional de La Plata
[3] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada, 18071, Spain

Palabras claves: Big Data, Preprocesamiento, Apache Spark.

1 Motivación

En los últimos años, se generan continuamente datos provenientes de distintas fuentes. Las aplicaciones de Internet (redes sociales, correo electrónico, comercio electrónico), los sensores de dispositivos de uso diario (relojes inteligentes, *smart homes*, *smart cities*), los sensores de experimentos científicos (CERN, Fermilab), son sólo algunos ejemplos [1]. A su vez, esta situación se ha visto exacerbada desde el comienzo de la pandemia COVID-19 con el notorio incremento del uso de Internet (a causa del crecimiento de las herramientas de comunicación online y del teletrabajo) y de la digitalización y toma de datos cotidiana.

Este enorme volumen de datos, realmente variado en su tipología, y que se genera y procesa a gran velocidad, es lo que se conoce como Big Data [2]. La importancia de los datos está en extraer conocimiento de ellos para predicción de comportamientos a futuro y toma de decisiones, lo cual se consigue mediante modelos de Machine Learning. Por lo anterior, Big Data y Machine Learning son tendencia en estos escenarios en donde se busca procesar grandes volúmenes de datos en tiempos razonables y con el fin de obtener conocimiento útil a partir de los mismos y hasta poder predecir comportamientos a futuro. Una búsqueda bibliográfica en *Web-of-Science*[1] revela que el

[1] https://www.webofscience.com/wos/woscc/basic-search

tópico central de esta tesis, <<Big Data>>, es un área de investigación muy popular, con más de 87000 publicaciones que contienen dicho término en el título, en el resumen o en las palabras claves del artículo, en los últimos 10 años (a fecha de 24 de noviembre de 2021). Además, hay que tener en cuenta que *Web-of-Science* ofrece una estimación conservadora; por lo tanto, es probable que el número real de publicaciones sea mucho mayor.

Estas disciplinas están siendo aplicadas cada vez más a distintas áreas. Se conoce su aplicación en la medicina, en la detección de fraudes electrónicos y seguridad informática en general, en el descubrimiento de nueva física, en problemas de bioinformática, en política, en deportes, entre tantos otros ejemplos. Tal es su relevancia que dichos campos de conocimiento no sólo forman parte de los planes de estudios en formación de postgrado de carreras afines, sino que últimamente se han comenzado a incorporar en la oferta de materias de grado en prestigiosas instituciones.

Por ser tendencia, hay que tener en cuenta que son áreas muy dinámicas y, en consecuencia, demandan una continua actualización. Además, para el análisis de Big Data basado en principios de Machine Learning, las técnicas de descubrimiento de conocimiento que hasta ahora ofrecían buenos resultados, no siempre soportan manejar grandes conjuntos de datos. Es por ello que necesitan ser adaptadas para trabajar en entornos distribuidos siguiendo un enfoque "divide y vencerás", o se deben crear nuevas técnicas o estrategias para lidiar con este nuevo escenario. Asimismo, Big Data supone restricciones en cuanto a capacidad de cómputo, de comunicación, de almacenamiento, entre otras cuestiones; por lo tanto, se debe tener acceso a la infraestructura adecuada, que provea los recursos computacionales necesarios [4].

A su vez, hay que tener en cuenta que los conjuntos de datos normalmente pueden tener ciertas características o complejidades no deseadas que interfieren en la efectividad del proceso de extracción del conocimiento [5]. La presencia de ellas puede deberse tanto a la naturaleza del problema que los datos representan, como a la forma en la que son capturados. Entre las complejidades, pueden encontrarse redundancia, ruido, valores faltantes, zonas ambiguas del problema, significativa desproporción de los datos que representan a distintos conceptos claves (desbalance ó desequilibrio), alta dimensionalidad en los datos, entre otras. Frente a la presencia de algunas de estas características, los datos deben ser tratados debido a que la mayoría de los modelos de aprendizaje asumen que los datos están libres de ellas. Las técnicas que se apliquen en cada caso, conforman la etapa de preprocesamiento del proceso de extracción del conocimiento. Ya en escenarios de datos de tamaño estándar, es conocido que esta tarea es una de las que más tiempo demanda para ser completada; por consiguiente, en escenarios Big Data la situación se intensifica. Sin embargo, preprocesar los conjuntos Big Data con el fin de obtener datos de buena calidad (ó también llamado Smart Data [6]), es un paso necesario para conseguir que los modelos alcanzados resulten de utilidad. En este contexto, Smart Data usualmente representa a un subconjunto del conjunto Big Data original, el cual no tiene por qué seguir siendo tan enorme, y a partir del cual será utilizado como entrada a los algoritmos para la extracción de conocimiento.

En consecuencia, esta línea de investigación aborda al preprocesamiento distribuido y escalable de conjuntos Big Data, con el fin de obtener Smart Data, es decir, datos

de mejor calidad. Particularmente se centra en los conjuntos de datos que representan un problema de clasificación, es decir, instancias de datos con sus respectivas etiquetas de clase, y a partir de las cuales se lleva a cabo un proceso de aprendizaje para obtener un modelo que represente a dichos datos. A continuación, y mediante la utilización de dicho modelo, poder clasificar nuevas instancias de datos de las cuales se desconoce su etiqueta de clase.

2 Objetivos y Aportes

El objetivo general de esta tesis doctoral es contribuir al área de preprocesamiento en el contexto de Big Data, dada la escasa cantidad de soluciones distribuidas en esta temática capaces de manipular estos tipos de datos. Puntualmente, proponer nuevas soluciones escalables con foco al tratamiento de las características intrínsecas de los datos que se encuentran más frecuentes en problemas Big Data. En concreto, atender al desbalance de datos, la redundancia y alta dimensionalidad, y al solapamiento de clases.

Para ello, se establecen los siguientes objetivos específicos:

- Habilitar que un algoritmo del estado del arte altamente empleado para el tratamiento de la distribución de clases en escenarios de datos tradicionales (Small Data), sea capaz de obtener resultados adecuados a partir de grandes conjuntos de datos de manera distribuida y en tiempos de ejecución razonables.
- Diseñar e implementar una metodología rápida y escalable para la reducción tanto en instancias como en atributos para los conjuntos Big Data que presentan alta redundancia y dimensionalidad, manteniendo la capacidad predictiva del conjunto de datos original.
- Diseñar e implementar una estrategia para la caracterización escalable de los datos en el contexto de clasificación en Big Data, con foco en las zonas ambiguas del problema.

En todos los casos, las implementaciones de nuestras propuestas se han llevado a cabo utilizando el framework de computación distribuida denominado Apache Spark a través del lenguaje de programación Scala. Las herramientas generadas son *open source* y se encuentran disponibles en sus respectivos repositorios.

3 Estado Actual y Trabajo Futuro

Con el fin de validar las propuestas antes mencionadas, se realizaron distintos experimentos sobre una amplia gama de conjuntos Big Data públicamente disponibles en repositorios de libre acceso. Los conjuntos de datos cubren una extensa variedad de escenarios Big Data, encontrándose diferentes cantidades de instancias, de atributos, como así también diversos grados de redundancia y de desproporción entre clases. A su vez, se evaluaron distintas métricas para medir el desempeño de aprender de los datos originales como de sus correspondientes versiones preprocesadas. En todos

los experimentos se ha llevado a cabo el método de validación cruzada para obtener los resultados.

De la experimentación se observa, por un lado, que el proceso de generación de instancias con el fin de balancear las clases de un problema se desempeña adecuadamente, alcanzando los mismos resultados que la solución secuencial en datasets de tamaño tradicional. Además, a partir de la estabilidad alcanzada en los resultados obtenidos de aplicar distinto grado de paralelismo (tanto en datos de tamaño tradicional como en Big Data), se observa que es un algoritmo totalmente escalable. Por otro lado, respecto a la condensación de datos, los resultados muestran la fortaleza de nuestra propuesta obteniendo valores de reducción elevados para la mayoría de los conjuntos de datos estudiados, tanto en lo que respecta a la dimensionalidad como a los porcentajes de reducción de instancias propuestos. Se alcanzó alrededor del 70 % de reducción de las características y 98 % de reducción de las instancia, para un umbral de pérdida predictiva máxima aceptada del 1 % del cual, en algunos casos, la calidad predictiva se mantuvo igual a la del conjunto original o incluso un poco por encima de la misma.

Como trabajo a futuro, la idea es extender todas nuestras propuestas para problemas multiclases. Además, en futuras investigaciones, sería de utilidad incluir un mayor número de métricas de complejidad de los datos, para buscar profundizar en la caracterización de los datasets de manera exhaustiva. Por último, y en relación a la gran reducción que se obtuvo en muchos dataset manteniendo el poder predictivo de los datos originales, sería de interés aplicarles además alguna otra técnica de preprocesamiento adicional con foco en mejorar su poder predictivo.

Referencias

1. Bousdekis, A. et al. (2021). A Review of Data-Driven Decision-Making Methods for Industry 4.0 Maintenance Applications. Electronics, 10(7), 828. https://doi.org/10.3390/electronics10070828
2. Marx, V. (2013). The big challenges of big data. Nature, 498(7453), 255–260. https://doi.org/10.1038/498255a
3. González, R. J. (2017). Hacking the citizenry?: Personality profiling, 'big data' and the election of Donald Trump. Anthropology Today, 33(3), 9–12. https://doi.org/10.1111/1467-8322.12348
4. Fernández, A. et al. (2014). Big Data with Cloud Computing: An insight on the computing environment, MapReduce, and programming frameworks: Big Data with Cloud Computing. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 380–409. https://doi.org/10.1002/widm.1134
5. Das, S. et al. (2018). Handling data irregularities in classification: Foundations, trends, and future challenges. Pattern Recognition, 81, 674–693. https://doi.org/10.1016/j.patcog.2018.03.008
6. García-Gil, D. et al. (2019). Enabling Smart Data: Noise filtering in Big Data classification. Information Sciences, 479, 135–152. https://doi.org/10.1016/j.ins.2018.12.002

Invarianzas en modelos de Aprendizaje Automático Profundo (Deep Learning). Aplicaciones en Visión por Computadora

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina

Investigador: Dr. Quiroga, Facundo Manuel[1,2]
Directora: Dra. Lanzarini, Laura Cristina[1]

[1] Instituto de Investigación en Informática III-LIDI, Facultad de Informática, Universidad Nacional de La Plata, Argentina
[2] Becario Posdoctoral UNLP
fquiroga@lidi.info.unlp.edu.ar

Palabras claves: Redes Neuronales, Visión por Computadora, Invarianza, Aprendizaje Automático Profundo, Deep Learning

1 Motivación

Las Redes Neuronales son los modelos de aprendizaje automático con mejor desempeño en la actualidad en una gran variedad de problemas. Son modelos generales y aproximadores universales. Con algoritmos de optimización basados en descenso de gradiente, pueden optimizar miles o millones de parámetros en base a una función de error. Se distinguen de otros modelos en que no requieren un diseño manual de características de los datos para funcionar; las características se aprenden automáticamente mediante el proceso de optimización, también llamado entrenamiento. Su diseño se organiza en capas que determinan su arquitectura. En los últimos años, se ha conseguido entrenar Redes Neuronales con múltiples capas mediante un conjunto de técnicas que suelen denominarse Aprendizaje Profundo (Deep Learning) [1].

En particular, las Redes Convolucionales, es decir, Redes Neuronales que utilizan capas convolucionales, son el estado del arte en la mayoría de los problemas de visión por computadora, incluyendo la clasificación de imágenes. Las capas convolucionales permiten aplicar convoluciones con filtros aprendidos para un mejor desempeño y eficiencia.

Muchos de los problemas para los cuales las Redes Convolucionales son el estado del arte requieren que los modelos se comporten de cierta manera ante transformaciones de su entrada. Existen dos propiedades fundamentales que capturan dicho requerimiento; la invarianza y la equivarianza. La invarianza nos dice que la salida del modelo no es afectado por las transformaciones. La equivarianza permite que la salida sea afectada, pero de una manera controlada y útil.

Si bien los modelos tradicionales de Redes Convolucionales son equivariantes a la traslación por diseño, no son ni invariantes a dicha transformación ni equivariantes a otras en los escenarios usuales de entrenamiento y uso. Existen dos opciones principales para otorgar invarianza o equivarianza a un modelo de red neuronal. La tradicional ha sido modificar el modelo para dotarlo de esas propiedades. La otra opción es entrenarlo con aumentación de datos utilizando como transformaciones el mismo conjunto al que se desea la invarianza o equivarianza [1].

Dotar con invarianza o equivarianza a los modelos tiene utilidades en varios dominios, como la clasificación de imágenes de galaxias, imágenes de microscopios o formas de mano. En particular, el reconocimiento de formas de mano en imágenes es una de las etapas más importantes de los sistemas de reconocimiento de lenguas de señas o gestos mediante imágenes o video. En muchos casos, la rotación, traslación o escalado de la mano en la imagen no afectan a su forma, y por ende se requiere dotar de invarianza a la red para mejorar el desempeño del sistema.

No obstante, no está claro cómo los modelos adquieren estas propiedades, tanto al usar aumentación de datos como al modificar el modelo. Tampoco está claro como las modificaciones de modelos afectan la eficiencia y el poder de representación de los mismos. Más aún, en los modelos tradicionales tampoco es conocido cómo se adquieren dichas propiedades con aumentación de datos, así como cuál es la mejor estrategia para aumentar los datos con este fin.

2 Objetivos y Aportes

El objetivo general en esta línea de investigación contribuir al entendimiento y mejora de la equivarianza de los modelos de redes neuronales, en particular aplicados a la clasificación de formas de mano para la lengua de seña y otros tipos de gestos mediante modelos de redes convolucionales.

Para ello, establecimos los siguientes objetivos particulares:

- Analizar los modelos específicos para equivarianza en CNNs.
- Comparar los modelos específicos y la aumentación de datos para obtener equivarianza. Evaluar estrategias de transferencia de aprendizaje para obtener modelos equivariantes a partir de modelos que no lo son.
- Desarrollar métricas de equivarianza para las activaciones o representaciones internas de las redes neuronales. Implementar las métricas en una librería de código abierto. Analizar el comportamiento de las métricas. Comparar con las métricas existentes.
- Caracterizar modelos de CNN para la clasificación de imágenes en términos de su equivarianza con las métricas propuestas.
- Comparar los modelos de CNN, con y sin equivarianza, para la clasificación de formas de mano.

3 Estado Actual y Trabajo Futuro

Analizamos diversas estrategias para obtener invarianza o equivarianza en modelos de clasificación de imágenes con redes neuronales. Comparamos los modelos tradicionales AllConvolutional y LeNet, y los modelos especializados Group CNN y Spatial Transformer Networks para determinar su desempeño. Realizamos experimentos con varios conjuntos de datos conocidos (MNIST y CIFAR10) utilizando aumentación de datos.

Los resultados arrojan evidencia en favor de la hipótesis de que aún con ingeniosas modificaciones de las redes convolucionales, la aumentación de datos sigue siendo necesaria para obtener un desempeño similar al de los modelos no invariantes. Más aún, en varios casos la aumentación de datos por si sola puede proveer un desempeño similar al de los modelos especializados, siendo al mismo tiempo más simples de entrenar y comprender [2].

Además, analizamos cómo re-entrenar una red previamente generada para convertirla en invariante, y encontramos que el entrenamiento de las últimas capas permite convertir un modelo no invariante en uno que si lo sea con un bajo costo computacional y leve pérdida de desempeño.

Si bien estos mecanismos permiten imbuir de invarianza o equivarianza una red, la forma en que la misma codifica o representa dichas propiedades no están claros. La comprensión de la invarianza o equivarianza de una red o cualquier sistema puede ayudar a mejorar su desempeño y robustez. Estas propiedades pueden estimarse midiendo los cambios en las salidas de la red en base a las transformaciones realizadas a su entrada [2].

Las metodologías actuales de evaluación y comprensión de la invarianza y equivarianza se enfocan solamente en las capas de salida de la red. No obstante, para poder comprender como se codifican, el análisis debe realizarse en base a toda la red, es decir, considerando las representaciones intermedias [1].

Desarrollamos métricas para medir la invarianza y equivarianza de las redes. Dichas métricas permiten cuantificar estas propiedades de forma empírica no solo en la salida de la red sino también en sus representaciones internas. De esta forma, podemos visualizar y cuantificar que tan invariante o equivariante es una red, ya sea en su totalidad, por capas, o por activaciones individuales. Las métricas son aplicables a cualquier red neuronal, sin importar su diseño o arquitectura, así como a cualquier conjunto de transformaciones. Realizamos una implementación de las métricas en una librería de código abierto, con soporte para la librería tensorial PyTorch. Las métricas fueron validadas para verificar su correcto funcionamiento y utilidad. Además,

estudiamos sus propiedades, como la variabilidad ante los conjuntos de datos, transformaciones, inicialización de los pesos, y otras [2].

Utilizando las métricas, también evaluamos modelos de redes neuronales convolucionales conocidos para caracterizarlos en términos de su invarianza o equivarianza. Asimismo, caracterizamos diversos tipos de capas como las de Batch Normalization, Max Pooling, diversas funciones de activación, capas convolucionales con distintos tamaños de filtro, y otros. Los resultados otorgan una primera mirada de los modelos de redes en términos de estas propiedades, y esperamos que puedan fomentar un mejora en ese área.

Por último, hacemos un tercer aporte al reconocimiento automático de lengua señas basado en video. El reconocimiento de señas es un subárea del reconocimiento de gestos o acciones. Tiene como objetivo traducir al lenguaje escrito un video en donde una persona se comunica mediante lengua de señas. Desde la aparición de tecnologías de captura de video digital existen intentos de reconocer gestos y señas con diferentes fines. Es un problema multidisciplinar complejo y no resuelto aún de forma completa.

Utilizando los conjuntos de datos de formas de mano LSA16 y RWTH-PHOENIX-Weather, realizamos experimentos con los modelos LeNet, VGG16D, ResNet, Inception y AllConvolutional para determinar su eficacia como clasificadores en este dominio. Los resultados indican que todos los modelos tienen un desempeño razonable en ambos conjuntos de datos, con resultados iguales o mejores que otros modelos diseñados específicamente para la tarea. No obstante, el modelo VGG16D obtuvo los mejores resultados. Incluimos también evaluaciones de transferencia de aprendizaje, con y sin re-entrenamiento de las capas; en ambos casos dichas estrategias obtuvieron un desempeño peor que los modelos entrenados sin transferencia de aprendizaje. Además, realizamos un estudio de varias estrategias de pre-procesamiento de las imágenes, encontrando que la segmentación de las manos del fondo otorga un incremento de desempeño significativo. Por último, también desarrollamos una librería de código abierto para facilitar el acceso y preprocesamiento de bases de datos de formas de manos.

Referencias

1. Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges.
2. Quiroga, F. M. (2020). Medidas de invarianza y equivarianza a transformaciones en redes neuronales convolucionales (Universidad Nacional de La Plata).

Representación multinivel para razonamiento por analogía utilizando factores clave en sistemas de aprovechamiento inteligente de datos

Doctorado en Tecnologías Informáticas Avanzadas, Universidad de Castilla La Mancha (UCLM), España.

Tesista: Antonio Lorenzo [1,2]
Director: José A. Olivas[2]

[1] Coordinador del Departamento de Business Intelligence, Gobierno de Castilla-La Mancha, Toledo, España.
alorenzo@jccm.es
[2] SMILe (**S**oft **M**anagement of **I**nternet and **L**earning). Escuela Superior de Informática, Universidad de Castilla-La Mancha, Ciudad Real, España.
JoseAngel.Olivas@uclm.es

Palabras claves: Predicción, Aprendizaje por analogía, Lógica Borrosa.

1 Motivación

Para los humanos ha sido una constante intentar anticiparse al futuro. Tradicionalmente se ha usado la experiencia y la intuición. En la historia más reciente de la predicción podemos distinguir dos periodos: el primero, durante el s. XX con el inicio de la estadística moderna en la cual se utilizaron diversas técnicas como las series temporales, las regresión o algoritmos (ARM, ARIMA…). Y una segunda aproximación, que comienza a mediados del s. XX con el inicio de la Inteligencia Artificial (IA), y alcanza su apogeo en el s. XXI. Se caracteriza por la disponibilidad de grandes volúmenes de datos, así como con el aumento de la capacidad de proceso, en las cuales se han desarrollado las herramientas de aprendizaje automático. Actualmente las técnicas de aprendizaje automático son las más utilizadas.

Las técnicas estadísticas para la predicción son adecuadas cuando hay una proyección de los datos al futuro, cuando se sigue la tendencia y hay una relación lineal entre los datos de entrada y los datos de salida, por ejemplo, podemos predecir la evolución de la población teniendo en cuenta la tendencia de los últimos años. En cambio, las técnicas de aprendizaje automático, dan buenos resultados cuando el modelo ha sido entrenado con el mismo tipo de casos que se desea predecir.

Pero no todos los eventos se pueden predecir aplicando la estadística o el aprendizaje automático. Hay eventos que no basta con procesar los datos históricos y actuales, se necesita aplicar conocimiento adicional y específico del evento a predecir. Solo con el procesamiento y análisis de los datos históricos no se pudo predecir que Donald Trump iba a ganar las Elecciones Presidenciales a Hillary Clinton en EEUU 2016, que iba resultar vencedor el Referéndum de la salida del Reino Unido de la

Unión Europea (Brexit) o que se iba a perder el Plebiscito sobre los acuerdos de paz de Colombia entre el Gobierno de Colombia y las FARC-EP. En todos ellos hubo factores claves que no fueron tenidos fallando las predicciones.

La predicción de eventos complejos se distingue porque hay componentes que juntos producen comportamientos no triviales y que no pueden explicarse analizando sus componentes por separado. Es una agregación cualitativa de componentes, no cuantitativa.

2 Objetivos y Aportes

Proponemos utilizar técnicas de Inteligencia Artificial (IA) sofisticadas apoyadas en la Ingeniería del Conocimiento para mejorar los resultados de las técnicas estadísticas y de aprendizaje automático habituales en la predicción de eventos complejos. Sobre la propuesta, se ha de delimitar el ámbito, el alcance y el contexto:

- El ámbito. De los escenarios de predicción (casi certeza, mantenimiento de tendencia, incertidumbre y azar) se aplica al escenario de incertidumbre en el cual solo se conoce parte de la información.
- El alcance. El comportamiento del evento a predecir se debe de materializar en una fecha concreta. No es válida para predecir cuándo sucederá un evento: cuándo ganará el Real Madrid 5-0 al FCB o cuándo alcanzará los 12.000 puntos el Ibex35.
- El contexto. El mismo caso a predecir, pero en diferentes contextos, puede producir resultados distintos. La predicción del caso “cuántos Diputados obtendrá cada partido político en las próximas Elecciones Generales” es diferente según el momento político, social, económico… en el que se produzcan.

3 Estado Actual y Trabajo Futuro

La propuesta está basada en conceptos de aprendizaje por analogía y más concretamente, en el Razonamiento Basado en Casos (CBR), que es adecuado para predecir eventos complejos debido a que no es sencillo formalizarlos mediante reglas, pero existen ejemplos anteriores en los que basarnos, que de alguna manera se pueden describir mediante características (*features*) permitiendo la comparación entre ellos. Los modelos clásicos de aprendizaje por analogía, y en particular, el Razonamiento Basado en Casos consta de 4 fases: recuperar, reutilizar, revisar y recordar.

Se tratar de encontrar el evento anterior más similar al nuevo evento a predecir, de tal forma que, si hay cierta analógica entre las causas del evento anterior con el evento futuro, por analogía con el nuevo evento, también se darán parte de los resultados. Tras seleccionar el evento anterior más similar, se adaptará la solución aplicando conocimiento experto al evento actual teniendo en cuenta los factores claves.

Los elementos que se proponen son los siguientes:

- Representación del conocimiento (Fig.1). Se representa en una estructura jerárquica multinivel en la cual se relacionan los eventos anteriores, sus características y sus resultados.

- La *Base de eventos* (*Events Base*) está compuesta por un conjunto finito de eventos, $E = (e_1, e_2\ldots, e_n)$.
- Un Evento (Event*, E*) se define como un vector que identifica el evento. Es una tupla con un identificador y unas descripciones del evento. Por ejemplo, si los eventos a representar fueran "Nº Diputados obtenidos por cada partido político en las Elecciones Generales", los *Eventos* serían E_1 (#1, DIC, 2015), E_2 (#2, JUN, 2016), E_3 (#3, ABR, 2019) ...
- Las *Características* (*Features, F*) describen el evento y representan las causas. Son aportadas por el dominio del conocimiento. Se representa con una tupla que contiene la denominación de la característica y una etiqueta lingüística borrosa que indica la influencia de esa característica en el evento. En el caso de las "Elecciones Generales", F_1 (#1, #1, Crisis Económica, Bajo), F_2 (#1, #2, Corrupción, Bajo), F_3 (#1, #3, Fragmentación izquierda, Medio), F_4 (#1, #4, Fragmentación derecha, Alto) …
- Los *Resultados* (*Results, R*) es un conjunto que representa los *efectos* de cada evento. Cada evento, está relacionado con un resultado a través de una combinación de características. Se representa con una tupla que contiene una denominación y un valor. En el ejemplo anterior: R_1(#1, #1, Partido Político1, Diputados, 90), R_2 (#1, #2, Partido Político 2, Diputados, 123) …
- Las *Relaciones*. Son las uniones entre los distintos componentes. Un *Evento* se describe por un conjunto de *Características,* y la combinación de varias características, obtienen unos *Resultados* determinados.

- Analogía entre eventos. La recuperación de un evento anterior más adecuado para el evento actual requiere realizar un proceso de "adaptación parcial" entre eventos. En general, este proceso no es totalmente preciso, porque las características que describen los eventos anteriores no son exactamente las mismas ni tiene la misma influencia que en el evento actual. Para ello se ha establecer una función de analogía local que permita comparar características de cada evento, de forma particular, y una función de analogía global que permita comparar eventos. Las características que describen los eventos son etiquetados con etiquetas borrosas (p.e. "muy bajo", "bajo", "medio", "alto", "muy alto") dependiendo de la influencia de la característica en el evento. Las etiquetas borrosas se transforman se valores numéricos a los cuales se les aplica las funciones de analogía.

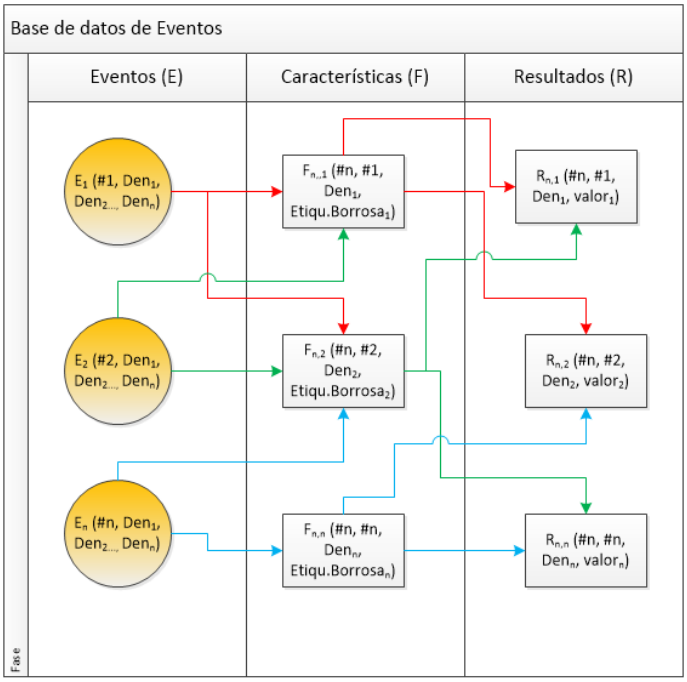


Fig. 1. Representación de los eventos

- Predicción del resultado del nuevo evento. La adaptación del resultado del evento recuperado se realiza mediante la identificación y cuantificación de factores clave específicos en el contexto del evento actual. Estos factores claves los especifica un experto en el dominio del conocimiento. Los nuevos factores claves se representan con una tupla y se cuantifican mediante etiquetas borrosas. Se hace inferencia borrosa a partir de los factores clave, con el fin de obtener la predicción del nuevo evento:
    - Para cada factor clave y resultado se determinan sus rangos numéricos y sus funciones de pertenencia.
    - Se construye la base de reglas. Se genera una tabla de reglas con los valores de todas las posibles combinaciones de los factores claves y sus resultados. Las reglas son del tipo "*IF* Fc_1= $ValBorroso_{1,1}$ *AND* Fc_2= $ValBorroso_{1,2}$ … *THEN* $Result_{1,1}$".
    - Borrosificar los factores clave. El experto determinará la combinación de valores más probables para los factores claves. Se determinan los grados de pertenencia para cada uno de los valores de los factores claves y se evalúan las reglas.
    - Razonamiento. Se disparan las reglas y se calcula el conjunto borroso de salida.
    - Desborrosificar usando alguno de los métodos habituales (centro de masas, criterio del máximo, media del máximo…).

Modelo de madurez para servicios de gobierno electrónico en el ámbito universitario

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina.

Tesista: Ariel Pasini [1]
Directoras: Patricia Pesado [1] Elsa Estévez [2]

[1] Instituto de Investigación en Informática LIDI (III-LIDI)*
Facultad de Informática – Universidad Nacional de La Plata 50 y 120 La Plata Buenos Aires
*Centro Asociado Comisión de Investigaciones Científicas de la Pcia. de Bs. As. (CIC)

{apasini, ppesado}@lidi.info.unlp.edu.ar

[2] Laboratorio de Ingeniería de Software y Sistemas de Información (LISSI)
Departamento de Ciencias e Ingeniería de la Computación – Universidad Nacional del Sur
Av. San Andrés 800 – Campus de Palihue - Bahía Blanca, Buenos Aires
Centro Asociado CIC

ece@cs.uns.edu.ar

Palabras claves: Gobierno electrónico; servicios públicos; gobierno universitario; servicios universitarios; madurez de servicios universitarios.

1 Motivación

La evolución del concepto de gobierno electrónico en los últimos veinte años ha sido notable desde todos los puntos de vista, pero los cambios más destacables se relacionan con: La mejora de procesos gubernamentales, mayor interacción con el ciudadano y la construcción de nuevos canales de comunicación con los diferentes receptores de los servicios de gobierno. Los tres aspectos tienen como objetivo común mejorar las relaciones entre las partes interesadas con el uso de la tecnología, logrando una prestación de servicios a los receptores mediante procesos más eficientes y una utilización eficaz de los recursos.

A lo largo de los años el gobierno electrónico ha pasado por diferentes instancias: desde fines de los 90 a mediados de los 2000, el concepto se basaba en una presencia e institucional en la web, con un sentido principalmente unidireccional, proporcionando información al ciudadano. Desde mediados de los 2000 a inicios del 2010, con el avance de la web 2.o y la creación de las redes sociales, se inicia una etapa de socialización de servicios públicos, donde aparece la comunicación bidireccional entre el gobierno y los ciudadanos, los receptores dejan de tener un rol pasivo que solo recibe información y pasan a tomar un rol más participativo en el uso de los servicios públicos en línea. En el mismo periodo crece el uso de dispositivos móviles, obligan-

do a los gobiernos a innovar en ese tipo de tecnologías para lograr una mayor captación de ciudadanos. Desde el 2010 las comunicaciones móviles, en particular el uso de datos móviles pasa a ser parte de la vida cotidiana de los ciudadanos. Lo que conllevo a una demanda importe de la comunidad para que los gobiernos aumenten la prestación de servicios públicos en línea. Todo el proceso evolutivo fue generando un alto nivel de información digitalizada, importantes bases de datos que con el tiempo se fueron convirtiendo en uno de los principales insumos para el análisis de datos que hoy en día utilizan los gobiernos para apoyar la toma de decisiones.

1.1 Evolución del conceto del gobierno electrónico

En [1] se presenta un modelo de 4 etapas que va creciendo en complejidad tecnológica y política.

1. Digitalización o Tecnológico
2. Transformación o Gobierno electrónico
3. Compromiso o gobernanza electrónica
4. Contextualización o gobernanza dirigida por políticas

La cuarta etapa apunta a identificar las necesidades puntuales de los sectores generando servicios acordes a los requisitos del contexto. Cada contexto limita el alcance de las políticas que periten el desarrollo las diferentes actividades de gobierno. El contexto puede ser regional, territorial, cultural, etc. Para cada uno de estos contextos podemos ver al gobierno digital como base para el desarrollo social, económico, político, cultural, educativo, etc. Donde las necesidades de cada uno de los sectores son adaptadas al contexto de pertenencia.

La idea de contextualización de gobernanza digital en el ámbito universitario se presenta como uno de los contextos de gobierno que se ve alcanzado por el gobierno electrónico. La Ley de Educación Superior (LES) [2], da políticas explicitas de cómo debe ser gobernada una universidad pública en nuestro país.

Las Universidades Nacionales, posee su propio gobierno democrático interno y gozan de autonomía del gobierno político del Estado. El gobierno universitario está compuesto por alumnos, docentes, no docentes, graduados, alumnos de posgrado, etc., dependiendo de la estructura definida en su estatuto y todos en su conjunto representan a los ciudadanos que desarrollan sus actividades en el marco de las reglamentaciones que dispone dicho gobierno.

Los gobiernos que dirigen las universidades, para cumplir con el objetivo principal de brindar una formación académica de alto nivel, prestan un amplio conjunto de servicios a su comunidad, dentro de los que se encuentra por ejemplo *registrar a un alumno en una asignatura* o *registra la nota de un alumno*. Estos servicios, casi transparentes para la comunidad de una universidad, son brindados a través del uso de las TICs. En línea con el concepto de gobierno electrónico presentado anteriormente, podemos inferir que los gobiernos universitarios, también pueden ser incluidos dentro del mismo concepto, incluyendo el modelo de madurez de servicios públicos propuesto por la ONU [3] donde se definen 4 niveles, basados en el nivel de automatización: 1) emergentes, 2) mejorado, 3) transaccional y 4) integrado.

2 Objetivos y Aportes

El objetivo principal de esta tesis es definir un modelo de evaluación, basado en los aplicados en estructuras de gobiernos masivas, que permita clasificar en niveles de madurez a los servicios prestados por las unidades académicas y, en consecuencia, permita definir un escalafón de unidades académicas en la prestación de servicios con el fin de ofrecer recomendaciones para mejorar la prestación de servicios a su comunidad.
Las principales contribuciones de la tesis es la vinculación de los conceptos de gobierno electrónico al ámbito del contexto de los gobiernos universitarios y la propuesta del modelo de evaluación de madurez para servicios de gobierno electrónico en el ámbito universitario

3 Estado Actual y Trabajo Futuro

Se desarrollo un modelo de evaluación que busca estandarizar la identificación de los servicios prestados por las universidades a fin de obtener una denominación común en todas las estructuras de unidades académicas analizadas. Para lo cual se tomaron como punto de partida 24 servicios: 8 correspondientes a los alumnos, 8 a los docentes, 4 a los graduados y 4 a los no docentes. La selección de los servicios se relacionó con las actividades básicas que tiene que realizar una unidad académica y que representen a gran parte de la comunidad universitaria.
Seleccionados los servicios se analizó el alcance de los servicios dentro de la unidad académica, y se los clasifico según los receptores de los servicios, tipo de servicios (informacional, autorización, certificación y control) y tipo de comunicación, identificando la combinación óptima para que cada servicio pueda alcanzar el máximo nivel de madurez. En base a los mismos criterios se le otorgará una calificación al servicio que permitirá realizar una evaluación cuantitativa de los mismos.
Por otro lado, se definió una escala formada por cinco niveles, que representan la evolución de la unidad académica en la prestación de servicios, dicho nivel se obtendrá en función de los valores cuantitativos obtenidos por cada uno de los servicios prestados. Permitiendo determinar el rango en la que se encuentra cada una de las unidades académicas.
Dado que el modelo se basa en 24 servicios, y el modelo de madurez apunta a un proceso de mejora continua, además se describen los pasos a seguir para una exención del modelo que permita asegurar el crecimiento de la unidad académica más allá de los servicios planteados en el modelo.

3.1 Validación del modelo

Se invito a participar en el proceso de validación a 20 unidades académicas, de las cuales aceptaron 18, la evaluación se realizó median un instrumento de recopilación de información, publicado en Google Forms, que contenía 115 preguntas agrupadas en 5 secciones.

3.2 Resultados Obtenidos

Las respuestas obtenidas de las 18 unidades académicas se procesaron según los tipos servicios universitarios prestados para cada grupo de receptores y por unidad académica.
El análisis por unidad académica se realizó desde 4 puntos de vista:

1. Alcance de los servicios
2. Receptores de los servicios
3. Utilización de los canales de comunicación
4. Clasificación de las UA

De las 18 UA, todas se encuentra en el rango satisfactorio, dentro de ese conjunto se clasifican de la siguiente forma:

- 7 UA se encuentran en el rango **Mínimamente Aceptable.**
- 9 UA se encuentran en el rango **Aceptable**.
- 2 UA se encuentran en **Rango Objetivo**.

Finalizado el proceso de evaluación, para cada una de las unidades académicas se les realizo una serie de recomendaciones sobre los servicios universitarios que debería mejorar elevar el nivel en la prestación de estos y en consecuencia el rango de la UA.

Referencias

[1] J. Bertot, E. Estevez, and T. Janowski, “Universal and contextualized public services: Digital public service innovation framework,” *Gov. Inf. Q.*, vol. 33, no. 2, pp. 211–222, 2016, doi: 10.1016/j.giq.2016.05.004.

[2] Republica Argentina, “Ley Nro 24521 - Ley de Educacion Superior,” 1995.

[3] G. Concha and A. Naser, “CEPAL - El desafío hacia el gobierno abierto en la hora de la igualdad,” 2012, [Online]. Available: http://www.eclac.cl/cgi-bin/getProd.asp?xml=/publicaciones/xml/9/46119/P46119.xml&xsl=/publicaciones/ficha.xsl&base=/publicaciones/top_publicaciones.xsl.

Metodología para la Evaluación de un Algoritmo Heurístico de Búsqueda basado en Muestreo y Agrupación

Doctorado en Informática, Universidad Autónoma de Barcelona, España.

Tesista: María de los Angeles Harita Rascón [1]
Director: Dolores Rexachs [1]
Co-Directores: Emilio Luque [1], Alvaro Wong González [1]

[1] Universidad Autónoma de Barcelona, Campus Bellaterra 08193, España

Palabras claves: Método heurístico, Agrupamiento, Optimización, Benchmarks.

1 Motivación

El trabajo de esta tesis se enmarca en el ámbito de la optimización y más concretamente en la optimización mediante los métodos heurísticos. Como antecedentes, partimos del diseño de un algoritmo heurístico [1] de búsqueda enfocado a un problema de optimización en el que se trata de encontrar la mejor configuración posible de staff médico capacitado para la sala de urgencias de un hospital en Sabadell, Barcelona.

Dado que estamos ante un problema de optimización naturalmente clasificado como complejo, según su dificultad y los recursos necesarios para resolverlo, el método Montecarlo [2] plus K-means (en adelante MCKM) se enfoca en la optimización no convencional con el planteamiento de que, ya que no es posible aplicar los métodos tradicionales para resolver este tipo de problemas por la cantidad de cómputo que sería necesario, una reducción de estos tiempos sólo podría ser posible aplicando técnicas heurísticas.

2 Objetivos y Aportes

La tesis se enfoca en la evaluación del método MCKM. Nuestro objetivo es el diseño y la evaluación del método heurístico basado en Montecarlo en conjunto con técnicas de agrupamiento utilizando las funciones tipo Benchmark [3] para analizar la calidad de los resultados y garantizar su eficiencia. Es decir, por un lado, hay que contrastar la calidad de los resultados, evaluar la capacidad del algoritmo y validar utilizando las funciones de Benchmark. Por otro lado, realizamos un análisis de las técnicas de agrupamiento que se adapten a nuestro modelo y a partir de aquí, proponemos una mejora en el diseño del método heurístico.

3 Estado Actual y Trabajo Futuro

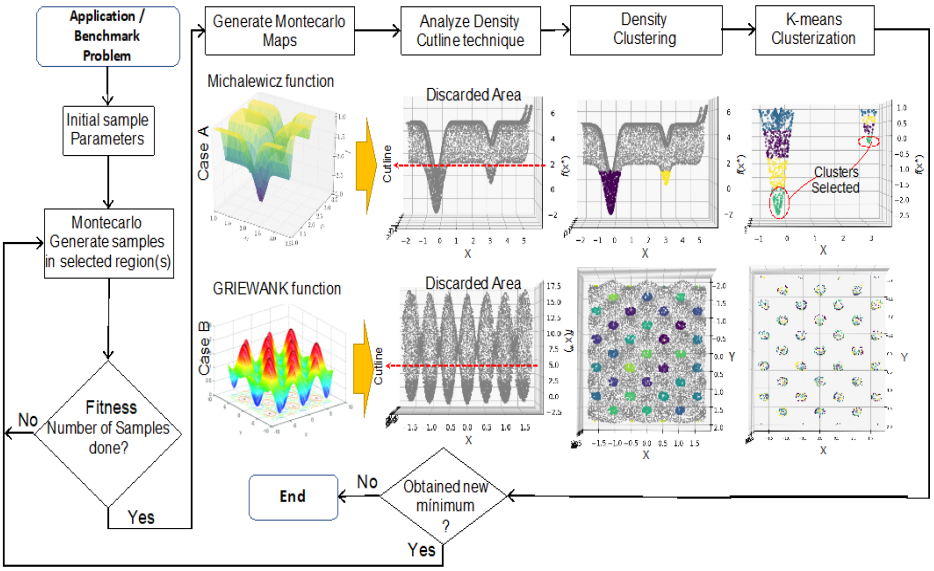


Figura 1 Visión global de la metodología.

Como se muestra en la Figura 1, nuestra propuesta comprende 4 etapas. En la primera de ellas se hace la implementación del problema como input y se llevan a cabo las técnicas de muestreo con los métodos de Montecarlo. Esto da lugar a la generación de los mapas de Montecarlo; Posteriormente, en la segunda etapa, se realiza el análisis de este mapa llevando a cabo una agrupamiento por densidad [4]. Al mismo tiempo que buscamos reducir el área de búsqueda, tenemos la posibilidad de encontrar zonas potenciales gracias a la propuesta en la etapa 3, de realizar un corte por el eje de la función *f(x),* con el fin de mejorar el funcionamiento del algoritmo de agrupamiento por densidad. Finalmente, se hace una nueva agrupación, como se puede ver en la etapa 4 con k-means, dando lugar a regiones factibles en las que es posible encontrar el valor más cercano al óptimo.

Una de las características de este método, es que hemos visto la necesidad de preparar los problemas. Partiendo de la base de que cada problema de optimización posee unas variables, en el caso más simple, su resolución consistiría en maximizar o minimizar una función real. Por tanto, esta preparación consistiría en definir su categoría, continua o discreta, los dominios del espacio de búsqueda, y otras limitaciones, como por ejemplo el número de dimensiones.

3.1 Resultados

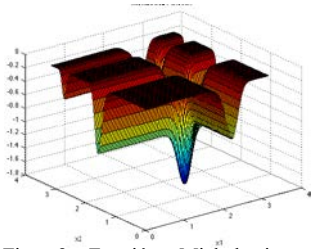


Fig. 2 Función Michalewicz mapa completo.

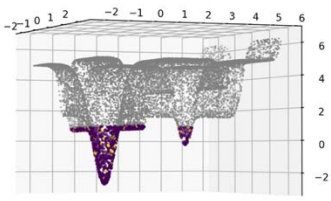


Fig. 3 Función Michalewicz: análisis de densidad y propuesta de corte

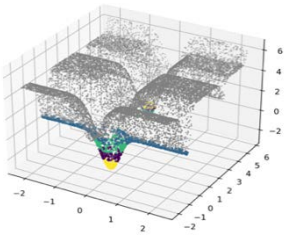


Fig. 4 Función Michalewicz: clustering con k-means. Localización de óptimo.

En las Figuras 2, 3, 4 se muestra la función de benchmark Michalewicz, la cual, fue resuelta utilizando el método heurístico habiendo obtenido resultados prometedores. En resumen, para *d=2*: *f(x*)* = -1.8013, el óptimo se localiza en x* = (2.20, 1.57). Nuestro resultado aplicando el método heurístico en *d=2*: *f(x**)* = -1.8012, y localizamos el óptimo en x** = (2.2024, 1.5709).

A continuación, en la Tabla 1 [5], se puede ver la comparativa en términos de calidad respecto a la aplicación del método heurístico para la resolución de un problema de regresión. Destacamos que, de todas las pruebas realizadas, los resultados fueron comparables a aquellos obtenidos utilizando una librería científica como scikit-learn [6].

Tabla 1 Resultados del método heurístico y scikit-learn para la regresión lineal.

	Execution Time (Sec)	Samples	Problem Size	Max. samples	Precision	Fitness	Range for $b0$	Range for $b1$	MSE	Equation	R^2
Fish Market											
Heuristic approach	31.8	8,606	3.80E+06	5.70E+05	3	1.00E-02	0 - 0.1	0 - 38	1.537	y=0.016*x+20.427	0.8775
scikit-learn approach	9.5	-	-	-	-	-	-	-	1.5297	y=0.0161*x+20.360	0.8781
Salt											
Heuristic approach	481.5	8,734	1.14E+06	7.61E+06	3	1.00E-03	-0.2 - 0.2	0 - 38	0.160	y=-0.050*x+ 34.376	0.2514
scikit-learn approach	10.2	-	-	-	-	-	-	-	0.1597	y=-0.0552*x+ 34.4407	0.2539
Stress Experiment I											
Heuristic approach	30.7	17,395	1.00E+10	1.50E+09	5	1.00E-04	0 - 0.1	-6 - 10	27.66690	y= 0.09765*x+ 0.02015	0.0119
Heuristic approach	16.4	6,273	1.00E+10	1.50E+09	3	1.00E-03	0 - 0.1	-6 - 10	27.626	y= 0.098*x+ 0.02	0.01167
scikit-learn approach	1.7	-	-	-	-	-	-	-	27.6668	y=0.09876*x+0.01524	0.0116
Stress Experiment II (Scattered data)											
Heuristic approach	362.9	97,832	2.00E+06	3.00E+05	3	1.00E-03	-0.02 - 0.2	-110 - 150	1525.412	y= -0.197*x+ 5.43	0.00150
scikit-learn approach	3.9	-	-	-	-	-	-	-	1561.5955	y=0.17665*x+5.18766	0.00066
Stress Experiment III (crossed Lines)											
Heuristic approach	21.9	13,279	2.00E+06	3.00E+05	3	1.00E-03	-0.02 - 0.2	-3 - 10	6.259	y= 0.003*x+ 4.222	-0.00005
scikit-learn approach	2.9	-	-	-	-	-	-	-	6.2833	y=0.000079*x+4.24524	0.0000

4 Conclusiones

A manera de conclusión, hemos comenzado la tesis basándonos en un método heurístico capaz de resolver un problema de optimización, y, con el fin de ser capaces de resolver otro tipo de problemas reales de optimización, como por ejemplo la regresión, hemos llevado a cabo un rediseño del método, mejorando las técnicas de agrupamiento, por un lado, y por otro, realizando una propuesta de línea de corte con la cual logramos reducir el área de búsqueda. Actualmente se está realizando un estudio de la aplicación del método heurístico en otro tipo de problemas y también se lleva a cabo un análisis de la complejidad de las etapas del método heurístico.

Como líneas de trabajo futuras, estamos en el camino de verificación de la escalabilidad del método dadas las propiedades de aleatoriedad del algoritmo, con lo cual creemos que el método es paralelizable.

Referencias

1. Cabrera, E., Taboada, M., Iglesias, M.L., Epelde, F., Luque, E.: “Optimization of healthcare emergency departments by agent-based simulation”. Procedia Computer Science 4, pp. 1880 – 1889, International Conference on Computational Science (ICCS-2011).
2. Liu, J., Qi, Y., Meng, Z.Y., Fu, L.: “Self-learning Monte Carlo method”. Journal of Physical Review B, number 4, vol. 95, ISSN 2469-9969 (Jan 2017).
3. Hussain, K., Salleh, M., Cheng, S., Naseem, R.: “Common benchmark functions for metaheuristic evaluation: A review”. International Journal on Informatics Visualization 1, pp. 218–223 (2017).
4. Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al., “A density-based algorithm for discovering clusters in large spatial databases with noise.,” in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), vol. 96, pp. 226–231 (1996).
5. M. Harita, A. Wong, D. Rexachs, E. Luque.: “Evaluation of a heuristic search algorithm based on sampling and clustering”. In Short Papers of the 9th Conference on Cloud Computing Conference, Big Data & Emerging Topics, pp 55-58 (2021).
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: “Sickit-learn: Machine Learning in Python”. Journal of Machine Learning Research, vol. 12, pp. 2825-2830 (2011).

Selección de características en entornos Big Data. Aplicación en Gene Signatures.

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina.

Tesista: Genaro Camele [1]
Director: Waldo Hasperué [1]

[1] III-LIDI, Facultad de Informática, Universidad Nacional de La Plata
{gcamele, whasperue}@lidi.info.unlp.edu.ar

Palabras claves: Feature Selection, Feature Extraction, Big Data, Data Mining, Bioinformatic.

1 Motivación

La medicina genómica es aquella que utiliza el conocimiento del genoma humano y de ciencias afines para identificar el riesgo de padecer una enfermedad, diagnosticarla precozmente y tratarla de forma personalizada. La medicina genómica ayuda a entender de forma más precisa por qué enfermamos, y el peso que tiene en una enfermedad la existencia de defectos genómicos frente a factores medioambientales que pueden desencadenar una enfermedad concreta.
En el ámbito de la genómica funcional, se destaca el análisis de perfiles de expresión génica; éstos tienen como objetivo principal la identificación de un grupo de genes, cuyo patrón de expresión se encuentren asociados a un fenotipo en particular, concepto conocido como gene signature [1]. Estos son un conjunto de genes que se sospecha, podrían ser marcadores de una patología en particular. Para evaluar la eficacia del mismo, se procesa un dataset que consta de pacientes que sufren la patología y la expresión de los genes que están especificados en el gene signature para cada una de estas personas, el tiempo transcurrido desde el último chequeo realizado y el estado vital del paciente.
Un objetivo particular de los signatures es su utilidad como biomarcador diagnóstico, pronóstico o predictivo de una patología en estudio. Los biomarcadores con valor pronóstico permiten una mejor estratificación de pacientes según su pronóstico de progresión de enfermedad independientemente de una terapia, abriendo el paso a investigaciones de tratamientos adecuados para cada categoría de paciente definida. Por otro lado, biomarcadores con valor predictivo permiten predecir si un tratamiento tendrá o no efecto en un paciente, logrando evitar tratamientos en pacientes para los cuales se supone no tendrán efecto positivo.

2 Objetivos y Aportes

Para llevar a cabo el descubrimiento de nuevos gene signatures es necesario un proceso de automatización que permita encontrar genes candidatos en base al conocimiento del experto. En la actualidad esta tarea es realizada de forma manual. Con la rápida acumulación de datos de expresión génica de diversas tecnologías, es posible aplicar algoritmos automáticos de reducción de dimensiones, con el objetivo de seleccionar aquellas que resulten más representativas del conjunto de características.

Los algoritmos de extracción de características tienen como tarea identificar un subconjunto de atributos del dataset, tal que con dicho subconjunto sea posible obtener modelos predictores con un poder pronóstico igual o similar a los que se consiguen con el conjunto de atributos completo. Los resultados obtenidos de esta identificación de atributos podrían ser interpretados como un posible gene signature.

El objetivo general de este plan de doctorado es el de contribuir con el desarrollo de algoritmos de extracción de características en entornos Big Data que permitan la identificación y la evaluación de posibles gene signatures.

Al inicio del doctorado se comenzó a trabajar en la localización, descarga y limpieza de las bases de datos más utilizadas y referenciadas en el tema de identificación de gene signatures. La preparación de estas bases de datos sirvió para la implementación de una plataforma web llamada Multiomix [2] que ofrece a usuarios inexpertos diferentes herramientas para el análisis de correlaciones entre genes y reguladores de expresión que pueden servir a la identificación de biomarcadores.

Esta plataforma se encuentra disponible de manera gratuita a través de internet, no se requiere experiencia previa ni la instalación de ningún tipo de dependencia o herramienta más que un navegador web para hacer uso de todas las funciones puestas a disposición.

Multiomix es un proyecto de código libre que ofrece a los usuarios la opción de consultar información extra de algunos genes y microARN (un regulador de expresión que forma parte de los datasets compatibles con Multiomix), fomentando la colaboración por parte de otros desarrolladores y permitiendo el despliegue privado de la plataforma para cubrir con los requerimientos de privacidad que los usuarios o instituciones establezcan.

Previo a Multiomix se desarrolló una herramienta de código libre llamada Modulector [3]. Esta herramienta ofrece una API de gran performance con funciones de filtración, paginación, ordenamiento y búsqueda a través de diferentes servicios que retornan información estandarizada obtenida de diferentes bases de datos biológicas. El uso de esta herramienta está incorporado en Multiomix para abstraer de su uso a los usuarios que no poseen conocimientos técnicos. También se puede realizar un despliegue en servidores privados para aquellos usuarios que así lo precisen.

Al mismo tiempo que se desarrollaban las plataformas Multiomix y Modulector, se llevó a cabo una comparación entre diferentes algoritmos de clasificación como Naïve Bayes (NB), SVM, Random Forest (RF) y redes neuronales (MLP) en un cluster de Apache Spark, framework para la ejecución de procesamiento distribuido. El objetivo de dicha experimentación consistió en medir el tiempo de ejecución de tales algoritmos en un cluster Spark conformado por tres nodos en total.

Dado que los algoritmos de extracción de características necesitan realizar varias pruebas seleccionando diferentes subconjuntos de atributos en cada una de ellas, resulta

de interés conocer el tiempo de ejecución de los algoritmos estudiados, según la cantidad de atributos que posea el dataset utilizado como entrenamiento. Para resolver este interrogante se llevaron a cabo diferentes ensayos con los cuatro algoritmos mencionados, donde en cada uno de ellos se variaba la cantidad de atributos del dataset utilizado como entrenamiento de los algoritmos estudiados.
Los resultados arrojaron que SVM y MLP son los algoritmos que se ven perjudicados a medida que el dataset posee más número de atributos. RF resultó el algoritmo con mejor balance entre el poco tiempo de ejecución que necesita y la evaluación de los modelos obtenidos (accuracy y f1-score). NB es el algoritmo que menos se ve perjudicado con el aumento de atributos, pero el poder pronóstico de los modelos obtenidos es muy bajo comparado con los obtenidos por los otros tres algoritmos [4].

3 Estado Actual y Trabajo Futuro

Actualmente se está estudiando la metaheurística denominada Binary Black Hole Algorithm que ha brindado excelentes resultados frente a metaheurísticas más asentadas como Binary Particle Swarm Optimization al realizar selección de características con datos biológicos o de supervivencia [5][6][7]. Esta metaheurística es estudiada en conjunto con varios algoritmos de regresión y clasificación sobre datos genómicos para el análisis de supervivencia de pacientes que padecen cáncer. En un algoritmo de optimización poblacional el costo más grande lo tiene la evaluación de fitness de cada individuo de la población. La tarea de evaluar un individuo, cuando se está llevando a cabo la tarea de selección de características, consiste en entrenar uno o más modelos de clasificación con el objetivo de medir el poder pronóstico de los modelos obtenidos con el subconjunto de atributos evaluados.
Teniendo en cuenta que el conjunto de datos utilizados para la obtención de modelos de clasificación es muy grande y que los algoritmos de optimización necesitan medir el fitness a sus individuos centenares o miles de veces, encontrar un mecanismo de ponderación que tenga en cuenta el número de atributos evaluados (como así también los modelos a utilizar), se torna una tarea compleja que amerita su estudio. La innovación que conlleva la presente tesis doctoral residirá en diferentes optimizaciones de la función de fitness utilizada por la metaheurística, como así también la propia metaheurística.
Cabe destacar que como objetivo se planea poner a disposición todos los algoritmos desarrollados como parte de la plataforma Multiomix, para que estén accesibles a todos los interesados. Esto abre las puertas al análisis exhaustivo con datos reales, impulsando una investigación abierta y ágil.

Referencias

1. Abba MC, Lacunza E, Butti M, Aldaz CM. **Breast cancer biomarker discovery in the functional genomic age: a systematic review of 42 gene expression signatures**. Biomarker Insights; 5:1-16. 2010.
2. Camele, G; Menazzi, S; Chanfreau, H; Marraco, A; Hasperué, W; Butti, MD; Abba, MC. **Multiomix: a cloud-based platform to infer cancer genomic and epigenomic events**

associated with gene expression modulation. Bioinformatics. ISSN 1367-4803. 2021a. https://doi.org/10.1093/bioinformatics/btab678

3. Marraco, Al; Camele, G.; Hasperué, W.; Menazzi, S.; Abba, M.; Butti, M. **Modulector: una plataforma como servicio para el acceso a bases de datos de micro ARNs**. Revista Innovación y Desarrollo Tecnológico y Social, Vol. 3 (1): 89-114. 2021. https://doi.org/10.24215/26838559e030
4. Camele, G; Hasperué, W; Ronchetti, F; Quiroga FM. **A Comparative Study of the Performance of the Classification Algorithms of the Apache Spark ML Library**. Congreso argentino de ciencias de la computación. Salta, Salta. 2021b.
5. Pashaei, Elnaz & Aydin, Nizamettin. **Binary black hole algorithm for feature selection and classification on biological data**. Applied Soft Computing. 56. 2017. https://doi.org/10.1016/j.asoc.2017.03.002.
6. Pashaei, E., Pashaei, E., & Aydin, N. **Gene selection using hybrid binary black hole algorithm and modified binary particle swarm optimization**. *Genomics*, *111*(4), 669-686. 2019. https://doi.org/10.1016/j.ygeno.2018.04.004
7. Pashaei, E. and Pashaei, E. **Gene Selection for Cancer Classification using a New Hybrid of Binary Black Hole Algorithm**. 28th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. 2020, https://doi.org/10.1109/SIU49456.2020.9302351.

Diseño de Sistemas Borrosos para el análisis de comportamientos específicos en Medios Sociales

Doctorado en Tecnologías Informáticas Avanzadas, Universidad de Castilla-La Mancha, España.

Tesista: Andres Montoro[1]
Director: Jose A. Olivas[1]
Co-Director: Adan Nieto[2]

[1] Departamento de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha, 13071 Ciudad Real, España
andres.montoro@alu.uclm.es, joseangel.olivas@uclm.es
[2] Departamento de Derecho Público y de la Empresa. Universidad de Castilla-La Mancha, 13071 Ciudad Real, España
adan.nieto@uclm.es

Palabras claves: Logica Borrosa, Computación Suave, Razonamiento Aproximado, Ingeniería del Conocimiento, Discurso de Odio, Radicalización.

1 Motivación

En la era del Big Data millones de personas están generando datos en toda clase de medios sociales. Este tipo de datos son desestructurados, contienen gran cantidad de ruido, es común que presenten una ausencia de formato unificado y son de longitud variable. Dentro del universo de los medios sociales, las relaciones entre entidades, también conocidas como redes sociales, se convierten en un vehículo extraordinario para la difusión masiva de mensajes. Un análisis eficaz de los medios sociales [8] requiere recopilar información sobre individuos o usuarios y entidades (redes sociales, sitios, etc.), analizar las interacciones entre ellos y descubrir patrones para comprender el comportamiento humano.

Las palabras desempeñan un papel fundamental en el análisis de los medios sociales y, en general, en el procesamiento del lenguaje natural. Cuando trabajamos con palabras nos enfrentamos a la imprecisión y la incertidumbre consustanciales a la forma de razonamiento humana. El concepto de computación con palabras fue desarrollado por Lofti A. Zadeh [6] como un campo en el que el elemento central a procesar son las palabras, frases o cualquier proposición extraída del lenguaje natural. El campo de la computación con palabras está estrechamente relacionado con otros grandes hitos de la inteligencia artificial teorizados por el profesor Zadeh como son la Lógica Borrosa [7] y la Computación Suave [5]. Lofti A. Zadeh definió la Computación Suave como un conjunto de metodologías con las que se pretende explotar la tolerancia a la imprecisión y la incertidumbre para lograr la trazabilidad, la solidez y el bajo coste de las soluciones. La Computación Suave es un conjunto de metodologías complementarias y no competitivas formadas por la Lógica Borrosa, la

teoría de Redes Neuronales y el Razonamiento Probabilístico, con este último subsumiendo las Redes de Creencias, los Algoritmos Genéticos y las Metaheurísticas, la Teoría del Caos y partes de la teoría del aprendizaje.

La futura tesis se centra en este campo, el empleo de la Computación Suave con la Lógica Borrosa como núcleo para analizar comportamientos en medios sociales.

2 Objetivos y Aportes

Dentro del universo de los medios sociales, existen múltiples comportamientos asociados a los usuarios que conforman este tipo de plataformas, desde conductas dirigidas a la promoción de productos, pasando por relaciones sociales entre usuarios, hasta el uso de estas plataformas para hacer política. El objetivo de la futura tesis es focalizar el análisis a comportamientos relacionados con disciplinas humanísticas como el derecho, criminología y sociología. Desde este punto de vista multidisciplinar se han detectado distintos comportamientos que atentan contra la libertad individual como es el discurso de odio y los comportamientos radicales que pueden derivar en acciones violentas e incluso, en casos extremos, terrorismo.

Internet y concretamente los medios sociales han modificado la forma de comunicación en sociedad, y se ha convertido, entre otras cosas por sus características de neutralidad y falta de censura en un nuevo ámbito de oportunidad delictiva que no preexistía [4]. Este inusual caldo de cultivo ha propiciado la aparición y sobre todo la universalización del discurso de odio gracias al anonimato que esta clase de medios puede proporcionar y el alcance del que se dispone a la hora de su difusión masiva, lo que conlleva un mayor efecto. El discurso y los delitos de odio envenenan a las sociedades al amenazar los derechos individuales, la dignidad humana y la igualdad, reforzando las tensiones entre los distintos grupos sociales, perturbando la paz y el orden público, lo que pone en peligro la convivencia pacífica. Pero en Internet, no se da únicamente este comportamiento indeseable y en ocasiones delictivo. Cuando el odio se lleva al extremo puede ser un indicador de radicalización en un individuo o conjunto de individuos.

La detección y evaluación de este tipo de perfiles es extremadamente compleja, e incluso los expertos juristas no se ponen de acuerdo en cómo medir la intensidad o gravedad de este tipo de comportamientos potencialmente delictivos. Para tratar esta cuestión en que la ambigüedad interpretativa, la imprecisión y la incertidumbre están presentes, se hace necesario el uso de técnicas de Inteligencia Artificial que sean tolerables a estas condiciones propias del dominio, como es la Computación Suave.

Abordar este tipo de disciplinas mediante el uso de Inteligencia Artificial, nos lleva a emplear técnicas que traten de imitar la forma de razonar que tenemos los humanos, por lo que únicamente el uso de técnicas como el Aprendizaje Automático no es suficiente, ya que este trata de buscar patrones en los datos, buscando correlaciones en

los mismos que en muchas ocasiones no implican causalidad. Esta es una de las grandes diferencias con el conocimiento humano que es en su mayoría, heurístico. La disciplina conocida como Ingeniería del Conocimiento [2] está centrada en el análisis y la propuesta de métodos para la adquisición de conocimiento, representación de conocimiento y su empleo. La Ingeniería del Conocimiento se ocupa del desarrollo de los Sistemas Basados en el Conocimiento o Sistemas Expertos que tratan de emular la forma de razonamiento de los humanos mediante la manipulación simbólica e inferencia para resolver problemas complejos.

En definitiva, el grueso de la futura tesis se centra en los conceptos previos expuestos, a saber, detectar y evaluar el discurso de odio criminalizado y la detección y evaluación de potenciales radicales. Para ello se hará uso de la Computación Suave para el análisis inteligente de datos guiado por Ingeniería del Conocimiento ya que para analizar y evaluar este tipo de comportamientos potencialmente delictivos es esencial el conocimiento experto, entendiendo este como conocimiento humano y bibliográfico, con el que comprender las particularidades del discurso de odio y del comportamiento radical en medios sociales.

El concepto de Inteligencia Artificial Explicable se torna imprescindible en este tipo de sistemas que hacen uso de Inteligencia Artificial, uno de los principales motivos se fundamenta en las posibles consecuencias derivadas de la evaluación de los comportamientos potencialmente delictivos en los medios sociales, porque puede conllevar penas de privación de libertad en los casos más extremos o atentar contra la libertad de expresión en los casos más leves. Y el último motivo, es normativo, ya que por parte de la Comisión Europea se han propuesto las nuevas normas y medidas para la implantación de la Inteligencia Artificial[1] y este tipo de sistemas (según la nueva norma) entrañaría un alto riesgo lo que implica que para su implantación es imprescindible que sea explicable.

3 Estado Actual y Trabajo Futuro

La tesis se encuentra en un estado avanzado con los prototipos para la detección y evaluación de discurso de odio y potenciales radicales implementados haciendo uso de Razonamiento Aproximado.

El trabajo futuro se centra en la mejora de los prototipos y perfilar los aportes científicos, a saber:

- Optimización de reglas borrosas haciendo uso de *data augmentation* a partir de la representación en dos tuplas de etiquetas lingüísticas propuesta por

[1] https://digital-strategy.ec.europa.eu/en/library/communication-fostering-european-approach-artificial-intelligence

Herrera y Martínez en [3] basada en la combinación convexa de etiquetas lingüísticas [1].

- Generación automática de reglas borrosas a partir de *data augmentation* con la menor pérdida de conocimiento posible.

- Aumentar el número de expertos disponibles para una mejor evaluación de los resultados.

- Desplegar los sistemas en un entorno real para validar el comportamiento, teniendo en cuenta los requisitos de un modelo de Inteligencia Artificial calificado con riesgo Alto dentro del nuevo reglamento para la implantación de la Inteligencia Artificial.

Referencias

1. Delgado, M., Verdegay, J. L., Vila, M. A.: On Aggregation Operations of Linguistic Labels. International Journal of Intelligent Systems 8(3), 351-370 (1993).
2. Feigenbaum, E. A.: Knowledge Engineering: The Applied Side of Artificial Intelligence. Stanford Heuristics Programming Project, pp. 1-14, Stanford (1980).
3. Herrera, F., Martinez, L.: A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on Fuzzy Systems 8(6), 746-752 (2000).
4. Miró-Llinares, F.: El cibercrimen: Fenomenología y criminología de la delincuencia en el ciberespacio. 1st edn. Marcial Pons, (2012).
5. Zadeh, L. A.: Fuzzy logic and soft computing: issues, contentions and perspectives. In 3rd Int. Conf. on Fuzzy Logic, Neural Nets and Soft Computing, pp. 1-2, Iizuka (1994).
6. Zadeh, L. A.: Fuzzy logic = computing with words. IEEE Transactions on Fuzzy Systems 4(2), 103-111 (1996).
7. Zadeh, L. A.: Fuzzy Sets. Information and Control 8(3), 338-353 (1965).
8. Zafarani, R., Abbasi, M. A., Liu, H.: Social Media Mining: An Introduction. 1st edn. Cambridge University Press, (2014).

Coplanificación de procesos maleables de aprendizaje automático mediante contenedores

Maestría en Cómputo de Altas Prestaciones, Universidad Nacional de La Plata, Argentina.

Tesista: Ing. Leandro Ariel Libutti [1]
Director: Dr. Francisco Igual [2]
Co-Director: Dra. Laura De Giusti [1]

[1] Instituto de Investigación en Informática LIDI, La Plata 1900, Argentina
[2] Departamento de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid, Madrid, España.
llibutti@lidi.info.unlp.edu.ar

Palabras claves: .coplanificación, tensorflow, aprendizaje automático, contenedores.

1 Motivación

El crecimiento exponencial en el interés por el aprendizaje automático en la última década está directamente relacionado con tres avances fundamentales:

- El desarrollo de mejores algoritmos con aplicaciones directas en muchos campos de la ciencia y la ingeniería;
- La disponibilidad de cantidades masivas de datos y la viabilidad de almacenarlos y analizarlos de manera eficiente;
- La aparición de arquitecturas de hardware novedosas, normalmente paralelas y / u homogéneas, que permiten una adecuada explotación de estos algoritmos sobre grandes conjuntos de datos en un tiempo asequible.

La aparición de nuevas arquitecturas y técnicas en HPC ha renovado el interés por el Machine Learning en una gran variedad de problemas, incluyendo aplicaciones de reconocimiento de imágenes, segmentación, reconocimiento de voz, procesamiento de lenguaje natural o traducción de idiomas, entre muchos otros. Tensorflow [1] , Caffe [2] , Keras [3] y PyTorch [4] son frameworks de ML que permiten ocultar detalles de implementación al usuario manteniendo un alto rendimiento tanto en el entrenamiento como en la inferencia de modelos.

Hoy en día, Tensorflow es uno de los frameworks más utilizados. Su diseño se basa es un grafo de ejecución en el cual las operaciones están representadas por nodos y el flujo de datos por tensores (matrices multidimensionales de datos). Para cada operación puede definirse el paralelismo (intra paralelismo). A su vez, se puede elegir la cantidad de operaciones simultáneas que pueden ejecutarse (inter paralelismo). El gran problema

que presentan ambos paralelismos es que son definidos antes de la ejecución del algoritmo sin posibilidad de modificarlos en tiempo de ejecución.

Actualmente, la ejecución de Tensorflow se puede realizar a través de contenedores. El uso de contenedores se ha popularizado durante los últimos años como una solución de virtualización ligera. Brevemente, este enfoque trae varios beneficios. Primero, el uso de contenedores permite incrustar una pila de software completa para ejecutar una aplicación en varios contextos para reforzar la portabilidad. Al igual que la virtualización clásica, la ejecución de aplicaciones dentro de contenedores permite el aislamiento del sistema host y de otros contenedores. Por lo tanto, los derechos de administrador se pueden asignar a los usuarios dentro de un contenedor sin afectar al host. Las principales ventajas que presenta el uso de contenedores en comparación con la virtualización clásica son las siguientes:

- Los contenedores son más livianos (ya que trabajan directamente sobre el Kernel) que las máquinas virtuales.
- No es necesario instalar un sistema operativo por contenedor.
- Menor uso de los recursos de la máquina.
- Mayor cantidad de contenedores por equipo físico.
- Mejor portabilidad.

Normalmente, los desarrolladores de software suelen utilizar contenedores para que todos los miembros de un proyecto puedan leer el código dentro del mismo entorno de software. Los contenedores también se utilizan con fines de demostración cuando las aplicaciones necesitan el despliegue de un ecosistema complejo y, más recientemente, con fines de producción, ya que permite el despliegue de varios servicios rápidamente y sin errores [6].

La planificación online de estos contenedores en plataformas HPC presenta grandes dificultades que deben ser atendidas por un sistema encargada de dicha tarea. La necesidad de abordar numerosos factores de programación conduce al desarrollo de algoritmos sofisticados que a menudo son difíciles de razonar y de implementar en sistema reales. La programación y planificación de trabajos en contenedores sigue siendo un tema abierto. La comunidad de código abierto propuso frameworks de planificación nativos de contenedores como Kubernetes, Mesos Marathon y Docker Swarm; los círculos académicos proponen algoritmos de programación creativos para buscar soluciones óptimas de equilibrio de carga o de eficiencia energética. Los esfuerzos conjuntos de ambas comunidades hacen que la tecnología de contenedores avance más hacia los usos prácticos y comerciales.

La mayoría de las aplicaciones/contenedores que se ejecutan en súper computadoras alcanzan sólo una fracción del rendimiento máximo teórico del sistema, aunque si se aplica un alto grado de mejoras pueden acercarse a ese límite. Aun cuando se realice una optimización al máximo, la evolución de las nuevas arquitecturas genera la búsqueda de constantes mejoras. Además, el aumento de los núcleos y de la heterogeneidad de las arquitecturas harán que sean aún más difícil explotar los recursos disponibles[7].

El rendimiento de los contenedores suele estar limitado por algunos recursos como, por ejemplo, el paralelismo utilizado y el ancho de banda de la memoria. Una posible solución sería ejecutar múltiples contenedores con diferentes demandas de recursos en

una misma máquina/nodo. Este enfoque generaría una caída en el rendimiento de cada contenedor, pero permitiría un aumento en el rendimiento general de todo el sistema aumentando la productividad del Sistema [5].

Por lo tanto, es necesario contar con un algoritmo de planificación eficaz para lanzar contenedores Docker a fin de mitigar dicha interferencia en el rendimiento y mejorar la utilización de recursos en entornos de ejecución de algoritmos de Machine Learning.

2 Objetivos y Aportes

El objetivo principal que se plantea es lograr una infraestructura que permita, de forma holística, realizar una coplanificación de múltiples instancias de procesos de entrenamiento/inferencia usando Tensorflow, así como una gestión dinámica tanto de los recursos expuestos a cada instancia como del grado de paralelismo que cada instancia puede explotar.

Este objetivo principal se divide en dos subobjetivos, cada uno con sus tareas asociadas:

- Modificación del esquema de paralelismo dentro de la infraestructura Tensorflow para permitir selección dinámica del paralelismo de las operaciones (maleabilidad intra):
- Diseño e implementación de un planificador de procesos que permita administrar el paralelismo de diferentes instancias Tensorflow maleables ejecutando en una misma máquina a través de contenedores.

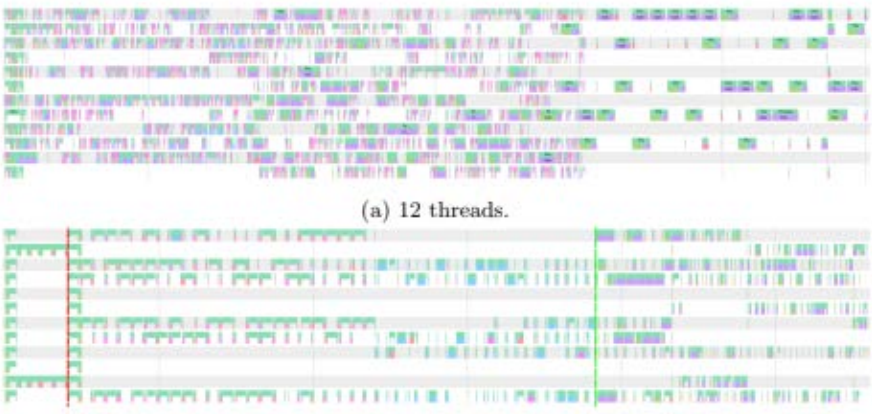
3 Estado Actual y Trabajo Futuro

3.1 Maleabilidad en Tensorflow

Para realilzar estos cambios, se modificó la librería Eigen [6] utilizada en Tensorflow para el manejo de los hilos y cálculos matemáticos. La figura 1 informa algunos resultados experimentales obtenidos en un Implementación maleable de TensorFlow. Estos resultados fueron extraídos en un sistema basado en un procesador Intel Core i7-8750H con 6 núcleos (12 núcleos lógicos a través de la tecnología HyperThreading), que se ejecutan a 2,2 GHz de frecuencia nominal. El sistema cuenta con 32 GB de memoria RAM DDR4.

Las trazas informan las líneas de tiempo de ejecución (una línea horizontal por hilo de trabajo) para un modelo de aprendiza automático (Resnet). Reportamos tres escenarios diferentes para el proceso de entrenamiento del modelo:

- La figura 1-A es una implementación típica de TensorFlow con 12 hilos de trabajo desde el principio hasta el final de la ejecución.
- La figura 1-B corresponde a una implementación de TensorFlow modificada para aportar maleabilidad. Se realizan dos cambios en la cantidad de hilos intervinientes: el primero disminuye la cantidad de hilos de 12 a 6 (línea roja vertical). Posteriormente, los hilos de los trabajadores se restauran nuevamente a 12 (punto marcado con una línea verde vertical).



(a) 12 threads.

(b) 12 + 6 threads.

Fig. 1. Ejecución de Tensorflow malleable.

3.2 Planificador de procesos Tensorflow

La figura 2 muestra cómo se ejecutan los procesos Tensorflow utilizando el planificador desarrollado. Cada aplicación/proceso de TF se despliega dentro de un contenedor Docker [9]. El proceso de planificación es el siguiente:

- Antes de su ejecución, el contenedor solicita al planificador la cantidad de recursos requeridos.
- El planificador recibe la petición y chequea los recursos disponibles en el sistema.
- Dependiendo la política de planificación, se definen la cantidad de recursos que utilizará el contenedor.
- Se asigna la cantidad de recursos al contenedor y se actualiza los recursos disponibles en el sistema.

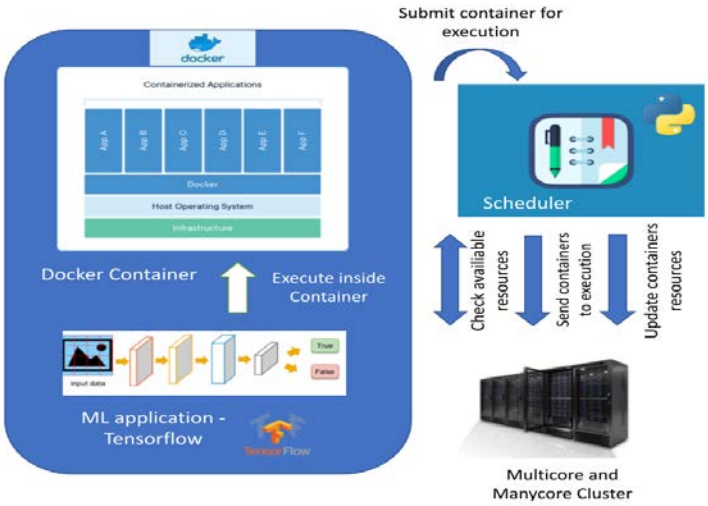


Fig. 2. Esquema de ejecución de procesos.

Si se encuentran contenedores activos que no tienen asignado la cantidad total de recursos solicitados, cuando finalizan otros contenedores se puede realizar el proceso de actualización en el caso que el usuario del planificador lo requiera. En este proceso se debe tener en cuenta la cantidad de recursos liberados para asignar y la política de distribución de nuevos recursos entre contenedores activos (mas antiguo tiene mayor prioridad, más nuevo tiene mayor prioridad, distribución equitativa, entre otras).

Actualmente estamos trabajando en la generación de reportes que permitan visualizar correctamente las trazas de ejecución del planificador.

Referencias

1. Mart´ın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, CraigCitro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghe-mawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, YangqingJia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dande-lion Man´e, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster,Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vin-cent Vanhoucke, Vijay Vasudevan, Fernanda Vi´egas, Oriol Vinyals, Pete Warden,Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow:Large-scale machine learning on heterogeneous systems, 2015. Software availablefrom tensorflow.org.
2. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long,Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional archi-tecture for fast feature embedding. InProceedings of the 22Nd ACM InternationalConference on Multimedia, MM ’14, pages 675–678, New York, NY, USA, 2014.ACM.
3. Fran ̧cois Chollet et al. Keras.https://keras.io, 2015
4. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, GregoryChanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Des-maison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, AlykhanTejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and SoumithChintala. Pytorch: An imperative style, high-performance deep learning library. InH. Wallach, H. Larochelle, A. Beygelzimer, F. Alch´e-Buc, E. Fox, and R. Garnett,editors,Advances in Neural Information Processing Systems 32, pages 8024–8035.Curran Associates, Inc., 2019.
5. Libutti, Leandro Ariel, et al. "Towards a Malleable Tensorflow Implementation." Conference on Cloud Computing, Big Data & Emerging Topics. Springer, Cham, 2020.
6. Pengfei Xu, Shaohuai Shi, and Xiaowen Chu. Performance evaluation of deep learning tools in docker containers. In 3rd International Conference on Big Data Computing and Communications, BIGCOM 2017, Chengdu, China, August 10-11, 2017, pages 395–403. IEEE Computer Society, 2017.
7. Docker y otros container: mas allá de la virtualización. Online: https://www.ionos.es/digitalguide/servidores/know-how/docker-container-las-ventajas-de-los-contenedores-web/ . 29/11/21
8. Gael Guennebaud, Beno ̂ıt Jacob, et al. Eigen v3. http://eigen.tuxfamily.org, 2010.
9. Docker. Online: https://www.docker.com/ . 29/11/21

Métodos escalables y rápidos guiados por datos para la sintonización de un simulador

Doctorado en Ciencias Informáticas, Universidad Nacional de La Plata, Argentina.

Tesista: Mariano Trigila [1]
Director: Emilio Luque [2]
Asesora científica: Adriana Gaudiani [3]

[1] Facultad de Ingeniería y Ciencias Agrarias, Pontificia Universidad Católica Argentina, Ciudad Autónoma de Buenos Aires, Argentina.
[2] Depto. de Arquitectura de Computadores y Sistemas Operativos, Universidad Autònoma de Barcelona, 08193 Bellaterra (Barcelona) España.
[3] Instituto de Ciencias, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina.

Palabras claves: Simulation, Optimization, Data assimilation, dynamics data-driven.

1 Motivación

Para resolver problemas complejos de la ingeniería, de la ciencia, de los negocios y del mundo real, a menudo es útil comprender la relación entre las variables de decisión y las variables de respuesta de un modelo para entender mejor el comportamiento del sistema real de interés [3]. Los modelos conceptuales del mundo real que representan la evolución de un sistema son desarrollados por expertos quienes crean e implementan un conjunto de algoritmos para simular el sistema sobre un dominio de interés. Este modelo computacional o simulador, está alimentado por un conjunto de parámetros y variables de entrada que definirán los escenarios de simulación cuando se corren (o lanzan) las ejecuciones en el simulador. En sistemas complejos y dinámicos, los simuladores tienden con frecuencia a perder la sintonización con respecto al comportamiento del sistema real, es aquí donde se requiere de un proceso metodológico de calibración que provea al simulador de un nuevo conjunto de valores iniciales de sus parámetros de entrada para lanzar la ejecución de la simulación y lograr como consecuencia sintonizar los datos de salida del simulador con los datos propios del sistema real que se está simulando. En la tesis doctoral que se está llevando adelante (ver propuesta en [7]) se investigan procedimientos metodológicos computacionales para encontrar los parámetros de ajuste de modelos que se encuentran implementados en algoritmos que conforman un simulador. Estos procedimientos deben cumplir con las propiedades de reducción de poder cómputo, reducción de tiempo de procesamiento y reducción de gasto energético.

2 Objetivos y Aportes

La tesis tiene como punto de partida los trabajos previos realizados por Gaudiani A., en su tesis doctoral [2]. En ella se expresa una metodología de sintonización de un

simulador de cauce del río Paraná de Argentina. Los trabajos (procesos metodológicos) que se presentan en este artículo que pertenecen a la tesis, son dos. El primer proceso metodológico desarrollado en la tesis fue denominado "Método de ajuste en pasos sucesivos" [1]. Es un método en donde al procedimiento se le agrega conocimiento sobre las características del dominio, información que lleva a reducir en gran magnitud el espacio de búsqueda en el proceso heurístico de ajuste. El segundo proceso metodológico actualmente se encuentra en investigación y desarrollo y se denomina "Método de ajuste basado en conocimiento" (*nombre aún no oficial*). Este método se basa en buscar en una base de datos, que contiene información y conocimiento del pasado del dominio, y en la cual se accede para buscar aquellos valores de parámetros de entrada del simulador que mejor ajusta la salida.

2.1 **Método de ajuste en pasos sucesivos**

Proponemos un proceso de calibración en sucesivos pasos de ajuste. En cada paso se va calibrando una posición física del dominio de interés, que en nuestro caso es el lugar donde se encuentra cada estación de monitoreo. En un procedimiento iterativo logramos obtener una sintonización aceptable del simulador.

2.1.1 **Premisas**

(*a*): El dominio del sistema se comporta como un sistema físico y continuo. (*b*): Los sitios físicos sucesivos o contiguos, sus valores de propiedades son similares.

2.1.2 **Pasos del proceso**

(1°): Identificar el sitio de interés a partir del cual se desea ajustar. *(2°)*: Proveer al simulador los valores iniciales de los parámetros para comenzar la ejecución de la simulación. Para la iteración inicial del sitio del dominio que se desea ajustar se tomarán los valores iniciales de los parámetros de aquel sitio sucesivo que ya se encuentra ajustado (por premisa a y b). *(3°)*: Arrancar de forma iterativa la simulación para los distintos valores de parámetros de entrada (distintos escenarios). Iterar hasta encontrar aquellos parámetros que producen el mejor ajuste del simulador. Comparando la serie de datos simulados con la serie con la serie de datos observados (o real) se determina un índice el cuál se utiliza para determinar que una salida del simulador es la que mejor representa la realidad.

2.1.3 **Resultados experimentales**

Se experimentó con un simulador del río Paraná (Arg.), que modela los niveles del río a lo largo del cauce. Los sitios físicos donde se obtienen diariamente las mediciones observadas reales (altura del nivel del río) se denominan estaciones. El simulador produce para una corrida de simulación una serie temporal para cada estación.
En Tabla 1 se puede observar dos estaciones que luego de aplicar el procedimiento de ajuste se obtuvo una mejora muy considerable con respecto a experiencias realizadas con los valores de los parámetros originales propuesto por los expertos del dominio.

Tabla 1 Ajuste realizado en dos estaciones contiguas *k* y estación *k + 1*, varios años.

Año	***Estación***	***ID Estación***	***Mejora***
2008	*k*	*ESQU*	*57 %*
1999	*k*	*ESQU*	39 %
1999	*k+1*	LAPA	45 %
2008	*k+1*	LAPA	24 %

2.2 **Método de ajuste basado en conocimiento**

Proponemos un segundo proceso de calibración el cual para resolver un estado de desajuste de un simulador se deberá contar con información almacenada de eventos pasados, que permitirá recuperar valores de parámetros que en el pasado han ajustado satisfactoriamente al simulador. Dichos valores podrán ser reutilizados en situaciones actuales similares para ajustar al simulador.

2.2.1 **Premisas, trabajos previos**

(*a*) Almacenar los valores de las series observadas y las series simuladas. Sincronizar ambas series. (*b*) Almacenar los parámetros de mejor ajuste del simulador y relacionarlos con la serie simulada. (*c*) Caracterizar y determinar los eventos disruptivos o de alta velocidad. Desarrollar un método de comparación. (*d*) Determinar el criterio de cuando el simulador está ajustado y cuando desajustado.

2.2.2 **Pasos del proceso**

(1°): Identificar un desajuste del simulador, y determinar que el desajuste producido es muy amplio o se produjo a muy alta velocidad (disrupción). *(2°)*: Con los datos que identifican y caracterizan al evento acceder a la base de datos y extraer un listado de los eventos similares con sus respectivos valores de los parámetros. *(3°)*: Determinar cuál conjunto de parámetros de los encontrados en el punto anterior es el de mejor ajuste al simulador. Si un conjunto de parámetros cumple con el criterio de ajuste (d), entonces pasar al quinto paso. Si ningún conjunto de parámetro cumple con el criterio de ajuste (d) entonces ir al cuarto paso. *(4°)*: Determinar cuál conjunto de parámetros es el mejor (de los obtenidos de la base de datos) para tomarlo como parámetro de inicialización del simulador y así comenzar un proceso de ajuste con métodos de pasos sucesivos. *(5°)*: Se ha encontrado el conjunto de parámetro de mejor ajuste. Almacenar el nuevo conjunto de parámetros y relacionarlo con la serie que llevó a determinar que el simulador se encontraba bajo un desajuste disruptivo.

3 Estado Actual y Trabajo Futuro

Con el “Método de pasos sucesivos” se realizaron experiencias y publicaciones [1]. El “Método de ajuste basado en el conocimiento” está aún en etapa de diseño y desarrollo

para experimentación. Conceptualmente promete ser muy efectivo en momentos de grandes desajustes. Requerirá un proceso previo de caracterización de eventos y sincronización de series reales y simuladas, requerirá mantener actualizada la base de datos con conocimiento de situaciones representando así el comportamiento del dominio. Uno de los grandes desafíos se centra en la caracterización de los eventos para poder realizar comparaciones de semejanzas y así lograr obtener para situaciones actuales parámetros de ajustes próximos al óptimo con datos registrados en el pasado almacenados en la base de conocimiento. Para finalizar, dejo una reseña con referencias de temas claves que abordan soluciones y técnicas similares a las planteadas en la tesis: Data Assimilation Methods (DA)[6][7]; Dynamic Data Driven Applications Systems (DDDAS) [4]; Cyber-Physical System (CPS) [5].

Referencias

1. Trigila, M., Luque, E., Gaudiani, A.: Agile tuning method in successive steps for a river flow simulator. Lecture Notes in Computer Science, vol 10862:639–646, (2018). https://doi.org/10.1007/978-3-319-93713-7_60
2. Gaudiani, A.: Simulación y optimización como metodología para mejorar la calidad de la predicción en un entorno de simulación hidrográfica. Diss. Universidad Nacional de La Plata (2015).
3. Wang, W.: A Dual Metamodeling Perspective for Design and Analysis of Stochastic Simulation Experiments. Diss. Virginia Tech, (2019).
4. Blasch, E., Ravela, S., Aved, A.: Handbook of Dynamic Data Driven Applications Systems. Springer, Cham, (2018).
5. Rai, R., Sahu C.K.: "Driven by Data or Derived Through Physics? A Review of Hybrid Physics Guided Machine Learning Techniques With Cyber-Physical System (CPS) Focus," in IEEE Access, vol. 8, pp. 71050-71073, (2020).
6. Gottwald G.A., Reich, S.: Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation, Physica D: Nonlinear Phenomena, Volume 423, (2021).
7. Trigila, M., Luque, E., Gaudiani, A., Naiouf, M..: Simulación computacional, ciencia de los datos, cómputo de alto rendimiento y optimización aplicados a mejorar la predicción de modelos de simulación que representan la evolución de sistemas complejos. In XX Workshop de Investigadores en Ciencias de la Computación (2018). http://sedici.unlp.edu.ar/handle/10915/67063

Modelo de analítica prescriptiva en tiempo real para negocios con grandes volúmenes de eventos

Doctorado en Ciencias de la Computación, Universidad Nacional de San Luis, Argentina.

Tesista: Esteban Alejandro Schab[2,3]
Director: María Fabiana Piccoli[1,3]
Co-Director: Carlos Antonio Casanova Pietroboni[2,3]

[1] LIDIC- Univ. Nacional de San Luis, San Luis
[2] Univ. Tecnológica Nacional, F. R. Concepción del Uruguay, Concepción del Uruguay
[3] Universidad Autónoma de Entre Ríos, Concepción del Uruguay
mpiccoli@unsl.edu.ar
{schabe, casanovac}@frcu.utn.edu.ar

Palabras claves: Inteligencia Computacional, Analítica Prescriptiva, Computación de Alto Desempeño.

1 Motivación

La mejora continua y adaptativa de los procesos de negocio es clave para las organizaciones que pretenden mantenerse competitivas. En este sentido, la digitalización de los procesos, así como el incremento en las tecnologías de monitoreo, han llevado a producir una gran cantidad de datos (Datos Masivos o Big Data), los cuales tienen un gran potencial para la mejora de los procesos conducida por analíticas [1] [2] [3].

Las analíticas buscan transformar los datos en conocimiento para la toma de decisiones [4], pudiendo distinguirse cuatro tipos de analíticas según su nivel de automatización de su proceso [5]. La más básica es la analítica descriptiva, la cual intenta responder qué ha pasado o qué está pasando; seguida por la analítica diagnóstica que apunta a por qué ha pasado o está pasando. Ambas analizan datos históricos. En un nivel más avanzado, la analítica predictiva busca responder qué sucederá; aplica el conocimiento para predecir nuevos datos sobre el presente o el futuro (pronósticos). Ninguno de estos enfoques sugiere acciones concretas, sino que descansan en el juicio subjetivo y las habilidades analíticas del usuario para deducir acciones de mejora. Finalmente, la analítica prescriptiva intenta responder qué debería hacer y cómo podría hacerse para que algo suceda; calcula acciones a ser ejecutadas en el momento (decisiones operativas) o en el futuro (decisiones tácticas para corto y mediano plazo o estratégicas para largo plazo).

A pesar de los avances tecnológicos y del continuo crecimiento del volumen y variedad de datos disponibles, en general, las analíticas de procesos existentes dentro de la industria actual no aprovechan completamente el conocimiento oculto en los grandes volúmenes de datos con los que cuentan [1] debido a las siguientes limitaciones:

- No hacen uso de técnicas prescriptivas para transformar los resultados del análisis en acciones de mejora concretas, dejando este paso completamente a criterio del juicio subjetivo del usuario.
- Hacen un uso intensivo de datos de sistemas en producción, generando un deterioro en el desempeño de las herramientas de software que soportan los procesos.
- La optimización es conducida a posteriori, después de completado el proceso, en contraste a la mejora proactiva durante la ejecución del proceso.

Un tópico emergente dentro del área de datos masivos (o Big Data) es el procesamiento de datastreams, también llamado Data Stream Mining [6] [7] [8]. Un datastream es una representación digital y transmisión continua de datos, los cuales describen una clase de eventos relacionada [9] [10]. Mediante el procesamiento de datastreams se puede lograr la respuesta en tiempo real a eventos en forma de toma de decisiones [11] [12], lo cual abre nuevas y amplias oportunidades para la creación de valor en las organizaciones.

Un caso de interés es el de los sistemas de atención al público en bancos, hospitales y comercios. En ellos la generación de datastreams es causada por los sistemas de gestión de la atención implementados. Estos sistemas en general utilizan modelos relacionales de bases de datos y, si bien permiten la elaboración de analíticas, puede resultar inapropiada su implementación debido a que el desempeño del sistema de atención, especialmente en un contexto de Software as a Service (SaaS), puede colapsar ante las continuas consultas para realizar el monitoreo. Por tal motivo, la generación y el procesamiento de los eventos en forma de datastreams como una componente paralela al sistema de atención puede ser la tecnología de base para permitir un monitoreo eficiente, sin deterioro del desempeño del sistema en general.

Otro caso de interés es el Enrutamiento de vehículos (VRP) [13] con suministro de información y reencaminamiento en tiempo real, orientado a la búsqueda de un paradigma de movilidad inteligente [14]. Dentro de este problema se pueden estudiar de forma particular o en conjunto la logística urbana, el transporte de personas y los conductores individuales. En este caso los datastreams son generados de forma distribuida por cada agente involucrado y pueden ser procesados de forma centralizada o distribuida dependiendo del esquema elegido y los recursos disponibles.

Otros casos de aplicación en estudio son las Smart Grids. Así como la generación de alertas tempranas para la detección de enfermedades en el cultivo de arroz en la zona de Entre Ríos. Principalmente la llamada “Quemado del Arroz” (Pyricularia oryzae).

2 Objetivos y Aportes

El objetivo general de esta tesis es “Diseñar un modelo para la recomendación automática y proactiva de acciones de mejora en tiempo real para sistemas con grandes volúmenes de eventos”, buscando superar los inconvenientes antes descritos. Estos modelos serán parte esencial de un proceso de mejora continua basado en la recomendación de acciones operativas y tácticas destinadas a mantener los indicadores de rendimiento del sistema dentro de valores deseados, en un contexto con grandes flujos de eventos.

Para la construcción de estos modelos prescriptivos, cuya principal función es la determinación de las acciones a llevar a cabo, se hace uso de modelos predictivos para explorar los futuros cercanos y modelos descriptivos para calcular la aptitud de dichos estados. Para ello se propone el uso de agentes basados en aprendizaje por refuerzo [15], junto con otras técnicas provenientes de la Inteligencia Computacional: redes neuronales como modelos, teoría de conjuntos difusos como lenguaje de especificación, y métodos numéricos y metaheurísticos para el entrenamiento de tales modelos [16] [17] [18].

Además, ante la necesidad de dar rápida respuesta a procesos de negocios dinámicos, y dadas las características propias de cada una de las técnicas de inteligencia computacional antes mencionadas, es mandatorio pensar en la aplicación de modelos/paradigmas de computación de alto desempeño [19] [20] [21], principalmente en el entrenamiento de los modelos.

3 Estado Actual y Trabajo Futuro

Se encuentra en desarrollo un modelo basado en agentes de aprendizaje por refuerzo para un caso de enrutamiento de vehículos con suministro de información y reencaminamiento en tiempo real. Particularmente un caso de logística urbana (distribución de productos).

Este modelo VRP cuenta con una planificación inicial del recorrido, estimaciones borrosas de la cantidad de productos requerida por cada cliente y de los tiempos de recorrido, un límite de capacidad en los vehículos, y objetivos múltiples: satisfacer la demanda real de los clientes, minimizar los tiempos de recorrido o los costos de combustible, minimizar los cambios de recorrido o las visitas múltiple a un mismo cliente. El agente decide si espera o no ante la demora en un cliente, que cantidad de productos entregar a un cliente y el próximo cliente a visitar (pudiendo cambiar la planificación).

Para la generación de un entorno y datos de prueba se desarrolló un modelo de simulación que permite entrenar y testear los modelos generados. Se implementó utilizando Python y la librería Simpy. En el diseño e implementación del agente, que se encuentra en desarrollo, se utilizan Python, TensorFlow y la librería TF-Agents. Para el procesamiento los primeros desarrollos se implementaron sobre una GPU Nvidia con CUDA. Para el procesamiento de datastreams se trabaja con Apache Spark.

Se debe completar el desarrollo del agente y realizar la validación del mismo. Junto con ello se analizará el desempeño de los algoritmos para realizar los ajustes necesarios. Finalmente se debe integrar el agente con el procesamiento de eventos en forma de datastreams para su funcionamiento en tiempo real.

Como trabajo futuro se prevé la aplicación a otros casos de interés mencionados.

Referencias

1. Christoph Gröger, Holger Schwarz, and Bernhard Mitschang. “Prescriptive analytics for recommendation-based business process optimization”. In International Conference on Business Information Systems, pages 25–37. Springer, 2014.

2. Mandeep Kaur Saggi and Sushma Jain. “A survey towards an integration of big data analytics to big insights for value-creation”. Information Processing & Management, 54(5):758-790, 2018.
3. Usarat Thirathon, Bernhard Wieder, Zoltan Matolcsy, and Maria-Luise Ossimitz. “Impact of big data analytics on decision making and performance”. In International Conference on Enterprise Systems, Accounting and Logistics, 2017.
4. Clyde Holsapple, Anita Lee-Post, and Ram Pakath. “A unified foundation for business analytics. Decision Support Systems”, 64:130-141, 2014.
5. Michael Minelli, Michele Chambers, and Ambiga Dhiraj. “Big data, big analytics: emerging business intelligence and analytic trends for today’s businesses”. Volume 578. John Wiley & Sons, 2013.
6. Albert Bifet and Jesse Read. “Ubiquitous artificial intelligence and dynamic data streams”. In Proceedings of the 12th ACM International Conferenceon Distributed and Event-Based Systems, DEBS ’18, Pp. 1–6, Association for Computing Machinery. New York, USA, 2018.
7. Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michal Wozniak, and Francisco Herrera. “A survey on data preprocessing for data stream mining: Current status and future directions”. Neurocomputing, 239:39 – 57, 2017.
8. Taiwo Kolajo, Daramola Olawande, and Adebiyi Ayodele. "Big data stream analysis: a systematic literature review". Journal of Big Data, vol. 6, no 1, pp. 1-30, 2019.
9. Federico Pigni, Gabriele Piccoli, and Richard Watson. “Digital data streams: Creating value from the real-time flow of big data”. California Management Review, 58(3):5–25, 2016.
10. Chris Wrench, et al. "Data stream mining of event and complex event streams: A survey of existing and future technologies and applications in big data." Enterprise Big Data Engineering, Analytics, and Management. IGI Global, 2016. pp. 24-47.
11. Uri Verner, Assaf Schuster, and Mark Silberstein. “Processing data streams with hard real-time constraints on heterogeneous systems”. In Proceedings of the international conference on Supercomputing, pages 120–129, 2011.
12. Uri Verner, Assaf Schuster, Mark Silberstein, and Avi Mendelson. Scheduling processing of real-time data streams on heterogeneous multi-gpu systems. In Proceedings of the 5th Annual International Systems and Storage Conference, pages 1–12, 2012.
13. Kris Braekers, Katrien Ramaekers, and Inneke Van Nieuwenhuyse, "The vehicle routing problem: State of the art classification and review", Computers & Industrial Engineering, vol. 99, pp. 300-313, 2016.
14. Sandra Melo, Joaquim Macedo, and Patrícia Baptista, "Guiding cities to pursue a smart mobility paradigm: An example from vehicle routing guidance and its traffic and operational effects", Research in transportation economics, vol. 65, p. 24-33, 2017.
15. Stuart J Russell and Peter Norvig. “Inteligencia Artificial: un enfoque moderno”. 2004.
16. A. Ebrahimnejad and J. L. Verdegay. “Fuzzy sets-based methods and techniques for modern analytics”. Springer International Publishing. 2018.
17. Nazmul Siddique and Hojjat Adeli. “Computational intelligence: synergies of fuzzy logic, neural networks and evolutionary computing”. John Wiley & Sons, 2013.
18. Lotfi A. Zadeh. “Fuzzy logic, neural networks, and soft computing”. Commun. ACM 37, 3, 77–84. DOI: https://doi.org/10.1145/175247.175255. March 1994.
19. Mercedes Barrionuevo, Mariela Lopresti, Natalia Miranda, and María Fabiana Piccoli. “Solving a big-data problem with gpu: the network traffic analysis”. Journal of Computer Science and Technology, 15(01): p.30–39, Apr. 2015.
20. S. Kurgalin and S. Borzunov, “A Practical Approach to High-Performance Computing”. Springer. 2019.
21. Peter Pacheco. “An Introduction to Parallel Programming”, 1st ed., San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

Modelado matemático para la generación de imágenes de venas de la palma para la evaluación de algoritmos de identificación biométrica

Doctorado en Modelamiento Matemático, Universidad Católica del Maule, Chile

Tesista: Edwin Hernando Salazar-Jurado [1]
Director: Karina Vilches-Ponce [1]
Co-Director: Ruber Hernández-García[1]
Co-Director: Ricardo J. Barrientos[1]

[1] Universidad Católica del Maule, Talca 3460000, Chile

Palabras claves: Biometría, venas de la palma, bases de datos a gran escala, imágenes sintéticas.

1 Motivación

Actualmente el reconocimiento de personas por medio de venas de la palma ha recibido la atención de la comunidad científica debido a la alta seguridad que proporcionan. Sin embargo, los algoritmos para el reconocimiento se validan con un número limitado de imágenes, lo que dificulta la implementación de métodos basados en *Deep Learning* y la evaluación de la escalabilidad para la identificación masiva. Esto se debe principalmente a la falta de bases de datos a gran escala de imágenes de venas de la palma, destinadas a proporcionar datos suficientes para evaluar algoritmos de reconocimiento. Crear una base de datos de imágenes reales de venas de la palma a gran escala es una tarea laboriosa en términos de tiempo, seguridad y costo. Por lo anterior, el objetivo del presente trabajo es generar imágenes sintéticas de venas de la palma mediante el modelado matemático de la estructura vascular y de los efectos producidos por los dispositivos de adquisición. Esto permitirá la elaboración de una base de datos a gran escala de imágenes sintéticas venas de la palma, favoreciendo la evaluación de algoritmos para dar soluciones fiables y escalables al reconocimiento biométrico de personas.

2 Objetivos y Aportes

Objetivo general: generar imágenes sintéticas de venas de la palma mediante el modelado matemático de la estructura vascular de la palma y de los efectos producidos por los dispositivos de adquisición, para la elaboración de una base de datos a gran escala que permita la evaluación de la fiabilidad y escalabilidad de algoritmos de reconocimiento biométrico de personas. Para llevar a cabo el objetivo general, nos planteamos los siguientes objetivos específicos.

- Desarrollar el modelo matemático para la simulación de la estructura venosa de la palma y de los efectos producidos por los dispositivos de adquisición.
- Crear una base de datos a gran escala de imágenes sintéticas de venas de la palma a partir de la implementación del modelo matemático anterior.
- Evaluar la similitud entre imágenes sintéticas generadas e imágenes reales de bases de datos públicas, a partir de métricas cualitativas y cuantitativas de visión por computadora para la comparación de imágenes y la evaluación de métodos biométricos.

3 Estado Actual y Trabajo Futuro

Un sistema de reconocimiento basado en las venas de la palma se puede resumir en cuatro procesos: la adquisición de la imagen, el preprocesamiento, la extracción de características y el reconocimiento [1]. Esta biometría es una tecnología relativamente nueva, por lo cual existen varios desafíos en cada uno de los procesos; en particular estudiar el proceso inicial del sistema de reconocimiento y las redes vasculares de la palma es esencial si el objetivo final es construir imágenes sintéticas, como es el caso de este trabajo.

3.1 Adquisición de venas de la palma

La adquisición de las imágenes de las venas de la palma se logra mediante el uso dispositivos infrarrojos (IR) que interactúan con la hemoglobina oxidada (HbO) y la hemoglobina desoxidada (Hb) de la red vascular [1]. En la ventana del infrarrojo cercano (NIR por sus siglas en ingles); cuando la longitud de onda de la luz está comprendida entre 720nm y 760nm la radiación es fuertemente absorbida por el Hb, lo que produce una sombra correspondiente al patrón de las venas. En 790nm existe un punto de intersección donde la Hb y la HbO presentan la misma absorción lo que permite la visualización de las venas y las arterias, para rangos superiores en el espectro de la ventana IR la HbO presenta un leve aumento en comparación con el Hb [2].

3.2 Estructura vascular de la palma

Las redes vasculares de la palma se dividen en venas y arterias; las venas drenan Hb y las arterias transportan HbO. Las redes vasculares de la palma presentan patrones complejos que permiten asegurar el flujo sanguíneo sin obstrucciones ante cualquier movimiento [3]. Estos patrones son precisamente los que permiten utilizar la red vascular para la identificación, particularmente se utilizan las redes superficiales dadas las limitaciones ópticas de los sistemas de adquisición [4]. La red arterial superficial de la palma de la mano está descrita en detalle; está formada por un arco palmar superficial que se forma alrededor de los metacarpianos y se ramifica en las arterias digitales palmares comunes para ascender a cada uno de los dedos [5]. En cuanto a la topología de las venas de la palma, es un tema poco estudiado, solo se describe como una red venosa aleatoria, prestando mayor atención a la funcionalidad; en especial se estudia como las venas pueden escapar a la presión durante el agarre [6].

3.3 Generación de imágenes sintéticas de patrones vasculares

La elaboración de imágenes de estructuras vasculares para el reconocimiento biométrico, según la revisión bibliográfica realizada, solo existe una base de datos sintética correspondiente a las venas de los dedos, la cual está formada por 50.000 imágenes correspondientes a 5.000 sujetos [7]. Aparte de esta base de datos, otros estudios han proporcionado metodologías para la generación de imágenes sintéticas de las venas del dorso de la mano y de la esclerótica de los ojos [4,8]. Con respecto a la simulación de imágenes de venas de la palma, aún no se conocen métodos que permitan reproducir la red vascular, esto se debe principalmente a los complejos patrones que se forman [9,10]. De hecho, los diferentes movimientos de la mano hacen que se requieran muchas estructuras trabajando en sincronía, organizando múltiples redes de vasos que aseguran el constante flujo sanguíneo. Para la simulación de la estructura venosa palmar se puede utilizar metodologías asociadas al estudio de diversas redes biológicas que se encuentran en la naturaleza. Estas redes han presentado ciertas similitudes con las redes sanguíneas que permiten compararlas [11-15].

3.4 Bases de datos de imágenes sintéticas de venas de la palma

En la Figura 1, se formaliza un esquema general para la generación de bases de datos sintéticas de venas de la palma a gran escala. El diagrama de flujo propuesto consta de tres procesos principales; el primero varía en función del método de generación implementado, mientras que los dos últimos permanecen constantes. El proceso GIV tiene como objetivo la generación de imágenes sintéticas aleatorias que simulan tanto los patrones venosos como las características visuales de las imágenes de las venas de la palma. El proceso anterior genera aleatoriamente imágenes sintéticas de venas de la palma, pero estas deben satisfacer el requisito de unicidad para formar parte de una base de datos biométrica. Por esto, el proceso DUI compara cada imagen recién generada con todas las imágenes de la base de datos sintética, garantizando este requisito. Este proceso necesita más tiempo a medida que la base de datos crece. Para reducir el tiempo de cómputo, se pude proponer un esquema basado en una estructura distribuida que permite repartir la carga de trabajo del DUI. Por otro lado, dado que una base de datos biométrica debe contener más de una muestra por individuo (muestras de galería y de prueba), el proceso AM tiene como objetivo obtener diferentes muestras para cada sujeto aplicando transformaciones aleatorias sobre la imagen generada, simulando las variaciones naturales que ocurren durante un procedimiento de adquisición sin contacto en el mundo real. El esquema anterior nos ha permitido desarrollar dos bases de datos sintéticas de venas de la palma. Por un lado, utilizamos redes GAN para la generación de las imágenes, y por otro lado, generamos un patrón venoso mediante un proceso de optimización y lo combinamos con huellas de la palma generadas con redes GAN. Las bases de datos se encuentran disponibles en https://www.litrp.cl/repository.html.

Entre los trabajos futuros están: la elaboración de un modelo matemático que tenga en cuenta la topología de la red venosa y la implementación de una estructura distribuida para minimizar el tiempo de computo en el discriminador de unicidad DUI.

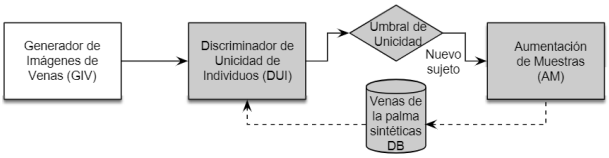


Figura1. Diagrama de flujo para la generación de imágenes sintéticas de venas de la palma para la creación de bases de datos sintéticas.

Referencias

1. Wu, W., Elliott, S., Lin, S., Sun, S. & Tang, Y. Review of palm vein recognition.IET Biom.9, 1–10 (2019).
2. Crisan, S. & Tebrean, B. Low cost, high quality vein pattern recognition device with liveness detection. workflow and implementations. Measurement 108, 207–216 (2017).
3. Nyström, Å., Fridén, J. & Lister, G. D. Superficial venous anatomy of the human palm. Scand. J. Plast. Reconstr. Surg. Hand Surg. 24, 121–127 (1990).
4. Crisan, S., Târnovan, I. G. & Crisan, T. A hand vein structure simulation platform for algorithm testing and biometric identification. In 16th IMEKO TC4 Symposium, Florence, Italy (2008).
5. Hansen, J. Netter's Anatomy Flash Cards E-Book. Netter Basic Science (Elsevier Health Sciences, 2017).
6. Botte, M. J. Surgical anatomy of the hand and upper extremity (Lippincott Williams & Wilkins, 2003).
7. Hillerström, F., Kumar, A. & Veldhuis, R. Generating and analyzing synthetic finger vein images. In 2014 International Conference of the Biometrics Special Interest Group (BIOSIG), 1–9 (IEEE, 2014).
8. Das, A., Mondal, P., Pal, U., Blumenstein, M. & Ferrer, M. A. Sclera vessel pattern synthesis based on a non-parametric texture synthesis technique. In Proceedings of international conference on computer vision and image processing, 241–250 (Springer, 2017).
9. Sukop, A. et al. Clinical anatomy of the dorsal venous network in fingers with regard to replantation. Clin. Anatomy: The Off. J. Am. Assoc. Clin. Anat. Br. Assoc. Clin. Anat. 20, 77–81 (2007).
10. Kiss, F. & Szentágothai, J. Atlas anatomiae corporis humani. Atlas d'anatomie du corps humain (Akadémiai Kiadó, 1974).
11. Aghamirmohammadali, S., Bozorgmehry Boozarjomehry, R. & Abdekhodaie, M. Modelling of retinal vasculature based on genetically tuned parametric l-system. Royal Soc. open science 5, 1–20 (2018).
12. Scianna, M., Bell, C. & Preziosi, L. A review of mathematical models for the formation of vascular networks. J. theoretical biology 333, 174–209 (2013).
13. Francis, C., Frederic, L., Sylvie, L., Prasanna, P. & Henri, D. Scaling laws for branching vessels of human cerebral cortex. Microcirculation 16, 331–344 (2009).
14. Runions, A. et al. Modeling and visualization of leaf venation patterns. In ACM SIGGRAPH 2005 Papers, 702–711 (2005).
15. Runions, A., Lane, B. & Prusinkiewicz, P. Modeling trees with a space colonization algorithm. Eurographics Work. on Nat. Phenom. 7, 63–70 (2007).

Modelamiento matemático de computación distribuida para clasificación multiclase en paralelo con bases de datos a gran escala basado en Extreme Learning Machine

Doctorado en Modelamiento Matemático Aplicado, Universidad Católica del Maule, Chile.

Tesista: Elkin Gelvez-Almeida[1,2]
Director: Ricardo J. Barrientos[1]
Co-Director: Karina Vilches-Ponce[1]
Co-Director: Marco Mora[1]

[1] Universidad Católica del Maule, Talca 3460000, Chile
[2] Universidad Simón Bolívar, San José de Cúcuta 540006, Colombia
elkin.gelvez@alu.ucm.cl

Palabras claves: Modelamiento matemático, álgebra computacional, computo distribuido y paralelo, redes neuronales de aprendizaje extremo.

1 Motivación

El gran avance de la tecnología y el internet, ha generado un aumento significativo en el tamaño de las bases de datos. Procesar esta gran cantidad de información es ahora un desafío, por los elevados tiempos de computo y las limitaciones del hardware. En este sentido, actividades como el análisis predictivo, reconocimiento de tendencias, detección de patrones de comportamiento, identificación de personas, visión por computadora, diagnostico, entre otras, requiere de tiempos elevados para ser ejecutadas. Lo anterior es un gran problema, especialmente cuando se intenta desarrollar aplicaciones en tiempo real. Esta situación ha llevado a la necesidad de mejorar los algoritmos actuales, y combinarlos con diferentes arquitecturas y tecnologías, con el propósito de disminuir los tiempos de ejecución sin comprometer la precisión y estabilidad.

2 Objetivos y Aportes

2.1 Objetivo general

Disminuir el tiempo de entrenamiento de las redes neuronales ELM en problemas de clasificación con bases de datos a gran escala utilizando modelos de distribución.

2.2 Objetivos específicos

1. Identificar modelos de distribución que se puedan implementar en redes secuenciales en linea OS-ELM para abordar problemas de clasificación con bases de datos a gran escala.
2. Diseñar una red neuronal secuencial en linea OS-ELM con modelos distribuidos que permita disminuir el tiempo de entrenamiento en problemas de clasificación con bases de datos a gran escala.
3. Analizar el rendimiento de la red neuronal secuencial en linea OS-ELM con modelo distribuido respecto a los algoritmos de clasificación existentes en la literatura.

2.3 Aportes

La presente investigación pretende desarrollar estrategias computacionales para abordar problemas de clasificación en tiempos razonables con bases de datos a gran escala. Esto significa un aporte valioso para la comunidad científica dado el incremento constante de tamaño en las bases de datos.

3 Estado Actual y Trabajo Futuro

Extreme Learning Machine (ELM) es un modelo propuesto por Huang et al. en 2004 [1], inicialmente para redes neuronales de alimentación directa de una sola capa oculta (SLFN), y luego extendido a redes de múltiples capas. Este modelo ha tenido gran aceptación en la comunidad científica por la simplicidad del modelo y su alta precisión. Las investigaciones en ELM han demostrado tener buena precisión, y tiempos de entrenamientos mas bajos en comparación con algoritmos clásicos como Back-Propagation (BP) y Support Vector Machines (SVM) [1]. El modelo estándar consiste en asignar aleatoriamente los pesos y sesgos de la capa oculta, y calcular los pesos de la capa de salida analíticamente por medio de la inversa generalizada de Moore-Penrose. En la Figura 1 se presenta la estructura básica de ELM estándar.

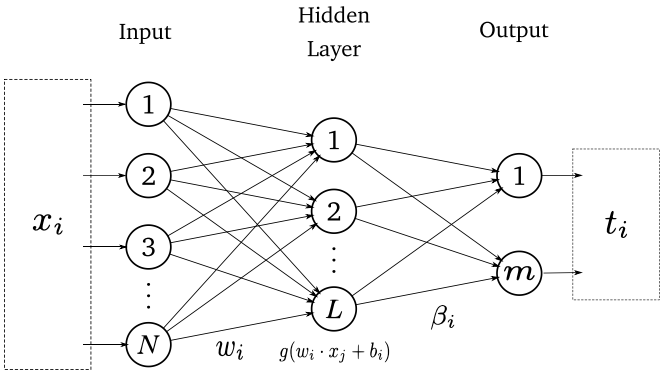


Figura 1. Estructura básica del modelo ELM estándar.

Sea un conjunto de entrenamiento arbitrario $\mathcal{N} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i) | \boldsymbol{x}_i \in \mathbb{R}^n, \boldsymbol{t}_i \in \mathbb{R}^m\}$, con $i = 1, \dots, N$, una función de activación $g(x)$: $\mathbb{R}^n \to \mathbb{R}^n$, y un número de neuronas en la capa oculta $L | L \leq N$, el algoritmo de entrenamiento de una SLFN está definido por

$$\sum_{i=1}^{L} \boldsymbol{\beta}_i g\left(\boldsymbol{\omega}_i \cdot \boldsymbol{x}_j + b_i\right) = t_j, \qquad j = 1, \dots, N, \tag{1}$$

Donde $\boldsymbol{\omega}_i$ y b_i son los i-ésimos pesos y sesgos de la capa oculta, respectivamente, $\boldsymbol{\beta}_i$ es el i-ésimo peso de la capa de salida, y $\boldsymbol{\omega}_i \cdot \boldsymbol{x}_j$ representa el producto interno de $\boldsymbol{\omega}_i$ y $\boldsymbol{x}_j$. La Ecuación anterior puede escribirse de forma matricial como $H\beta = T$, donde H es llamada la matriz de salida de la capa oculta. Teniendo en cuenta lo anterior, el algoritmo de entrenamiento de ELM se puede resumir en los siguientes tres pasos:

1. Asignar valores aleatorios a los pesos $\boldsymbol{\omega}_i$ y los sesgos b_i de la capa oculta;
2. Calcular la matriz $\boldsymbol{H}$ de salida de la capa oculta;
3. Computar los pesos $\boldsymbol{\beta}_i$ de la capa de salida con $\boldsymbol{\beta} = \boldsymbol{H}^{\tau}\boldsymbol{T}$, donde H^{τ} es la inversa generalizada de Moore-Penrose de la matiz H.

Con el propósito de mejorar la precisión y estabilidad del algoritmo, y responder a diferentes aplicaciones en particular, se han propuesto algunas variantes de ELM [2]. Una variante que ha llamado mucho la atención por la comunidad científica, y que en particular nos interesa para procesar grandes volúmenes de datos es Online Sequential Extreme Learning Machine (OS-ELM), propuesto por Liang et al. en 2006 [3]. El algoritmo permite entrenar con el conjunto de datos por bloques o uno a uno. Este método es ideal para aquellos casos en los que los datos no se obtienen al mismo tiempo, ya que al momento de tener nuevos datos, no es necesario realzar un nuevo entrenamiento completo, por el contrario, permite actualizar los pesos $\boldsymbol{\beta}_i$ de la capa de salida en base a los resultados anteriores. Esta estrategia permite realizar varios entrenamientos con diferentes bloques de datos de forma independiente, con lo que se puede utilizar la computación paralela, y de esta forma, aprovechar al máximo la arquitectura con la que se cuente.

En este orden de ideas, también es importante resaltar las tecnologías para el computo distribuido y paralelo que se han desarrollado. Entre las tecnologías de mayor interés para nuestra investigación se encuentra MapReduce, propuesta por Dean y Ghemawat en 2004 [4]. Este modelo puede procesar grandes cantidades de datos, que de otro modo serían procesados en cientos o miles de máquinas para tener tiempos de procesamiento razonables. Por otro lado, Apache Spark es un modelo propuesto por Zaharia et al. en 2010[5]. Este modelo permite abordar problemas que no son posibles con MapReduce, mientras se preserva la escalabilidad y la tolerancia a fallas. Finalmente, la programación en Graphics Processing Unit (GPU) ha sido una herramienta útil para acelerar los tiempos de ejecución de programas que procesan grandes volúmenes de datos en diferentes aplicaciones. Este modelo de programación en GPU está organizado por tres niveles jerárquicos correspondientes a hilos, bloques y redes [6].

En este sentido, este proyecto pretende diseñar un modelo de computo distribuido que permita resolver problemas de clasificación en paralelo con bases de datos a gran escala, utilizando Extreme Learning Machine. Los desafíos de esta propuesta se pueden enmarcar en dos principales aspecto: Por un lado, el poder procesar bases de datos a gran escala, y por otro lado, realizar entrenamiento en tiempos razonables. Para abordar estos desafíos, las tecnologías de computo distribuido combinadas con las variantes ELM, se presentan como un escenario adecuado.

Referencias

1. Huang, G. B., Zhu, Q. Y., Siew, C. K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541), vol. 2, pp. 985–990. IEEE, Budapest (2004).
2. Wang, J., Lu, S., Wang, SH., et al.: A review on extreme learning machine. Multimedia Tools and Applications (2021).
3. Liang, N. Y., Huang, G. B., Saratchandran, P., et al.: A fast and accurate online sequential learning algorithm for feedforward networks. IEEE Transactions on Neural Networks 17(6), 1411–1423 (2006).
4. Dean, J., Ghemawat, S.: MapReduce: Simplified data processing on large clusters. In: 6th Symposium on Operating Systems Design and Implementation, vol 6, pp. 137–149. USENIX Association (2004).
5. Zaharia, M., Chowdhury, M., Franklin, M. J., et al.: Spark: Cluster Computing with Working Sets. In: Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, pp. 10. USENIX Association, Boston (2010).
6. Navarro, C. A., Carrasco, R., Barrientos. R. J., et al.: GPU Tensor Cores for Fast Arithmetic Reductions. IEEE Transactions on Parallel and Distributed Systems 32(1). 72–84 (2021).