Sistemas inteligentes en el uso de aplicaciones de bioinformática y sistemas embebidos

W. Hasperué[1], C. Estrebou[1], G. Camele[1,3], E. Rucci[1], F. Ronchetti[1], D. Castillo [2,4], G. Reyes Zambrano[2,5], L. Lanzarini[1], A. Fernandez Bariviera[6]

[1] Instituto de Investigación en Informática LIDI*, Facultad de Informática, UNLP, La Plata, Argentina
[2] Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina
[3] Becario postgrado UNLP.
[4] Facultad de Ciencias de la Educación, Universidad Tecnológica Indoamérica, Ambato, Ecuador
[5] Facultad de Ciencias Físicas y Matemáticas, Universidad de Guayaquil, Guayaquil, Ecuador
[6] Dpto. de Economía, Universitat Rovira i Virgili, Reus, España
* Centro asociado de la Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC)

{whasperue, cesarest, gcamele, erucci, fronchetti,laural}@lidi.info.unlp.edu.ar,
{david.castillos, gary.reyesz}@info.unlp.edu.ar, aurelio.fernandez@urv.net

CONTEXTO

Esta presentación corresponde a las tareas de investigación que se llevan a cabo en el III-LIDI en el marco del proyecto "Sistemas inteligentes. Aplicaciones en reconocimiento de patrones, minería de datos y big data" perteneciente al Programa de Incentivos (2018-2022).

RESUMEN

Esta línea de investigación se centra en el estudio y desarrollo de Sistemas Inteligentes para la resolución de problemas de Big Data y Minería de Datos utilizando técnicas de Aprendizaje Automático. Los sistemas desarrollados se aplican particularmente al procesamiento de grandes volúmenes de información y al procesamiento de flujo de datos.

Las investigaciones correspondientes al procesamiento de datos masivos están enfocadas en el estudio y desarrollo de técnicas de selección de características, donde el foco está puesto en la reducción de los tiempos de cómputo. La optimización puede realizarse tanto en la mejora de la ejecución en un entorno distribuido, como en la propuesta de técnicas metaheurísticas que permitan obtener un subconjunto óptimo de atributos.

Por otro lado, y desde el punto de vista de la salud, se está trabajando con investigadores del CENEXA (CONICET-UNLP-CIC) en la obtención de modelos de predicción de diabetes y prediabetes.

Las investigaciones relacionadas con el análisis de flujos de datos se centran en la construcción de modelos dinámicos descriptivos. En particular el énfasis está puesto en la resolución de dos problemas de sumo interés en distintas áreas: el análisis de trayectorias GPS a fin de identificar congestiones en el tránsito y la identificación temprana de patrones de movimiento en pacientes con Alzheimer.

Palabras clave: Big Data, Minería de Datos, Diabetes, Análisis de flujos de datos, Reducción de características, Tiny ML, GPS.

1. INTRODUCCION

El Instituto de Investigación en Informática LIDI tiene una larga trayectoria en el estudio, investigación y desarrollo de Sistemas Inteligentes basados en distintos tipos de estrategias adaptativas. Los resultados obtenidos han sido aplicados en la solución de problemas de distintas áreas. A continuación, se detallan las investigaciones realizadas durante el último año.

1.1. BIG DATA

Selección de características

El objetivo de los algoritmos de selección de características es el de reducir las entradas a un tamaño apropiado para su procesamiento y

análisis. La selección de características implica la elección de ciertos atributos, tal que, con ese subconjunto, las "propiedades naturales" de los datos no sean alteradas.

Actualmente, en el III LIDI se están realizando tareas de investigación que incluyen el desarrollo de algoritmos de selección de características que puedan ser utilizados en bases de datos con información génica. Un objetivo de la medicina genómica es identificar un grupo de genes, cuyo patrón de expresión se encuentre asociado a un fenotipo en particular: concepto conocido como *gene signature* (biomarcador diagnóstico, pronóstico o predictivo de una patología en estudio).

Cuando el volumen de información a procesar crece, la ejecución de los algoritmos de selección de atributos convencionales incrementa notablemente su tiempo de procesamiento. En la actualidad se cuenta con herramientas que, al distribuir el cómputo entre diferentes nodos que conforman un cluster de computadoras, hacen posible el procesamiento de grandes volúmenes de datos de una manera eficiente. En este aspecto Apache Spark es uno de los frameworks más utilizados. En particular, su librería Spark ML contiene la implementación de muchos algoritmos de machine learning.

En [1] y en [2] se llevó a cabo una comparación entre cuatro algoritmos de clasificación implementa-dos en MLlib: Random Forest, Support Vector Machine, Naïve Bayes y MultiLayer Perceptrón. Se analizaron cuatro métricas de poder pronóstico de los modelos junto con los tiempos de ejecución requeridos. Los experimentos están enfocados a comparar las métricas de los cuatro algoritmos estudiados en función del número de atributos seleccionado de una base de datos.

Identificación de Personas con Riesgo de Diabetes y Prediabetes

La Diabetes Tipo 2 (DT2) es una enfermedad crónica caracterizada por una disminución precoz y progresiva de la masa y de la función de las células beta del páncreas. Debido a su creciente prevalencia en combinación con su elevado costo de atención, constituye un serio problema de salud pública, por lo que se han realizado grandes esfuerzos por desarrollar estrategias efectivas para su prevención y tratamiento, así como para evitar sus complicaciones crónicas. En ese sentido, resulta importante reconocer que las consecuencias negativas de esta enfermedad comienzan en una etapa previa conocida como prediabetes, la cual implica un riesgo elevado de desarrollar DT2 en los siguientes años.

El desarrollo de la DT2 es un proceso lento y progresivo condicionado por factores genéticos, ambientales y de comportamiento. Aunque no existe una cura definitiva para esta enfermedad, varios estudios han demostrado que se puede prevenir o demorar su aparición en personas con prediabetes a través de la adopción de un estilo de vida saludable y/o asociado con la ingesta de diversos fármacos. En Argentina, existe la iniciativa Programa PPDBA desarrollado por el CENEXA (CONICET-UNLP-CIC) [4].

La detección de DT2 y prediabetes representa un verdadero desafío para la medicina debido a la ausencia de síntomas patogenómicos y/o la falta de conocimiento de los factores de riesgo asociados. Los modelos existentes para predicción de diabetes y prediabetes no necesariamente aplican a la población argentina [3] y, hasta donde llega nuestro conocimiento, no existe modelo ni herramienta similar disponible en nuestro medio que permita identificar personas con alta probabilidad de tener estas enfermedades. Brevemente, esta línea propone desarrollar y validar modelos predictivos de diabetes y prediabetes específicos para la población argentina utilizando técnicas de Aprendizaje Automático. Se cuenta con acceso a la base de datos del PPDBA y apoyo del equipo médico de CENEXA. La concreción de esta línea representaría un avance en el conocimiento y un instrumento útil para los sistemas de salud de Argentina.

1.2. ANALISIS DE FLUJOS DE DATOS

Patrones de tránsito vehicular

El volumen de tráfico vehicular de las grandes ciudades se ha incrementado en los últimos años originando problemas de movilidad; por

ello el análisis de los datos del flujo vehicular son de suma importancia. Los Sistemas Inteligentes de transporte realizan el monitoreo y control vehicular recolectando trayectorias GPS. Las técnicas de agrupamiento permiten identificar patrones sobre el flujo vehicular.

En este sentido se ha desarrollado en el III-LIDI una metodología capaz de identificar, de manera dinámica, las velocidades más representativas del flujo vehicular en un período de tiempo. Para ello, el flujo de datos GPS entrada es representado utilizando un reticulado de celdas correspondientes a pequeñas zonas geográficas. Luego, las características de las celdas son agrupadas dinámicamente. La repetición de estos pasos en forma periódica utilizando períodos de tiempo cortos fueron representados en un mapa interactivo facilitando de esta manera la interpretación del desplazamiento de los vehículos a distintas velocidades. La validación del método propuesto fue realizada con dos conjuntos de datos de Roma y Guayaquil y los resultados obtenidos fueron satisfactorios. Los resultados de estas investigaciones se publicaron en [5].

Actualmente se está completando este procesamiento con la incorporación de índices de congestión con el objetivo de identificar los cambios en el flujo vehicular, analizando las zonas geográficas en un período dado de tiempo, y estimar si se trata de una congestión o no.

Detección de poses en pacientes con Alzheimer

Según la OMS, la demencia es una de las principales causas de discapacidad y dependencia entre las personas mayores. En todo el mundo, más de 55 millones de personas viven con demencia y se calcula que esta cifra aumentará a 78 millones para 2030. Demencia es un término general para varias enfermedades que generalmente son de naturaleza crónica, que resultan en deterioros cognitivos e interfieren con la capacidad para realizar las actividades de la vida diaria. La enfermedad de Alzheimer es la forma más común de demencia contribuyendo al 60-70% de los casos. Está comprobado que un diagnóstico precoz mejora la calidad de vida del paciente y del familiar, aumenta o mantiene su autonomía personal y mantiene sus capacidades cognitivas. Las personas que padecen Alzheimer suelen manifestar poco interés en las cosas o en el medio que lo rodea y presentan cambios conductuales variables desarrollando movimientos corporales relacionados de forma directa con su estado anímico. Esta línea de investigación se centra en la identificación automática de los movimientos corporales a partir de imágenes y videos.

En [6] se comenzó trabajando con el análisis de la marcha de pacientes con Alzheimer. Luego en [7] se trabajó sobre las poses: "De pie" y "Sentado" en pacientes con Alzheimer a partir de imágenes obtenidas de centros de atención al adulto mayor del cantón Ambato, Ecuador. Se trabajó con una población de 45 personas de ambos sexos diagnosticadas con Alzheimer con edades entre 75 y 89 años. Estas poses identificadas fueron utilizadas posteriormente en un análisis exploratorio relacionado con las categorías de deambulación, nervioso, deprimido, desorientado o aburrido arrojando resultados satisfactorios.

En [8] se buscó clasificar automáticamente el estado de ánimo de un paciente con Alzheimer en una de las siguientes categorías: deambulante, nervioso, deprimido, desorientado, aburrido o normal a partir de videos obtenidos en hogares de ancianos del cantón Ambato, Ecuador. Se trabajó con una población de 39 personas de ambos sexos diagnosticadas con Alzheimer y cuyas edades oscilaban entre los 75 y 89 años. Los métodos utilizados fueron detección de poses, extracción de características y clasificación de poses. Se utilizaron redes neuronales, el clasificador walk y la métrica de distancia Levenshtein. Como resultado del análisis de cada video se generó automáticamente una secuencia de estados de ánimo interpretable por el experto humano. Se pudo afirmar satisfactoriamente que el software de visión artificial facilita el reconocimiento de los estados de ánimo de los enfermos de Alzheimer durante los cambios de pose a lo largo del tiempo ayudando al profesional médico en su diagnóstico.

1.3. TINYML

La combinación de las áreas del aprendizaje automático y de los sistemas embebidos da lugar al campo de estudio conocido como TinyML. En esta línea de investigación, desde 2021 se lleva adelante un proyecto de investigación [9, 10] con el objetivo de explorar y adaptar técnicas y modelos de aprendizaje automático para dispositivos con importantes limitaciones de hardware como los microcontroladores (MCU).

Uno de los principales aportes del proyecto es el desarrollo de EmbedIA, un framework que transforma modelos de redes neuronales Tensorflow/Keras (TF/Keras) en código compatible con C/C++/Arduino. Esta característica permite realizar inferencia de modelos en cualquier MCU sin requerimientos específicos de hardware, a diferencia de otros frameworks o bibliotecas que demandan dispositivos de 32 bits (mayoritariamente basados en arquitectura ARM) o soporte para instrucciones DSP o SIMD. Además, se han generado y adaptado modelos TF/Keras para su ejecución en diversos MCUs. De la comparación de éstos con otros frameworks [10,11,12], se concluyó que, en general, los modelos ocupan memos memoria y se ejecutan más rápido en EmbedIA.

Entre las tareas que se están desarrollando actualmente, se encuentra la integración a EmbedIA de nuevas funcionalidades. Por un lado, se está trabajando en la integración de soporte para modelos de aprendizaje automático Scikit-Learn basados en árboles de decisión. Por otro lado, se está trabajando en la compatibilidad con modelos de redes neuronales en formatos como tflite, onnx y torch, además del ya soportado por TF/Keras. También se están desarrollando y evaluando modelos convolucionales reducidos para detección de fruta en mal estado, detección de personas e identificación de palabras clave en audio para placas de desarrollo ESP32-CAM, Raspberry Pi Pico y ARM Stm32f411.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Medición de performance de algoritmos de machine learning en entornos distribuidos.
- Desarrollo y mantenimiento de una herramienta para el análisis de progenie, basada en Spark.
- Identificación de Personas con Riesgo de Diabetes y Prediabetes
- Preprocesamiento de trayectorias vehiculares. Técnicas de segmentación.
- Agrupamiento dinámico de flujos de datos.
- Reconocimiento de poses humanas en imágenes y en videos
- Desarrollo de una BBDD de poses de movilidad.
- Estudio de técnicas de compresión de modelos para microcontroladores.
- Análisis de bibliotecas y frameworks de aprendizaje automático para microcontroladores.
- Desarrollo de un Framework de código abierto que transforma modelos desarrollados en Tensorflow/Keras y Scikit-Learn para ejecutarlos en microcontroladores.

3. RESULTADOS OBTENIDOS

- Medición de tiempos de ejecución de algoritmos de ML en un entorno Spark.
- Modelos de predicción de diabetes y prediabetes
- Diseño e implementación de un nuevo método de detección de rangos de velocidad agrupamiento de trayectorias GPS aplicable a la predicción de congestiones vehiculares.
- Desarrollo de una metodología capaz de clasificar automáticamente el estado de ánimo de un paciente con Alzheimer.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo de la línea de I/D aquí presentada está formado por: 4 profesores doctores con dedicación exclusiva, un profesor con dedicación exclusiva, 3 tesistas de Doctorado en Cs. Informáticas (1 con beca de postgrado de la UNLP) y un profesor extranjero.

Dentro de los temas involucrados en esta línea de investigación, en los últimos 3 años se han finalizado 4 tesis de doctorado, 3 tesis de especialista, 5 tesinas de grado de Licenciatura y 7 prácticas profesionales supervisadas.

Actualmente se están desarrollando 5 tesis de doctorado, 3 tesis de maestría y 5 prácticas profesionales supervisadas. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] Camele, G.; Hasperué, W.; Ronchetti, F.; Quiroga, F.; A comparative study of the performance of four classification algorithms from the Apache Spark ML library. CACIC. Salta. 2021.

[2] Camele, G.; Hasperué, W.; Ronchetti, F.; Quiroga, F.M. Statistical analysis of the performance of four Apache Spark ML algorithms. Journal of Computer Science and Technology 22 (2), e14-e14, 2022.

[3] Choudhury A., Gupta D. A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques. In: Recent Developments in Machine Learning and Data Analytics. Advances in Intelligent Systems and Computing, vol 740. Springer, Singapore. 2019.

[4] Gagliardino J. J., Etchegoyen G., Bourgeois M., Fantuzzi G., García S., González L., Elgart J. F., Ré M., Ricart A., Ricart J. P., Spinedi E., "Prevención primaria de diabetes tipo 2 en argentina: estudio piloto en la provincia de buenos aires," Revista Argentina de Endocrinología y Metabolismo, vol. 53, no. 4, pp. 135 – 141, 2016.

[5] Data stream processing method for clustering of trajectories. Reyes G., Lanzarini L., Estrebou C., Fernández-Bariviera A. Communications in Computer and Information Science. Serie CCIS. Springer International Publishing. ISSN 1865-0929. Pags.151-163. 2022.

[6] Using Kinect to Detect Gait Movement in Alzheimer Patients. Castillo-Salazar D., Lanzarini L., Guevara C., Alvarado H.G. In: Trends and Applications in Information Systems and Technologies. WorldCIST 2021. Advances in Intelligent Systems and Computing, vol 1365, pp.14-28. ISBN 978-3-030-72656-0. Springer. 2021.

[7] Castillo Salazar, D.R., Lanzarini, L., Gómez Alvarado, H.F., Cabrera López, J. The Detection, Extraction, and Classification of Human Pose in Alzheimer's Patients. In: Troiano, L., Vaccaro, A., Kesswani, N., Díaz Rodriguez, I., Brigui, I. (eds) Progresses in Artificial Intelligence & Robotics: Algorithms & Applications. Lecture Notes in Networks and Systems, vol 441. Springer, Cham. ISBN 978-3-030-98531-8. 2022.

[8] Castillo Salazar, D., Lanzarini, L., Alvarado, H., Varela-Aldás, J. Artificial vision system to detect the mood of an Alzheimer's patient. (2022). In: Tareq Ahram, Jay Kalra and Waldemar Karwowski (eds) Artificial Intelligence and Social Computing (AHFE 2022) International Conference. AHFE Open Access, vol 28. USA. ISSN 2771-0718.

[9] Estrebou C., Feming M., Saavedra M. D., Adra F. MbedML: A Machine Learning Project for Embedded Systems. IX Jornadas de Cloud Computing, Big Data & Emerging Topics. ISBN: 978-950-34-2016-4 Pp. 25-28. Facultad de Informática. UNLP. Junio 2021.

[10] Estrebou, C., Saavedra, M. D., Adra, F., Fleming, M. TinyML for Small Microcontrollers. X Jornadas de Cloud Computing, Big Data & Emerging Topics. ISBN 978-950-34-2126-0. Pp 42-46. Facultad de Informática. UNLP. Mayo 2022.

[11] Estrebou C., Feming M., Saavedra M. D., Adra F. A Neural Network Framework for Small Microcontrollers. XXVII Congreso Argentino de Ciencias de la Computación. ISBN: 978 -987-633-574-4. Pp. 51-60. Univ. Nacional de Salta. Octubre 2021.

[12] Estrebou C., Feming M., Saavedra M. D., Adra F., De Giusti, A. E. Lightweight Convolutional Neural Networks Framework for Really Small TinyML Devices. Second International Conference on Smart Technologies, Systems and Applications. SmartTech-IC 2021. Quito, Ecuador. Diciembre 2021.