Procesamiento inteligente de la información. Aplicaciones en bioinformática, trayectorias vehiculares, mantenimiento preventivo industrial y sistemas embebidos

W. Hasperué[1], C. Estrebou[1], G. Camele[1,3], P. López[2], M. Peña[2], G. Reyes Zambrano[4]
L. Lanzarini[1], A. Fernandez Bariviera[5], M. Cerrada[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 Físicas y Matemáticas, Universidad de Guayaquil, Guayaquil, Ecuador
[5] Dpto. de Economía, Universitat Rovira i Virgili, Reus, España
[6] GIDTEC, Universidad Politénica Salesiana, Cuenca, Ecuador
*** Centro asociado de la Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC)**

{whasperue, cesarest, gcamele, pdlopez, laural}@lidi.info.unlp.edu.ar, mario.pena@info.unlp.edu.ar, gary.reyesz@ug.edu.ec, martinezvictor@hotmail.com, aurelio.fernandez@urv.net, mcerrada@ups.edu.ec

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 dos temas: el estudio y desarrollo de técnicas de selección de características y el diseño de librerías que faciliten el procesamiento masivo de datos. En lo referido a selección de características, 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 nuevas técnicas que permitan obtener un subconjunto óptimo de atributos.

Por otro lado, se está desarrollando una librería con el objetivo de facilitar el manejo de bases de datos de progenie en entornos de Big Data. El objetivo de esta librería es brindar una API simple para que cualquier investigador con pocos conocimientos de programación pueda utilizarla de manera simple.

En cuanto a las investigaciones relacionadas con el análisis de flujos de datos se centran en la construcción de modelos que faciliten la interpretación de los patrones obtenidos y la posterior extracción del conocimiento. En particular el énfasis está puesto en la resolución de dos problemas de sumo interés en distintas áreas: el mantenimiento de máquinas industriales basado en su condición de funcionamiento y el análisis de trayectorias GPS a fin de identificar las características del flujo vehicular en un período de tiempo dado.

Palabras clave: Big Data, Minería de Datos, Minería de procesos, Análisis de flujos de datos, Reducción de características.

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

En el área de la minería de datos y su aplicación con técnicas de machine learning, los algoritmos de selección de características juegan un papel muy importante. El objetivo de estos algoritmos es el de reducir las entradas a un tamaño apropiado para su procesamiento y análisis. La selección de características en una base de datos implica la elección de ciertos atributos, tal que, con ese subconjunto de atributos, las "propiedades naturales" de los datos pertenecientes a un dataset 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 signatura* (biomarcador diagnóstico, pronóstico o predictivo de una patología en estudio [1]).

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 [2] se llevó a cabo una comparación entre cuatro algoritmos de clasificación implementados 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.

Por otro lado, en un trabajo conjunto con el Centro de Altos Estudios en Tecnología Informática (CAETI) de la Facultad de Tecnología Informática (UAI) y el Centro de Investigaciones Inmunológicas Básicas y Aplicadas (CINIBA) de la Facultad de Ciencias Médicas (UNLP), se desarrollaron plataformas que permiten la consulta de bases de datos con información génica, en particular con microARNs, moléculas de gran relevancia en estudios multi-ómicos que tienen por objetivo descifrar los mecanismos de desregulación de expresión génica en enfermedades complejas, multigénicas y multifactoriales, como es el caso de la mayoría de los tipos de cáncer.

En [3] se presenta Modulector, una plataforma que integra la información de las bases de datos públicas más relevantes de microARNs y las disponibiliza mediante una plataforma web simplificando el acceso a información actualizada. Esta plataforma puede ser accedida vía web a través de Multiomix [4], desde otras herramientas o plataformas que la integren, o incluso realizar un despliegue en un entorno privado, mediante el uso de las distintas APIs que provee.

Procesamiento de progenie en Big Data

Cuando en una base de datos, la información tiene naturaleza de un árbol n-ario, su análisis se dificulta, ya que para cualquier consulta deben realizarse varias operaciones *Join*. En particular, una operación por cada nivel del árbol que se desea explorar. El uso de múltiples *joins* incrementa la complejidad de la consulta a realizar, dificultando el análisis a quienes no tienen experiencia en programación.

En relación a esta línea, uno de los desarrollos que se están llevando a cabo es la implementación de una librería que permita el fácil tratamiento de los datos cuando estos están organizados y relacionados como un árbol, como por ejemplo, los datos de progenie.

En [6] presentamos TreeSpark, una herramienta basada en Spark, que permite realizar un

análisis de progenie. TreeSpark posee una API que permite un acceso simple a la información de los individuos y a sus relaciones tanto de progenie como de progenitores.

1.2. ANALISIS DE FLUJOS DE DATOS

Identificación de patrones de velocidad

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 toma importancia para los investigadores. Los Sistemas Inteligentes de transporte realizan el monitoreo y control vehicular recolectando trayectorias GPS, información que brinda en tiempo real la ubicación geográfica de los vehículos. Por medio de técnicas de agrupamiento es posible 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 características del flujo vehicular en un período de tiempo. Para ello utiliza una adaptación original de una técnica de agrupamiento dinámico y un mapa interactivo para analizar el flujo del tráfico en sectores específicos facilitando la identificación de zonas de posibles atascos.

Los resultados obtenidos sobre un conjunto de datos de la ciudad de Guayaquil-Ecuador son satisfactorios y representan claramente la velocidad de desplazamiento de los vehículos identificando de manera automática los rangos más representativos para cada instante de tiempo. Como resultado del agrupamiento se obtuvieron diferentes grupos que dinámicamente cambian de una evolución a otra, identificando velocidades comunes en diferentes instantes de tiempo lo que permite tomar decisiones con respecto al tráfico de la ciudad. Para más detalle consultar [5] [7].

Deriva de concepto (concept drift)

El mantenimiento preventivo industrial se ha convertido en un tema importante y ha captado la atención de los investigadores en los últimos años. Las señales recogidas, como las vibraciones, la corriente o las emisiones acústicas, son series temporales que suelen estar asociadas a una gran cantidad de flujos de datos. El análisis de estos flujos de datos se enfrenta a varios retos debido a que son potencialmente ilimitados, generados a ritmos muy rápidos, con limita-ciones de tiempo de procesamiento y memoria y, además, pueden evolucionar con el tiempo.

La extracción de conocimiento a partir de estas grandes cantidades de datos asume que los datos son estáticos y generados por una distribución de probabilidad desconocida pero estacionaria. Sin embargo, en las aplicaciones de la vida real, la distribución de los flujos de datos cambia con el tiempo, ya que los entornos subyacentes son naturalmente dinámicos y se modifican constantemente. Este fenómeno se conoce como deriva de conceptos (concept drift). Los modelos que describen estos fenómenos deben actualizarse constantemente para reflejar el concepto actual. En el campo del mantenimiento preventivo, la deriva de concepto puede asociarse con los cambios en el estado de la máquina o en las condiciones de trabajo.

Recientemente, en el marco de una tesis de Doctorado en Cs. Informáticas de la UNLP se ha propuesto una metodología para detectar la evolución del concepto mediante el uso de medidas de divergencia para cuantificar las desviaciones. Fue utilizada en la clasificación de la severidad de los fallos de una caja de cambios helicoidal sometida a diversas cargas y velocidades. Los resultados muestran que el algoritmo propuesto detecta adecuadamente la evolución de concepto, que es un indicador de degradación de la máquina. La divergencia Kullback-Leibler ha resultado ser la métrica correcta para determinar la discrepancia entre las diferentes distribuciones de probabilidad empíricas dadas por las condiciones de fallo de la máquina.

Por otro lado, la selección de características en los flujos de datos sigue siendo un tema de investigación abierto. No hay muchos trabajos en torno al análisis de cómo evoluciona la importancia de las características con respecto al tiempo, por lo que se continuará trabajando

en esta importante parte de la metodología propuesta. Para más detalle consultar [8].

1.3. TINYML

TinyML es un campo de estudio que abarca al aprendizaje Automático y a los Sistemas Embebidos en la exploración de alternativas para adaptar modelos tradicionales para ejecutarlos en dispositivos con restricciones de memoria. En esta línea se inició un proyecto dedicado a la investigación de técnicas de Aprendizaje Automático aplicadas a microcontroladores (MCU) [10].

En el marco de este proyecto se construyeron modelos de redes neuronales convolucionales para analizar el funcionamiento de los frameworks Tensorflow Lite, Microtensor y Eloquent TinyML en dispositivos con diferentes características de hardware. Como resultado de este análisis se encontró que varios de estos frameworks requieren arquitecturas ARM, datos de 32 bits o soporte para instrucciones DSP o SIMD, dejando de lado a una cantidad importante de MCUs. Además, según estimaciones realizadas, se encontró que pequeños modelos requieren una cantidad de memoria mayor a la necesaria. En consecuencia, se decidió iniciar el desarrollo EmbedIA, un framework de código abierto compatible con C/C++ para realizar inferencia sobre modelos de redes neuronales convolucionales. Además de las funciones de inferencia se implementaron varias optimizaciones como aritmética de punto fijo para 32, 16 y 8 bits y estrategias de alocación de memoria para evitar la fragmentación. Se midió el desempeño de EmbedIA con los frameworks investigados en cinco microcontroladores diferentes [11][12]. En particular la implementación de punto fijo de 16 bits alcanzó, en promedio, una mejora de 5 a 10 veces en el tiempo de inferencia, de unas 3 veces en los requerimientos de memoria de datos y de 3 a 7 veces en los requerimientos de memoria de programa.

Actualmente, se está investigando la adaptación de otras técnicas de aprendizaje automático y posibles optimizaciones en la implementación de redes neuronales convolucionales para reducir el impacto de la memoria requerida en microcontroladores.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Medición de performance de algoritmos de machine learning en entornos distribuidos.
- Desarrollo de una plataforma pública de acceso web para ejecutar análisis de correlación entre grandes bases de datos de genes y moduladores de expresión.
- Desarrollo de una herramienta para el análisis de progenie, basada en Spark.
- Estudio de métricas para comparar distribuciones de probabilidad.
- Medidas de divergencia como herramienta para detectar la deriva de concepto.
- Preprocesamiento de trayectorias vehiculares. Técnicas de segmentación.
- Agrupamiento dinámico de flujos de datos.
- Estudio de técnicas de compresión de modelos para microcontroladores.
- Análisis de bibliotecas y frameworks de aprendizaje automático para microcontroladores.
- Desarrollo de un Framework que transforma modelos desarrollados en Tensorflow/Keras a código C/C++ para ejecutarlos en microcontroladores.

3. RESULTADOS OBTENIDOS

- Implementación de una plataforma web que integra información de bases de datos genómicas, con bases de datos de micro-ARNs, para simplificar el acceso a las distintas dimensiones de información.
- Implementación de una plataforma interactiva en la nube que permite identificar eventos (epi)genéticos asociados a la modulación transcripcional de genes relacionados con el cáncer a través del análisis de datos multiómicos.

- Implementación de una librería que facilita el desarrollo de pequeñas aplicaciones en Spark, para el tratamiento de bases de datos con información de progenie.
- 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 para detectar la evolución del concepto mediante el uso de medidas de divergencia para cuantificar las desviaciones.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo de la línea de I/D aquí presentada está formado por: 2 profesores doctores con dedicación exclusiva, 3 tesistas de Doctorado en Cs. Informáticas (1 con beca de postgrado de la UNLP), 1 tesista de grado y 2 profesores extranjeros.

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 3 prácticas profesionales supervisadas.

Actualmente se están desarrollando 3 tesis de doctorado, 3 tesis de maestría, 4 tesis de especialista y 4 tesinas de grado de Licenciatura. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] Abba, M. C.; Lacunza, E.; Butti, M. 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.; 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.

[3] Marraco, A. D., Camele, G., Hasperué, W., Menazzi, S., Abba, M., & Butti, M. (2021). Modulector: una plataforma como servicio para el acceso a bases de datos de micro ARNs. Innovación y Desarrollo Tecnológico y Social, 3(1), 89-114.

[4] Camele, G., Menazzi, S., Chanfreau, H., Marraco, A., Hasperué, W., Butti, M. D., & Abba, M. C. (2022). Multiomix: a cloud-based platform to infer cancer genomic and epigenomic events associated with gene expression modulation. Bioinformatics, 38(3), 866-868.

[5] Reyes G., Lanzarini L., Fernández Bariviera A., Hasperué W. A proposal for a pivot-based vehicle trajectory clustering method. Transportation Research Record, Journal of the Transportation Research Board. ISSN: 0361-1981.

[6] López, P. D.; Hasperué, W.; Quiroga, F.; Ronchetti, F. TreeSpark: A Distributed Tool for Progeny Analysis based on Spark. CACIC. Salta. 2021.

[7] Reyes G., Lanzarini L., Estrebou C. (2021). Vehicular Flow Analysis Using Clusters. XVIII Workshop Base de Datos y Minería de Datos (WBDMD) en CACIC. Univ. Nacional de Salta. Octubre 2021.

[8] Peña M., Lanzarini L., Cerrada M., Sánchez RV. (2021) Data-Driven Gearbox Fault Severity Diagnosis Based on Concept Drift. IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), 1-6

[10] 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.

[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.

Waldo Hasperué: 0000-0002-9950-1563
César Estrebou: 0000-0001-5926-8827
Genaro Camele: 0000-0001-6979-9103
Mario Peña: 0000-0002-3986-7707
Gary Reyes: 0000-0002-3711-1906
Laura Lanzarini: 0000-0001-7027-7564
Aurelio Fernandez Bariviera: 0000-0003-1014-1010
Mariela Cerrada: 0000-0003-4379-8836