Procesamiento inteligente de grandes volúmenes de información y de flujos de datos

W. Hasperué[1], C. Estrebou[1], G. Camele[1,3], P. López[2], P. Jimbo Santana[4], G. Reyes Zambrano[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 Administrativas, Universidad Central del Ecuador, Quito, 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, pdlopez, laural}@lidi.info.unlp.edu.ar
prjimbo@uce.edu.ec, gary.reyesz@ug.edu.ec, 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 (2018-2021).

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 reducción de características y el diseño de estrategias que faciliten el procesamiento masivo de datos a usuarios no informáticos. En lo referido a reducción de características, dado que se está trabajando con bases de datos genómicas, el foco está puesto en las estrategias de selección de atributos. El análisis a realizar sobre estos datos tiene por objetivo identificar grupos de genes cuyos patrones de expresión se encuentren asociados a fenotipos específicos. Por otro lado, se está desarrollando una librería con el objetivo de facilitar el manejo de bases de datos en contextos Big Data. Esto tendrá un impacto directo en el trabajo conjunto que se viene desarrollando junto con la Facultad de Ciencias Veterinarias de la UNLP en relación al análisis de datos de progenie de distintas especies animales.

En cuanto a las investigaciones relacionadas con la Minería 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: las técnicas de agrupamiento aplicables a flujos de datos y la generación de reglas de clasificación.

Palabras clave: Big Data, Minería de Datos, Reducción de características, Flujos de datos, técnicas de optimización, Redes Neuronales.

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 medidos en la solución de

problemas pertenecientes a distintas áreas. A continuación, se detallan las investigaciones realizadas durante el último año.

1.1. BIG DATA

Reducción de características

En el área de la minería de datos y su aplicación con técnicas de machine learning, los algoritmos de reducción de características juegan un papel muy importante. El objetivo de esos algoritmos es el de reducir las entradas a un tamaño apropiado para su procesamiento y análisis. La reducción de características en una base de datos implica la elección de ciertos atributos y/o creación de nuevos atributos en función de los existentes, tal que, con ese subconjunto de atributos, las "propiedades naturales" de los datos pertenecientes a un dataset no sean alteradas (o lo sean con una pequeña pérdida de información).

Cuando el volumen de información a procesar crece, la ejecución de los algoritmos de selección de atributos convencionales incrementa notablemente su tiempo de procesamiento. Si bien puede considerarse la separación o el análisis independiente de cada atributo, muchas veces resulta útil poder analizar correlaciones entre dos o más variables. Por ello contar con algoritmos que puedan realizar este tipo de análisis en grandes volúmenes de datos resulta de mucho interés.

Actualmente, en el III LIDI se están realizando tareas de investigación con bases de datos genómicas. 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 esta área se destaca el análisis de perfiles de expresión génica que tienen como objetivo principal la identificación de un grupo de genes, cuyo patrón de expresión se encuentre asociado a un fenotipo en particular: concepto conocido como *gene signature*.

Un objetivo particular de los *gene signatures* es su utilidad como biomarcador diagnóstico, pronóstico o predictivo de una patología en estudio [1]. Los biomarcadores con valor pronóstico permiten una mejor estratificación de pacientes. En la actualidad la tarea del descubrimiento de nuevos *gene signatures* es realizada mayormente de manera manual por expertos. Es por ello que se están desarrollando estrategias de soporte automático que permita seleccionar aquellos genes que resulten más representativos y por ende ser interpretados como un posible *biomarcador con poder pronóstico*. Las estrategias que se están desarrollando están basadas en algoritmos de selección de características.

En relación a las técnicas de selección de atributos, y en entornos de procesamiento de flujos de datos, se están desarrollando estrategias basadas en técnicas de Aprendizaje Automático que permitan la selección de los atributos más relevantes, brindando resultados en tiempos de respuestas cortos los cuales se adaptan de manera dinámica a la llegada de nuevos datos.

Acceso al procesamiento de Big Data

Uno de los frameworks más utilizados en entornos Big Data es Spark. Este framework brinda facilidad en el análisis de grandes bases de datos ya que ofrece una capa de alto nivel para el procesamiento distribuido y paralelo de los datos. El análisis de los datos se puede realizar a través de su propia API la cual trabaja con sus bases de datos distribuidas internas (RDDs). Sobre esta API han aparecido luego más abstracciones que permiten el análisis usando DataFrames e incluso el lenguaje de consultas SQL.

Cuando los datos a analizar pueden ser almacenados en bases de datos equivalentes a tablas relacionales de cualquier motor de bases de datos, entonces no hay mayores problemas para su uso, pero cuando la información está organizada en forma de árbol, el análisis se vuelve un tanto complicado, ya que para cualquier consulta deben realizarse varias operaciones *Join*. En

particular, se requerirá una operación por cada nivel del árbol que se desea explorar. El uso de múltiples *joins* incrementa la complejidad de la consulta a realizar, aún incluso en SQL y dificulta el análisis posterior de los resultados

En relación a esta línea, uno de los desarrollos que se están llevando a cabo actualmente consiste en 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 de cualquier especie animal. Esta librería puede ser incluida en cualquier desarrollo realizado en Spark y su objetivo es permitir a los investigadores de distintas áreas realizar análisis entre individuos de distintas generaciones de una manera sencilla y amena [2].

1.2. MINERÍA DE DATOS

Agrupamiento de flujos de datos

El avance tecnológico ha dado lugar a la generación de datos masivos en tiempo real en áreas muy diversas: consultas en la web, video vigilancia, flujos en redes sociales, redes de sensores, análisis de mercado de valores, supervisión de tráfico, etc.

La minería de flujos de datos posee múltiples aplicaciones en la vida real y ha cobrado fuerza debido a su capacidad de extraer patrones ocultos en dichos flujos. Los algoritmos involucrados deben responder a diversos desafíos para satisfacer restricciones tales como: memoria limitada, paso único de los datos, respuesta en tiempo real, adaptación y clasificación de la deriva de concepto (concept drift) y manejo de datos multidimensionales.

Los desarrollos tecnológicos han cambiado la forma en que la gente almacena, comunica y procesa los datos. Para procesarlos, es preciso utilizar algoritmos capaces de generar, de manera incremental, modelos que incorporen la nueva información de los datos más recientes mientras eliminan los efectos de los datos antiguos.

Por lo tanto, el procesamiento de flujos de datos presenta varios retos. El hecho de reentrenar un modelo con nuevos ejemplos es ineficaz e inadecuado en función del volumen y la velocidad con que se generan.

Un área clave de la minería de flujos de datos es el uso de técnicas de agrupamiento. Dada la necesidad de mantener un modelo dinámico, las estrategias partitivas basadas en centroides requieren de una estructura adicional para conformar cada grupo. En cuanto a la representación interna, deben utilizarse estructuras que resuman el flujo preservando el significado de los datos originales sin la necesidad de guardarlos. Esto puede observarse en [3] donde, además de utilizar una estructura de datos particular, emplea un factor de olvido para controlar el dinamismo del modelo y una estructura en forma de árbol para reunir las distintas partes que conforman un mismo grupo.

En esta dirección se han realizado diferentes investigaciones en el framework Spark [4] [5] [6]. Actualmente, se está trabajando en una adaptación del algoritmo de procesamiento de trayectorias definido en [7], incorporándole una estrategia de procesamiento incremental, capaz de operar con flujos de datos, basada en el tiempo de registro de cada ubicación dentro de la trayectoria original.

Extracción de Reglas de Clasificación

Cuando se busca construir un modelo predictivo para resolver un problema de clasificación a partir de datos estructurados, las reglas de clasificación resultan sumamente atractivas por su capacidad explicativa. La literatura muestra distintas alternativas de generación basadas en árboles de clasificación construidos total o parcialmente. En ambos casos, el antecedente de cada regla se forma a partir de la conjunción de ítems conformados por el valor que aparece en el nodo y el valor indicado en la rama correspondiente del árbol. Por su parte el consecuente estará dato por la clase mayoritaria de ejemplos presentes en la hoja [8] [9]. Otra forma de construcción consiste en generar las reglas a través de un proceso iterativo que, en cada paso, analiza

cuál es el valor de clase a utilizar en el consecuente, construye de manera incremental el antecedente y retira del conjunto de ejemplos de entrada los correctamente cubiertos. Este proceso de construcción incremental da lugar a una lista de reglas de clasificación que debe ser aplicado en el orden en que fueron generadas [10].

A diferencia de estas alternativas, en el III-LIDI se trabaja en la generación de reglas de clasificación utilizando técnicas de optimización por su habilidad para considerar la conformación del antecedente completo en lugar de hacerlo en pasos sucesivos. Dado que se utilizan estrategias poblacionales computacionalmente costosas, se ha buscado reducir el tiempo necesario para realizar el proceso de búsqueda utilizando alguna técnica de clustering. En esta dirección se han efectuado pruebas con redes neuronales competitivas supervisadas y no supervisadas tanto sobre datos de repositorio como reales [11]. Como resultado de estas investigaciones en julio de 2020 se ha defendido una tesis de doctorado en cotutela entre la UNLP y la URV (España) donde se detallan los resultados obtenidos al aplicar estas estrategias sobre datos reales correspondientes a distintas bases de datos de instituciones financieras en el Ecuador [12].

Además, con el objetivo de ayudar al usuario en el momento de ponderar la credibilidad de la regla, se ha incorporado un factor de confianza que le permite ponderar el riesgo de utilizar dicha regla al momento de tener que tomar una decisión. Esto es algo interesante ya que pueden existir condiciones adicionales, ajenas a los atributos considerados durante el proceso de construcción, que justifiquen este tipo de acciones. Los resultados obtenidos con esta nueva variante pueden consultarse en [13].

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Diseño e implementación de un algoritmo de selección de atributos capaz de operar en batch y en streaming.
- Estudio de algoritmos de selección de atributos para la detección de *gene signatures*.
- Estudio e implementación de técnicas inte-ligentes en el framework Spark Streaming.
- Análisis de bases de datos con información de progenie en entornos Big Data.
- Modelización de trayectorias espacio-temporales con capacidad para establecer características comunes y detectar situaciones anómalas.
- Estudio de técnicas de clustering dinámico basadas en densidad para modelar trayectorias GPS e identificar sectores de posible congestionamiento.
- Estudio de técnicas de optimización poblacionales y redes neuronales artificiales para la obtención de reglas difusas de tipo IF-THEN.

3. RESULTADOS OBTENIDOS

- 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 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 una técnica de agrupamiento dinámico para flujos de datos basada en densidad.
- Diseño e implementación de un nuevo método de agrupamiento de trayectorias GPS aplicable a la predicción de congestiones vehiculares.
- Desarrollo de un método de obtención de reglas de clasificación difusas que incorpora un factor de confianza que ayuda al usuario a ponderar el riesgo de su uso.

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), 2 tesistas de grado y 1 profesor extranjero.

Dentro de los temas involucrados en esta línea de investigación, en los últimos 3 años se han finalizado 2 tesis de doctorado, 1 tesis de especialista y 5 tesinas de grado de Licenciatura.

Actualmente se están desarrollando 3 tesis de doctorado, 2 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.; Aldaz, C. 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] López, P. D.; Hasperué, W.; Rearte, R.; De La Sota, R. L. Herramienta informática para el análisis de Progenie. nnovación y Desarrollo Tecnológico y Social.La Plata: Universidad Nacional de La Plata. Vol.2 n°1. pp 25-54. ISSN 2683-8559. 2020.

[3] Barbosa, N.; Travé-Massuyès, L.; Grisales-Palacio, V. DyClee: Dynamic clustering for tracking evolving environments, Pattern Recognition. Volume 94, Pages 162-186, ISSN 0031-3203. 2019.

[4] Molina, R.; Hasperué, W. D3CAS: un Algoritmo de Clustering para el Procesamiento de Flujos de Datos en Spark. XXIV CACIC. UNS. Tandil. 2018

[5] Molina, R.; Hasperué, W.; Villa Monte, A.; D3CAS: Distributed Clustering Algorithm Applied to Short-Text Stream Processing. Communications in Computer and Information Science. : Springer. p211 - 220. ISBN 978-3-030-20786-1. 2019

[6] Reyes-Zambrano, G.; Lanzarini, L.; Hasperué, W.; Fernández-Bariviera, A. GPS trajectory clustering method for decision making on intelligent transportation systems. Journal of Intelligent & Fuzzy Systems, vol. Pre-press, pp. 1-6. ISSN 1064-1246. 2020.

[7] Liu, L. X.; Song, J. T.; Guan, B.; Wu, Z. X.; He, K. J. Tradbscan: a algorithm of clustering trajectories," in Applied Mechanics and Materials, vol. 121. Trans Tech Publ. pp. 4875–4879. 2012.

[8] Frank, E.; Witten, I. H. Generating accurate rule sets without global optimization. In Proceedings of the Fifteenth International Conference on Machine Learning, pp. 144–151, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 1998.

[9] Quinlan, J. R. C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann. 1992.

[10] Clark, P; Niblett, T. The CN2 induction algorithm. Machine learning journal. 3:4, pp. 261-283. 1989.

[11] Jimbo, P.; Lanzarini, L.; Fernandez-Bariviera, A. Fuzzy Classification Rules with FRvarPSO Using Various Methods for Obtaining Fuzzy Sets. Journal of Avdances in Information Technology, 11(4), 233-240. ISSN 1798-2340. 2020.

[12] Jimbo, P. Obtención de reglas de clasificación difusas utilizando técnicas de optimización. Caso de estudio Riesgo Crediticio. Tesis de doctorado en Ciencias Informáticas realizada en la UNLP en cotutela con la Universitat Rovira i Virgili (URV) (España). http://sedici.unlp.edu.ar/ handle/10915/101163. 2020.

[13] Jimbo, P.; Lanzarini, L. Fernandez-Bariviera, A. FRvarPSO as an alternative

to measure credit risk in financial institutions. Serie Springer Advances in Intelligent Systems and Computing. Sep. 2021.