Sistemas Inteligentes. Aplicaciones en Minería de Datos y Big Data

L. Lanzarini[1], W. Hasperué[1], C. Estrebou[1], A. Villa Monte[1,3], P. Jimbo Santana[4], G. Reyes Zambrano[5], G. Camele[1,3], P. López[2], J. Corvi[2], A. Fernandez Bariviera[6], J. A. Olivas[7]

[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
[7] Dpto. Tecnología y Sistemas de la Información, Universidad de Castilla-La Mancha, Ciudad Real, España
*** Centro asociado de la Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC)**

{laural, whasperue, cesarest, avillamonte, gcamele, pdlopez}@lidi.info.unlp.edu.ar
prjimbo@uce.edu.ec, gary.reyesz@ug.edu.ec, julieta.corvi@gmail.com , aurelio.fernandez@urv.net, joseangel.olivas@uclm.es

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) y del proyecto PITAP-BA "Computación de Alto Desempeño, Minería de Datos y Aplicaciones de Interés Social en la Provincia de Bs.As." evaluado y subsidiado por la Comisión de Investigaciones Científicas de la Provincia de Bs.As. (2017-2019).

RESUMEN

Esta línea de investigación se centra en el estudio y desarrollo de Sistemas Inteligentes para la resolución de problemas de Minería de Datos y Big Data 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.

En el área de la Minería de Datos se está trabajando, por un lado, en la construcción de conjuntos de reglas de clasificación difusas que faciliten y permitan justificar la toma de decisiones y, por otro lado, en el diseño de técnicas de agrupamiento para flujos de datos con aplicación al análisis de trayectorias vehiculares para predecir congestión de tránsito.

Con respecto al área de Big Data se está trabajando en el diseño y desarrollo de algoritmos de selección de características en grandes bases de datos de muchas columnas. Las implementaciones que se están llevando a cabo serán utilizadas en problemas de genómica molecular con el objetivo de determinar *gene signatures*. En esta misma línea se está trabajando sobre algoritmos de selección de características para el tratamiento de flujos de datos.

Por otro lado y como transferencia tecnológica concreta, se efectuó un análisis sobre la producción de leche en ganado bovino a partir de la base de datos de la Asociación de la Región Pampeana de Entidades de Control Lechero (ARPECOL).

En el área de la Minería de Textos se han desarrollado estrategias para resumir documentos a través de la extracción de los párrafos más representativos utilizando métricas de selección y técnicas de optimización.

Palabras clave: Minería de Datos, Minería de Textos, Big Data, Redes Neuronales, Resúmenes Extractivos, Stream Processing. Selección 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 medidos en la solución de problemas pertenecientes a distintas áreas. A continuación se detallan los resultados obtenidos durante el último año.

1.1. MINERÍA DE DATOS

Extracción de Reglas de Clasificación

Las técnicas de minería de datos son de gran interés para las organizaciones, ya que facilitan la adopción de decisiones tácticas y estratégicas, generando una ventaja competitiva. En el caso especial de las organizaciones que conceden créditos, es importante definir claramente los criterios de rechazo/aprobación. En este sentido, las reglas de clasificación son un instrumento adecuado, siempre que el conjunto de reglas sea fácil de comprender y de aplicar y además posea una tasa de acierto razonable. La simplicidad de la regla se logra generando antecedentes cortos, formados por pocas condiciones.

En el III LIDI se analizan y diseñan soluciones basadas en técnicas de optimización para construir conjuntos de reglas de clasificación con las características antes mencionadas. Los modelos construidos son capaces de operar con atributos nominales y numéricos y emplean redes neuronales competitivas como punto de partida para reducir el tiempo de búsqueda. Además, para facilitar la comprensión del modelo, se incorpora la lógica difusa en la construcción del antecedente. En esta dirección, el énfasis está puesto en la importancia del uso de conjuntos difusos al expresar las condiciones que implican atributos numéricos. Por su intermedio, no sólo se facilita la comprensión y la aplicación de las reglas sino que también se consigue una precisión significativamente mayor respecto de las condiciones *crisp* o no difusas.

En el caso particular de las reglas aplicables a riesgo crediticio se utiliza la información del prestatario y el entorno macroeconómico en el momento de conceder el préstamo.

Los resultados obtenidos de aplicar las técnicas propuestas a distintas bases de datos de instituciones financieras en el Ecuador han sido publicados en [1-3]. Se trata de instituciones que se especializan en la colocación masiva de créditos así como en créditos de consumo y líneas de crédito empresarial. En todos los casos se ha observado que la incorporación de la lógica difusa genera conjuntos de reglas con mayor precisión.

Agrupamiento de flujos de datos

Las técnicas de agrupamiento han sido ampliamente utilizadas a la hora de construir modelos capaces de resolver tareas descriptivas, es decir, modelos que buscan establecer similitudes y diferencias entre las distintas situaciones que pueden ocurrir a fin de identificar cuáles son las características más importantes que deben ser tenidas en cuenta a la hora de explicar el comportamiento de un proceso.

Si se analiza la evolución de este tipo de técnicas puede verse que, en un principio, los modelos han operado buscando establecer similitudes en un conjunto de datos conocido a priori. Utilizando el enfoque tradicional, dicho conjunto puede ser analizado todas las veces que sea necesario hasta determinar las características más representativas del problema. Esta forma de trabajo, si bien sigue siendo válida en numerosas situaciones, presenta problemas cuando deben procesarse volúmenes de información sumamente grandes. En estos casos puede ocurrir que no se disponga de espacio suficiente para su almacenamiento o incluso que no resulte conveniente guardarlos por la velocidad con la que la información se actualiza.

Por lo antes expuesto, se está trabajando con técnicas de agrupamiento para flujos de datos capaces de establecer similitudes analizándolos una única vez [4]. Por lo tanto,

la información no será almacenada sino que será procesada inmediatamente al momento de arribar. Procesar un flujo de datos implica realizar modificaciones sobre el modelo de manera automática porque, como se dijo previamente, ya no se tendrán los datos originales para efectuar modificaciones sino que deben ir efectuándose registros más genéricos a medida que se ingresa la información. Este tipo de entrada se ajusta perfectamente al procesamiento de información temporal.

En esta línea de investigación interesa especialmente considerar la modalidad de recolección de los datos, la cantidad de veces que los mismos pueden ser utilizados y los parámetros de configuración que debe indicar el usuario.

Los resultados obtenidos hasta el momento se están aplicando en el análisis de trayectorias GPS. Los avances tecnológicos facilitan el registro de información sobre las trayectorias GPS de los vehículos en las carreteras públicas. El análisis inteligente de estos datos permite identificar patrones extremadamente útiles para la toma de decisiones en situaciones relacionadas con el urbanismo, el tráfico y la congestión de las carreteras, entre otros. En [5] se presentó un nuevo método de agrupamiento de trayectorias GPS que utiliza información angular para segmentar los recorridos y una función de similitud guiada por un pivote. El proceso de adaptación inicia distribuyendo los centroides de manera uniforme en la región a analizar formando un reticulado. Los resultados obtenidos luego de aplicar el método propuesto sobre una base de datos de trayectorias reales fueron satisfactorios y muestran una mejoría significativa en comparación con los métodos publicados en la bibliografía.

1.2. BIG DATA

Aplicaciones en Big Data

En esta línea se trabaja sobre el procesamiento, en streaming y en batch, de grandes volúmenes de datos. Para el procesamiento en streaming se están desarrollando estrategias basadas en técnicas de Aprendizaje Automático que permitan la selección de los atributos más relevantes de un flujo de datos, brindando resultados en tiempos de respuestas cortos los cuales se adaptan de manera dinámica a la llegada de nuevos datos [6].

Estas técnicas dinámicas se están implementando en el framework Spark Streaming, adecuado para procesamiento paralelo, distribuido y online.

Los algoritmos de selección de atributos sobre flujos de datos representan un desafío interesante ya que se deben tratar con cuidado el problema de *concept drift* para no perder información relevante.

Por otro lado y como una transferencia tecnológica concreta se ha trabajado en el tratamiento de la información proveniente de ARPECOL, una asociación que nuclea entidades de control lechero de la provincia de Buenos Aires. Las entidades de control lechero son organizaciones que brindan el servicio de medición de la producción de leche individual a los productores. Estas entidades toman muestras de la leche de las vacas para realizar análisis de laboratorio de la calidad de la leche producida (porcentaje de grasa, de proteínas, de solidos totales, conteo de células somáticas). En esta línea de investigación se colaboró con un proyecto del INIRA de la Facultad de Veterinaria de la UNLP que tiene por objetivo determinar factores genéticos para la identificación de las principales enfermedades reproductivas, mastitis y cojeras que afectan la lactancia de las vacas de tambo.

1.3. MINERIA DE TEXTOS

Hoy en día, la información que nos rodea lo hace en su gran mayoría en forma de texto. El volumen de información no estructurada crece continuamente de tal manera que resulta necesario separar por medio de técnicas de procesamiento de texto lo esencial de lo que no lo es. Los instrumentos de resumen

automático de textos tienen un gran impacto en muchos campos, como la medicina, el derecho y la investigación científica en general. A medida que aumenta la sobrecarga de información, los resúmenes automáticos permiten manejar el creciente volumen de documentos, generalmente asignando pesos a las frases extraídas en función de su importancia en el resumen previsto. La obtención del contenido principal de un documento determinado en menos tiempo del que llevaría hacerlo manualmente sigue siendo una cuestión de interés. En [7] se presentó un nuevo método capaz de generar automáticamente resúmenes extractivos de documentos mediante la ponderación adecuada de las características de puntuación de las frases utilizando optimización por cúmulo de partículas. El método propuesto es capaz de identificar las características que más se aproximan al criterio utilizado por el individuo al hacer el resumen. Para ello, combina una representación binaria y otra continua, utilizando una variación original de la técnica desarrollada por los autores de este documento. Las experimentaciones realizadas confirman que el uso de la información etiquetada por el usuario en el conjunto de entrenamiento ayuda a encontrar mejores métricas y pesos. Los resultados empíricos difundidos en [7] reflejan una mayor precisión en comparación con los métodos anteriores utilizados en este campo.

Por otro lado, en [8] se efectuó un estudio similar a partir de un modelo basado en una red neuronal que, a partir de la puntuación calculada con distintas métricas, es capaz de predecir la importancia de la sentencia dentro del documento. Una vez entrenada la red, este criterio puede aplicarse a otros documentos para obtener, como resultado, un resumen similar al que el usuario habría hecho manualmente.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Estudio de técnicas de optimización poblacionales y redes neuronales artificiales para la obtención de reglas difusas de tipo IF-THEN.
- Modelización de trayectorias espacio-temporales con capacidad para establecer características comunes y detectar situaciones anómalas.
- Métodos estructurados y no estructurados aplicables a la representación de documentos.
- Representación de documentos de texto utilizando métricas.
- Obtención de resúmenes automáticos de texto.
- Implementación de técnicas inteligentes en el framework Spark Streaming
- Implementación de un algoritmo de selección de atributos en batch y en streaming.
- Uso de algoritmos de selección de atributos para la detección de gene signatures.
- Análisis de la base de datos de ARPECOL para la identificación de características genéticas que mejoren la producción de leche de las vacas de tambo.

3. RESULTADOS OBTENIDOS

- Desarrollo de un método de obtención de reglas de clasificación difusas con énfasis en la reducción de la complejidad del modelo aplicable a riesgo crediticio.
- Diseño e implementación de técnicas de agrupamiento para flujos de datos considerando especialmente los parámetros de configuración que debe indicar el usuario.
- Diseño e implementación de un nuevo método de agrupamiento de trayectorias GPS aplicable a la predicción de congestiones vehiculares.
- Desarrollo de un algoritmo de clustering que selecciona el número de clusters de manera dinámica implementado en el framework Spark streaming aplicado a

flujos de textos cortos.

- Identificación de las partes relevantes de un documento. Propuesta de distintas métricas y una representación vectorial de oraciones de diferentes longitudes.
- Análisis y comparación de resúmenes extractivos de documentos.

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, 4 tesistas de Doctorado en Cs. Informáticas (2 con beca de postgrado de la UNLP), 2 tesistas de grado y 2 profesores extranjeros.

Dentro de los temas involucrados en esta línea de investigación, en los últimos 2 años se han finalizado 1 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 3 tesinas de grado de Licenciatura. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] Jimbo P., Lanzarini L., Fernandez-Bariviera A. (2019) Variations of Particle Swarm Optimization for Obtaining Classification Rules Applied to Credit Risk in Financial Institutions of Ecuador. Journal Risks,8,1,1-14, MDPI, Open Access Journal

[2] Jimbo P., Lanzarini L., Fernandez-Bariviera A. (2019). Fuzzy Classification Rules with FRvarPSO Using Various Methods for Obtaining Fuzzy Sets. International Journal of Machine Learning and Computing.:International Association of Computer Science & Information Technology. issn 2010-3700.

[3] Jimbo P., Lanzarini L., Fernandez-Bariviera A. (2019). Particle Swarm Optimization for Obtaining Classification Rules Applied to Credit Risk in Financial Institutions of Ecuador. Risks 2020, Vol.8, Issue 1 doi:10.3390/risks8010002.

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

[5] Reyes-Zambrano G., Lanzarini L., Hasperué W., Fernández-Bariviera A (2020) 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. DOI: 10.3233/JIFS-179644

[6] Molina, R, Hasperué, W. Villa Monte, A. *D3CAS: Distributed Clustering Algorithm Applied to Short-Text Stream Processing. Communications in Computer and Information Science. Springer. 2019. pp. 211-220.*

[7] Villa-Monte A., Lanzarini L., Fernández-Bariviera A., Olivas-Varela J.A. (2019) User-Oriented Summaries Using a PSO Based Scoring Optimization Method. Journal Entropy. 21 (6), 617. ISSN 1099-4300. https://doi.org/10.3390/e21060617

[8] Villa-Monte A., Lanzarini L., Corvi J., Fernández-Bariviera A.(2020). Document summarization using a structural metrics based representation. Journal of Intelligent & Fuzzy Systems. 1-10. 10.3233/JIFS-179648.

Laura Lanzarini:
0000-0001-7027-7564

Waldo Hasperué:
0000-0002-9950-1563

César Estrebou:
0000-0001-5926-8827

Augusto Villa Monte:
0000-0002-9854-3083

Aurelio Fernandez Bariviera:
0000-0003-1014-1010

José Ángel Olivas Varela:
0000-0003-4172-4729