Minería de Datos, Minería de Textos y Big Data

L. Lanzarini[1], W. Hasperué[1], A. Villa Monte[1,3], P. Jimbo Santana[4], G. Reyes Zambrano[5], 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, avillamonte}@lidi.info.unlp.edu.ar
prjimbo@uce.edu.ec, gary.reyesz@ug.edu.ec, julieta.corvi@gmail.com , aurelio.fernandez@urv.net, cristina.puente@icai.comillas.edu, 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 textos 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 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 una técnica de clustering dinámico que se ejecuta de manera distribuida. Esta implementación se está llevando a cabo utilizando el framework Spark Streaming.

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 ARPECOL.

En el área de la Minería de Textos se han desarrollado estrategias para resumir documentos a través de la extracción utilizando métricas de selección y técnicas de optimización de los párrafos más representativos. Además, se han desarrollado métodos capaces de determinar la subjetividad de oraciones escritas en español.

Palabras clave: Minería de Datos, Minería de Textos, Big Data, Redes neuronales, Resúmenes extractivos, Sentencias causales temporales, Stream processing.

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

Esta línea de trabajo comenzó hace varios años con el diseño de nuevos algoritmos para la obtención de conjuntos de reglas de clasificación haciendo énfasis en la simplicidad del modelo y la facilidad de interpretación por parte de quien debe tomar decisiones. En esa línea se definió oportunamente un algoritmo que demostró tener la capacidad de generar conjuntos de reglas adecuados [1]. Luego, con el objetivo de llevar a cabo una transferencia tecnológica en el área de Riesgo Crediticio se estudiaron consideraciones especiales basadas en información de expertos en otorgamiento de créditos. Se llegó a la conclusión de que era factible incorporar conjuntos difusos para tratar los atributos numéricos como variables lingüísticas y así simplificar la interpretación de las reglas por parte del oficial de crédito [2]. El nuevo método desarrollado se denomina FRvarPSO y tiene la capacidad de obtener de reglas de clasificación difusas operando sobre atributos nominales y numéricos [3]. Utiliza una técnica de optimización de población variable inicializada por medio de una red neuronal competitiva. Luego, por medio de los conjuntos difusos asociados a las variables lingüísticas, incorpora un criterio de votación que guía la búsqueda de las partículas facilitando la identificación de los valores a utilizar en la construcción las condiciones que darán lugar a los antecedentes de las reglas. Esta nueva propuesta fue medida en tres bases de datos de entidades financieras del Ecuador: una Cooperativa de Ahorro y Crédito y dos Bancos que otorgan diferentes tipos de crédito, con resultados satisfactorios [4]. Las mediciones realizadas utilizando 12 bases de datos del repositorio UCI y su comparación con otros métodos existentes en la literatura también han arrojados buenos resultados evidenciando la factibilidad de aplicar este nuevo método en distintos contextos [5].

Actualmente se sigue trabajando en el área de riesgo crediticio buscando mejorar dos aspectos: la incorporación de información macroeconómica y la incorporación de un factor de certeza a la recomendación dada por la regla. El primer aspecto debería impactar en la precisión de la regla y el segundo en la toma de decisiones por parte del agente de crédito.

Análisis de trayectorias GPS

El avance tecnológico facilita el registro y recolección de información de trayectorias GPS de vehículos en la vía pública. El análisis inteligente de estos datos lleva a identificar patrones sumamente útiles a la hora de tomar decisiones en situaciones relacionadas con urbanismo, circulación y congestionamiento, entre otras. En esta dirección se ha trabajado en el diseño e implementación de 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 machine learning que presenten la característica de ser iterativas, operando sobre el conjunto completo de los datos de un flujo,

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 temas que se abordan en esta línea abarcan la implementación de técnicas de clustering para el tratamiento de flujos de datos, la detección de tópicos, el análisis de sentimiento y el procesamiento de datos relacionados al comercio realizado con criptomonedas [7].

Por otro lado y como una transferencia tecnológica concreta se está trabajando 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 muestra 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 está colaborando 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 así como distinguir proposiciones subjetivas de las objetivas.

Resumen Automático de Documentos

Esta línea de investigación se centra en la generación automática de resúmenes. Entre los enfoques existentes se ha puesto el énfasis en el extractivo cuyo resumen está formado por un subconjunto de sentencias de un documento seleccionadas apropiadamente. Actualmente, a partir del trabajo realizado en [8] se están analizando en la construcción de distintos tipos de resúmenes (1) el impacto de varias tareas de preprocesamiento de textos, (2) la participación de un conjunto amplio de métricas y (3) la incorporación de semánticas en el análisis [9]. Para llevar a cabo estos experimentos se desarrolló una herramienta de manipulación de documentos científicos programada en Python con MySQL utilizando las librerías NLTK, urllib y bs4, entre otras. Los experimentos están siendo realizados sobre artículos científicos publicados en PLOS ONE hasta tanto se consiga el acceso a las colecciones DUC.

Por otro lado, en [10] se estudió la relación entre algunos tipos de resúmenes extractivos y los formados únicamente por las sentencias causales detectadas en un documento. Este tipo de sentencias son de suma utilidad para analizar documentos clínicos por ser una componente principal de toda explicación médica. Ellas expresan, por ejemplo, las causas de las enfermedades o muestran los efectos de cada tratamiento. Actualmente, se están investigando las restricciones temporales asociadas a relaciones causales.

Clasificación de oraciones

Con el objetivo de analizar la subjetividad u objetividad de un texto se desarrolló una representación de oraciones escritas en español en formato vectorial que permite etiquetarlas. Esta representación utiliza distintas métricas lingüísticas para convertir una oración a una matriz numérica. Dado que la cantidad de filas de estas matrices depende de la longitud de la oración se realiza una normalización que convierte dicha matriz en un vector de longitud fija para poder comparar los vectores de distintas oraciones.

Se han utilizado las redes neuronales y las máquinas de soporte vectorial para entrenar modelos que permitan clasificar una oración en objetiva o subjetiva [11].

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Estudio de técnicas de optimización poblaciones 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 clustering dinámico en Spark streaming.
- 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 un nuevo método de agrupamiento de trayectorias GPS aplicable a la predicción de congestiones vehiculares.
- Desarrollo de una representación de términos que, junto con un modelo de clasificación, permite identificar palabras clave en un documento.
- Desarrollo de un algoritmo de clustering que selecciona el número de clusters de manera dinámica implementado en el framework Spark streaming.
- 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.
- Aplicación de las sentencias causales en el desarrollo de un sistema que asista en la administración de medicamentos mediante el control de intervalos de tiempo.

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 becario doctoral UNLP), 1 tesista de grado y 2 profesores extranjeros.

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

Actualmente se están desarrollando 4 tesis de doctorado, 1 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] Lanzarini L., Villa Monte A., Fernandez Bariviera A., Jimbo Santana P. *Simplifying Credit Scoring Rules using LVQ+PSO.* Kybernetes. Emerald Group Publishing Limited. vol. 46. pp 8-16. ISSN 0368-492X. 2017.

[2] Jimbo Santana P., Villa Monte A., Rucci E., Lanzarini L, and Fernández Bariviera A. *Analysis of Methods for Generating Classification Rules Applicable to Credit Risk*., Journal of computer science &

technology (ISSN 1666-6038), vol. 17, num. 1, págs. 20-28, abril de 2017.

[3] Jimbo Santana P., Lanzarini L., Fernández-Bariviera A. *Fuzzy Credit Risk Scoring Rules using FRvarPSO*. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems (ISSN 0218-4885). Vol 26. Nro1. pp. 39-57. World Scientific. 2018

[4] Jimbo Santana P., Lanzarini L., Fernández-Bariviera A. *Extraction of knowledge with population-based metaheuristics fuzzy rules applied to credit risk.* Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science. pp 153-163. Vol 10942. Springer, Cham. 2018.

[5] Jimbo Santana P., Lanzarini L., Fernández-Bariviera A. *FRvarPSO a method for obtaining fuzzy classification rules using optimization techniques*. International Conference on Modeling and Simulation in Engineering, Economics and Management (MS'2018). Girona. España (en prensa).

[6] Molina, R, Hasperué, W. *D3CAS: un Algoritmo de Clustering para el Procesamiento de Flujos de Datos en Spark*. XXIV Congreso Argentino de Ciencias de la Computación (CACIC 2018). Octubre 2018.

[7] Bariviera, A. F., Basgall, M. J., Hasperué, W., & Naiouf, M. (2017). *Some stylized facts of the Bitcoin market*. Physica A: Statistical Mechanics and its Applications, 484, 82–90. https://doi.org/10.1016/j.physa.2017.04.159

[8] Villa Monte A., Lanzarini L., Rojas Flores L., Olivas Varela J. A.: *Document summarization using a scoring-based representation.* XLII Conferencia Latinoamericana en Informática (CLEI 2016). ISBN 978-1-5090-1633-4, pp. 1-7. Octubre de 2016.

[9] Villa-Monte A., Lanzarini L., Fernández-Bariviera A. and Olivas J. A. *Obtaining and evaluation of extractive summaries from stored text documents*. III Conference on Business Analytics in Finance and Industry. Enero 2019.

[10] Puente C., Villa Monte A., Lanzarini L., Sobrino A. and Olivas Varela J. Á. *Evaluation of causal sentences in automated summarie*. Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), págs. 1-6, doi. 10.1109/FUZZ-IEEE.2017.8015666, 2017.

[11] Coria, J.M. *Clasificación de Subjetividad utilizando Técnicas de Aprendizaje Automático*. Tesis de grado. Facultad de Informática, UNLP. Febrero 2018.