Minería de Datos y Big Data. Aplicaciones en riesgo crediticio, salud y análisis de mercado

L. Lanzarini[1], W. Hasperué[1], A. Villa Monte[1,3], M. J. Basgall[1,4], R. Molina[1,5], L. Rojas Flores[6], J. Corvi[2], P. Jimbo Santana[7], A. Fernandez Bariviera[8], C. Puente[9], J. A. Olivas[10]

[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] UNLP, CONICET, III-LIDI, La Plata, Argentina [5] Becario CIN
[6] Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Ushuaia, Argentina
[7] Facultad de Ciencias Administrativas, Universidad Central del Ecuador, Quito, Ecuador
[8] Dpto de Economía, Universitat Rovira i Virgili, Reus, España
[9] Escuela Técnica Superior de Ingeniería ICAI, Universidad Pontificia Comillas, Madrid, España
[10] 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, mjbasgall}@lidi.info.unlp.edu.ar
{luisf.09, julieta.corvi}@gmail.com, pjimbo@pcpsolutions.com, aurelio.fernandez@urv.net, cristina.puente@icai.comillas.edu, joseangel.olivas@uclm.es

CONTEXTO

Esta presentación corresponde al proyecto "Sistemas inteligentes. Aplicaciones en reconocimiento de patrones, minería de datos y big data" (Periodo 2018–2021) del Instituto de Investigación en Informática LIDI.

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 textos y reconocimiento de patrones en imágenes.

En el área de la Minería de Datos se está trabajando, por un lado, en la generación de un modelo de fácil interpretación a partir de la extracción de reglas de clasificación que permita justificar la toma de decisiones y, por otro lado, en el desarrollo de nuevas estrategias para tratar grandes volúmenes de datos.

Con respecto al área de Big Data se están realizando diversos aportes usando el framework Spark Streaming. En esta dirección, se está investigando en una técnica de clustering dinámico que se ejecuta de manera distribuida. Además se ha implementado en Spark Streaming una aplicación que calcula el índice de Hurtz de manera online, actualizándolo cada pocos segundos con el objetivo de estudiar un cierto mercado de negocios.

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.

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

Obtención de Reglas de Clasificación

Esta línea de investigación está centrada en el diseño de nuevos algoritmos para la obtención de conjuntos de reglas de clasificación con tres características principales: precisión adecuada, baja cardinalidad y facilidad de interpretación. Esto último está dado por el uso de un número reducido de atributos en la conformación del antecedente que, sumada a la baja cardinalidad del conjunto de reglas, permite distinguir patrones sumamente útiles a la hora de comprender las relaciones entre los datos y tomar decisiones [1]. La aplicación de estos métodos en la predicción de riesgo crediticio ha arrojado resultados satisfactorios [2,3].

Actualmente se está trabajando en la fuzificación de las reglas con el objetivo de facilitar aún más su comprensión por parte del agente que debe decidir el otorgamiento del crédito. Se ha comprobado que con sólo fuzificar el antecedente de la regla se obtienen conjuntos de reglas de clasificación con un incremento significativo en la precisión en relación a lo publicado en [4].

A futuro se incorporará, a la recomendación dada por la regla, un factor de confianza que ayude a discernir entre posibles recomendaciones. Este es un aspecto importante ya que además de las características propias del solicitante del crédito existen condiciones macroeconómicas que condicionan la respuesta.

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 en formato texto. Para esto se están desarrollando estrategias que aplican 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 [5, 6].

Estas técnicas dinámicas se están implementando en el framework Spark Streaming, adecuado para procesamiento paralelo, distribuido y online. En este framework se desarrolló una aplicación que permite el cálculo del índice de Hurtz de manera online y dinámica, esto es, cada cierto tiempo la aplicación usa los nuevos datos recolectados y los procesa con aquellos que habían sido procesados previamente para poder hacer un seguimiento online de un cierto mercado de negocios [7].

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 [8].

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 [9] 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 [10] y (3) la incorporación de semánticas en el análisis [11]. 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 [12] 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. [13]

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.
- Métodos estructurados y no estructurados a la representación de documentos.
- Problemas de clasificación con desbalance de clase severo.
- Representación de documentos de texto utilizando métricas.
- Obtención de resúmenes automáticos de texto.
- Implementación de técnicas en el paradigma de MapReduce
- Implementación del índice de Hurtz en Spark streaming.
- Implementación de un algoritmo de clustering dinámico en Spark streaming.
- Propuesta de una representación vectorial de oraciones de longitud variable.
- Desarrollo de modelos que permiten clasificar oraciones en subjetivas u objetivas.

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.
- 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 frameworks Spark streaming.
- Implementación en Spark Streaming de una aplicación que calcula de manera online el coeficiente de Hurst y lo actualiza cada un cierto tiempo.
- 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.
- Implementación de modelos usando redes neuronales para la determinación de

subjetividad en oraciones extraídas en textos escritos en español.

- 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 con dedicación exclusiva, 1 becario doctoral UNLP, 1 becario doctoral CONICET, 1 becario CIN, 1 doctorando, 2 tesistas de grado y 3 profesores extranjeros.

Dentro de los temas involucrados en esta línea de investigación, en el último año se han finalizado 1 tesis de doctorado y 2 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., Aquino G., De Giusti A. *Obtaining classification rules using lvqPSO*. Advances in Swarm and Computational Intelligence. Lecture Notes in Computer Science. Vol 6433, pp. 183-193. ISSN 0302-9743. Springer-Verlag Berlin Heidelberg. Junio 2015.

[2] Jimbo P., Villa Monte A., Rucci E., Lanzarini L., Fernández A. *An exploratory analysis of methods for extracting credit risk rules*. XIII Workshop Bases de Datos y Minería de Datos (WBDDM). XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016), ISBN: 978-987-733-072-4, págs. 834-841, octubre de 2016

[3] 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.

[4] Lanzarini L., Villa Monte A., Fernandez Bariviera A., Jimbo Santana P. *Simplifying Credit Scoring Rules using LVQ+PSO.* : The International Journal of Systems& Cybernetics. Emerald Group Publishing Limited. vol. 46. Pp 8-16. ISSN 0368-492X. 2017.

[5] Basgall, M. J., Hasperué, W., Estrebou C., Naiouf M. *Clustering de un flujo de datos usando MapReduce*. XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016). Pp 682-691. ISBN 978-987-733-072-4. Octubre 2016.

[6] Basgall, M. J., Hasperué, W., Estrebou C., Naiouf M. *Data stream treatment using sliding windows with MapReduce.* Journal of Computer Science & Technology. Vol. 16. ISSN 1666-6038. Pp. 76-83. 2016.

[7] Basgall, M. J., Hasperué, W., Naiouf, M., & Bariviera, A. F. (2017). *Cálculo del exponente de Hurst utilizando Spark Streaming: enfoque experimental sobre un flujo de transacciones de criptomonedas*. XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017).

[8] 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

[9] Villa Monte A., Lanzarini L., Rojas L., Oliva Varela J.. *Document Summarization using a Scoring-Based Representation*. 2016 XLII Latin American Computing Conference (CLEI), págs. 1-7, doi. 10.1109/CLEI.2016.7833396, octubre de 2016.

[10] 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.

[11] 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 2018. *En prensa*.

[12] 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.

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