Minería de Datos y Big Data. Aplicaciones en Señales y Textos

L. Lanzarini[1], W. Hasperué[1], C. Estrebou[1], F. Ronchetti[1,2], A. Villa Monte[1,2], G. Aquino[1,3], F. Quiroga[1,2], M. J. Basgall[1,3], L. Rojas[4], J. Corvi[1], C. Luna[1], P. Jimbo[5], A. Fernandez[6], C. Puente[7], J. A. Olivas[8], A. Rosete[9]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, UNLP, La Plata, Argentina
Facultad de Informática, UNLP, La Plata, Argentina
[2] Becario postgrado UNLP [3] UNLP, CONICET, III-LIDI, La Plata, Argentina
[4] Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Ushuaia, Argentina
[5] Dpto. Ciencias de la Computación, Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
[6] Dpto de Economía, Universitat Rovira i Virgili, Reus, España
[7] Escuela Técnica Superior de Ingeniería ICAI, Universidad Pontificia Comillas, Madrid, España
[8] Dpto de Tecnología y Sistemas de la Información, Universidad de Castilla-La Mancha, Ciudad Real, España
[9] Universidad Tecnológica de La Habana "José Antonio Echeverría" (CUJAE), La Habana, Cuba

{laural, whasperue, cesarest, fronchetti, avillamonte, gaquino, fquiroga, mjbasgall}@lidi.info.unlp.edu.ar
{luisf.09, julieta.corvi, carla.lunagennari}@gmail.com, pjimbo@pcpsolutions.com, aurelio.fernandez@urv.net, cristina.puente@icai.comillas.edu, joseangel.olivas@uclm.es, rosete@ceis.cujae.edu.cu

CONTEXTO

Esta presentación corresponde al Subproyecto "Sistemas Inteligentes" perteneciente al proyecto "Cómputo paralelo de altas prestaciones. Fundamentos y evaluación de rendimiento en HPC. Aplicaciones a sistemas inteligentes, simulación y tratamiento de imágenes" (Periodo 2014–2017) 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 señales y textos.

Con respecto al procesamiento de Señales el énfasis está puesto en el análisis de videos con el objetivo de identificar acciones humanas que faciliten la interfaz hombre/máquina y en la detección de patrones de movimiento de los objetos presentes.

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 a Minería de Textos se han desarrollado métodos capaces de extraer las palabras clave de documentos independientemente del lenguaje. Además, se han desarrollando estrategias para resumir documentos a través de la extracción de párrafos.

Palabras clave: Estrategias adaptativas, Reconocimiento de Patrones, Minería de Datos, Minería de Textos, Big Data.

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. PROCESAMIENTO DE SEÑALES

En el III LIDI, desde hace varios años se viene trabajando en el procesamiento de señales de audio y video. Como resultado de estas investigaciones se han diseñado e

implementado técnicas originales aplicables al reconocimiento tanto de gestos dinámicos como de medidas biométricas. En relación a esta línea, actualmente se están desarrollando los siguientes temas:

Reconocimiento de gestos

El reconocimiento de lengua de señas es un campo de investigación relativamente nuevo cuyo objetivo final es traducir de la lengua de señas a una lengua escrita. Esto implica poder tomar un video en donde una persona habla en lengua de señas, y reconocer la posición de la persona, de su cara y sus manos, la expresión de su rostro, la forma de sus manos y también la de sus labios si la seña requiere pronunciar la palabra para desambiguar. Con esa información, se debe reconocer la seña realizada, para luego con la información de una secuencia de señas generar una traducción a una lengua escrita.

En esta área, se publicó un método para clasificar señas en videos pre-segmentados que abarca todas las etapas del reconocimiento [1, 2]. Éste método no utiliza la información secuencial de la seña, es decir, no utiliza la información temporal. No obstante, los resultados de los experimentos muestran que aún con esa dificultad se pueden clasificar correctamente el 96% de las señas del conjunto de prueba. De esta forma se pudo determinar que la información temporal no es tan importante para el reconocimiento de señas, al menos en bases de datos de estas características.

Para los experimentos se utilizaron dos conjuntos de datos recolectados por nuestro grupo, LSA16 y LSA64. El primero, LSA16, contiene 800 imágenes con 16 clases de formas de mano y el segundo, LSA64, está formado por 3200 videos de 64 clases de señas dinámicas. Los detalles de la base de señas dinámicas LSA64 también han sido publicados [3].

Actualmente se está trabajando en mejorar las etapas de detección y segmentación de la mano y el reconocimiento de su forma aplicando redes convolucionales profundas y otras técnicas relacionadas.

Detección de patrones de movimiento en video

El análisis automático de video con el objetivo de detectar patrones de movimiento es de suma utilidad en distintas áreas. Este tema combina el procesamiento de imágenes digitales con la minería de datos ya que se deben analizar automáticamente la estructura de la escena, las actividades que en ella se están desarrollando y los patrones de movimiento de los objetos involucrados con el propósito de detectar situaciones anómalas.

Se espera poder contribuir al diseño y desarrollo de nuevas estrategias adaptativas aplicables al análisis de videos. Los resultados de esta investigación pueden aplicarse en distintas áreas tales como seguridad, a través de la detección automática de situaciones de riesgo o amenazas en escenas captadas a través de sistemas de video-vigilancia o salud a través de la identificación de comportamientos en personas que padecen enfermedades que alteran su movimiento corporal.

1.2. MINERÍA DE DATOS

Obtención de Reglas de Clasificación

Esta línea de investigación está centrada en la obtención de un conjunto de reglas de clasificación con tres características principales: precisión adecuada, baja cardinalidad y facilidad de interpretación [4,5]. 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.

Como resultado se ha desarrollado un nuevo método de extracción de reglas de clasificación que hace uso de una variante original de la técnica de optimización basada en cúmulos de partículas PSO inicializada a través de una red neuronal competitiva LVQ. Los resultados de su aplicación a un conjunto de 13 bases de datos de repositorio han sido satisfactorios.

Como área de transferencia tecnológica se ha analizado la situación de dos compañías financieras al momento de determinar el riesgo en una operación de otorgamiento de crédito para consumo. Se trata de operaciones con montos muy inferiores a los préstamos hipotecarios que requieren tomar decisiones rápidas ya que generalmente son acordados con los clientes a través de un servicio en línea. En ambos casos se han obtenido conjuntos de reglas con una precisión aceptable y una cardinalidad sensiblemente menor a los métodos convencionales. Volviendo a la necesidad de tomar una decisión rápida, este modelo ofrece una gran ventaja con respecto al mecanismo habitual. Para más detalles consultar [6,7].

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 presentan la característica de ser iterativas, operando sobre el conjunto completo de los datos ó sobre 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.

Estas técnicas dinámicas se están empleando bajo el paradigma MapReduce, adecuado para procesamiento paralelo y distribuido. Para el tratamiento de un flujo de datos, se utiliza el enfoque de ventana deslizante temporal manejando el tamaño de la misma de manera dinámica en función de la frecuencia de llegada de los datos y el tiempo de respuesta de la tarea iterativa a realizar sobre ellos, permitiendo que cada dato sea utilizado por el proceso iterativo la mayor cantidad de veces posibles [8,9].

Los temas que se abordan en esta línea abarcan el procesamiento del lenguaje natural, la detección de tópicos, el análisis de sentimiento y procesamiento de datos relacionados al comercio realizado con criptomonedas.

1.3. MINERIA DE TEXTOS

Extracción de palabras clave en documentos de texto

Esta línea de investigación tiene su eje central en el estudio y aplicación de distintos métodos de representación de documentos así como de distintas técnicas adaptativas aplicables en la resolución de problemas de extracción de palabras clave, tarea de sumo interés ya que permite caracterizar un documento facilitando su búsqueda y clasificación [10].

En esta línea se está trabajando en un método de identificación de palabras clave a partir de documentos de texto en español utilizando redes neuronales como estimadores de probabilidad [11].

Este método define una representación vectorial para los términos, para luego aplicar un proceso de filtrado gramatical con el fin de remover términos inválidos. Es importante destacar que ésta es la única parte del método que es dependiente del idioma en cuestión [12].

Una vez obtenida la representación se utiliza un *ensemble* de redes neuronales para construir un modelo de clasificación. Dado un documento a ser analizado, el modelo determina para cada término la probabilidad de ser palabra clave. Estas probabilidades son utilizadas para construir un ranking de términos, lo que proporciona la flexibilidad de seleccionar los mejores N términos.

Actualmente se continúa el desarrollo del método para mejorar su precisión y ampliar su dominio de aplicación.

Síntesis automática de documentos

En esta línea de investigación se desarrollan técnicas capaces de representar y modelizar uno o varios documentos de texto con el objetivo de construir una versión más corta de ellos en forma automática preservando su información. Esto resulta de sumo interés ya que permite obtener el contenido principal de un documento en menos tiempo del que

llevaría hacerlo en forma manual a partir del texto completo.

En [13] se diseñó una estrategia para extraer el criterio utilizado por una persona al resumir un texto. Luego, se lo aplicó a otros documentos logrando obtener un resumen similar al que se hubiera conseguido en forma manual. Para ello, se utilizaron los capítulos de una tesis escrita en LaTex a los cuales se les calculó un conjunto de métricas conocidas. Luego a través de una técnica de optimización basada en cúmulos de partículas se identificó el aporte de cada métrica en la construcción del resumen esperado.

Actualmente se están llevando a cabo dos trabajos utilizando un conjunto de artículos científicos de acceso libre. Por un lado se está trabajando en el agrupamiento de los documentos para descubrir relaciones entre las métricas que los representan. Por otro se está realizando la comparación de los resúmenes formados por las sentencias causales de un documento y los resúmenes obtenidos de aplicar cada una de las métricas por separado.

2. TEMAS DE INVESTIGACIÓN Y DESARROLLO

- Estudio de técnicas de optimización y redes neuronales artificiales para la obtención de reglas de tipo IF-THEN.
- Estudio de técnicas de segmentación de objetos en movimiento presentes en un video.
- Estudio de técnicas de agrupamiento aplicables a la detección de patrones de movimiento.
- Representación y clasificación de configuraciones de manos para el lenguaje de señas.
- Clasificación de señas dinámicas.
- Problemas de clasificación con severo desbalance de clases y métodos. Algoritmos aplicables a los mismos.
- Estudio de distintos métodos de caracterización de textos haciendo énfasis en su estructura, longitud, idioma y formalidad en la redacción.
- Métodos estructurados y no estructurados aplicables a la representación de documentos. Representación de documentos de texto utilizando métricas.
- Estudio de técnicas para resumen automático de documentos.
- Implementación de técnicas en el paradigma de MapReduce
- Estudios de performance de los algoritmos desarrollados

3. RESULTADOS OBTENIDOS

- Desarrollo de un método de extracción de reglas de clasificación con énfasis en la reducción de la complejidad del modelo aplicable a riesgo crediticio.
- Desarrollo de una representación de términos y un modelo de clasificación con el fin de identificar palabras clave en un documento.
- Desarrollo de un modelo de clasificación de señas segmentadas y comparación de su desempeño con otros modelos del estado del arte.
- Desarrollo de una estrategia dinámica empleando el paradigma MapReduce para aplicar algoritmos iterativos sobre los datos, los cuales arrojan resultados muy similares a los que se obtienen con las mismas tareas ejecutadas de manera secuencial pero utilizando el conjunto de datos completo.
- Determinación de coeficiente de Hurst en transacciones de Bitcoins.
- Identificación de las partes relevantes de un documento.
- Análisis y comparación de resúmenes extractivos de documentos.
- Caracterización de documentos por medio de su agrupamiento.

4. FORMACIÓN DE RECURSOS HUMANOS

Dentro de los temas involucrados en esta línea de investigación, en los últimos 5 años se han finalizado 4 tesis de doctorado, 2 tesis de maestría, 3 tesis de especialista y 9 tesinas de grado de Licenciatura.

Actualmente se están desarrollando 5 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] Ronchetti F., Quiroga F., Estrebou C., Lanzarini L., Rosete A. *Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language*. Publicado en Ibero-American Conference on Artificial Intelligence IBERAMIA 2016 (pp. 338-349)

[2] Ronchetti, F., Quiroga, F., Estrebou, C.A., Lanzarini. *Handshape recognition for Argentinian Sign Language using ProbSom*. Journal of Computer Science & Technology, vol. 16, N° 1, págs. 1-5, ISSN 1666-6038, 2016.

[3] Ronchetti, F., Quiroga, F., Estrebou, C.A., Lanzarini, L.C., Rosete, A. . *LSA64: An Argentinian Sign Language Dataset*, publicado en el XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016) (pp. 794-803).

[4] Lanzarini, L., Villa Monte, A., Ronchetti, F.: SOM+PSO. *A Novel Method to Obtain Classification Rules*. Journal of Computer Science & Technology (JCS&T), Vol. 15, No 1, pp. 15-22. ISSN 1666-6038. Abril 2015.

[5] 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.

[6] Lanzarini L., Villa Monte A., Fernandez Bariviera A., Jimbo Santana P. *Obtaining Classification Rules Using LVQ+PSO: an application to Credit Risk*. Scientific Methods for the Treatment of Uncertainty in Social Sciences. Advances in Intelligent Systems and Computing. Springer-Verlag Berlin Heidelberg. vol. 377. pp 383-391. ISSN 2194-5357. 2015.

[7] 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.

[8] 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.

[9] 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.

[10] Aquino G., Lanzarini L. *Keyword identification in spanish documents using neural networks*.. Journal of Computer Science and Technology, JCS&T. ISSN: 1666-6046. Volumen 15. Número 2. pp. 55-60. Noviembre 2015.

[11] Aquino G., Hasperué W., Lanzarini L. *Keyword Extraction using Auto-associative Neural Networks*. XX Congreso Argentino de Ciencias de la Computación (CACIC 2014) - ISBN 978-987-3806-05-6. pp.562-570. 2014.

[12] Aquino, G, Hasperué, W, Estrebou, C, Lanzarini, L. *A Novel Language-Independent Keyword Extraction Method*. Publicado en el Libro Computer Science & Technology Series – XIX Argentine Congress of Computer Science - Selected Papers., 2014. pp.221-232

[13] Villa Monte A., Lanzarini L., Rojas L., Olivas Varela J.A. *Document summarization using a scoring-based representation.* 2016 XLII Latin American Computing Conference (CLEI). 2016, pp. 1-7.