ANÁLISIS DE DATOS SIMBÓLICOS PARA DATA SCIENCE

Adriana Mallea[1], Jorgelina Carrizo[1], Leonel Ganga[2], Cecilia Martínez[2], Andrea Salas[1]

[1]Departamento de Matemática, FFHA, Universidad Nacional de San Juan
[2]Departamento de Informática, FCEFN, Universidad Nacional de San Juan
lamallea@ffha.unsj.edu.ar

RESUMEN

La ciencia de datos, considerada como una ciencia en sí misma, es en términos generales, la extracción de conocimiento de los datos. Data Mining es una poderosa tecnología con gran potencial para extraer tal conocimiento. Sin embargo, desde el punto de vista estadístico, sus herramientas sólo han sido desarrolladas para trabajar con matrices de datos clásicas, es decir, donde cada unidad es individual y las variables toman un único valor para cada individuo. El análisis de datos simbólicos (SDA, por sus siglas en inglés) brinda una nueva forma de pensar en Data Science al extender la entrada estándar a un conjunto de clases de entidades individuales. Por lo tanto, las clases de una población dada se consideran unidades de una población de nivel superior a estudiar. Tales clases a menudo representan las unidades reales de interés. Para tener en cuenta la variabilidad entre los miembros de cada clase, las clases se describen por intervalos, distribuciones, conjunto de categorías o números que a veces se ponderan. De esa manera, obtenemos nuevos tipos de datos, llamados "simbólicos", ya que no se pueden reducir a números sin perder información sobre la variabilidad interna. SDA es un nuevo paradigma que abre un vasto dominio de investigación y aplicaciones al proporcionar resultados complementarios a los métodos clásicos aplicados a los datos estándar.
A lo largo de las últimas tres décadas se han extendido distintos métodos del análisis clásico de datos al simbólico, la mayoría de ellos descriptivos. Esto fundamenta la necesidad de continuar investigando sobre la modelización e inferencia en el contexto de datos de naturaleza simbólica. El presente proyecto pretende responder a esta necesidad. Las metodologías se aplicarán a problemas reales o simulados.

Palabras clave: Data Science, Simbólico, Análisis

CONTEXTO

El proyecto *Análisis de Datos Simbólicos para Data Science* propone continuar con la investigación y desarrollo de nuevas metodologías, cuyo estudio se inició en el proyecto CICITCA 2018-2019, 21/F1085, sobre todo referidas a modelación e inferencia en el marco del SDA, que permitan la extracción de conocimientos. Es un proyecto cuyo tipo de actividad de I+D es investigación básica, presentado para su acreditación en diciembre de 2019, inició en 2020 y es de carácter bi-anual, financia la UNSJ. Tiene como unidad ejecutora el Departamento de Matemática de la FFHA

y sus integrantes desarrollan sus tareas de docencia e investigación en las áreas de matemática e informática. La línea de investigación corresponde a minería de datos, en un marco más general de Ciencia de los Datos.

1. INTRODUCCIÓN

El volumen de datos que circula en la Web, o almacenado por las empresas, está creciendo constantemente. Para explotar esta riqueza, es necesario extraer conocimiento de grandes volúmenes de información. El dominio que apunta a resolver esta problemática es la ciencia de los datos (Data Science). Data Science tiene como objetivo extraer conocimiento de todo tipo de datos (estructurados o no, de fuentes homogéneas o heterogéneas, etc.). Representa la intersección de varias disciplinas como estadística, matemática, inteligencia artificial y minería de datos (Data Mining). Data Mining ofrece métodos de análisis muy útiles para extraer conocimiento. Sin embargo, desde el punto de vista estadístico, sus herramientas sólo han sido desarrolladas para trabajar con matrices de datos clásicas, es decir, donde cada unidad es individual y las variables toman un único valor para cada individuo. Para que el estudio de los datos sea accesible en varios niveles de agregación, ha surgido el campo del análisis de datos simbólicos (SDA). Desde entonces, este campo se ha desarrollado proponiendo varios métodos específicos de análisis de datos simbólicos. Estos métodos se han implementado en herramientas como Sodas [S3], Syr [S1] y bibliotecas de R [S5, S4, S2] que permiten sus pruebas y sus aplicaciones en nuevas bases de datos. SDA en esencia se construye sobre la noción de que las inferencias estadísticas se requieren comúnmente a nivel de grupo en lugar de a un nivel individual (Billard, 2011, Billard y Diday, 2006). Por ejemplo, en los exámenes de prueba estandarizados, el rendimiento de la escuela. y las unidades de nivel superior suelen ser de interés más que el rendimiento de los estudiantes individuales. SDA adopta explícitamente esta idea al agregar datos de nivel individual (los microdatos) en resúmenes distribucionales a nivel grupo (es decir, los símbolos), y luego construir modelos para inferencia directamente a nivel de grupo basado en estos resúmenes (Billard, 2011, Billard y Diday, 2006). La elección más común de estos resúmenes es el intervalo aleatorio (o el equivalente d-dimensional: el hipercubo aleatorio) pero existen otro tipo de símbolos que incluyen histogramas aleatorios (Dias y Brito, 2015, Le-Rademacher y Billard, 2013) y variables categóricas multivaluadas (Billard y Diday, 2006).

Este enfoque es extremadamente atractivo dadas las tendencias tecnológicas actuales que requieren análisis de conjuntos de datos cada vez más grandes y complejos. Al agregar los microdatos a un número mucho menor de símbolos a nivel de grupo, los análisis de “big data” se pueden realizar de manera económica y efectiva en los dispositivos informáticos de baja gama. Más allá de la agregación de datos, las observaciones de valor distribucional pueden surgir naturalmente a través del proceso de registro de datos.

Desde sus inicios en la década de los ochenta, se han desarrollado muchas técnicas del SDA para analizar variables aleatorias con valores de distribución, incluyendo modelos de regresión (Irpino y Verde, 2015, Dias y Brito, 2015, Giordani, 2015), análisis de componentes principales (Kosmelj et al., 2014, Le-Rademacher y Billard, 2013, Ichino, 2011), series de tiempo (Lin y Gonzalez-Rivera, 2016, Wang et al., 2016, Arroyo et al., 2011), clustering (Brito et al.,

2015), análisis discriminante (Silva y Brito, 2015) y modelado jerárquico bayesiano (Lin et al., 2017). Sin embargo, si bien ha habido muchos avances en el análisis de datos simbólicos, desde una perspectiva estadística, la mayoría de las técnicas de SDA son descriptivas y no permiten la inferencia estadística sobre los parámetros del modelo, lo que impide que sus métodos realicen su potencial en el conjunto de herramientas del estadístico moderno.

Existen pocos trabajos en la literatura que emplean inferencia: la inferencia basada en la verosimilitud fue introducida por Le-Rademacher y Billard (2011), con mayor desarrollo y aplicación por Brito y Duarte Silva (2012). Los trabajos más recientes en este sentido se deben a Zhang y Sisson (2016) y B. Beranger, H. Lin and S. A. Sisson (2018).

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

Las líneas de investigación, tal como se mencionan en el resumen, se enmarcan dentro de Data Science y Data Mining. Debido a que las herramientas desarrolladas en esta última sólo sirven para trabajar con matrices de datos clásicas, ha surgido en la década de 1980 el análisis de datos simbólicos que brinda una nueva forma de pensar en Data Science, al extender la entrada estándar a un conjunto de clases de entidades individuales. En muchas ocasiones tales clases son el objeto de estudio y para tener en cuenta la variabilidad entre los miembros de cada clase, las mismas se describen por datos distribucionales. De esa manera, obtenemos nuevos tipos de datos, llamados "simbólicos", ya que no se pueden reducir a números sin perder mucha información. El primer paso en SDA es construir la tabla de datos simbólicos donde las filas son clases y las columnas son variables que pueden tomar valores simbólicos. El segundo paso es estudiar y extraer nuevos conocimientos de estos nuevos tipos de datos mediante al menos una extensión de Estadística Computacional y Data Mining a datos simbólicos.

SDA es un nuevo paradigma que abre un vasto dominio de investigación y aplicaciones al proporcionar resultados complementarios a los métodos clásicos aplicados a los datos estándar. SDA también brinda respuestas a los grandes volúmenes de datos (big data) y datos complejos, ya que los primeros se pueden reducir y resumir por clases y los datos complejos, con múltiples tablas de datos no estructurados y las variables no apareadas se pueden transformar en una tabla de datos estructurada con variables apareadas de valores simbólicos.

En este proyecto trabajamos con ambos enfoques, Data Mining y SDA para la extracción de conocimientos.

3. RESULTADOS OBTENIDOS/ESPERADOS

El presente proyecto tiene como finalidad investigar sobre metodologías del Análisis de Datos Simbólicos en el contexto de Data Science. El mismo sigue la línea de investigación del proyecto acreditado desarrollado en el período 2018-2019. En él se ha trabajado fundamentalmente con el estudio de Series Simbólicas de intervalo y de Regresión Lineal Simbólica. En el proyecto actual se ha profundizado, en el primer año de trabajo, el estudio y aplicación de técnicas del SDA referentes a Clustering, Regresión y Series Temporales. Algunas de las metodologías propuestas en regresión simbólica de intervalo se han aplicado a datos en un

contexto biométrico. Por otra parte se ha trabajado con datos de COVID-19 publicados en el sitio https://github.com/owid/covid-19-data. Para estos datos se han empleado técnicas del SDA para describir los países de América respecto a características de la evolución de la pandemia y posteriormente hacer una clasificación supervisada que evidencia el posicionamiento de cada país frente a la pandemia, de acuerdo a variables tales como los valores de los casos confirmados acumulados, el nuevo aumento diario de casos confirmados y los relativos por millón de habitantes. Los resultados obtenidos se han presentado y publicado en congresos nacionales e internacionales. Además se comenzó el estudio de papers recientemente publicados que abordan el problema de la modelización e inferencia en datos de naturaleza simbólica.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de investigación está formado por docentes investigadores de dos facultades de la UNSJ, algunos de ellos son jóvenes investigadores. En el marco del proyecto desarrolla su segundo año de beca de iniciación a la investigación una egresada de Licenciatura en Matemática, que actualmente cursa las últimas materias en la carrera Maestría en Matemática de la Universidad Nacional de San Luis y está escribiendo su trabajo de tesis sobre Series Simbólicas de Intervalo. Entre los integrantes del proyecto hay además dos maestrandos, que aplicarán en sus trabajos de tesis las herramientas objeto de la presente investigación.

5. BIBLIOGRAFÍA

Books

[B1] Afonso, F., Diday, E., Toque, C. (2019) "Data Science par Analyse Des Données Symboliques". Editions TECHNIP

[B2] Arroyo, J. (2008) “Métodos de predicción para series temporales de intervalos e histogramas”. Ph. D. Dissertation, Universidad Pontificia Comillas, Madrid.

[B3] Billard, L., Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.

[B4] Diday, E. and Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley.

Papers

[P1] Arroyo, J., R. Espínola, and C. Maté (2008). “Diferent approaches to forecast interval time series: a comparison in finanance.” Computation Statistics and Data Analysis (submitted).

[P2] Arroyo, J. and C.Maté (2008). “Forecasting time series of observed distributions with smoothing methods based on the barycentric histogram”. In Computational Intelligence in Decision and Control. Proceedings of the 8th International FLINS Conference, pp. 61.66. World Scienti.c.

[P3] Berenger B., Lin H., Sisson S.A. (2018) “New models for symbolic data analysis”. arXiv: 1809.03659v1[stat. CO]

[P4] Billard, L. (2011). “Brief overview of symbolic data and analytic issues”.

[P5]Brito, P., A. P. D. Silva, and J. G. Dias (2015). “Probabilistic clustering of interval data. Intelligent Data Analysis”. 19, 293-313

[P6] Brito, P., Duarte Silva, A. P. (2012): "Modelling Interval Data with Normal and Skew-Normal Distributions”. Journal of Applied Statistics, Volume 39, Issue 1, 3-20.

[P7] Brito, P. (2007): "Modelling and Analysing Interval Data". In: "Advances in Data Analysis", Decker, R., Lenz, H.-J. (Eds.), Series "Studies in Classification, Data Analysis and Knowledge Organization", Springer, Berlin, Heidelberg, New-York, 197-208.

[P8] Giordani, P. (2015).”Lasso-constrained regression analysis for interval-valued data”. Advances in Data Analysis and Classification 9, 5-19

[P9] González-Rivera, G. and Arroyo, J. (2012). “Time series modeling of histogram-valued data: The daily histogram time series of S&P500 intradaily returns”, International Journal of Forecasting, 28 (1), 20–-33.

[P10] Han, A., Hong, Y., Lai, K. and Wang, S. (2008). “Interval time series analysis with an application to the Sterling-Dollar exchange rate”, Journal of Systems Science and Complexity, 21 (4), 558-573.

[P11] Irpino, A. and R. Verde (2015).”Linear regression for numeric symbolic variables: a least squares approach based on Wasserstein distance”. Advances in Data Analysis and Classi_cation 9, 81-106.

[P12] Kosmelj, K., J. Le-Rademacher, and L. Billard (2014).”Symbolic covariance matrix for interval-valued variables and its application to principal component analysis: a case study”. Metod. Zvezki 11, 1-20.

[P13] Le-Rademacher, J. and L. Billard (2011). “Likelihood functions and some maximum likelihood estimators for symbolic data”. Journal of Statistical Planning and Inference” 141, 1593-1602.

[P14] Lin, H., M. J. Caley, and S. A. Sisson (2017).”Estimating global species richness using symbolic data meta-analysis”. arXiv:1711.03202

[P15] Lin, W. and G. González-Rivera (2016).”Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data”. Comp. Stat. Dat. An 100, 694-711.

[P16] Teles, P. and Brito, P. (2015).” Modelling Interval Time Series with Space-Time processes”, Communications in Statistics: Theory and Methods, Volume 44, Issue 17. DOI: 10.1080/03610926.2013.782200

[P17] Wang, X., Z. Zhang, and S. Li (2016).”Set-valued and interval-valued stationary time series”. Journal of Multivariate Analysis 145, 208-223.

[P18] Zhang, X. and S. A. Sisson (2016).”Constructing likelihood functions for interval-valued random variables”. arXiv:1608.00945 .

Software

[S1] F. Afonso, E. Diday, and R. Haddad (2012) Latest developments of the syr software for symbolic data analysis of complex data. 3rd Workshop on Symbolic Data Analysis.

[S2] V. Batagelj and N. Kejzar (2010) Clamix clustering symbolic objects. Computer software R program. Vienna : R Foundation for Statistical Computing. Available : https ://r-forge. r-project.org/projects/clamix.

[S3] E. Diday, M. Noirhomme-Fraiture, et al.(2008) Symbolic data analysis and the SODAS software. J.Wiley & Sons.

[S4] A. Irpino and R. Verde (2015) Basic statistics for distributional symbolic variables : a new metric-based approach. Advances in Data Analysis and Classification, 9(2) :143–175.

[S5] R. Oldemar, J. Murillo, and J. Villalobos (2012) An r package for symbolic data analysis. XVIII SIMMAC.