Herramientas estadísticas para data mining y modelos dinámicos

Adriana Mallea[1], Myriam Herrera[2], Jorgelina Carrizo[1], Andrea Salas[3] Cecilia Martínez[3], Leonel Ganga[3]

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

RESUMEN

En los últimos años surgió el término Big Data, en el contexto de Data Mining, para referirse a conjuntos de datos tan grandes y complejos que se vuelven difíciles de procesar en un tiempo razonable con aplicaciones tradicionales de análisis de datos. En general, los analistas necesitan analizar datos con variabilidad, por ej., datos que resultan de la agregación de registros individuales en grupos de interés, o datos que representan entidades abstractas como especies biológicas, o regiones como un todo. El Análisis Simbólico de Datos, al ofrecer la posibilidad de agregación de datos al nivel de granularidad elegido por el usuario, mientras se mantiene la información sobre la variabilidad intrínseca, puede desempeñar un papel importante en este contexto. Esta metodología resulta particularmente interesante para el estudio de Economía y Gestión, Marketing, Ciencias Sociales, Geografía, estadísticas sobre datos oficiales, así como para Biología y análisis de datos Geológicos.

Por otra parte, existen campos de la actividad humana; tales como economía, meteorología, administración de empresas; donde parte de los problemas y su solución se presentan en un contexto dinámico. Esto requiere analizar las variables relevantes del problema a resolver a partir de datos recogidos secuencialmente a intervalos regulares de tiempo.

El presente proyecto tiene como finalidad investigar sobre las metodologías estadísticas apropiadas para el tratamiento de grandes bases de datos; en particular del Análisis Simbólico de Datos; como así también estudiar modelos aleatorios para el tratamiento de variables indexadas en el tiempo. Es claro que existen problemas provenientes de diversas áreas del conocimiento donde se necesita de ambos tipos de herramientas, por ejemplo cuando aparecen series temporales o de espacio-tiempo que toman valores en intervalos o incluso distribución de valores. Las metodologías se aplicarán a datos reales.

.

Palabras clave: Data Mining, Modelos, Dinámicos

CONTEXTO

El proyecto *Herramientas estadísticas para Data Mining y modelos dinámicos* se encuentra en su segundo año de ejecución. Es un proyecto acreditado, de carácter bi-anual y financiado por la UNSJ. Tiene como unidades ejecutoras en primer lugar el departamento de Matemática de la FFHA y en segundo

lugar el Instituto de Informática de la FCEFN de la UNSJ. La línea de investigación corresponde a minería de datos, en un marco más general de Ciencia de los Datos.

1. INTRODUCCIÓN

En la actualidad, gracias a los avances en tecnología y recursos informáticos, es sencillo almacenar información, lo que lleva a tener grandes bases de datos. Respuestas referentes a cuestiones formuladas sobre estas bases se pueden dar mediante *Data Mining.*

Data Mining [13] es una poderosa tecnología con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Datos. Las herramientas de minería de datos predicen futuras tendencias y comportamientos, permitiendo tomar decisiones proactivas en los negocios y conducidas por un conocimiento acabado de la información. Permiten responder a preguntas de negocios que tradicionalmente consumen demasiado tiempo para ser resueltas y a las que los usuarios de esta información casi no están dispuestos a aceptar. Estas herramientas exploran las bases de datos en busca de patrones ocultos, encontrando información predecible que un experto no puede llegar a encontrar porque se encuentra fuera de sus capacidades debido a la complejidad dada por el volumen de información.

Las técnicas de Data Mining se pueden implementar rápidamente en plataformas ya existentes de software y hardware para acrecentar el valor de las fuentes de información existentes y se pueden integrar con nuevos productos y sistemas.

Data Mining está soportado por tres tecnologías:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

En los últimos años surgió el término "Big Data", en el contexto de Data Mining, para referirse a conjuntos de datos tan grandes y complejos que se vuelven difíciles de procesar en un tiempo razonable con aplicaciones tradicionales de análisis de datos. El Análisis Simbólico de Datos, al ofrecer la posibilidad de agregación de datos, desempeña un rol importante, tal como se comentó en el resumen.

Por otra parte, al analizar datos, éstos pueden aparecer secuencialmente, lo que lleva a considerar variables que dependen del tiempo y, por ende, surge la necesidad de trabajar con modelos dinámicos. En este tipo de situaciones, las propiedades evolutivas que muestran las distintas variables son muy diferentes y, en consecuencia, son también muy distintos los modelos que pueden determinar estas variables en función de variables explicativas.

El conocimiento y caracterización diferenciada de la evolución de las diversas variables relevantes, así como el conocimiento y asimilación de los modelos (probabilísticos) que las pueden explicar, constituyen información imprescindible a la hora de tomar decisiones.

Lo antes mencionado justifica la necesidad de investigar sobre las metodologías estadísticas apropiadas para el tratamiento de grandes bases de datos (Big Data); en particular del Análisis

Simbólico de Datos; como así también estudiar modelos de naturaleza estocástica. Existen problemas provenientes de diversas áreas del conocimiento donde se necesita de ambos tipos de herramientas, por ejemplo evolución de datos financieros, de variables meteorológicas, evolución de indicadores socio- demográficos, es decir variables que toman valores en intervalos o incluso distribución de valores.

1. LINEAS DE INVESTIGACIÓN Y DESARROLLO

La ciencia de datos, considerada como una ciencia en sí misma, es en términos generales, la extracción de conocimiento de los datos [16]. 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) [9, 10] 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 y similares. 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 variables 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.

1. RESULTADOS OBTENIDOS/ESPERADOS

El presente proyecto tiene como finalidad investigar sobre las metodologías estadísticas apropiadas para el tratamiento de grandes bases de datos; en particular del Análisis Simbólico de Datos; como así también estudiar modelos aleatorios para el tratamiento de variables indexadas en el tiempo y aplicarlas a datos reales.

En el primer año de desarrollo del proyecto se ha trabajado fundamentalmente con el estudio de Series Simbólicas de intervalo y de histograma, como así también de Regresión Lineal Simbólica, en seminarios internos de investigación

desarrollados por los integrantes del proyecto.

Las metodologías propuestas en papers relacionados se han aplicado a datos provenientes de Economía y Finanzas. En particular, para el caso de series simbólicas de intervalo se han usado técnicas de pronóstico de indicadores macroeconómicos. Se han aplicado distintos métodos de regresión simbólica de intervalo y comparado sus performances, usando medidas de error adaptadas al caso simbólico, en un problema de costos.

Por otra parte se han aplicado técnicas del Análisis Simbólico de Datos a datos provenientes de la Encuesta Permanente de Hogares del Gran San Juan, correspondientes al tercer trimestre del año 2016, con el objetivo de caracterizar a éstos hogares y analizar la situación laboral en el Gran San Juan en relación al Nivel de Estudio.

Los resultados obtenidos se han presentado en congresos nacionales e internacionales.

En este segundo año de ejecución se espera profundizar el estudio de herramientas del SDA (Resultados del Workshop Advances in Data Science for Big and complex data, University Paris Duphine, January 2019) y aplicar me todologías de Clustering Simbólico, Regresión Simbólica y Series de Tiempo simbólicas a datos reales o simulados.

1. 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 beca de iniciación a la investigación una egresada de Licenciatura en Matemática, que fue alumna adscripta en el proyecto anterior y actualmente cursa su segundo año en la carrera Maestría en Matemática de la Universidad Nacional de San Luis. Entre los integrantes del proyecto hay cuatro maestrandos, que aplicarán en sus trabajos de tesis las herramientas objeto de la presente investigación.

1. BIBLIOGRAFÍA

2. Arroyo, J. (2008) “Métodos de predicción para series temporales de intervalos e histogramas”. Ph. D. Dissertation, Universidad Pontificia Comillas, Madrid.
3. 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).
4. Arroyo, J; Gonzales Rivera, G; Maté, C. (2010) “Forecasting with interval and histogram data Some financial applications”. Handbook of empirical economics and finance, 247-280
5. Billard, L., Diday, E. (2007). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley.
6. Brito, P. (2014): "Symbolic Data Analysis: another look at the interaction of Data Mining and Statistics". WIREs Data Mining and Knowledge Discovery, Volume 4,

Issue 4, July/August 2014, 281–295. DOI: 10.1002/widm.1133

7. 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.
8. 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.
9. Brito, P. (2007): "On the Analysis of Symbolic Data". In: "Selected Contributions in Classification and Data Analysis", Brito, P., Bertrand, P.,
10. Diday, E (2016) "Thinking by classes in data Science: the symbolic data analysis paradigm". WIREs Comput Stal,8:172-205,doi: 0.1002/wics.1384.
11. Diday, E. and Noirhomme-Fraiture, M. (2008). Symbolic Data Analysis and the SODAS Software. Wiley.
12. Gallardo, V; Mallea, L. A. (2016). "Una introducción a Datos Simbólicos". Servicio de Publicaciones de la FFHA-UNSJ. ISBN 978-950-605-837-1
13. 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.
14. Han, J. Kamber, M.(2006), Data Mining. Concepts and Techniques, 2.a ed. Morgan Kaufman.
15. Maia, A.L.S., De Carvalho, F.A.T. and Ludermir, T.D. (2008). "Forecasting models for interval-valued time series", Neurocomputing, 71 (16-18), 3344-3352.
16. Mallea, L.A.; Martínez, E.; Salas, A. (2017) "Series Temporales Simbólicas de Intervalo". Servicio de Publicaciones de la FFHA-UNSJ. ISBN 978-950-605-851-7
17. Taylor, J. W. "Exponential smoothing with a damped multiplicative trend. International Journal of Forecasting", vol. 19, páginas 715_725, 2003.
18. Teles, P. y Brito, M. P. (2005) "Modelling interval time series data". En Proceedings of the 3rd IASC World Conference on Computational Statistics & Data Analysis.
19. Teles, P. and Brito, P. (2015)." Modelling Interval Time Series with Space-Times processes", communications in Statistics: Theory and Methods, Volume 44, Issue 17. DOI: 10.1080/03610926.2013.782200