XXIII Workshop de Investigadores en Ciencias de la Computación

ANÁLISIS DE DATOS SIMBÓLICOS PARA DATA SCIENCE

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

1. CONTEXTO

El proyecto Análisis de Datos Simbólicos (SDA) 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.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

Las líneas de investigación, 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.

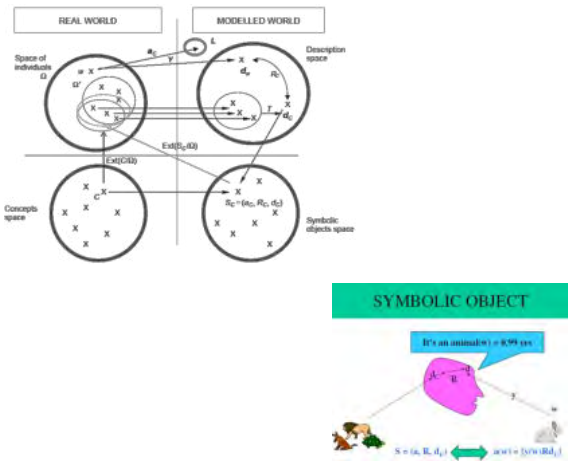

Figura 1: *Modelización por un objeto simbólico de un concepto conocido por su extensión*

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. En el proyecto se utilizan dos softwares para construir tablas simbólicas o aplicar métodos de SDA: SODAS (Analysis System of Symbolic Official Data) es el resultado del proyecto ASSO y su arquitectura se muestra en Figura 2 y R+R Studio con sus paquetes específicos (RSDA, iRegression, entre otros)

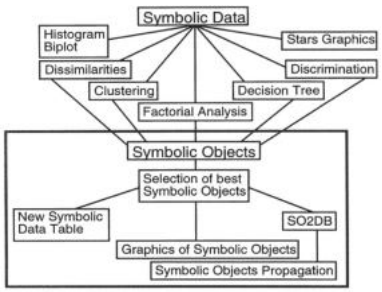

Figura 2: *Investigación y desarrollo de software del proyecto SODAS*

Figura 3: *Softwares estadísticos utilizados para la obtención de los resultados*

3. RESULTADOS OBTENIDOS/ESPERADOS

El presente proyecto tiene como finalidad investigar sobre metodologías del Análisis de Datos Simbólicos en el contexto de Data Science. Durante el primer año de trabajo se profundizó 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.

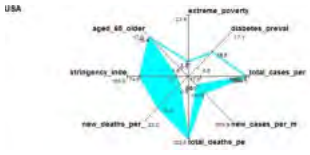

Figura 4: *Visualización simbólica de Estados Unidos*

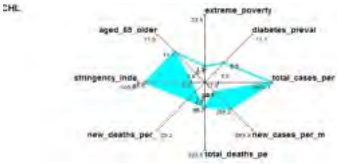

Figura 5: *Visualización simbólica de Chile*

Figura 6:*Visualización simbólica de Argentina*

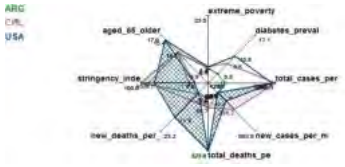

Figura 7: *Superposición de los objetos simbólicos EE UU, Chile y Argentina*

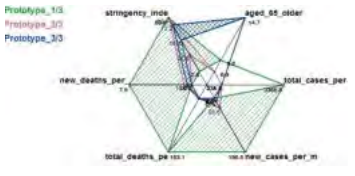

Figura 8: *Superposición de Prototipos de las tres clases obtenidas con SCLUST*

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.