Práctica 4: Calidad de los Datos

Introducción

Cuando se quiere resolver un problema de minería de datos en un dominio específico, el principal insumo son los datos. El problema es que en la mayoría de las bases de datos (independientemente del formato) existen errores, que pueden ser datos incorrectos, faltantes o inconsistentes.

Para intentar solucionar estos inconvenientes, existe la etapa de preparación o preprocesado de los datos.

Exploración de los datos

Para poder observar estos “errores” de los datos, primero se deben detectar o reconocer y para ello se deben utilizar las herramientas conocidas para tal fin: análisis multivariado.

La exploración consiste entonces en la observación de las características de los datos (dimensión y tipos de datos) y las medidas estadísticas de resumen (de los atributos). Además, se debe realizar alguna visualización para entender los datos.

Procesamiento de los datos

El procesamiento dependerá de los datos en sí, pero las tareas más comunes que se utilizan son la eliminación (ruido, duplicados, etc.), muestreo, selección, discretización, transformación y creación de atributos.

En el caso de valores faltantes, su tratamiento dependerá del dominio del problema, las opciones pueden ser: dejar los valores faltantes, eliminar los individuos/registros con valores faltantes, filtrar el atributo (opción menos recomendable), reemplazar el valor faltante (manualmente o predecirlo por algún método) o podemos intentar obtener el valor (no siempre es posible).

Implementación en R

La mayoría de las herramientas mencionadas están disponibles en la instalación base de R, aunque existen paquetes específicos que tienen funciones más potentes que las básicas. Tal es el caso de los paquetes **dplyr** [2] y **ggplot2** [3], para manejo de datos y gráficos respectivamente.

Para las medidas estadísticas básicas, con las funcionen *min()*, *max()*, *mean()*, *median() y var()* podemos obtener el mínimo, máximo, valor medio, mediana y varianza de un vector.

Otra función de la instalación base que podemos utilizar para obtener medidas de resumen es *summary()*. Para el caso de un vector numérico, nos va a devolver el mínimo, mediana, promedio, máximo, primer y tercer cuartil. El resultado que nos entrega dependerá del objeto que usemos como argumento.

Para las gráfica exploratorias, se puede utilizar desde la función básica *plot()* y las opciones del paquete de la instalación base graphics; hasta la función *ggplot()* y sus opciones más complejas de geometría.

Actividades

1 – Cargue los datos contenidos en el archivo “Entrenamiento_ECI_2020.csv”, utilizados en la actividad 2 de la guía anterior.

a) Indique la dimensión de los datos y tipos de los atributos. Obtenga las medidas de resumen de cada uno. ¿La forma en que se cargan los datos se corresponde con el tipo de dato que debería tener el atributo? ¿Hay datos faltantes?
b) Transforme los datos de la variable objetivo (Stage) al tipo adecuado y discretícela de acuerdo con las necesidades del problema planteado.

Exploración de los datos:

c) Obtenga los boxplot de la variable “Total_Amount” para cada una de las clases obtenidas en el punto anterior (que se visualicen de forma adecuada).
d) Obtenga una gráfica de dispersión de las variables “TRF” , “Total_Amount” y “Total_Taxable_Amount”, según la clase de la variable objetivo.
e) Grafique las frecuencias de las variables “Region” y “Bureaucratic_Code”.
f) Si tuviera que completar los valores faltantes de la variable “Total_Amount” de la actividad anterior, teniendo en cuenta la distribución de los datos, que valor estadístico utilizaría?

2 – Transformación e integración de datos.

Se dispone de un conjunto de datos de 200 señales de electrocardiogramas (ECGs) en archivos *.mat (MATLAB), que contienen las muestras de los valores de tensión (en mV) de los ECGs. Además, se dispone de un archivo con las medidas estadísticas básicas de los intervalos RR de dichas señales.

Considerando que las señales de ECG tienen todas duraciones distintas, se necesita agregar al dataset con las estadísticas de los intervalos RR, las siguientes variables de las señales:

- Tiempo total de la señal en segundos (la frecuencia de muestreo es 300 Hz).
- Los valores: mínimo, máximo, medio, mediana y desvío estándar de la amplitud en mV.

El conjunto de señales de ECG se encuentra en el archivo ECGs.RAR. Para abrir los archivos *.MAT utilice la función *read.mat()* del paquete **rmatio** [].

Una vez calculadas las variables necesarias y agregadas a los datos originales, guardar el nuevo conjunto de datos en un archivo CSV.

Referencias

1. Orallo, J. et all. “Introducción a la Minería de Datos” (2004).
2. Paquete dplyr. Disponible en: https://cran.r-project.org/web/packages/dplyr/dplyr.pdf
3. Paquete ggplot2. Disponible en: https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
4. Paquete rmatio. Disponible en: https://cran.r-project.org/web/packages/rmatio/rmatio.pdf