Data Mining

Datos Faltantes

Outline

- Tratamiento de datos faltantes.
  - Origen de los faltantes.
  - Tratamientos:
    - Eliminación de filas o columnas.
    - Imputaciones puntuales (imputación por media, moda, etc.).
    - Métodos cold/hot deck
    - Imputaciones múltiples (MICE).

Tratamiento de datos faltantes

Datos Faltantes

- El problema de los ***missing data*** está presente en el análisis de datos desde los **orígenes del almacenamiento**.

- Este problema de missing data se vuelve incluso un dilema propio del **proceso de KDD** donde los volúmenes de datos se incrementan y con ello la probabilidad de inconsistencias y faltantes.

Razones de la inconsistencia

- **Factores propios del procedimiento**.

  Formularios mal diseñados, errores de programación, etc.

- **Negativa a responder**.

  Ej: Cuestiones relacionadas con la edad, ¿cuánto gana? afiliación política, religión, etc.

- **Respuestas inaplicables**.

  Ej: ¿Cuánto gasto en juguetes para sus hijos el último año?...no tengo hijos cuack!

Problemas de trabajar con faltantes

- Los datos faltantes (DF) dificultan el **análisis de datos**.
- Un manejo inapropiado de DF en el análisis puede introducir sesgos y puede resultar en **conclusiones engañosas**.
- También pueden **limitar la generalización** del conocimiento alcanzado.

Hay tres tipos de problemas que se asocian con los DF en minería de datos:

1. Pérdida de eficiencia.
2. Complicaciones en el análisis y el manejo de datos
3. Sesgos que resultan de las diferencias entre los datos completos y los faltantes.

Tipos de datos faltantes

Existen diferentes mecanismos de faltantes, los tipos estándar de DF son:

- **Outliers tratados como datos faltantes:** Cuando se conocen los límites de las diferentes variables del dataset, los datos que caen fuera del rango definido se deben etiquetar como faltantes.

- **Datos faltantes al azar**. *Missing At Random* ***MAR***
  - La probabilidad de que la variable Y tenga un dato faltante depende de X, pero no de Y.
  - Es decir, el patrón de los datos faltantes se puede predecir a partir de otras variables de la base de datos.

Tipos de datos faltantes

- Datos faltantes completamente al azar. *Missing completely at random* ***MCAR***
  - La probabilidad de que la variable Y tenga un dato faltante es independiente de X.
  - Los datos existentes en Y son una muestra al azar de los valores de Y

- Datos faltantes que dependen de un predictor no observado. ***Non-Ignorable*** *missing data.*
  - El dato faltante se podría estimar a partir de otra variable, pero que no fue registrada.
  - Se puede intentar modelar la variable con datos faltantes a partir de las presentes en el dataset.

Métodos para tratar con datos faltantes

- **Utilizar solo registros completos**: Esto va a depender de cómo es el origen del faltante.
  - Se recomienda que se utilice solo en los casos en que el mecanismo de faltante sea MCAR.
- **Borrar casos seleccionados o variables**: La eliminación de registros ante la presencia de faltantes puede utilizarse cuando hay un **patrón no aleatorio de datos faltantes**.
  - Si la eliminación de un subconjunto disminuye significativamente la utilidad de los datos, la eliminación del caso puede no ser efectiva.
- **Imputación de datos**: Son métodos de relleno de faltantes.
- Aproximaciones basadas en modelado: Múltiples imputaciones

Imputación de datos

- Son métodos de relleno de datos ante la presencia de faltantes.
- Según (Hair, 1998), *"es el proceso de estimar datos faltantes de una observación a partir de valores válidos de otras variables"*
- Se debe tener precaución al emplear métodos de imputación, ya que pueden generar sesgos importantes entre los datos reales y los imputados.
- Los métodos de imputación generalmente usados son:
  - Sustitución de casos
  - Sustitución de medias
  - Imputaciones: Hot deck y Cold deck
  - Imputaciones utilizando regresiones
  - Imputaciones múltiples

Imputación de Datos Faltantes

- **Sustitución de casos.** Se reemplaza con valores no observados. Debería ser realizado por un experto en esos datos.

- **Sustitución de Medias**. Se reemplaza utilizando el promedio calculado de los valores presentes. Debe verificarse que los datos ajusten a una distribución normal, si los datos están sesgados es mejor utilizar la mediana.

Hay tres desventajas en el uso de la media:

- La varianza estimada de la nueva variable no es válida porque está atenuada por los valores repetidos
- Se distorsiona la distribución
- Las correlaciones que se observen estaran deprimidas debido a la repetición de un solo valor constante.

Imputación: Cold Deck

Cold Deck

- Selecciona valores o usar relaciones obtenidas de fuentes distintas de la base de datos actual.
- Se **sustituye un valor constante** derivado de fuentes externas o de investigaciones previas.
- Tiene las mismas desventajas de aplicar la media.

Imputación: Hot Deck

Hot Deck: Reemplaza los faltantes con valores obtenidos de registros que son los más similares.

Ventajas de hot deck:

- ❏ Conceptualmente simple
- ❏ Conserva los niveles de medición adecuados para las variables
- ❏ Finaliza el proceso de imputación con un conjunto completo de datos.

Desventajas:

- ❏ la dificultad para definir qué es similar.

Table I
Illustration of hot deck imputation: incomplete data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	???
4	2	5	10	2

Table II
Illustration of hot deck imputation: imputed data set

Case	Item 1	Item 2	Item 3	Item 4
1	10	2	3	5
2	13	10	3	13
3	5	10	3	13
4	2	5	10	2

Imputación utilizando regresiones

- Se utiliza el análisis de regresión para para predecir valores faltantes a partir variables relacionadas en el conjunto de datos.
- Se pueden utilizar regresiones simples o múltiples.
- Se identifican las variables independientes y dependiente.

```
# Imputación por Regresión
income = read.csv("missing_income.csv", sep = ';', dec = ',')
          ln(income ~ (age + years_of_collage), data=income[1:17,])
f_income = lm(income~age + years_of_college, data=income[1:17,])
summary(f_income)

income$imputado = 33912.1 + 300.9*income$age + 1554.2*income$years_of_college
```

Filter

	caso	income	age	years_of_college
1	1	45251.25	26	4
2	2	62498.27	45	6
3	3	49350.32	28	5
4	4	46424.92	28	4
5	5	56077.27	46	4
6	6	51776.24	38	4
7	7	51410.97	35	4
8	8	64102.33	50	6
9	9	45953.96	45	3
10	10	50818.87	52	5
11	11	49078.98	30	0
12	12	61657.42	50	6
13	13	54479.90	46	6
14	14	64035.71	48	6
15	15	51651.50	50	6
16	16	46326.93	31	3
17	17	53742.71	50	4
18	18	59786.8	55	6
19	19	50660.4	35	4
20	20	53418.2	39	5

Imputaciones Múltiples. MICE

MICE

Es una técnica de imputación y su acrónimo significa: Multivariate imputation by chained equations (MICE).

MICE opera bajo el supuesto de que el origen de los faltantes es Missing At Random (MAR).

La probabilidad de que falte un valor depende solo de los valores observados y no de los valores no observados.

La aplicación de MICE a datos que no son MAR puede generar resultados sesgado.

MICE Steps

El proceso de *Chained Equation* se puede dividir en cuatro pasos generales:

Paso 1: Imputación Simple

A partir de un reemplazo por medias o modas u otra imputación se completan todos los faltantes.

Paso 2: Place Holders

Se toma una de las variables (**X**) con faltantes y se la vuelve a poner NA en sus faltantes originales.

Paso 3: Modelado

Se ajusta un modelo que permita predecir a partir de todas (o algunas de las variables completas) los NA de **X**.

Paso 4: Imputaciones

Los faltantes de **X** son reemplazados por las predicciones. Luego, X será usada como predictor para completar otra variable.

Los **Pasos 2 a 4** son repetidos para cada variable con datos faltantes.

Una vez que se procesan todas las variables con faltantes se cumple un **ciclo o iteración**.

Estos ciclos pueden repetirse varias veces, según la bibliografía 10 veces es suficiente.

MICE: Ejemplo

Conjunto inicial con datos faltantes

Age	Income	Gender
33	N.A.	F
18	12,000	N.A.
N.A.	13,542	M

Paso 1: Imputación puntual

Age	Income	Gender
33	12,771	F
18	12,000	F
25.5	13,542	M

A pesar de la imputación del paso 1 vamos a mantener una máscara de faltantes

Age	Income	Gender
33	[redacted]	F
18	12,000	[redacted]
[redacted]	13,542	M

Paso 2: Pasamos a faltante uno de los casos imputados en 1

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

Paso 3: Ajustamos un modelo con los casos completos

Age	Income	Gender
33	12,771	F
18	12,000	F
N.A.	13,542	M

Paso 4: Imputamos con el valor que surja de aplicar el modelo

Age	Income	Gender
33	12,771	F
18	12,000	F
35.3	13,542	M

MICE: Ejemplo

Paso 2: Pasamos a faltante uno de los casos imputados en 1

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

Paso 3: Ajustamos un modelo con los casos completos

Age	Income	Gender
33	N.A	F
18	12,000	F
35.3	13,542	M

Paso 4: Imputamos con el valor que surja de aplicar el modelo

Age	Income	Gender
33	13,103	F
18	12,000	F
35.3	13,542	M

Age	Income	Gender
33	13,103	F
18	12,000	M
35.3	13,542	M

Por último se modela Gender y se completa la primer iteración.

Con los valores imputados en este 1er ciclo y utilizando la máscara de faltantes se vuelve a iterar.

MICE Ventajas y Desventajas

Ventajas

No produce sesgo (Esto dependerá del modelo de imputación)

Puede ser utilizado para cualquier tipo de análisis.

Es fácil de usar.

Desventajas

Hay que pensar en el modelo de imputación además del modelo de análisis.

Puede ser costoso computacionalmente

Genera un dataset completo por cada iteración

Bibliografía

- Jiawei Han, Micheline Kamber, Jian Pei. 2012. Tercera edición. Data Mining: Concepts and Techniques. Cap. 3
- García, S., Luengo, J., & Herrera, F. (2016). *Data preprocessing in data mining*. Springer. (Cap. 5)
- Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. Korean journal of anesthesiology, 70(4), 407-411. [pdf]
- Brown, M. L., & Kros, J. F. (2003). Data mining and the impact of missing data. Industrial Management & Data Systems, 103(8), 611-621. [pdf]
- Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non‐response. International statistical review, 78(1), 40-64. [pdf]
- Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40-49. [pdf]
- https://technofob.com/2018/05/30/mice-is-nice-but-why-should-you-care/