



Preprocesamiento de datos

Wilmer Gonzalez

6213 - Minería de datos

Facultad de ciencias Universidad Central de Venezuela

21 de mayo de 2024

Contenido

- 1 Objetivo del preprocesamiento
- 2 Preprocesamiento estructural
- 3 Preprocesamiento funcional

"Notoriamente no hay clasificación del universo que no sea arbitraria y conjetural."^a

^aEl idioma analítico de John Wilkins -Jorge Luis Borges



Objetivo del preprocesamiento

Representar los datos de tal manera que cumplan con los requerimientos funcionales y estructurales de las técnicas de minería de datos a aplicar.

Reg. Estructurales

La representación física de los datos es consistente con el significado que se desea documentar. (Wickham, 2014)

Req. Funcionales

La representación física de los datos contiene la información suficiente para su análisis.



Objetivos del preprocesamiento estructural (Codd, 1971)

- ► Cada fenómeno está encapsulado en una tabla
- Cada observación está encapsulada



Objetivos del preprocesamiento estructural (Codd, 1971)

- Cada fenómeno está encapsulado en
- Cada variable está documentada en una columna
- Cada observación está encapsulada

Tip

Es más sencillo describir relaciones funcionales entre variables que entre filas

Objetivos del preprocesamiento estructural (Codd, 1971)

- Cada fenómeno está encapsulado en una tabla
- Cada variable está documentada en una columna
- Cada observación está encapsulada en una fila

Tip

Es más sencillo realizar comparaciones entre grupos de obaservaciones que entre columnas

Datos de ejemplo

Pacientes de un tratamiento médico

	tratamientoA	tratamientoB
Cornelle Mycroft	_	16
Shawn Opdenort	18	5
Irma Clowney	3	7

¿Cúal es la mejor representación?

	Cornelle Mycroft	Shawn Opdenort	Irma Clowney
tratamientoA	-	18	3
tratamientoB	16	5	7

Pacientes de un tratamiento médico

	tratamiento	resultado
Cornelle Mycroft	Α	_
Cornelle Mycroft	В	16
Shawn Opdenort	Α	18
Shawn Opdenort	В	5
Irma Clowney	Α	3
Irma Clowney	В	7

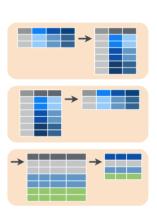
Problemas estructurales comunes

- ► Los descriptores de las columnas son posibles valores, no variables.
- ► Las columnas contienen múltiples variables.
- Las variables están presentes en filas y columnas.
- Muchos fenómenos están documentados en la misma tabla.
- ▶ Un fenómeno está compartido en distintas tablas.



Técnicas disponibles 1 2

- ► Transponer columnas/filas.
- ► Descomponer columnas en distintas columnas.
- Reagrupar columnas/filas como sea necesario.



¹Rstudio Cheat sheet

²Pandas Cheat sheet

Preproc. funcional

En general este tipo de preprocesamiento trata de aumentar la **usabilidad** y **fiabilidad** de los datos.

Algunos criterios dinámicos:

- Existe una alta cobertura de los datos posibles en un dominio objetivo.
- ► Los valores de cada variables son consistentes con el significado esperado.
- ► Los datos disponibles son vigentes y/o coinciden con la venta de tiempo que se desea analizar.

En general este tipo de preprocesamiento trata de aumentar la **usabilidad** y **fiabilidad** de los datos.

- ► Existen pocos valores faltantes.
- Los valores anómalos han sido documentados.
- Las variables usadas son las más informativas.
- Las variables están representadas de la manera más informativa

- Podría reducir la cantidad de datos disponibles.
- Su análisis podría ayudar a corregir la razón de valores faltantes.
- ► Podría introducir sesgos durante el análisis.

En general este tipo de preprocesamiento trata de aumentar la **usabilidad** y **fiabilidad** de los datos.

- Existen pocos valores faltantes.
- Los valores anómalos han sido documentados.
- Las variables usadas son las más informativas
- Las variables están representadas de la manera más informativa

- Puede ser producto de errores de imputación.
- ► Podría introducir sesgos durante el modelado

En general este tipo de preprocesamiento trata de aumentar la **usabilidad** y **fiabilidad** de los datos.

- Existen pocos valores faltantes.
- Los valores anómalos han sido documentados.
- Las variables usadas son las más informativas.
- Las variables están representadas de la manera más informativa

- Disminuye los recursos computacionales necesarios para el análisis.
- ► Incrementa la información contenida en cada variable.

En general este tipo de preprocesamiento trata de aumentar la usabilidad y fiabilidad de los datos.

Preproc. funcional

- Existen pocos valores faltantes.
- Las variables usadas son las más
- ▶ Las variables están representadas de la manera más informativa

- ► Facilita el análisis de las variables.
- ► Mejora la compatibilidad de los datos con los modelos a aplicar.

Técnicas disponibles³

- ► Representar variables categóricas en texto a arreglos n-dimensionales.
- ► Standardization, Normalization, reescalamiento, Transformaciones no-lineales.
- Discretización, Binarización.
- Inserción en valores faltantes.
- Reducción de dimensionalidad.
- ► (Datos en lenguaje natural) Tokenización, conteo, normalización, vectorización.
- ► (Datos en imágenes) Extracción de parches, compresión, aumento de datos, transformación de colores.

Referencias



Codd, E. F. (1971). "Normalized data base structure: a brief tutorial". En.



Wickham, Hadley (2014). "Tidy Data". En.

Referencias

¡Gracias!

github.com/ucvia/dm

