Preprocesamiento de datos

Wilmer Gonzalez

6213 - Minería de datos

Facultad de ciencias
Universidad Central de Venezuela

20 de junio de 2024

Contenido

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

[a] El idioma analítico de John Wilkins - Jorge Luis Borges

Objetivo

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.

Req. 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.

Preproc. estructural

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

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 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 observaciones 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

Datos de ejemplo

Pacientes de un tratamiento médico

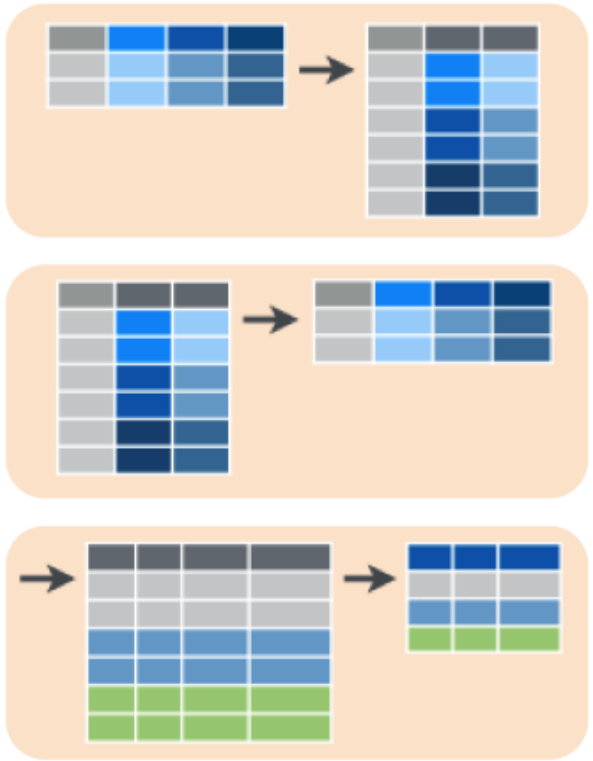
	tratamiento	**resultado**
Cornelle Mycroft	A	–
Cornelle Mycroft	B	16
Shawn Opdenort	A	18
Shawn Opdenort	B	5
Irma Clowney	A	3
Irma Clowney	B	7

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.



[1] Rstudio Cheat sheet
[2] Pandas Cheat sheet

Preproc. funcional

Objetivos del preprocesamiento 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.

Objetivos de preprocesamiento funcional

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

Algunos criterios estáticos:

- 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.

Objetivos de preprocesamiento funcional

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

Algunos criterios estáticos:

- 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.

Objetivos de preprocesamiento funcional

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

Algunos criterios estáticos:

- 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.

Objetivos de preprocesamiento funcional

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

Algunos criterios estáticos:

- 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.

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

Técnicas disponibles[3]

- 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.

[3] scikit-learn.org

Reducción de dimensionalidad

Planteamiento del problema

- La disponibilidad de los datos no implica su utilidad al hacer MD[4].
- Es posible que existan datos desconocidos que pueden ser más útiles de estar presente para comprender un fenómeno.
- Con más variables numéricas, la influencia de cada una disminuye, con respecto al análisis de agrupación o en operaciones propias de un espacio euclídeo (Leskovec, Rajaraman y Ullman, 2014).
- La correlación entre variables numéricas dificulta la atribución de comportamiento de los datos de dichas variables. Esto hace inestable a algunas técnicas de modelado.
- $+$ datos $\rightarrow$ (- velocidad de iteración, $+$ recursos computacionales)

[4] Minería de datos

Reducción de dimensionalidad

Algunas técnicas disponibles

Para reducir la dimensionalidad pueden aplicarse técnicas especializadas en el dominio del problema o hacer una revisión manual determinada por conocimiento profundo de las variables u observaciones que mejor describen el fenómeno de interés, a este tipo de técnicas las denominaremos reducción por selección.

También contamos con técnicas genéricas que independientemente del contexto del problema sirvan para proyectar los datos disponibles en un espacio más pequeño, estas técnicas las denominaremos reducción por proyección.

Reducción de dimensionalidad

Reducción por selección (Han, Kamber y Pei, 2012)

Selección de observaciones

Estas técnicas nos ayudan a establecer criterios con los cuales podemos reducir la cantidad de observaciones a usar en un problema dado.

- Muestreo sesgado: Selección de observaciones más representativas basado en nuestro conocimiento del problema, por ejemplo, en un caso dónde la vigencia sea de particular interés podemos dar prioridad a las observaciones más recientes.
- Muestreo estratificado[5]: Muestreo de observaciones garantizando mantener la proporción de los datos disponibles dado un aspecto de interés, por ejemplo, al trabajar con información demográfica puede ser necesario garantizar las proporciones por grupos etários al reducir a una muestra de las observaciones disponibles.

[5] Típicamente aplicado en problemas de clasificación usando la variable objetivo

Reducción de dimensionalidad

Reducción por selección - Selección de variables

Problemas supervisados

Estas técnicas nos ayudan a usar el aprendizaje supervisado para deterctar las variables más relevantes del problema.

- Indicadores de filtrado: Estos modelos anteceden a los modelo de aprendizaje supervisado con el fin de utilizar la distribución de los valores de cada atributo con respecto a la variable objetivo y medir su influencia, algunos indicadores: Gini, Entropía, puntaje de Fisher.
- Modelos de empaquetado: Se utilizan modelos que optimizan la selección de variables dadas al modelo de aprendizaje al hacer cambios incrementales a la selección de variables.
- Modelos incrustados: Algunos modelos de aprendizaje, indican ciertas propiedades de las variables al finalizar el ciclo de entrenamiento, por lo que pueden ser utilizados cómo un medio para evaluar la influencia de éstas, algunos modelos: Arboles de decisión, Regresión lineal.[6]

[6] Profundizaremos en estas técnicas en el tema 5

Reducción de dimensionalidad

Reducción por selección - Selección de variables

Problemas no supervisados

Estas técnicas nos ayudan a usar identificar que variables ayudan a explicar mejor la muestra disponible, por ejemplo, midiendo la calidad de una agrupación.

- Modelos de filtrado: Uso de funciones personalizadas de similitud para evaluar la utilidad de variables independientes en crear grupos. Una estrategia de este tipo consiste en definir modelos de aprendizaje supervisado sobre las variables disponibles para evaluar que tanto pueden correlacionarse las variables entre sí mediante la precisión del modelo resultante. Otros indicadores incluyen la entropía e indicador de Hopkins.

Reducción por dimensionalidad

Reducción por proyección - Análisis de componentes principales (PCA)

Objetivo

La idea principal de PCA, consiste en utilizar la covarianza entre las variables para conseguir una menor cantidad de variables, resultante de un nuevo sistema de coordenadas, que logren maximizar la varianza mantenida durante el proceso.

$$cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Referencias

- Codd, E. F. (1971). "Normalized data base structure: a brief tutorial". En.
- Han, Jiawei, Micheline Kamber y Jian Pei (2012). *Data mining concepts and techniques, third edition*.
- Leskovec, Jure, Anand Rajaraman y Jeffrey David Ullman (2014). *Mining of Massive Datasets*.
- Wickham, Hadley (2014). "Tidy Data". En.

¡Gracias!

github.com/ucvia/dm