

Sistemas Inteligentes para Gestión en la Empresa

Máster en Ingeniería Informática

02. Preprocesamiento de datos



UNIVERSIDAD
DE GRANADA



Preprocesamiento de datos

Bibliografía

S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Springer.

<https://link.springer.com/book/10.1007%2F978-3-319-10247-4>

M. Kuhn, K. Johnson (2019) *Feature Engineering and Selection: A Practical Approach for Predictive Models*. <https://bookdown.org/max/FES/>

G. Grolemund, H. Wickham (2017) *R for Data Science*. O'Reilly.

Tidyverse <https://www.tidyverse.org>

P. Casas (2018) *Data Science Live Book* <https://livebook.datascienceheroes.com>

H. Wickham (2016) *Elegant Graphics for Data Analysis*. Springer. <https://ggplot2-book.org/>

S. van Buuren (2018) *Flexible Imputation of Missing Data*. CRC Press <https://stefvanbuuren.name/fimd/>

Índice

- ▶ 1. Introducción
- 2. Integración, limpieza y transformación
- 3. Reducción de datos
- 4. Datos imperfectos
- 5. Resumen

Preprocesamiento de datos

Introducción

Preprocesamiento

Conjunto de tareas destinadas a la preparación de los datos previas al uso de algoritmos de extracción de conocimiento.

Dificultad

Proceso manual – consume el 60-80% del tiempo dedicado al análisis de datos (Adriaans & Zantinge, 1996)

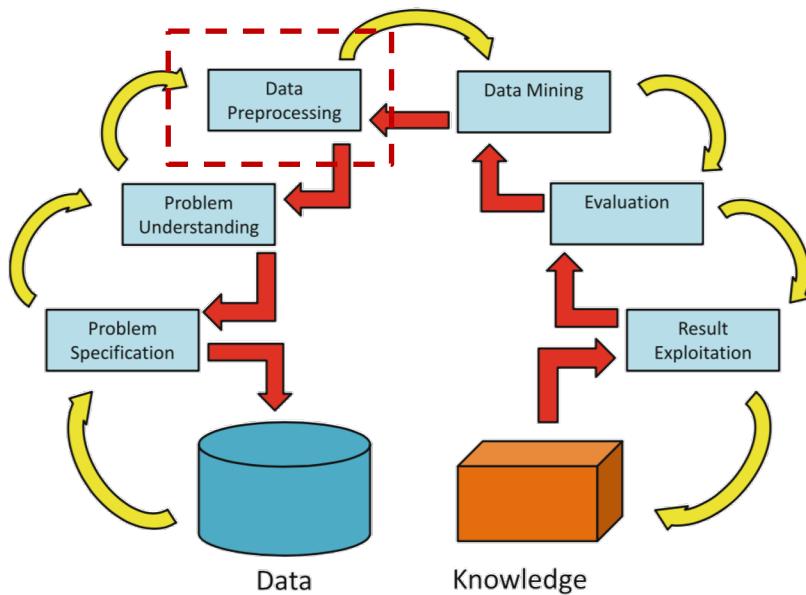
Objetivos

- Mejora de la calidad
 - incompletos, ruido, inconsistentes
- Reducción del tamaño
 - selección, eliminación, discretización



Preprocesamiento de datos

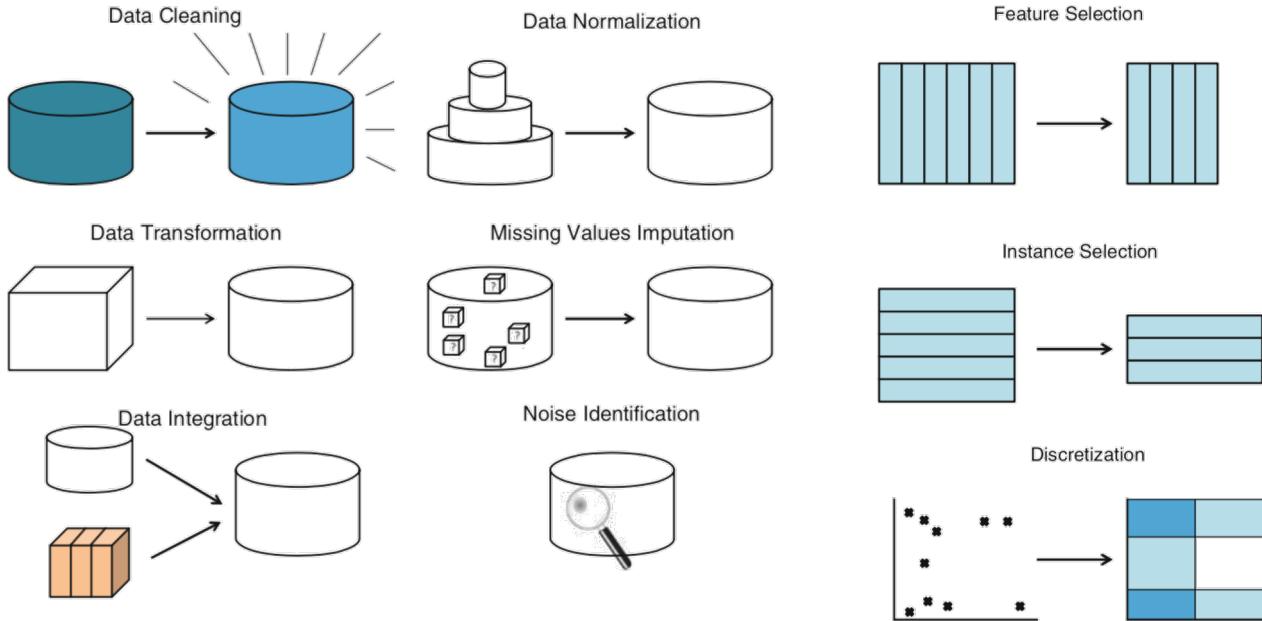
Introducción



S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. [Capítulo 1, Sección 1.6](#).
Springer.

Preprocesamiento de datos

Introducción

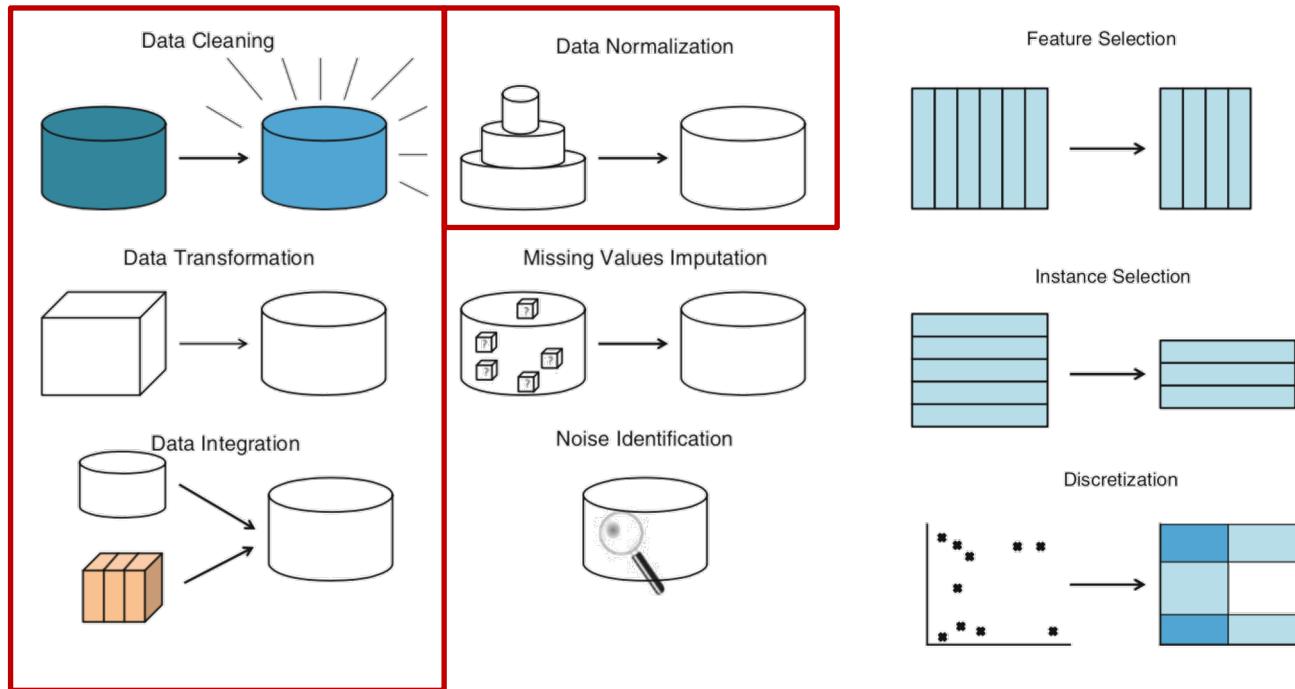


Índice

1. Introducción
- ▶ 2. Integración, limpieza y transformación
3. Reducción de datos
4. Datos imperfectos
5. Resumen

Preprocesamiento de datos

Integración, limpieza y transformación



Preprocesamiento de datos

Integración, limpieza y transformación

Integración

Combinación de datos de diferentes fuentes

Similar a ETL (*extraction, transformation & load*)

Tareas

Integración de esquema

Unificación de la codificación

Detección de duplicados e inconsistencias

Redundancias

Análisis de correlaciones

Preprocesamiento de datos

Integración, limpieza y transformación

Limpieza

Modificaciones para conseguir datos de más calidad

Tareas

Resolver inconsistencias

Rellenar valores perdidos (*)

Suavizar ruido (*)

Identificar *outliers*

(*) se estudiarán más adelante en el tema

W. Kim, B. Choi, E.-D. Hong, S.-K. Kim (2003) A taxonomy of dirty data. *Data Mining and Knowledge Discovery* 7, 81-99.

Preprocesamiento de datos

Integración, limpieza y transformación

Transformación

Convertir, derivar, resumir...

Tareas

Agregación

Generalización

Normalización

Otras transformaciones

T. Y. Lin (2002) Attribute Transformation for Data Mining I: Theoretical Explorations. *International Journal of Intelligent Systems* 17, 213-222.

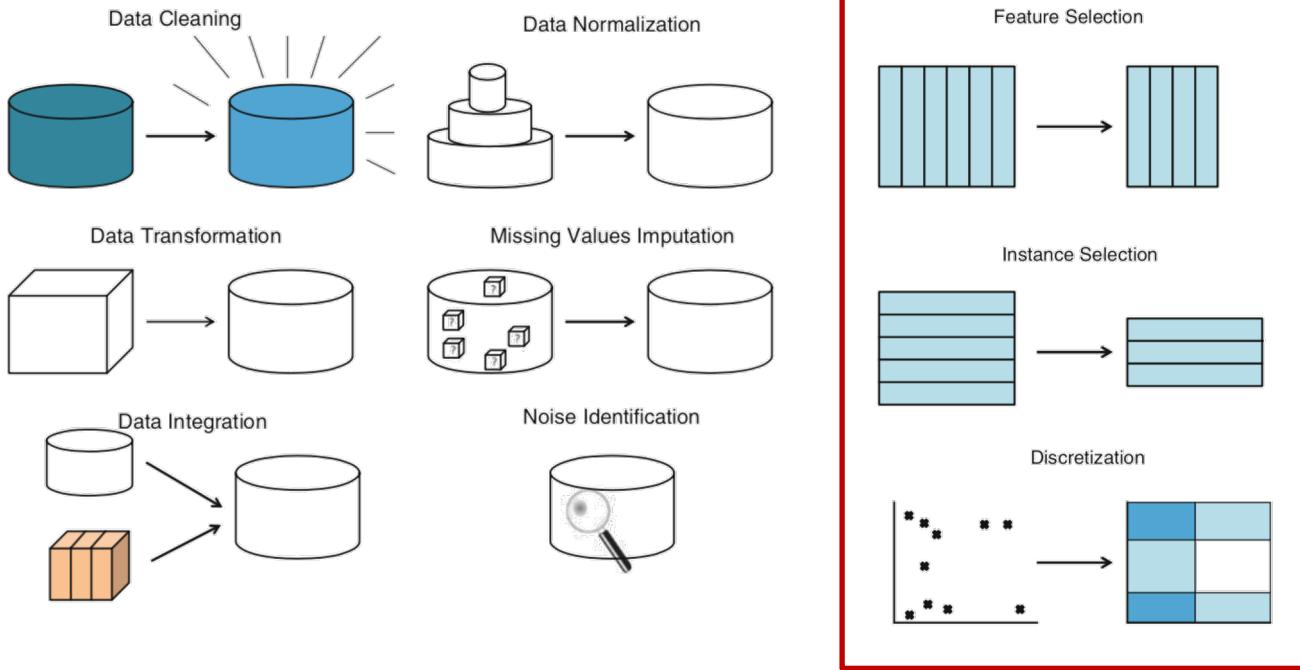
S. García, J. Luengo, F. Herrera (2015) *Data Preprocessing in Data Mining*. Capítulo 3 Data Preparation Basic Model. Springer.

Índice

1. Introducción
2. Integración, limpieza y transformación
- ▶ 3. Reducción de datos
4. Datos imperfectos
5. Resumen

Preprocesamiento de datos

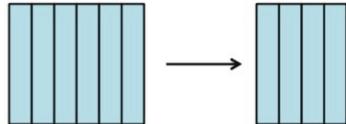
Reducción de datos



Preprocesamiento de datos

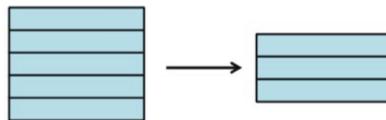
Reducción de datos

Feature Selection



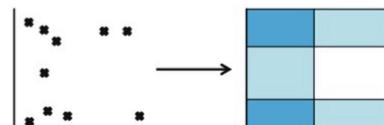
Reducir dimensionalidad >
Selección de características

Instance Selection



Eliminar muestras redundantes y
conflictivas >
Selección de ejemplos

Discretization



Simplificar dominio de una
variable >
Discretización



S. García, J. Luengo, F. Herrera (2015) Data Preprocessing in Data Mining. [Capítulo 7](#). Springer.

Preprocesamiento de datos

Reducción de datos ► Selección de características

Selección de características

Seleccionar un subconjunto de variables del problema que optimiza la creación un modelo de predicción correcto

- Menos datos → *Algoritmos más rápidos*
- Mayor precisión → *Más generalización*
- Resultados más simples → *Más interpretables*

Puede verse como un problema de búsqueda:

Encontrar el subconjunto de variables más pequeño que *optimice* la creación del modelo

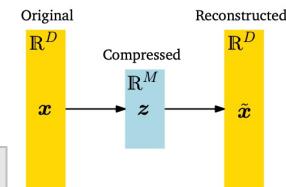
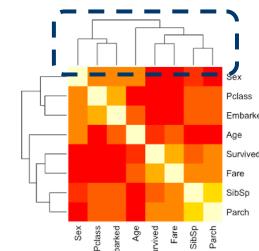
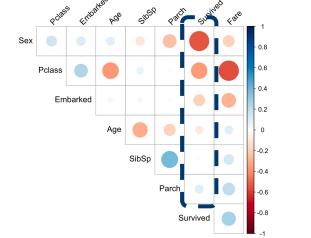
¿Cómo se evalúa qué subconjunto de variables es óptimo?

Preprocesamiento de datos

Reducción de datos ► Selección de características

Ideas básicas

- Quedarnos con variables que son buenos “predictores” de la variable objetivo
 - Variables con **alta correlación** con la variable objetivo
titanic : Sex, Pclass, Age, etc.
- No quedarnos con variables que ofrecen “la misma información”
 - De un grupo de variables con **alta correlación entre sí, quedarnos solo con una**
titanic : Pclass, Fare
- Descubrir la geometría de los datos y proyectarlos a un espacio de dimensión menor



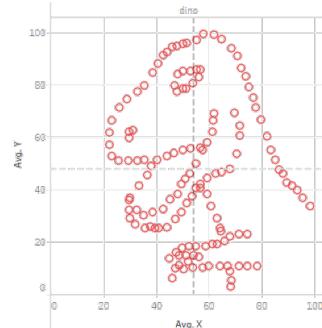
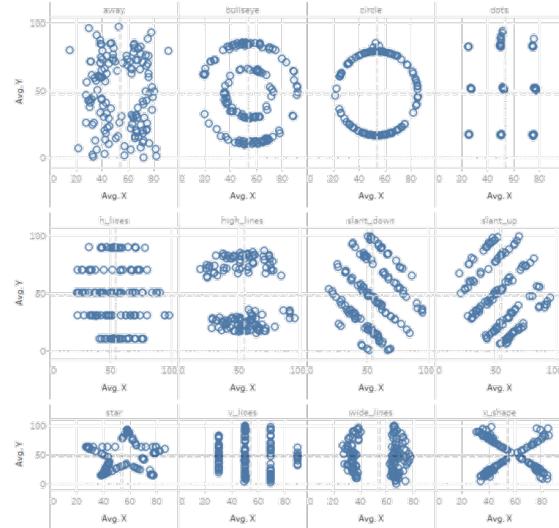
M.P. Deisenroth, A.A. Faisal, C.S. Ong (2020) *Mathematics for Machine Learning*. Cambridge University Press.

Preprocesamiento de datos

Reducción de datos ► Selección de características

Ojo con las correlaciones y otros estadísticos

Datasaurus Dozen



Todas estas distribuciones, incluida el dinosaurio, tienen los mismos parámetros estadísticos: media, varianza, etc.

J. Matejka, G. Fitzmaurice (2017) Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. ACM SIGCHI Conference on Human Factors in Computing Systems. (Online: <https://www.autodeskresearch.com/publications/samestats>)

Preprocesamiento de datos

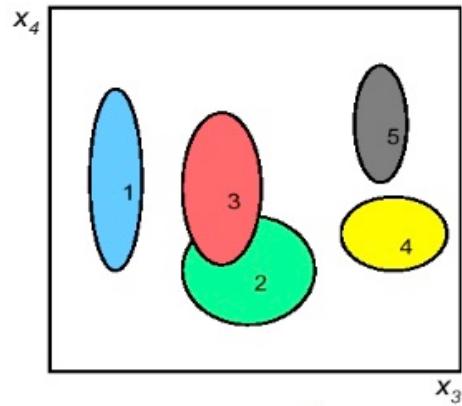
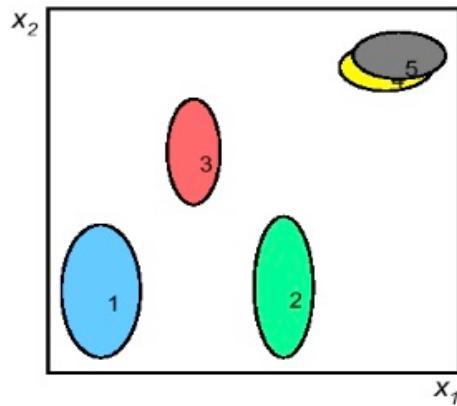
Reducción de datos ► Selección de características

¿Qué subconjunto de variables debería seleccionarse?

Supongamos un problema de clasificación con 4 variables $\{x_1, x_2, x_3, x_4\}$

Supongamos que la variable objetivo tiene 5 valores posibles $\{1, 2, 3, 4, 5\}$

Representamos los valores clasificados por grupos de dos variables



Preprocesamiento de datos

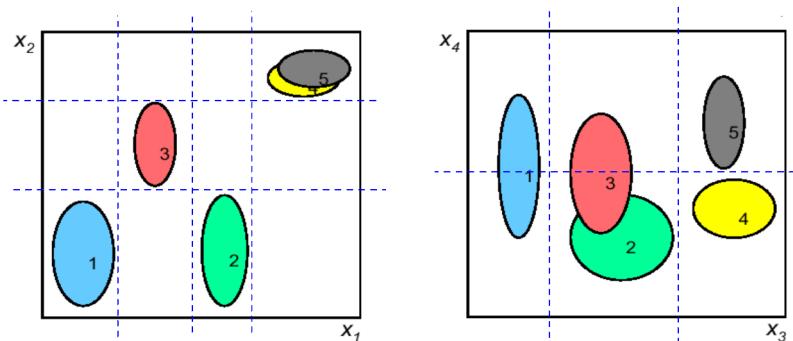
Reducción de datos ► Selección de características

¿Qué subconjunto de variables debería seleccionarse?

Si estudiamos las variables una por una en orden de “separabilidad”:

- x_1 es mejor que x_2 en x_1 se pueden definir 4 intervalos que separan las clases en 4 grupos: {1}, {2}, {3}, {4, 5} ; en x_2 con 3 intervalos hay 3 grupos
- x_3 es mejor que x_4
- x_1 es mejor que x_3

Por lo tanto, si seleccionamos dos variables solamente, tomaríamos $\{x_1, x_2\}$



Ejemplo detallado:

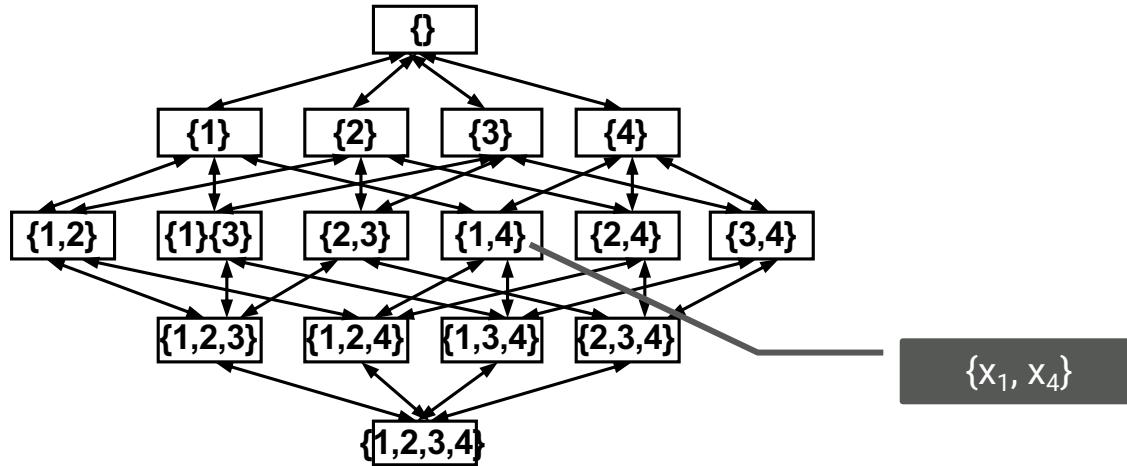
M. Kuhn, K. Johnson (2019) Feature Engineering and Selection: A Practical Approach for Predictive Models. [\[link\]](#)

Sin embargo, la mejor selección de variables sería $\{x_1, x_4\}$, pues se pueden definir intervalos en ellas para separar las 5 clases

Preprocesamiento de datos

Reducción de datos ► Selección de características

Búsqueda en el espacio de todas las posibles selecciones



¿Cómo se evalúa qué un subconjunto de variables es *bueno*?

Preprocesamiento de datos

Reducción de datos ► Selección de características

Filtro (*filter*)

Se realiza una evaluación $U(X)$ de un conjunto de variables X a partir de parámetros estadísticos sobre los valores de las variables en el conjunto de datos

- Distancias
- Correlaciones  Prácticas > lending_club_seleccion.Rmd
- Teoría de la información
- Geometría (PCA, descomposición espectral, tSNE, distancia)  Prácticas > titanic_seleccion.Rmd

Envolvente (*wrapper*)

Se realiza una evaluación $U(X)$ de un conjunto de variables X aprendiendo un modelo de aprendizaje con ellas y viendo si funciona mejor que los construidos con otros conjuntos

Mixta

Combinación de las anteriores

Preprocesamiento de datos

Reducción de datos ► Selección de características

Métodos de búsqueda

Cómo realizar la expansión del árbol de búsqueda

Selección hacia adelante

Comienza con un conjunto vacío, al que va añadiendo secuencialmente el atributo que maximiza $U(S \cup x_i)$

Selección hacia atrás

Comienza con el conjunto de todos las variables, del que se va eliminando secuencialmente el atributo que minimiza $U(S - x_i)$

Enfoques mixtos

Selección l-más r-menos

Repite l cálculos de x_+ y r cálculos de x_-

Selección bidireccional

Ejecución paralela de adelante y atrás

Selección flotante

Selección l-más r-menos que no fija a priori l y r

Construcción de árboles de decisión

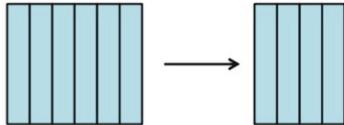
Identifican los predictores en orden de prioridad

Todos estos algoritmos **son lineales**: añaden o eliminan atributos de uno en uno y pueden caer en óptimos locales, como en el ejemplo anterior >> Añadir aleatoriedad

Preprocesamiento de datos

Reducción de datos

Feature Selection



Reducir dimensionalidad >
Selección de características

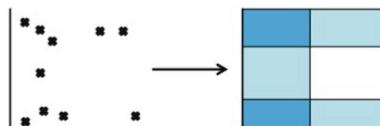
Instance Selection



Eliminar muestras redundantes y
conflictivas >

Selección de ejemplos

Discretization



Simplificar dominio de una
variable >

Discretización



S. García, J. Luengo, F. Herrera (2015) Data Preprocessing in Data Mining. [Capítulo 8](#). Springer.

Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos

Selección de ejemplos

Seleccionar un subconjunto de instancias del problema sin perder precisión en el modelo

Muestreo

Seleccionar ejemplos según una distribución de probabilidad (normal, aleatoria, etc.)

Selección de prototipos

Seleccionar los ejemplos más representativos de agrupaciones sobre los datos (con k-means u otros)

Aprendizaje activo

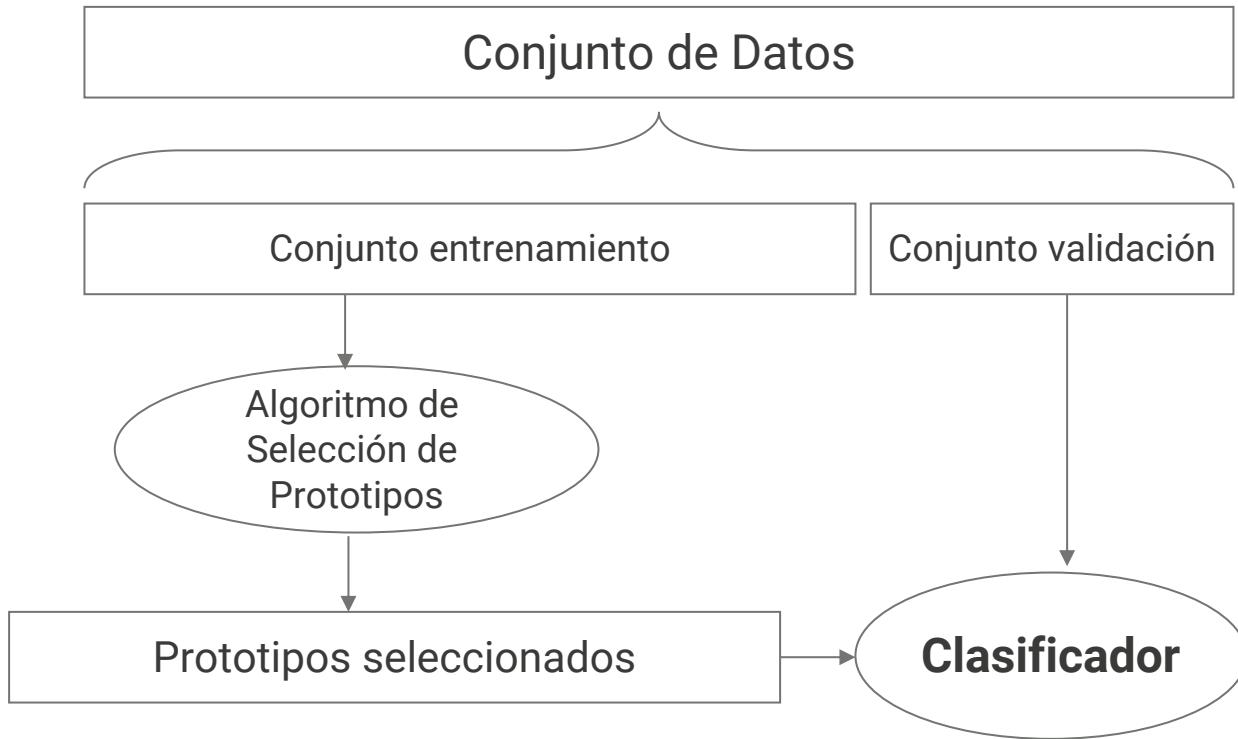
Seleccionar ejemplos según resulte el modelo predictivo aprendido

Por ejemplo, empezar con pocos ejemplos e ir subiendo mientras el proceso no tarde demasiado y se sigan mejorando los resultados

T. Reinartz (2002) A unifying view on instance selection. *Data Mining and Knowledge Discovery* 6, 191-210.

Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos



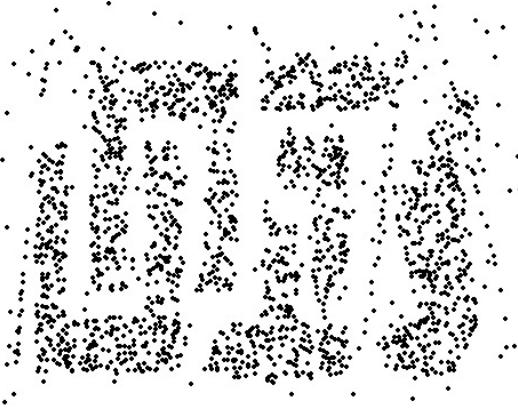
Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos

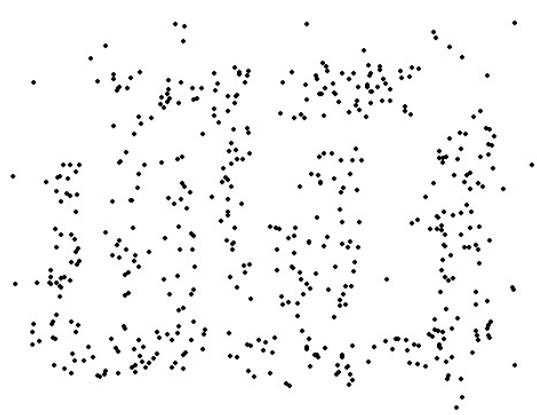
Algoritmo de
Selección de
Prototipos



8000 puntos



2000 puntos

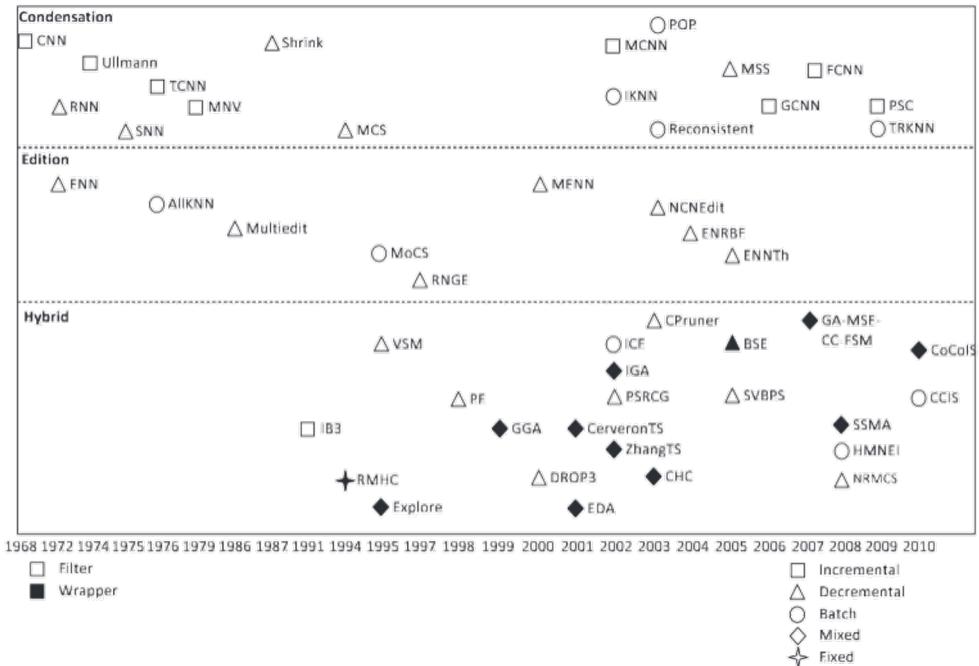


500 puntos

Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos

Algoritmo de Selección de Prototipos

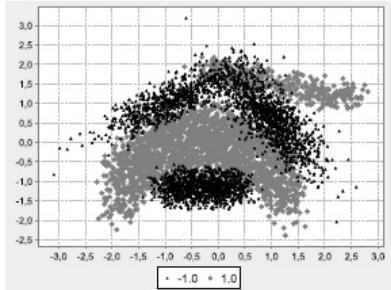


S. García, J. Luengo, F. Herrera (2015) Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(3), 417-435.

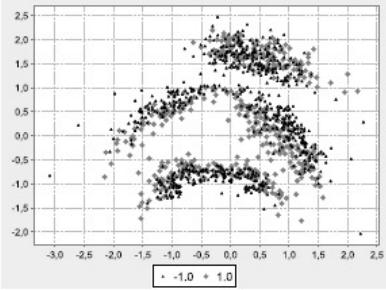
Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos

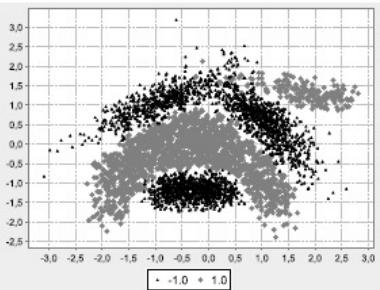
Algoritmo de
Selección de
Prototipos



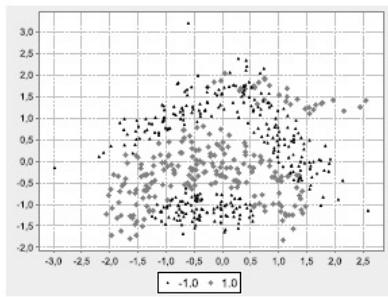
(a) Banana
(0.8751, 0.7476)



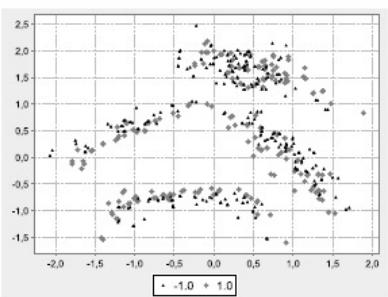
(b) CNN (0.7729, 0.8664, 0.7304)



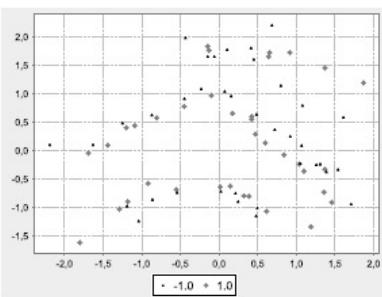
(h) AllKNN (0.1758, 0.8934, 0.7831)



(k) RMHC (0.9000, 0.8972, 0.7915)



(e) DROP3
(0.9151, 0.8696, 0.7356)



(l) SSMA (0.9879, 0.8964, 0.7900)

Preprocesamiento de datos

Reducción de datos ► Selección de ejemplos

Conjuntos de datos “no balanceados”

Problemas con presencia de clases desigual

Diagnóstico médico, e-commerce, ciberseguridad

Es sencillo obtener un clasificador con un alto porcentaje de clasificación correcta, pero no son útiles Si el 90% de los datos son de la clase A, un modelo de predicción que siempre diga A tendrá un 90% de acierto

Aproximaciones relacionadas con la selección de ejemplos

Reducción de datos de las clases mayoritarias

Sobremuestreo de las clases minoritarias

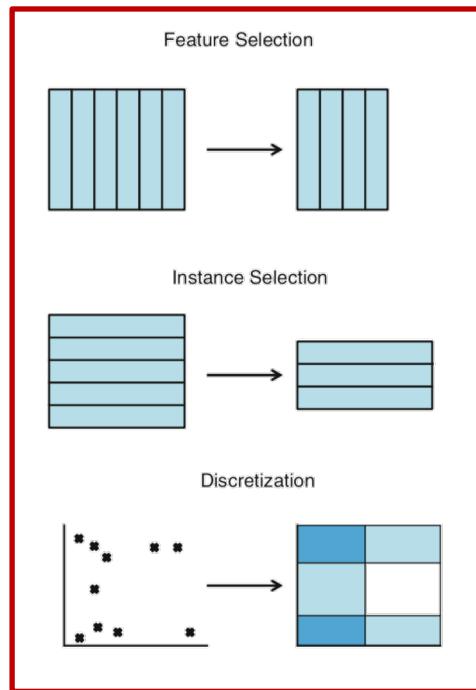
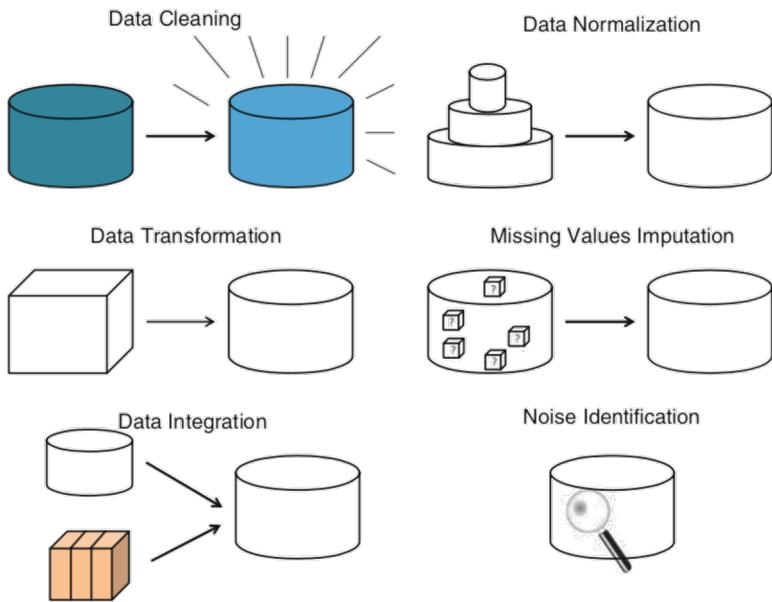
Generación de instancias artificiales (SMOTE)

Hibridación entre selección de instancias y características

Aproximaciones relacionadas con el proceso de aprendizaje [Tema 3]

Preprocesamiento de datos

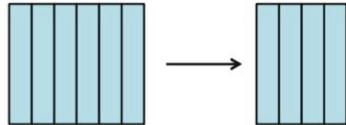
Reducción de datos



Preprocesamiento de datos

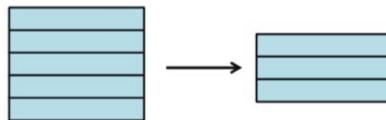
Reducción de datos

Feature Selection



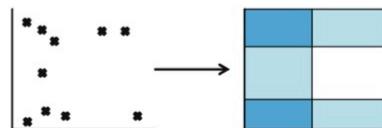
Reducir dimensionalidad >
Selección de características

Instance Selection



Eliminar muestras redundantes y
conflictivas >
Selección de ejemplos

Discretization



Simplificar dominio de una
variable >
Discretización

Preprocesamiento de datos

Reducción de datos ► Discretización

Discretización de valores

Transformar valores ordenados (numéricos) en valores nominales (categorías o intervalos)
Los valores discretos son más fáciles de manejar en aprendizaje automático

Manual

Dirigido por el experto y el analista de datos

Algoritmos no supervisados

Intervalos de igual amplitud

Intervalos de igual frecuencia

Clustering

Algoritmos supervisados

Entropía

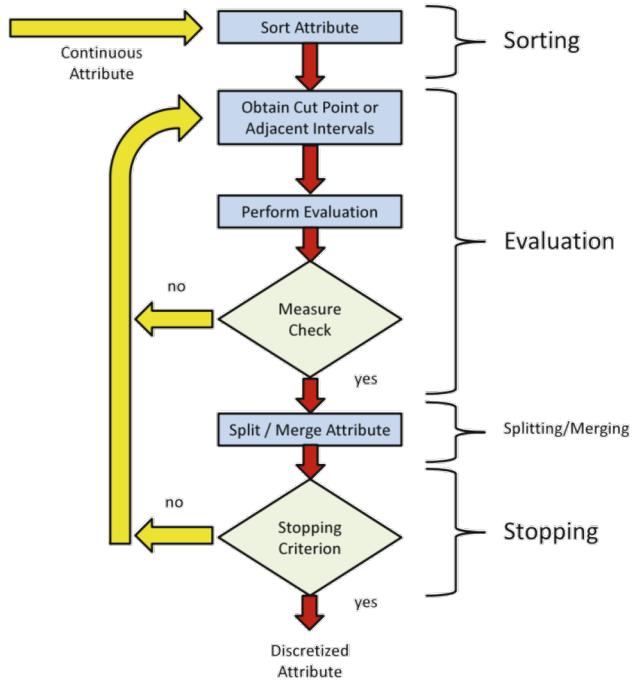
Chi-square

¿Cuál es mejor? Igual que con selección de variables e instancias: los que permitan trabajar con menos datos sin perder calidad del clasificador

Preprocesamiento de datos

Reducción de datos ► Discretización

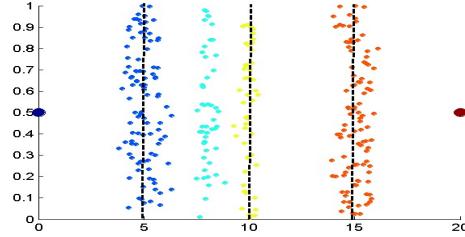
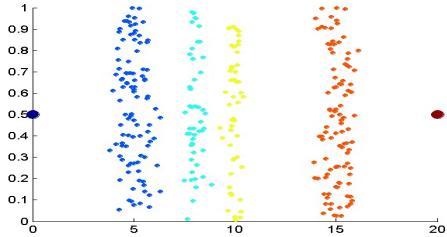
Proceso general



Preprocesamiento de datos

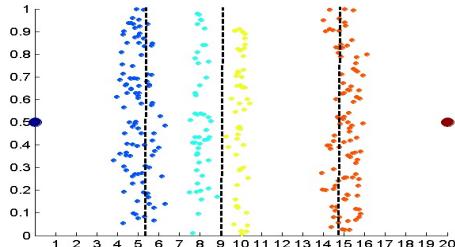
Reducción de datos ► Discretización

Algoritmos no supervisados



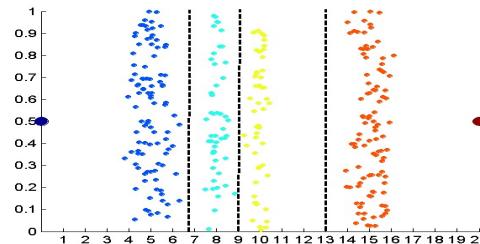
Igual anchura de intervalo

Separar valores en n intervalos de igual longitud



Igual frecuencia de valores

Separar valores en n intervalos con el mismo número de valores



K-medias

Aplicar k -medias con n clusteres

Preprocesamiento de datos

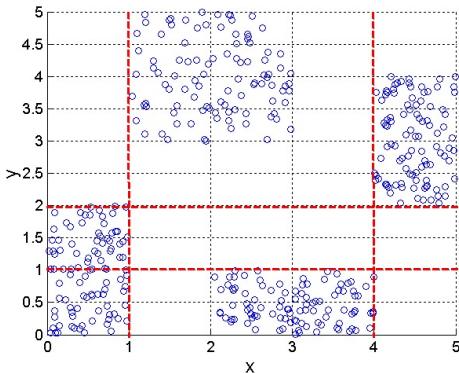
Reducción de datos ► Discretización

Algoritmos supervisados

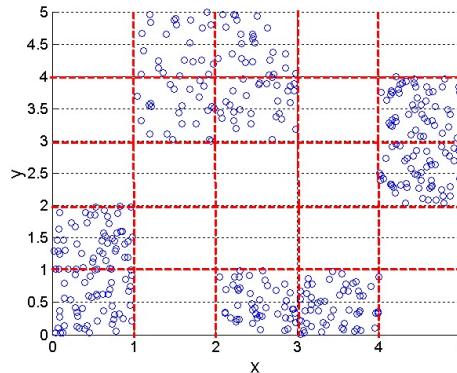
Se tienen en cuenta los valores de la clase objetivo

Se definen intervalos de forma que se agrupan o separan los valores de clasificación

Pueden utilizarse algoritmos de clasificación (por ejemplo, CART)



3 intervalos



5 intervalos

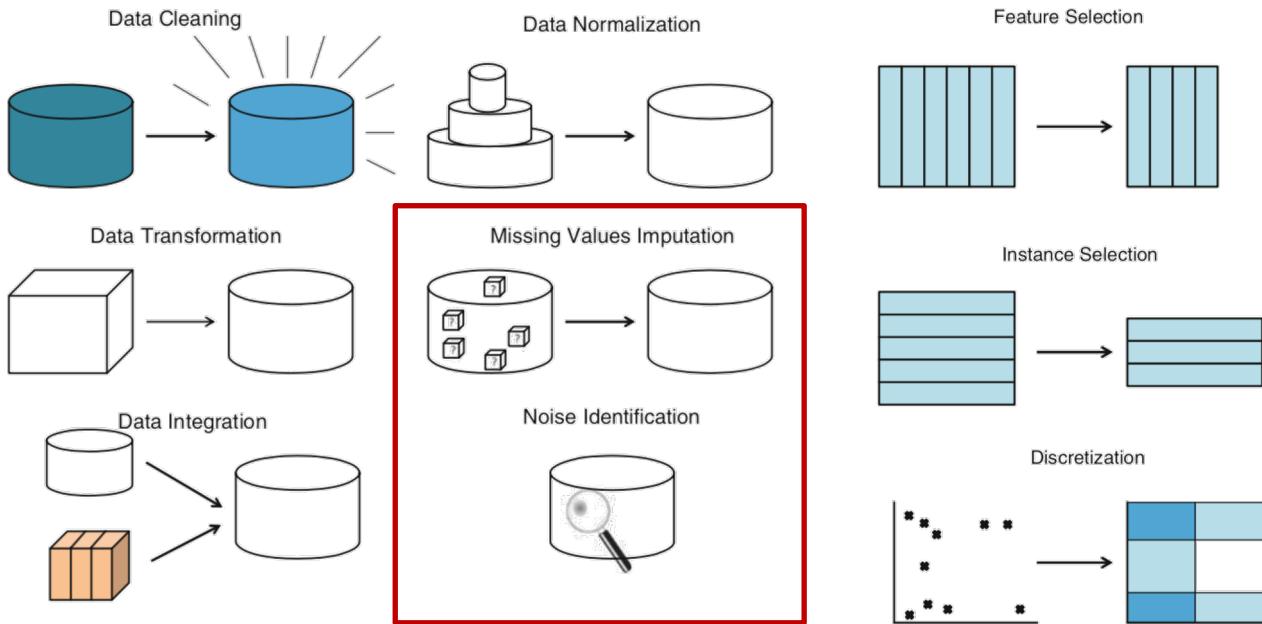
Los valores **O** son la clase positiva (en el ejemplo se omite la clase negativa)

Índice

1. Introducción
2. Integración, limpieza y transformación
3. Reducción de datos
- ▶ 4. Datos imperfectos
5. Resumen

Preprocesamiento de datos

Datos imperfectos



Preprocesamiento de datos

Datos imperfectos ► Valores perdidos

Valores perdidos

Valores no disponibles por diversos motivos

Pueden repartirse de forma aleatoria o no

sensor de temperatura falla aleatoriamente

sensor de temperatura falla en el 0.1% de las veces

sensor de temperatura falla por encima de 45°C

sensor de temperatura no funciona cuando la humedad está por encima del 75%

Procesamiento

Eliminar

Asignar manualmente

Asignar valor global

Rellenar con media/desviación

Rellenar con valor más “probable” << **Imputación de valores perdidos**

Preprocesamiento de datos

Datos imperfectos ► Valores perdidos

MICE

Multivariate Imputation by Chained Equations

Software para imputar valores perdidos o incompletos basado en FCS (*fully conditional specification*)

Sustituir valores perdidos por "valores más probables", estimados mediante inferencia a partir del resto del dataset

Facilita:

Crear modelos predictivos para imputar valores perdidos

Utilizar datasets con diferentes tipos de imputaciones para el análisis (*pooling*)

Alternativas

- Amelia, caret (kNN y otros)



S. van Buuren (2018) *Flexible Imputation of Missing Data*. Capítulos [1.1](#), [1.3](#). CRC Press.

Preprocesamiento de datos

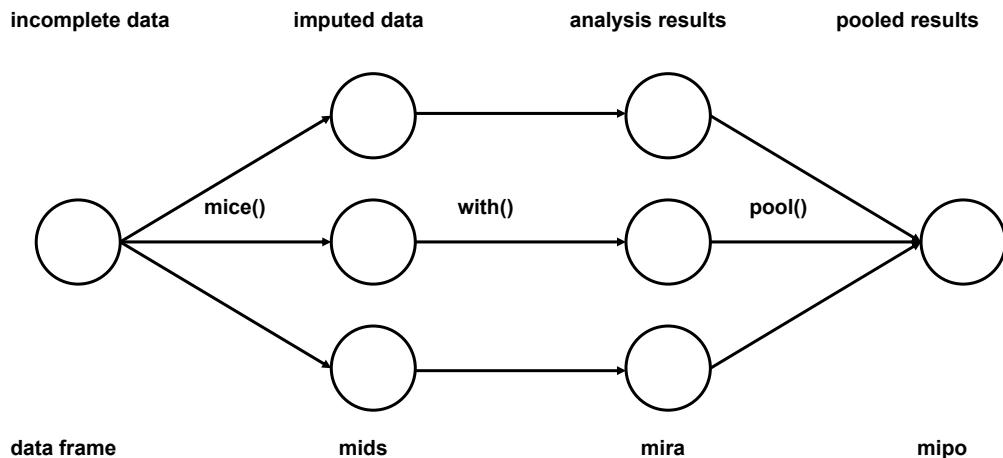
Datos imperfectos ► Valores perdidos

MICE

Imputación múltiple

Se realizan varias imputaciones alternativas a la vez

Después se puede trabajar con los *datasets* resultantes de cada imputación por separado o de forma conjunta



S. van Buuren (2018) Flexible
Imputation of Missing Data.
Capítulo 1.4. CRC Press.

Preprocesamiento de datos

Datos imperfectos ► Valores perdidos

MICE

Procedimiento para realizar la imputación

1. MAR (*Missing at Random*) vs MNAR (*Missing Not at Random*)
2. Forma del modelo de imputación (estructura, distribución del error)
3. Conjunto de variables que se usarán como predictores
4. Imputar variables derivadas de variables incompletas
5. Orden de imputación de las variables
6. Imputaciones iniciales y número de iteraciones
7. Número de datasets a la salida



titanic-missing-noise.Rmd

Preprocesamiento de datos

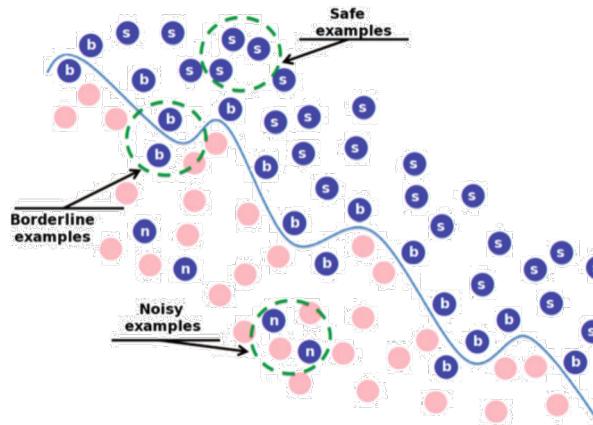
Datos imperfectos ► Valores con ruido

Valores con ruido

Valores incorrectos

sensor de temperatura da lectura incorrecta aleatoriamente

... (se aplican las mismas consideraciones que con los valores perdidos)



Preprocesamiento de datos

Datos imperfectos ► Valores perdidos

NoiseFiltersR

Implementación de algoritmos de preprocesamiento para tratamiento de ruido de clase en problemas de clasificación

Eliminan las observaciones identificadas como ruidosas o modifican la clase asignada

Métodos basados en distancia (vecindario) o clasificación

Sintaxis

Dataset

Fórmula describiendo la etiqueta con ruido y las clases que se utilizarán para calcular la probabilidad de ruido

Salida

Datos sin ruido

Vector de índices eliminados

J. Luengo (2016) *NoiseFiltersR: Label Noise Filters for Data Preprocessing in Classification* [[link](#)]

Preprocesamiento de datos

Datos imperfectos ► Valores perdidos

NoiseFiltersR

- AENN: All-k Edited Nearest Neighbors
- BBNR: Blame Based Noise Reduction
- C45ensembles: Classical Filters based on C4.5
- CNN: Condensed Nearest Neighbors
- CVCF: Cross-Validated Committees Filter
- DROP: Decremental Reduction Optimization Procedures
- dynamicCF: Dynamic Classification Filter
- edgeBoostFilter: Edge Boosting Filter
- EF: Ensemble Filter
- ENG: Editing with Neighbor Graphs
- ENN: Edited Nearest Neighbors
- EWF: Edge Weight Filter
- GE: Generalized edition
- HARF: High Agreement Random Forest

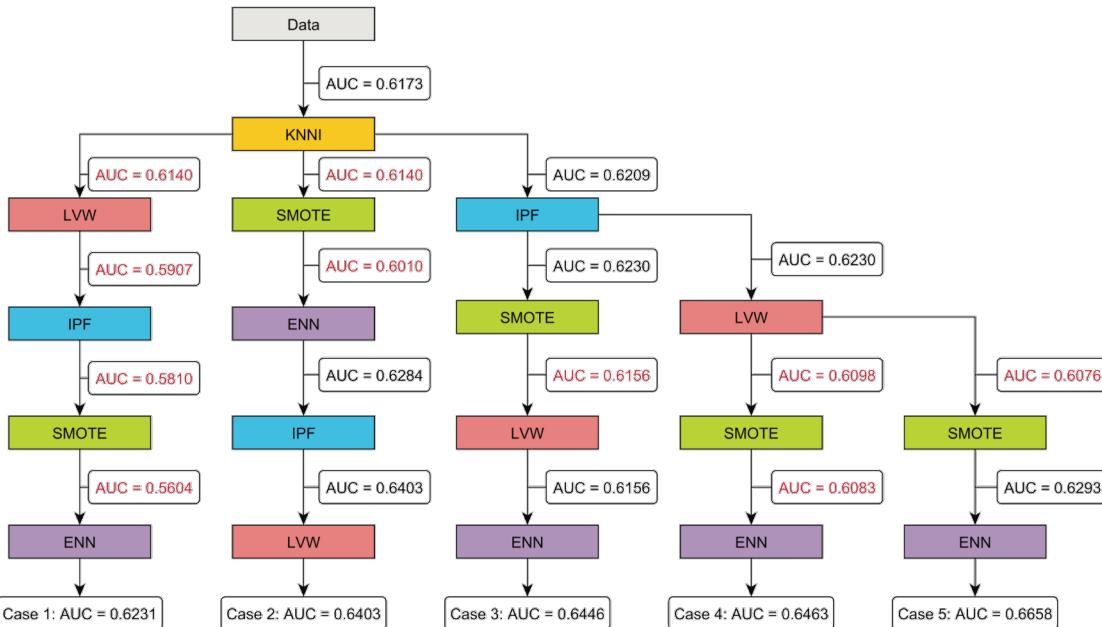
- hybridRepairFilter: Hybrid Repair-Remove Filter
 - INFCC: Iterative Noise Filter based on the Fusion of Classifiers
 - IPF: Iterative Partitioning Filter
 - ModeFilter: Mode Filter
 - ORBoostFilter: Outlier Removal Boosting Filter
 - PF: Partitioning Filter
 - PRISM: Preprocessing Instances that Should be Misclassified
 - RNN: Reduced Nearest Neighbors
 - saturationFilter: Saturation Filters
 - TomekLinks: TomekLinks
-
- summary: Summary method for class filter

Índice

1. Introducción
2. Integración, limpieza y transformación
3. Reducción de datos
4. Datos imperfectos
- ▶ 5. Resumen

Preprocesamiento de datos

Resumen



S. García, J. Luengo, F. Herrera (2016) Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge Based Systems 98, 1-29.