Tema 5 - Preprocesado de datos

▼ Integración

- Obtención de datos de diferetes fuentes de información → emparejamiento correcto
- Resolución de problemas de representación y codificación

▼ Integración de datos desde diferentes tablas para crear información homogénea

- Detección de datos duplicados
- Eliminación de redudancia mediante análisis de correlaciones para selección de atributos → medición de la fuerza con la que un atributo implica a otro

▼ Limpieza

- Resolver inconsistencias
- Rellenar valores perdidos
- Suavinzar ruido
- Identificar y eliminar outilers

▼ Transformación

- Transformación de datos para mejorar aplicación de algoritmo de minería de datos

▼ Operaciones

- Agregación
- Generalización
- Transformación de variables

▼ Normalización

- Utililzada para algoritmos basados en instancias
- Muy sensible a outilers (no necesariamente a ruido)

- Datos imperfectos
  - Valores perdidos
    - Soluciones
      - Ignorar la tupla → útil cuando la variable no tiene valor
      - Rellenar manualmente la tupla → impracticable
      - Usar constante global para sustitución (“desconocido”)
      - Rellenar usando media/desviación típica del resto de tuplas (de la misma clase o no)
      - Utilizar técnicas de inferencia (árboles de decisión) para rellenar con el valor más probable
  - Ruido
    - Tipos de ruido
      - Ruido de clase
        - Ejemplos contradictorios
        - Ejemplos sin etiquetar
      - Ruido de atributos
        - Valores erróneos
        - Valores incompletos/vacíos
        - Valores sin importancia
    - Técnicas de eliminación de ruido
      - EF
        - Uso de diferentes algoritmos de aprendizaje
        - Para cada algoritmo
          - Validación cruzada pra marcar cada ejemplo de validación como correcto o mal etiquetado (en función de si la predicción coincide con la etiqueta)
        - Esquema de votación
          - Por consenso

- Elimina ejemplo si es mal clasificado por todos los clasificadores

▼ Por mayoría

- Elimina ejemplo si es mal clasificado por más de la mitad de los clasificadores

▼ CVCF

- Similar a EF pero con matices
- Mismo algoritmo de aprendizaje → árboles de decisión (C4.5)
- Validación cruzada para todos los ejemplos de entrenamiento (no sólo ejemplos de validación)

▼ IPF

- Elimina datos ruidosos en múltiples iteraciones con CVCF hasta alcanzar criterio de parada
- Proceso detenido si durante varias iteraciones consecutivas el nº de ejemplos ruidosos es inferior a un porcentaje del CE

▼ Reducción de datos

▼ Discretización

▼ Técnicas

▼ Igual frecuencia

▼ Ventajas/Desventajas

Ventajas	Desventajas
Evita desequilibrios en balanceo de valores	Obliga a cajas para valores especiales
Puntos de corte más intuitivos	Obliga a puntos de corte interpretables

▼ Selección de características

▼ Características

- Menos datos
- Mayor exactitud

- Resultados más simples

▼ Estructura de algoritmo de SC

- Estrategia de búsqueda para selccionar subconjuntos candidatos

▼ Función objetivo que evalúe esos subconjuntos

▼ Enfoques

▼ Filtro

- Evaluación de subconjuntos según la información que contienen

▼ Medidas

- Separabilidad
- Correlaciones
- Consistencia

▼ Envolvente

- Aplicación de técnica de aprendizaje que se utilizará finalmente sobre la proyección de los datos al subconjunto candidato
- Devuelve el acierto del clasificador construido

▼ Ventajas/Desventajas

	Ventajas	Desventajas
Filtro	- Velocidad - Generalidad	- Tendencia a incluir muchas variables
Envolvente	- Exactitud (eficacia) - Regulación de sobreajuste	- Muy costosos - Pérdida de generalidad

▼ Algoritmos de selección

▼ Hacia adelante

- Comienza con el conjunto vacío y se van añadiendo atributos al subconjunto actual

        - Funciona mejor cuando el subconjunto óptimo tiene pocas variables
        - No elimina variables
    - ▼ Hacia atrás
        - Comienza con el conjunto completo y se van eliminando atributos del subconjunto actual
        - Funciona mejor cuando el subconjunto óptimo tiene muchas variables
        - Problema al reevaluar la utilidad de algunos atributos previamente descartados
    - ▼ l-más r-menos
        - Generalización hacia adelante y hacia detrás
    - ▼ Bidireccional
        - Implementación paralela de forward y backward
        - Hay que asegurar que atributos eliminados no son introducidos por el otro algoritmo
- ▼ Selección de instancias
    - Elección de ejemplso relevantes para una aplicación
    - ▼ Características
        - Menos datos
        - Mayor exactitud
        - Modelos generados más simples
    - ▼ Técnicas
        - ▼ Procesamiento de clases no balanceadas
            - Reduciendo las clases mayoritarias
            - Sobremuestro de clases minoritarias
    - ▼ Algoritmos
        - ▼ CNN
            - Incremental

- Inserta sólo las instancias mal clasificadas
- Dependiente del orden de presentación
- Retiene puntos pertenecientes al borde

▼ ENN

- Por lotes
- Se eliminan aquellas instancias mal clasificadas
- Suaviza fronteras pero retiene resto de puntos