Minería de Datos

Tema 1 - Preprocesamiento.

Nombre: Alejandro pérez Belando

---

1. Introducción al Preprocesamiento

La calidad y cantidad de los datos son cruciales para la minería de datos. Los datos se representan mediante objetos descritos por atributos.

2. Limpieza de Datos

Los datos del mundo real pueden contener errores debido a fallos en la adquisición de datos, errores humanos o computacionales. Estos errores se manifiestan como datos incompletos, ruidosos, atípicos, duplicados o inconsistentes.

2.1. Datos Ausentes

Problemas: Pueden reducir la eficacia del modelo, complicar el análisis y sesgar los resultados.
Detección: Los datos ausentes pueden representarse como valores nulos o estar camuflados. Es crucial comprender la causa de estos datos faltantes.
Soluciones:

- Ignorar los datos ausentes si el método es robusto.
- Filtrar atributos o eliminar objetos con valores nulos (puede introducir sesgos).
- Reemplazar por valores que conserven media o varianza (para datos numéricos) o la moda (para nominales). Otra opción es sustituir mediante imputación (valor medio, más probable o predecirlo).
- Crear un atributo adicional que indique que el dato estaba ausente.

2.2. Datos con Ruido

Definición: Se define como varianza o error aleatorio en las mediciones.
Soluciones:

- Discretización: Suaviza los datos consultando su vecindad. Los valores ordenados se distribuyen en categorías con, o bien: mismo N° de elementos (eq. frequency) o con el mismo tamaño (eq. width). Otra opción es sustituir los valores de cada categoría por la media, mediana o extremo más cercano.
- Regresión: Ajusta una función y sustituir los valores por los predichos.
- Clustering: Identifica valores atípicos.

2.3. Datos Inconsistentes y Discrepancias

Causas: Errores en la entrada de datos, obsolescencia o inconsistencia en los formatos.
Detección: Se utilizan metadatos y reglas de unicidad.
Herramientas: Herramientas de depuración (detectar errores con el conocimiento del tema) y auditoría (encontrar discrepancias con un análisis que nos de las reglas y relaciones entre datos) de datos.

2.4. Variables con Varianza Cercana a Cero (tienen un solo valor)

Problema: Pueden causar problemas en ciertos modelos.
Detección:

- Ratio de frecuencia $= \dfrac{\text{Frecuencia variable más frecuente}}{\text{Frecuencia variable que estudio}}$
  Si $Ratio \approx 1$, variable balanceada. Si $Ratio \gg 1$, variable mal balanceada (varianza cercana a 0).
- Porcentaje de valores únicos $= \left(\dfrac{N^\circ\text{valores únicos}}{\text{Total entradas en la variable}}\right)$

Solución: Eliminar estas variables.

3. Transformación de Datos

Objetivo: Prepara los datos para las técnicas de minería de datos.

3.1. Técnicas

la mayor parte de estas técnicas es de aplicación sobreyectiva: un valor transformado se puede generar a través de uno o varios valores originales.

- Suavizado: Elimina el ruido.
- Agregación: Resume o agrega datos.
- Generalización: Transforma datos de bajo nivel a un nivel más alto.
- Creación de atributos: A partir de los existentes.
- Normalización: Escala los datos a un rango específico, como $[0, 1]$ o $[-1, 1]$. Evita que las variables con rangos mayores dominen.

3.1.1. Métodos de Normalización

Normalización Min-Max: Transformación lineal de una variable A cuyo rango es $[min_A, max_A]$. Los nuevos valores v' en el nuevo rango $[min'_A, max'_A]$:

$$v' = \frac{(v - \text{mín}_A)}{(\text{máx}_A - \text{mín}_A)} \times (\underset{A}{\text{máx}}' - \underset{A}{\text{mín}}') + \underset{A}{\text{mín}}' \tag{1}$$

Esta transformación mantiene la relación de los datos originales.
Normalización Z-score: Los valores de A son normalizados en función de su media ($\bar{A}$) y desviación típica (σ_A):

$$v' = \frac{(v - \mu_A)}{\sigma_A} \tag{2}$$

Se emplea cuando los rangos de las variables son desconocidos o hay valores atípicos.
Escalado Decimal: Desplazamiento del punto decimal de los valores del atributo. Las posiciones que se desplaza depende del valor absoluto máximo de la variable A.

$$v' = \frac{v}{10^j} \tag{3}$$

3.2. Discretización

Convierte valores numéricos en nominales ordenados. Los tipos de discretización:

- Supervisada: si la técnica usa la información sobre la clase.
- No supervisada: en caso contrario.
- Global: los métodos aplican los mismos puntos de corte a todas las instancias.
- Local: los métodos aplican distintos puntos de corte a distintos conjuntos de instancias.
- Descendente (Top-Down) (splitting): se dividen los rangos hasta que no se puede separar más.
- Ascentente (Bottom-Up) (merging): se fusionan puntos cercanos entre sí para formar intervalos.

Técnicas:

- Binning (descendente, no supervisada):
  - Por intervalos de la misma longitud (equal-width): se divide el rango de valores en intervalos de igual longitud:
    $$w = \frac{V_{max} - V_{min}}{N}$$
    Los límites de los intervalos: $V_{min} + w,\ V_{min} + 2w,\ V_{min} + (N-1)w$.
  - Por intervalos de la misma amplitud (equal-depth, frecuency): se divide el rango de valores en iuntervalos que contengan aproximadamente el mismo número de elementos:
    $$N^{\circ}\ Elementos = \frac{N^{\circ}\ Instancias}{N^{\circ}\ Intervalos}$$
    Para determinar los valores que hacen la partición se usa el punto medio entre los dos extremos de los intervalos.
- Histogramas (descendente, no supervisada): muestra la frecuencia de cada uno de los posibles valores del atributo agrupando en intervalos, o *buckets* (pares valor-frecuencia). Los *buckets* pueden ser de la misma longitud, misma frecuencia, de varianza óptima (se consideran todas las posibilidades de agrupación en *buckets* y se selecciona la de menor varianza) o de máxima diferencia (Si deseamos β intervalos, sus límites se establecen entre valores consecutivos con las $\beta - 1$ mayores distancias).

  Criterios [1]:
  - **Raíz cuadrada**:
    $$N_{intervalos} = \sqrt{N} \quad ; \quad w = \frac{max(x) - min(x)}{\sqrt{(n)}}$$
  - **Sturges**:
    $$N_{intervalos} = 1 + log_2(N) \quad ; \quad w = \frac{max(x) - min(x)}{1 + log_2(N)}$$
  - **Rice**:
    $$N_{intervalos} = 2\sqrt[3]{N} \quad ; \quad w = \frac{max(x) - min(x)}{2\sqrt[3]{N}}$$
  - **Scott**:
    $$N_{intervalos} = \frac{3{,}5 \cdot \sigma}{\sqrt[3]{N}} \quad ; \quad w = \frac{max(x) - min(x)}{\frac{3{,}5 \cdot \sigma}{\sqrt[3]{N}}}$$

[1] En los dos últimos lo he puesto como yo creo que es. Me da que el profesor se lió a la hora de hacer la presentación.

- **Freedman-Diaconis**:

$$N_{intervalos} = \frac{2 \cdot IQR(x)}{\sqrt[3]{N}} \quad ; \quad w = \frac{max(x) - min(x)}{\frac{2 \cdot IQR(x)}{\sqrt[3]{N}}}$$

- Entropía (descendente supervisada). Proceso:

  1. Cálculo de la entropía: $H(s) = -\sum p_i \log_2 p_i$ ($p_i \equiv$ proporción de valores del atributo de la clase i)
  2. Selección de puntos clave: se divide el conjunto de datos en dos subconjuntos (S_{izq} y S_{der})
  3. Cálculo de la Ganancia de Información:

$$IG(T) = H(S) - \left(\frac{|S_{izq}|}{|S|} H(S_{izq}) + \frac{|S_{der}|}{|S|} H(S_{der}) \right)$$

  4. Elegir el punto que maximiza la ganancia de información.
  5. Repetir hasta alcanzar un criterio de parada:
     - La ganancia de información es menor que un umbral.
     - Se alcanza el número mínimo de instancias por intervalo.

- Fusión de intervalos mediante análisis χ^2 (ascendente, supervisado): fusiona intervalos adyacentes que tengan distribuciones de clases similares. proceso:

  1. Inicialización: ordenar los valores de característica contínua.
  2. Bining inicial: cada valor único es un bin separado.
  3. Cálculo χ^2 para cada pareja de bins adyacentes.
  4. Fusión de pares de bins adyacentes con el menor χ^2.
  5. Condición de parada: Nº intervalos, umbral χ^2...
  6. Bins finales: los que mejor preservan la relación con la variable objetivo.

- Análisis de Clusters (ascendente o descendente, no supervisado): para discretizar un atributo numérico, asociando una categoría a cada grupo (cluster). Hay que tener en cuenta la distribución del atributo y la distancia entre los datos.

3.3. De Variables Categóricas a Numéricas

Variable categórica: aquella cuyo dominio lo forman un número finito de categorías. existen las nominales (categorías no relacionadas) y ordinales (categorías ordenadas) **Técnicas de codificación**:

- **Codificación Ordinal**: Asigna enteros a cada categoría manteniendo el orden. Hay que ser cuidadosos al aplicarla a la variable a predecir, normalmente debe mantenerse como categoría.
- **One-Hot**: se aplica a las variables categóricas ordinales. Trata de crear una nueva variable binaria para cada categoría. Cada nueva variable toma el valor 1 si presenta la categoría, y 0 en caso contrario. Sin embargo, puede introducir información redundante.
- Variables Dummy: Similar a One-Hot, pero para N categorías, se crean $N-1$ variables. Esa categoría excluida se codifica tenniendo un 0 en el resto de variables.

4. Datos Desbalanceados

Problema: Ocurre cuando una clase tiene una proporción mucho menor que las otras, afectando el entrenamiento y la evaluación del modelo. Algunas soluciones pueden ser: usar otras técnicas de muestreo, otras medidas de rendimiento en la evaluación del modelo u otros modelos

4.1. Técnicas de Muestreo

- Downsampling: Seleccionar aleatoriamente un subconjunto de todas las clases para que sus frecuencias se ajusten a la clase minoritaria (reduce la clase mayoritaria).
- Upsampling: de la misma forma pero hace que las frecuencias se ajusten a la clase mayoritaria.
- SMOTE (Synthetic Minority over-sampling Téchnique): Genera muestras sintéticas de la clase minoritaria basadas en vecinos cercanos.
- ROSE (Random Over-Sampling Examples): Genera nuevas muestras en la vecindad de las existentes para equilibrar las frecuencias de las clases. Se emplea para problemas de clasificación binaria.

4.2. Datos desbalanceados: medidas de Rendimiento

- Matriz de confusión.
- Precisión (valor predictivo positivo): Exactitud de la predicción de la clase minoritaria. $\frac{VP}{VP+FP}$
- Recall (sensibilidad): capacidad del modelo de predecir la clase minoritaria. $\frac{VP}{VP+FN}$
- F1 Score: media ponderada de Precisión y Recall
- Área bajo la curva ROC: miede la capacidad del modelo de diferenciar observaciones entre clases.

Las combinaciones entre Precisión y Recall nos aportan información:

- **Precisión alta y Recall alto**: la clase es perfectamente detectada.
- **Precision alta y Recall bajo**: el modelo no puede detectar a clase, pero cuando lo hace es muy fiable.
- **Precisión baja y Recall alto**: la clase es detectada aceptablemente pero también incluye muestras de otras clases.
- **Precisión baja y Recall bajo**: el modelo no puede detectar la clase.

4.3. Datos desbalanceados: modelos

- Algoritmos optimizados para datos desbalanceados: SVM Y KNN.
- Aprendizaje sensitivo al costo: se puede dar mayor importancia a los falsos positivos de la clase mayoritaria o a los verdaderos positivos de la clase minoritaria. Puede causar sobreajuste.
- Métodos Ensemble: para reducir la varianza de la clasificación (SMOTEBoost, RUSBoost, ...)
- One-Class learning: el modelo es entrenado para representar adecuadamente la clase desbalanceada.