Reconocimiento de Formas

Reducción de la dimensionalidad

Índice:

- Selección de características.
- Estrategias de búsqueda.
- Función criterio:
  - Wrapper: Error de clasificación.
  - Filter: Separación entre clases, medidas de dependencia y medidas de información.
- Ponderación de características.
- Extracción de características.

Diferentes Perspectivas

- Selección: obtener un subconjunto del conjunto inicial de atributos.
- Ponderación: aplicar un peso a cada una de las características iniciales.
- Extracción: obtener un conjunto a partir de la transformación o combinación de los atributos originales.

Problema de la Selección

- Buscar un subconjunto reducido de *d* características a partir de los *D* atributos iniciales que componen el vector de características de las muestras.
  - Reducción de la dimensionalidad del espacio de características.

Objetivos

- Reducir el coste computacional asociado a una regla de clasificación.
  - Eliminando atributos irrelevantes.
  - Eliminando atributos redundantes.
- Aumentar la precisión (tasa de aciertos) de una regla de clasificación.
  - Eliminando atributos “dañinos”.
  - Reduciendo el número de atributos cuando se dispone de pocas muestras de entrenamiento:
    - Ratio entre número de muestras y número de características.

Posibles Aplicaciones

- Fusión de la información de múltiples sensores (Information Fusion).
- Minería de datos (Data Mining).
- Análisis de imágenes de percepción remota (Remote Sensing).
- Otras con grandes volúmenes de datos.

Formulación Matemática

- Problema de optimización combinatoria:
  - Dado un conjunto Y de D características, escójase un subconjunto $X \subseteq Y$ de talla d que optimice una cierta función criterio $J(X)$.

$$J(X) = \max_{\substack{Z \subseteq Y \\ |Z|=d}} J(Z)$$

Solución Trivial

- Realizar una búsqueda exhaustiva del mejor subconjunto de $d \leq D$ características.
  - Examinar todos los posibles subconjuntos de *d* características:

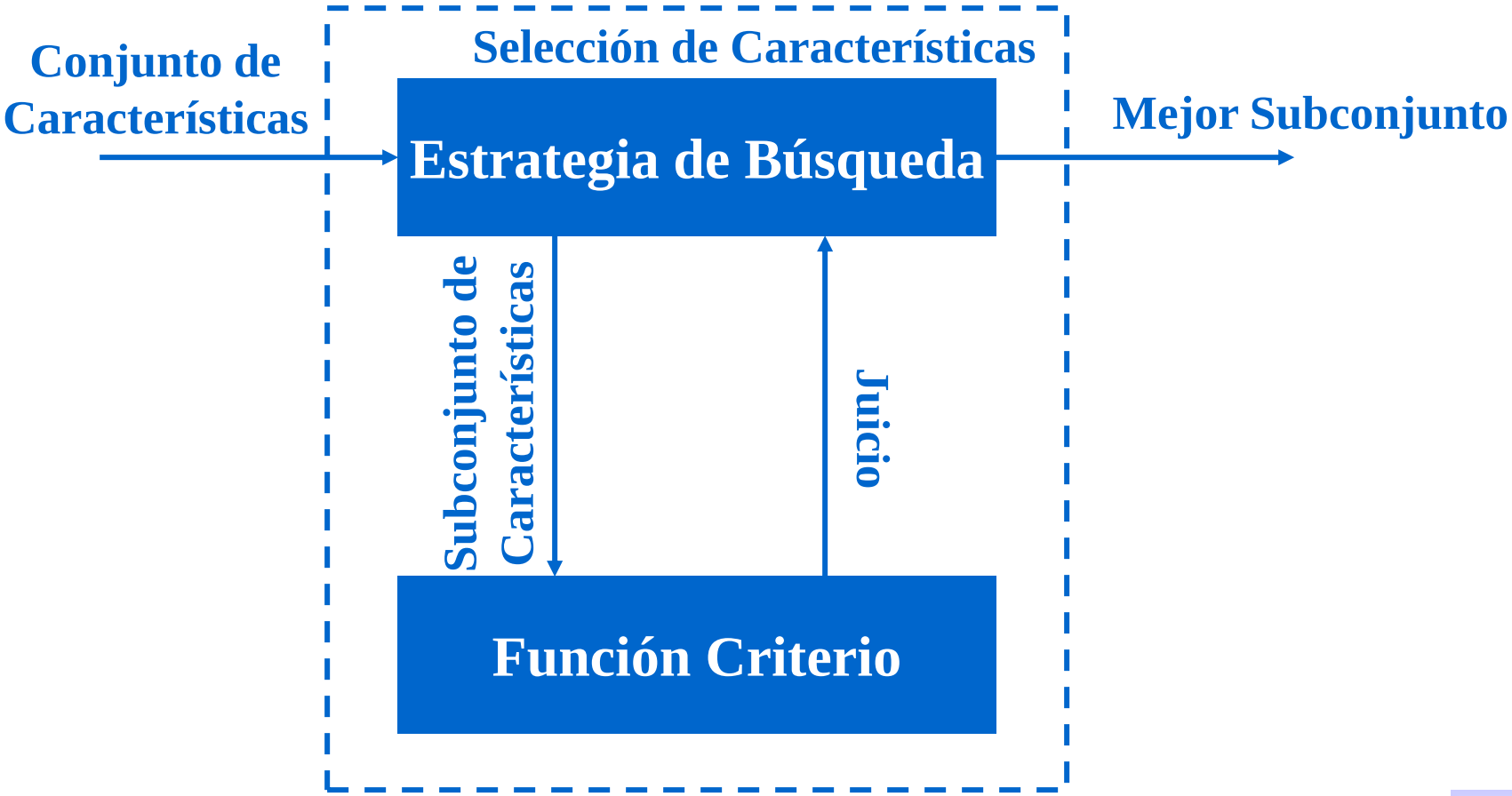
$$\binom{D}{d} combinaciones \Rightarrow crece\, exponencialmente$$

- Si en cambio buscamos todos los posibles subconjuntos de características entonces el problema tiene $2^D - 1$ combinaciones.

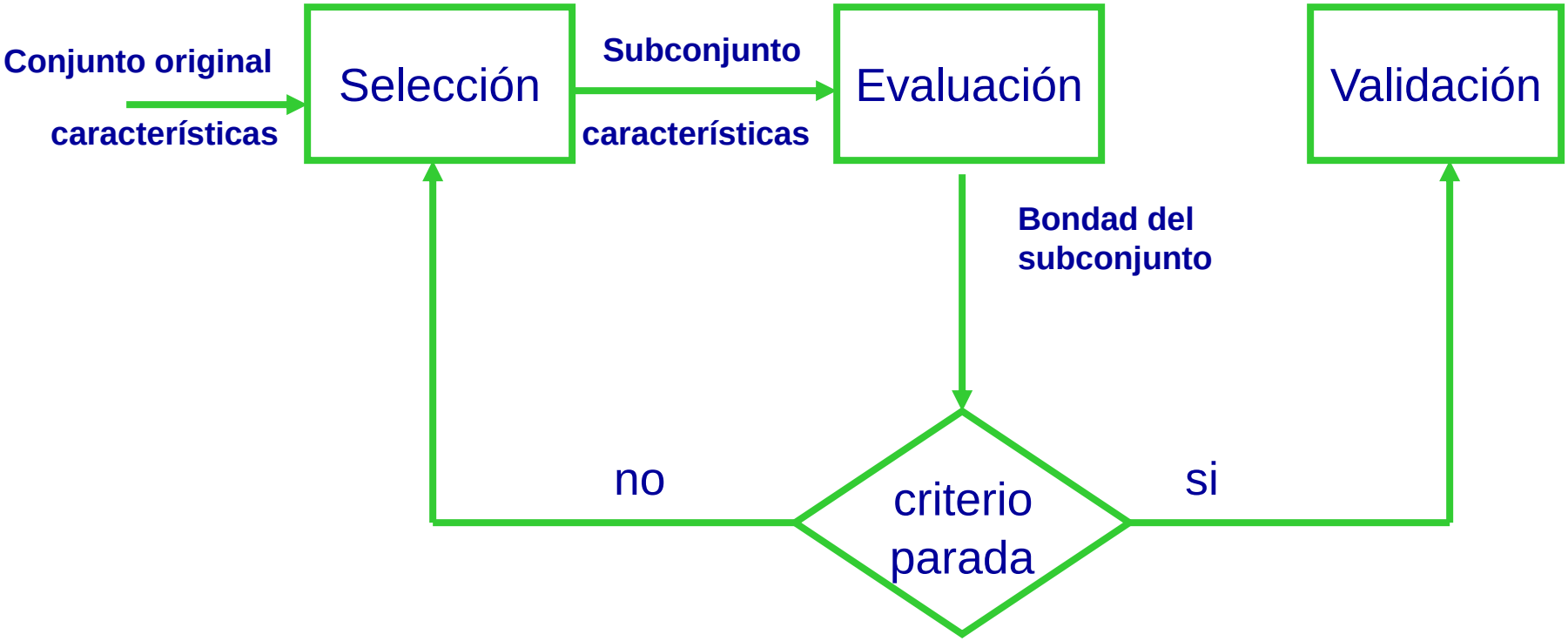
Solución Alternativa

- Algoritmos de selección de características:
  - Se basan en el uso de una estrategia de búsqueda y una función criterio que permita evaluar la calidad de cada subconjunto seleccionado.
  - En general, la función criterio trata de medir la capacidad discriminatoria de un atributo o de un subconjunto de características.

Esquema


Conjunto de Características
Selección de Características
Estrategia de Búsqueda
Mejor Subconjunto
Subconjunto de Características
Juicio
Función Criterio

Proceso de selección de características con validación



Factores de Influencia

- En general, el algoritmo de búsqueda utilizado no resulta tan determinante como la función criterio: distintos algoritmos obtienen la misma solución.
- Lo realmente fundamental en el resultado es la función criterio empleada.

Clasificación por Estrategia usando el propio clasificador

- Solución óptima
  - Algoritmos de Ramificación y Poda.
- Solución subóptima
  - Algoritmos Secuenciales.
  - Algoritmos Genéticos.

Algoritmos Secuenciales

- Búsqueda Secuencial Hacia Delante (Forward Sequential Search, FSS).
- Búsqueda Secuencial Hacia Atrás (Backward Sequential Search, BSS).
- Búsqueda Secuencial Flotante (Sequential Floating Search, SFS).

Algoritmo FSS

- Parte de un conjunto de características vacio.
- En cada iteración, añade al conjunto la “mejor” característica.

Ejemplo FSS (Datos)

- Conjunto inicial: $Y = \{y_1, y_2, y_3, y_4, y_5\}$.
- Objetivo: seleccionar un conjunto X con las 2 mejores características ($|X| = 2$).

Ejemplo FSS (Ejecución)

- $X = \varnothing$
- Seleccionar el mejor subconjunto:
  $X = \{y_1\}$, $X = \{y_2\}$, $X = \{y_3\}$, $X = \{y_4\}$, $X = \{y_5\}$
- Seleccionar el mejor subconjunto:
  $X = \{y_2, y_1\}$, $X = \{y_2, y_3\}$,
  $X = \{y_2, y_4\}$, $X = \{y_2, y_5\}$.

Algoritmo BSS

- Parte de un conjunto formado por todas las características disponibles.

- En cada iteración, elimina del conjunto la “peor” característica.

Ejemplo BSS (Datos)

- Conjunto inicial: $Y = \{y_1, y_2, y_3, y_4, y_5\}$.
- Objetivo: seleccionar un conjunto X con las 2 mejores características ($|X| = 2$).

Ejemplo BSS (Ejecución)

- $X = Y = \{y_1, y_2, y_3, y_4\}$.
- Seleccionar la mejor combinación:
  $X = \{y_1, y_2, y_3\}$, $X = \{y_1, y_2, y_4\}$,
  $X = \{y_1, y_3, y_4\}$, $X = \{y_2, y_3, y_4\}$.
- Seleccionar el mejor subconjunto:
  X = $\{y_1, y_3\}$, $X = \{y_1, y_4\}$.

Inconvenientes de FSS y BSS

- No pueden “corregir” adiciones o eliminaciones anteriores.
- Pueden dar lugar a conjuntos no óptimos: por ejemplo, en el BSS, el mejor X de 3 características era $X = \{y_1, y_3, y_4\}$ y, sin embargo, el mejor X de 2 elementos podría haber sido $X = \{y_1, y_2\}$.

Algoritmo SFS

- Dos variantes: “forward” y “backward”.
- Una mejora sobre FSS (o BSS): mediante la inclusión (o eliminación) condicional de características.

Algoritmo SFS

- Después de cada iteración hacia delante (o hacia atrás), se vuelve hacia atrás (o hacia delante) para comprobar si existe alguna combinación mejor.

- En la práctica, se aplica directamente el FSS (o BSS) en las 2 primeras iteraciones.

Ejemplo SFS (Datos)

- Conjunto inicial: $Y = \{y_1, y_2, y_3, y_4, y_5\}$.
- Objetivo: seleccionar un conjunto X con las 3 mejores características ($|X| = 3$).

Ejemplo SFS (Ejecución 1)

- Suponemos que el FSS ya ha seleccionado el mejor subconjunto de 2 elementos: $X = \{y_2, y_3\}$.

- El SFS selecciona ahora el mejor conjunto con 3 características; supongamos que éste es $X = \{y_2, y_3, y_4\}$.

Ejemplo SFS (Ejecución 2)

- La vuelta atrás consiste en ver si existe algún subconjunto de 2 características mejor que el seleccionado anteriormente ($X = \{y_2, y_3\}$): en este caso, sólo podría ser el conjunto $X = \{y_3, y_4\}$.

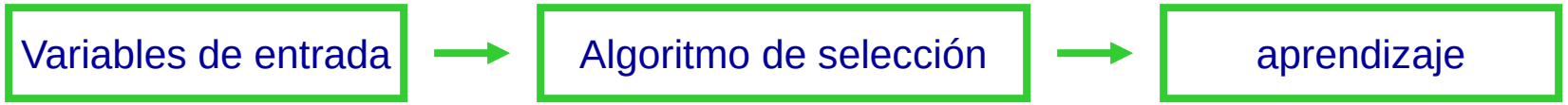
- Si se cumple, se cambia y se vuelve hacia delante con el nuevo subconjunto.

Clasificación por Función

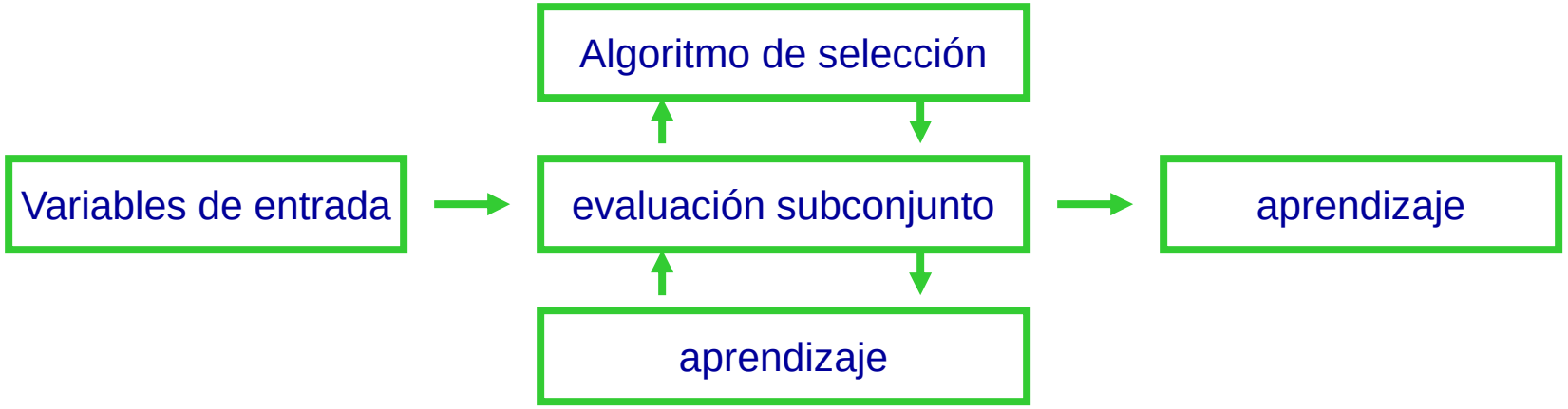
- Filters: la función criterio utilizada es independiente del clasificador a emplear en la fase de clasificación.

- Wrappers: la función criterio utilizada es la propia regla que posteriormente se empleará para la clasificación de nuevas muestras.

Métodos filter y wrapper



Método filter



Método wrapper

Funciones Criterio

- Medidas de distancia o separabilidad: p.e., distancia euclidea, “city-block”.
- Medidas de información: p.e., entropía.
- Medidas de dependencia: p.e., coeficiente de correlación.
- Medidas de la tasa de error: clasificador.

Medidas de Distancia

- Para un problema de 2 clases, un atributo X es preferible a otro Y si X induce una mayor diferencia entre las probabilidades condicionales de las dos clases.
- Por ejemplo, se puede utilizar la distancia entre los centroides de las dos clases.

Medidas de Información

- Miden la “ganancia de información” debida a cada atributo.

- Esta ganancia de información se puede definir a partir de la entropía.

Entropía

$$Entropía\ (S) = -\sum_{i=1}^{c} p_i \log_2 (p_i)$$

- donde p_i es la proporción de muestras de la clase i en el conjunto S.

$$p_i = \frac{|S_i|}{|S|}$$

- Mide el "grado de impureza" de un cierto conjunto de muestras:
  - Será máxima cuando todas las clases están representadas en la misma proporción.

Ganancia de Información

- Entonces, la ganancia de información de un atributo *A* será:

$$Ganancia\,(S, A) = Entropía\,(S) - \sum_{i=1}^{n} \frac{|Ss_i|}{|S|} Entropia\,(S_i)$$

Medidas de Dependencia

- Cuantifican la capacidad para predecir el valor de una variable a partir del valor de otra variable: coeficiente de correlación.
- Coeficiente de correlación (entre 0 y 1): mide el grado de relación lineal entre dos variables. Un valor igual a 0 indica que no existe relación entre ellas.

Medidas de Dependencia

- Para detectar atributos redundantes, se puede determinar el coeficiente de correlación de un atributo *X* con otro *Y*.

$$r^2 = \frac{s_{XY}^2}{ss_{XX}\, ss_{YY}}$$

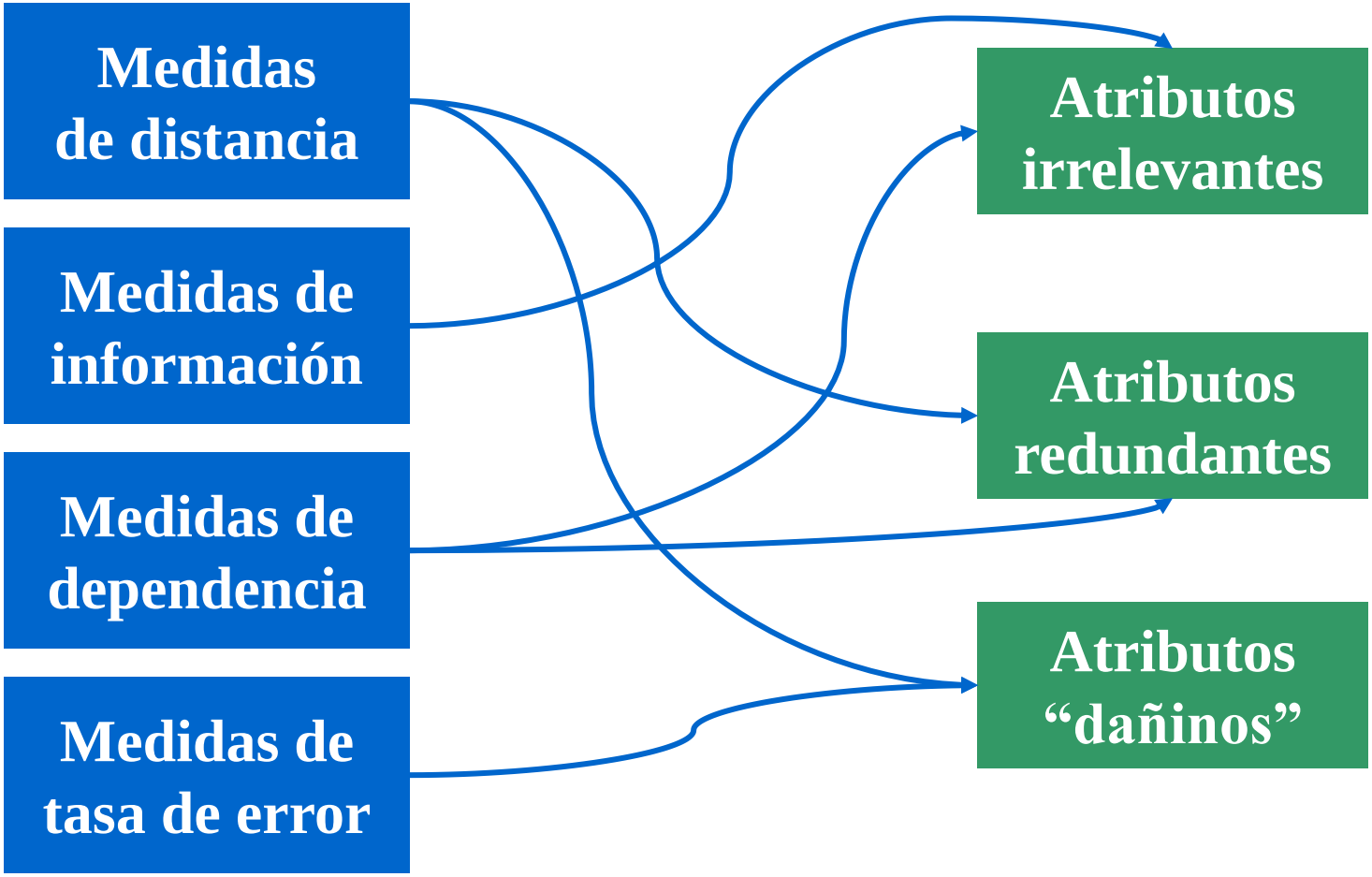
$$ss_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \quad ss_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad ss_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Medidas de Tasa de Error

- Básicamente, se corresponden con los métodos "wrapper".

- Como función criterio se utiliza el mismo clasificador que posteriormente vaya a emplearse para la clasificación de nuevas muestras.

Función Criterio - Problema



Situación Actual

- En general, las propuestas actuales están dirigidas a resolver un único problema: atributos irrelevantes, atributos redundantes, atributos "dañinos".
- El problema a resolver viene determinado, básicamente, por la función criterio utilizada.

Ponderación de Características

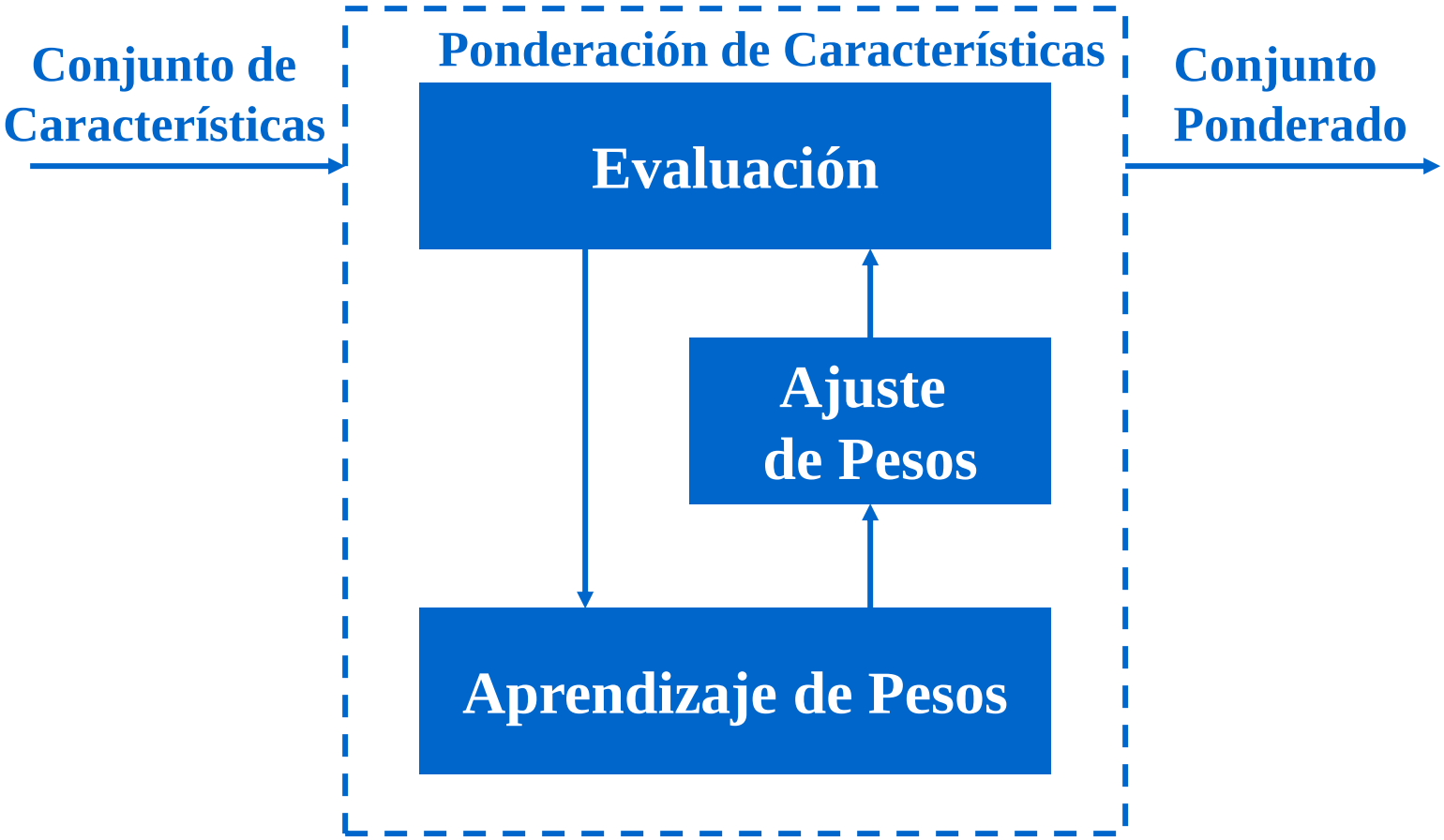
- Ponderar, en vez de seleccionar, cada uno de los atributos: asignar un peso a cada atributo en función de su importancia.
- Esta opción se centra en aumentar la tasa de aciertos de la regla de decisión, no en reducir el coste computacional.

Problema de la Ponderación

- Transformar el conjunto Y de los D atributos iniciales en un nuevo conjunto X de D características con distintos pesos:

$$X = \{w_1 y_1, w_2 y_2, \ldots, w_D y_D\}$$

Esquema


Conjunto de Características
Ponderación de Características
Evaluación
Ajuste de Pesos
Aprendizaje de Pesos
Conjunto Ponderado

Base del Problema

- Idea: los atributos irrelevantes tendrán muy poca influencia sobre el resultado global de la clasificación.

- Solución: ponderar la importancia de los atributos en función de su contribución al problema.

Fundamentos

- Las estrategias de ponderación tratan de:
  - Premiar a los atributos responsables de clasificaciones correctas.
  - Penalizar a los atributos culpables de clasificaciones erróneas.

Estrategias de Ponderación

- Ajustar los pesos en función del resultado de las predicciones.
- Ajustar los pesos en función de la clase de los vecinos más próximos.
- Ajustar los pesos en función de la probabilidad condicional de cada clase.

Ponderación — Selección

- La selección de características puede verse como un caso particular de la ponderación de características.
  - El peso asignado a cada atributo sólo puede tomar los valores 0 (mínima relevancia) y 1 (máxima relevancia).

Extracción de características

- Transformar el espacio de entrenamiento E en otro espacio E* donde los datos estén menos correlacionados.
- Encontrar un nuevo conjunto de ejes ortogonales en el que la varianza de los datos se máxima.
- En ese nuevo espacio los ejes principales serán los que más información aporten y los ejes secundarios no aportarán casi información por lo que se pueden eliminar y reducir la dimensionalidad.
- En la matriz de covarianza de los datos la dirección de máxima varianza es la del eje principal de la elipse.

- **Análisis de componentes principales:**
  - Se determina la matriz de covarianza de los datos.
  - A continuación se diagonaliza la matriz calculando sus autovalores y sus autovectores.
  - El resultado es una rotación rígida donde el autovector asociado al mayor autovalor está en la dirección de máxima varianza.

- Solución:
  - Se ordenan los autovalores.
  - Se suma la traza de la matriz λ_T que contendrá la suma de los autovalores y se normaliza cada autovalor por la suma:

$$\lambda_i(\%) = 100 \cdot \frac{\lambda_i}{\sum_{i=1}^{d} \lambda_i}$$

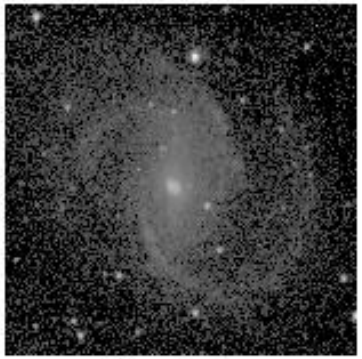
  - De esta forma conocemos el grado de influencia de cada autovalor en la varianza total.
  - Iremos acumulando de mayor a menor analizando cuanto porcentaje del total estamos acumulando hasta un umbral. Ej. 90 %.

- Mediante los autovectores Σ_X podemos calcular las coordenadas de los puntos del conjunto de entrenamiento **x** en el espacio transformado **x**':

$$\Sigma_X = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \quad \mathbf{x}' = \begin{bmatrix} x_1' \\ \vdots \\ x_d' \end{bmatrix}$$

$$\begin{bmatrix} x_1' \\ \vdots \\ x_d' \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{dd} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

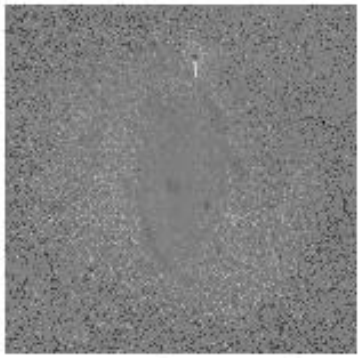
- Si sólo utilizamos los autovalores asociados a los autovalores más altos tendremos conjunto de entrenamiento en un espacio transformado *más reducido*.



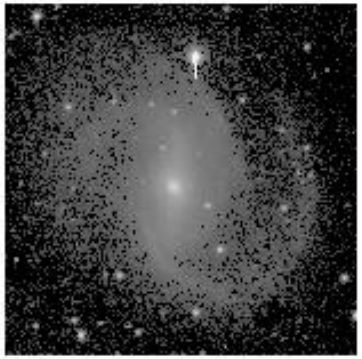
Banda 1

C.P. 1 (92.626 %)

Banda 2



C.P. 2 (4.949 %)



Banda 3

C.P. 3 (2.425 %)

Conclusiones

- **Selección:** determinar la función criterio y el algoritmo de búsqueda.
- **Ponderación:** determinar la estrategia de ajuste de los pesos.
- **Extracción:** Pasar a un espacio transformado quedándonos con los ejes que más decorrelacionan los datos.