Tema 12: Selección de Variables

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Introducción

Selección del subconjunto óptimo de variables predictoras: FSS (*feature subset selection*)

- *No monotocidad* de la bondad de un clasificador con respecto al número de variables predictoras
- *Variables irrelevantes*: el conocimiento de su valor no aporta nada a la variable C
  - En naive Bayes: $p(X = 0|C = 0) = 0{,}7$ y $p(X = 0|C = 1) = 0{,}7$
- *Variables redundantes*: su valor puede ser determinado a partir de otras variables
  - En naive Bayes: $p(X_1 = 0|C = 0) = p(X_2 = 1|C = 0)$ y $p(X_1 = 0|C = 1) = p(X_2 = 0|C = 1)$

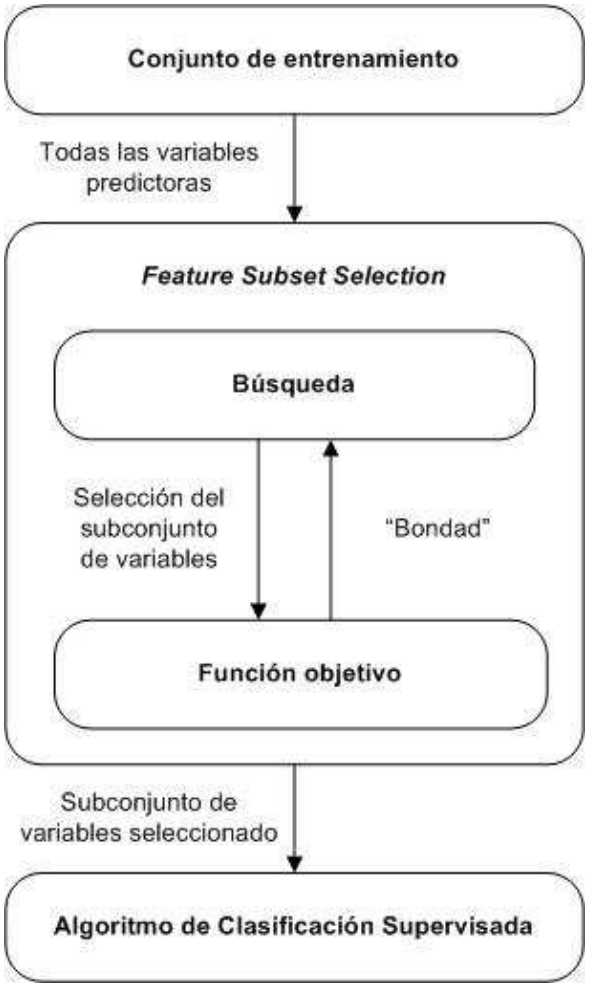
Introducción

- *Modelos parsimoniosos*: ante dos modelos que explican lo suficientemente bien los datos, se debe de escoger el modelo mas simple
- Estadística: *Occam's razor*
- Aprendizaje automático: *Keep It as Simple as possible, Stupid (KISS)*

Beneficios del FSS

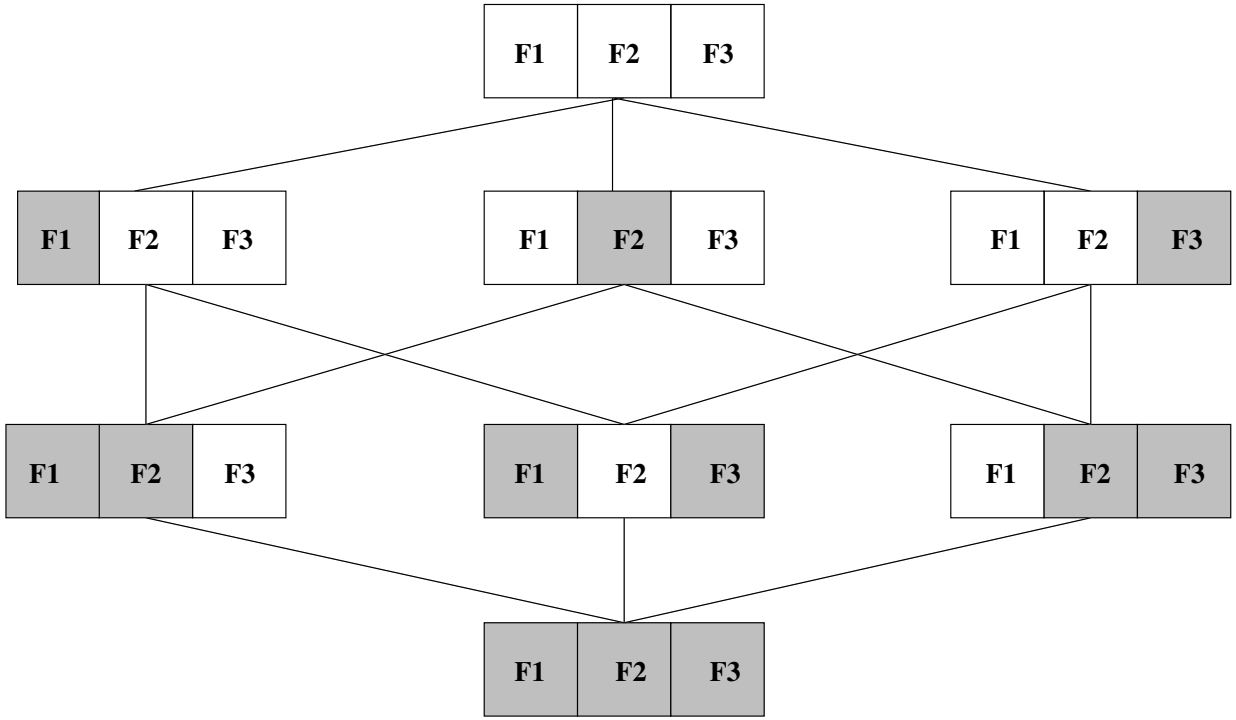
- Reducción del coste de adquisición de los datos
- Mejora en la comprensión del modelo clasificatorio
- Inducción mas rápida del modelo clasificatorio
- Mejora en la bondad

Esquema básico del FSS



Esquema del procedimiento de selección de subconjuntos de variables para problemas de clasificación supervisada

Esquema básico del FSS



FSS como un problema de búsqueda

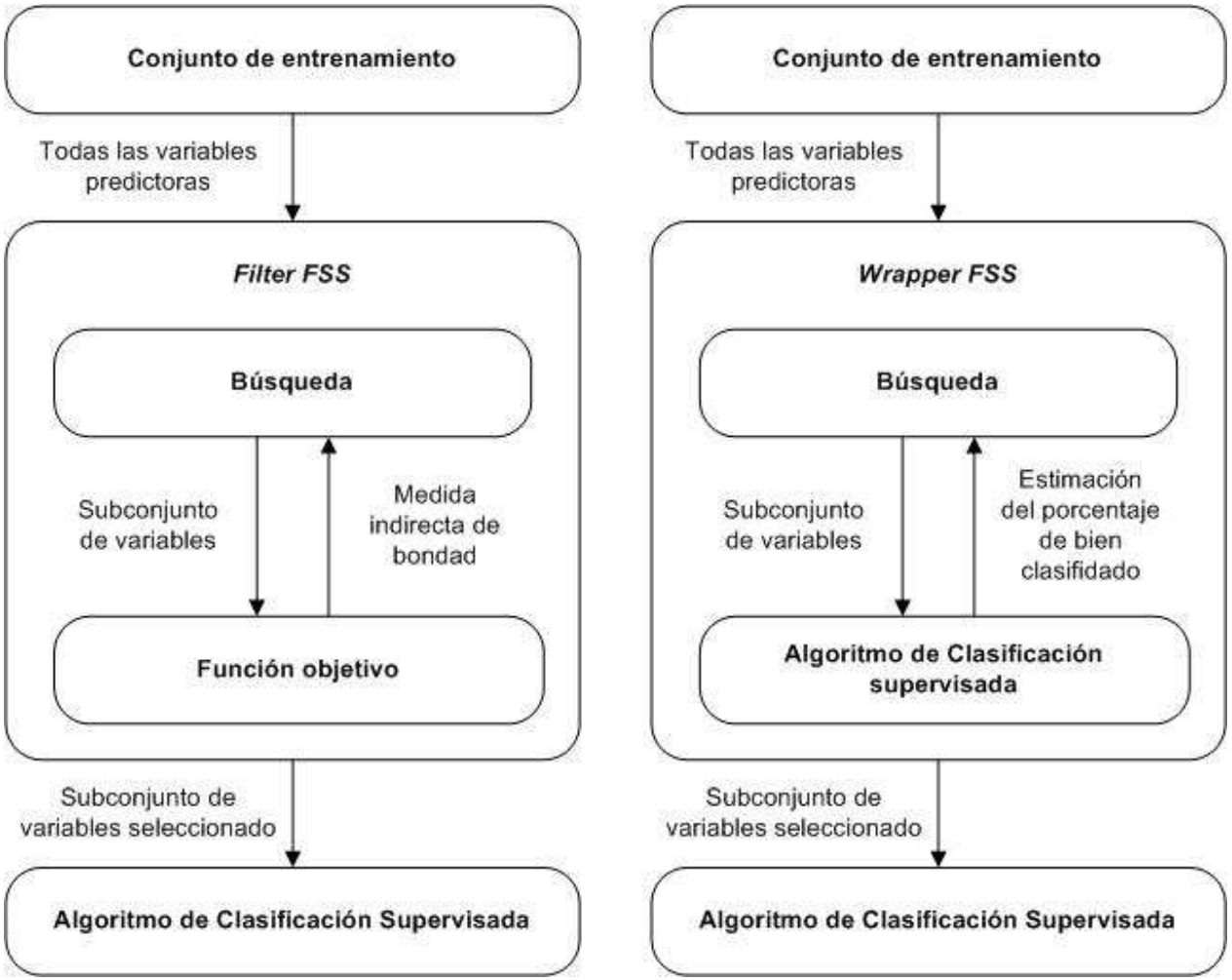
Aproximación indirecta (*filter*)

- Se establece una medida indirecta (cantidad de información mutua, incremento en la verosimilitud del modelo, etc.) de la aportación de cada variable a la clasificación
- Las variables se ordenan según dicho criterio, seleccionándose las k mejores de las n variables predictoras
- La medida indirecta (o criterio) no tiene en cuenta el paradigma clasificador a utilizar posteriormente

Aproximación directa (*wrapper*)

- Cada subconjunto de variables candidato es evaluado directamente (porcentaje de bien clasificados, área bajo la curva ROC, etc.) en el modelo clasificatorio construido con dicho subconjunto
- Problema de búsqueda de cardinalidad 2^n
- Abordar el problema por medio de cualquier heurístico de optimización

Filter versus *wrapper*



Diferencias entre las aproximaciones *filter* y *wrapper* a la clasificación supervisada