Selección de variables (Feature selection)

Abdelmalik Moujahid

Grupo de Inteligencia Computacional
Universidad del País Vasco UPV/EHU
Curso 2014-2015

Índice

Índice

Feature subset selection

Introducción

- El problema de selección de variables consiste en hallar el **subconjunto óptimo de variables** (atributos) de un conjunto de datos adoptando algún criterio,
- De manera que el clasificador que se induce basandose en ese subconjunto óptimo sera el clasificador con mayor precisión posible.

Feature subset selection

Motivación

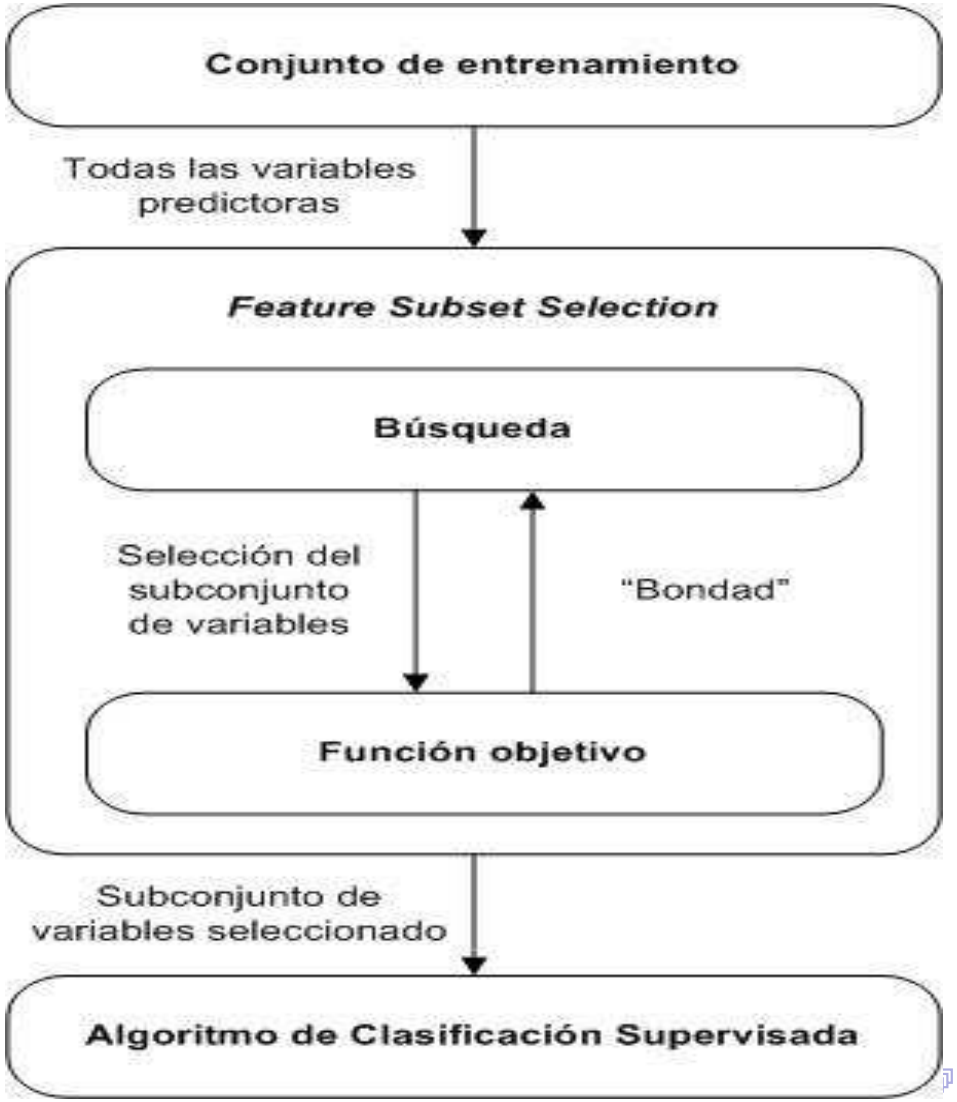
- *No monotocidad* de la bondad de un clasificador con respecto al número de variables predictoras
- *Variables irrelevantes*: el conocimiento de su valor no aporta nada a la variable C
- *Variables redundantes*: su valor puede ser determinado a partir de otras variables

Feature subset selection

Beneficios

- 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



Feature subset selection

Formulación

Formalmente, podemos definir el problema de selección de variables como:

- Dado un conjunto de variables predictoras $X = (X_1, ..., X_n)$ con valores $x = (x_1, ..., x_n)$, y una variable clase C con valores $c_1, ..., c_k$
- Sea G cierto subconjunto de X y x_G los valores del vector en G
- Entonces, el objetivo de la selección de variables es encontrar el subconjunto mínimo G tal que:

$$P(C|G = x_G) \simeq P(C|X = x)$$

Llamamos a este subconjunto mínimo el **subconjunto óptimo**

Índice

Feature subset selection

Aproximación 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
- Ignora los posibles efectos que puede tener el subconjunto óptimo de atributos sobre el rendimiento del clasificador inducido.

Índice

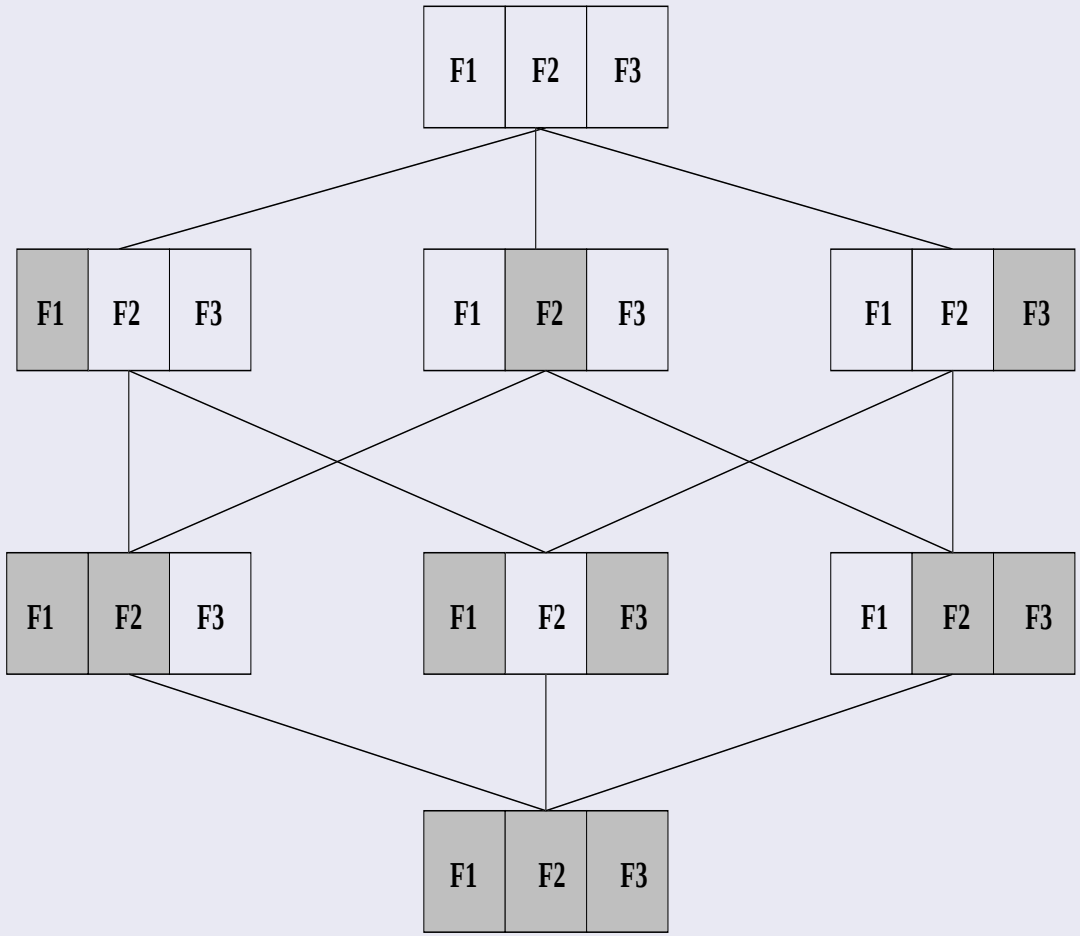
Feature subset selection

Aproximación Wrapper

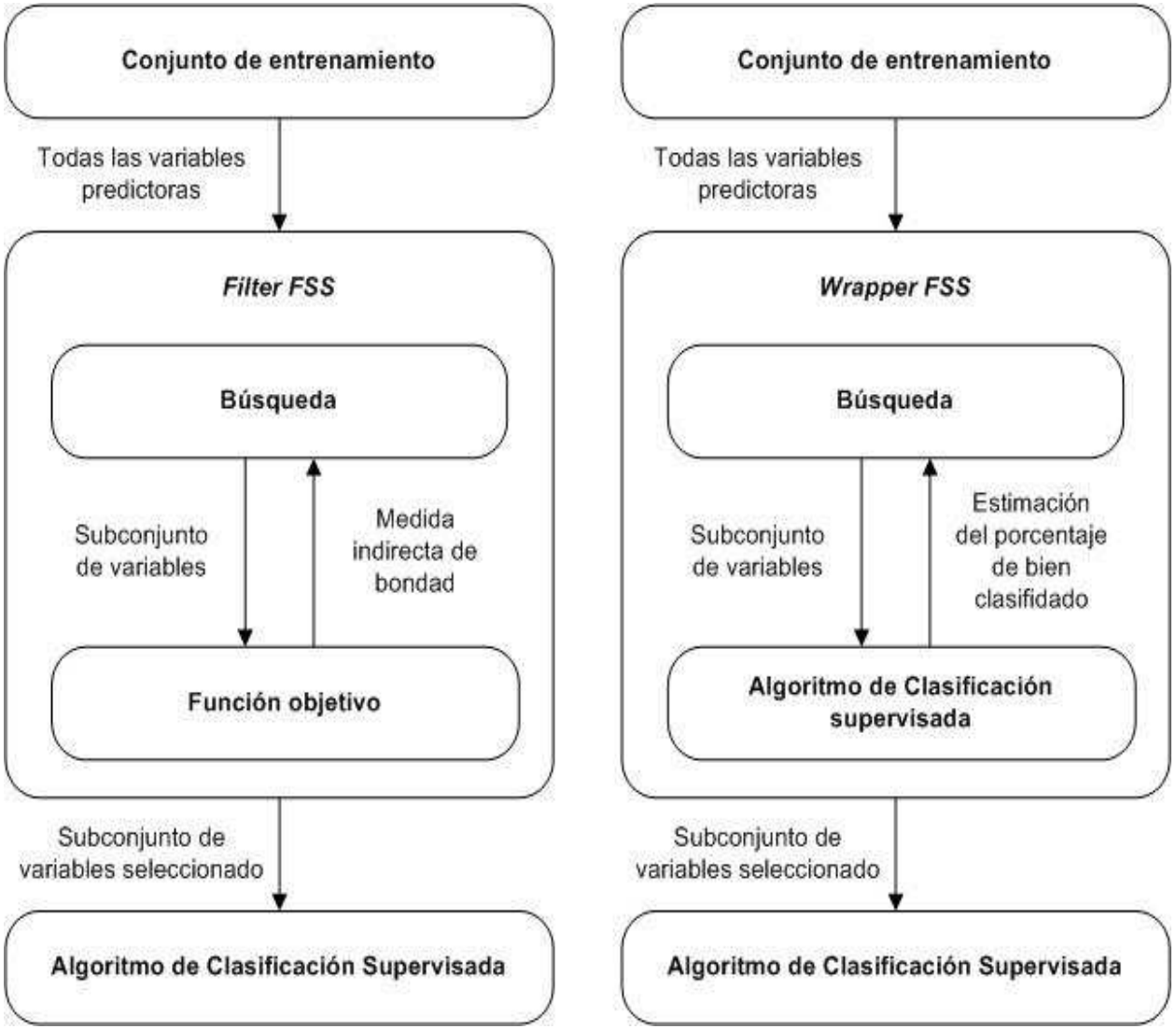
- Cada subconjunto de variables candidato es evaluado directamente (porcentaje de bien clasificados, área bajo la curva ROC, etc.) en el modelo clasificatorio construido con dicho subconjunto
- Problema de búsqueda de cardinalidad 2^n
- Abordar el problema por medio de cualquier heurístico de optimización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

Feature subset selection

FSS como un problema de búsqueda



Filter versus Wrapper



Filter versus Wrapper

Ejemplo

El dataset hepatitis.arff contiene 155 ejemplos cada uno descrito por 19 atributos y un atributo clase. Una selección de atributos mediante los método *filter* y *Wrapper* nos proporciona los siguientes resultados:

- **Information Gain Ranking Filter**: Selected attributes: 17,14,12,11,19,5,6,13,18,1,2,10,4,3,7,8,9,16,15 : 19
- **Gain Ratio feature evaluator**: Selected attributes: 12,14,17,13,11,1,5,18,19,6,2,10,4,7,3,8,9,16,15 : 19
- **Wrapper Subset Evaluator** (NaiveBayes as Learning algorithm): Selected attributes: 1,12,19 : 3
- **Wrapper Subset Evaluator** (J48 as Learning algorithm): Selected attributes: 1,12: 2

Filter versus Wrapper

Ejercicio

La tabla adjunta contiene 20 ejemplos que constituyen el fichero de casos para un árbol de clasificación. Determinar cual es el atributo que utilizaría cada uno de los algoritmos ID3 y C.45 como nodo raíz. Los atributos A1 y A2 son discretos con 4 y 2 posibles valores respectivamente.

Filter versus Wrapper

Caso	X_1	X_2	C
1	1	1	1
2	1	1	1
3	2	1	1
4	2	1	1
5	1	1	1
6	3	1	1
7	3	1	1
8	1	1	1
9	2	2	1
10	2	2	1
11	2	2	2
12	4	2	2
13	4	2	2
14	2	2	2
15	1	2	2
16	2	2	2
17	1	2	2
18	2	1	2
19	2	1	2
20	2	1	2