Tema 12. Selección de Variables

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid
Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco–Euskal Herriko Unibertsitatea

12.1 Introducción

En este tema se va a presentar la problemática relacionada con la selección de subconjuntos de variables adecuados para inducir modelos clasificatorios.

Una vez enunciados los beneficios de realizar la tarea de seleccionar el subconjunto de variables adecuado, veremos dos filosofías distintas con las que abordar dicho problema. Dichas aproximaciones se conocen bajo el nombre de indirecta (*filter*) y directa (*wrapper*).

12.2 Beneficios de la selección de variables

El problema de la selección del subconjunto de variables para la inducción de un modelo clasificador, se denomina FSS (*Feature Subset Selection*) y surge motivado por la *no monotocidad* de los modelos clasificatorios en relación con el número de variables predictoras, asi como por la existencia de ciertas variables predictoras que pueden llegar a ser *irrelevantes* o incluso *redundantes*.

Veamos con más detalle los motivos anteriores:

- la *no monotocidad* de la probabilidad de éxito de un sistema clasificador se debe al hecho –constatado empíricamente, y demostrado matemáticamente para algunos paradigmas– de que no por construir un modelo clasificatorio con una variable añadida a las ya existentes, la probabilidad de éxito que se va a obtener con este nuevo modelo clasificador deba superar a la del modelo actual
- se considera que una variable predictiva es *irrelevante* cuando el conocimiento del valor de la misma no aporta nada para la variable C
- una variable predictiva se dice *redundante* cuando su valor puede ser determinado a partir de otras variables predictivas

El objetivo que uno se plantea al tratar de resolver el problema FSS es el de detectar aquellas variables que son irrelevantes y/o redundantes para un problema clasificatorio dado.

Con esto lo que se pretende es crear *modelos parsimoniosos*, guiados por el principio que en Estadística se denomina *Occam's razor* y que en Aprendizaje Automático se conoce como KISS (*Keep It as Simple as possible, Stupid*). Es decir la idea subyacente a la modelización parsimoniosa es que si uno tiene dos modelos que explican suficientemente bien los datos, se debe de escoger con el modelo mas simple de los dos.

Veamos a continuación algunos beneficios derivados del FSS:

- *reducción en el costo de adquisición de los datos*, debido a que el volumen de información a manejar para inducir el modelo es menor

- *mejora en la comprensión del modelo clasificatorio*, ya que no vamos a inducir el modelo con un gran número de variables
- *inducción más rápida del modelo clasificatorio*, derivada de que el algoritmo de inducción del modelo va a trabajar con menos variables
- *mejora en la bondad del modelo clasificatorio*, derivado de la no monotocidad explicada anteriormente.

12.3 *Filter versus Wrapper*

La Figura 1 muestra el esquema básico relativo al procedimiento de selección de subconjuntos de variables para problemas de clasificación supervisada.

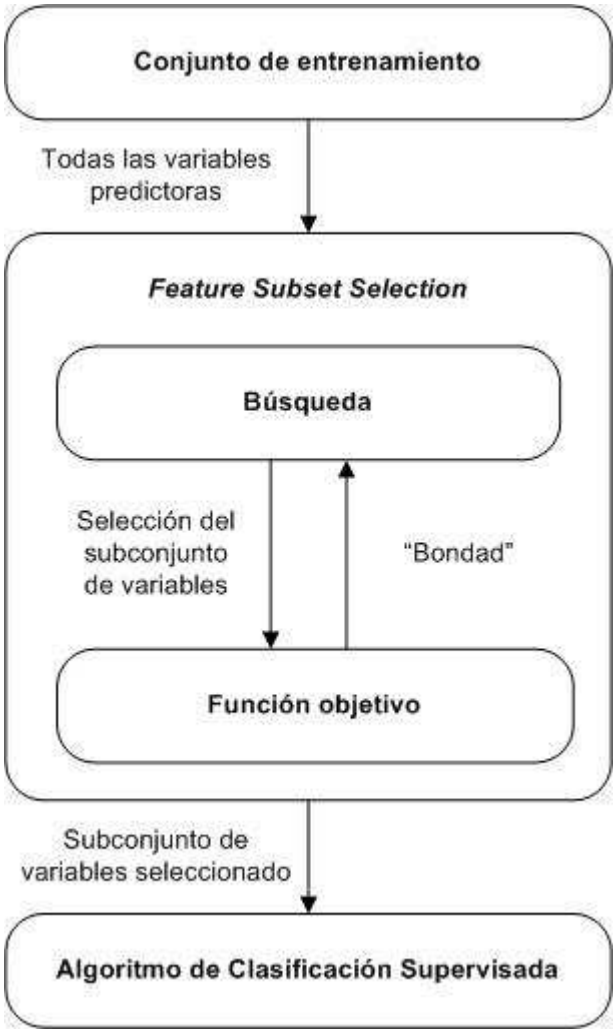

Figura 1: Esquema del procedimiento de selección de subconjuntos de variables para problemas de clasificación supervisada

Dos son básicamente las aproximaciones al problema FSS: indirecta o *filter* y directa o *wrapper*.

12.3.1 Aproximación indirecta o *filter*

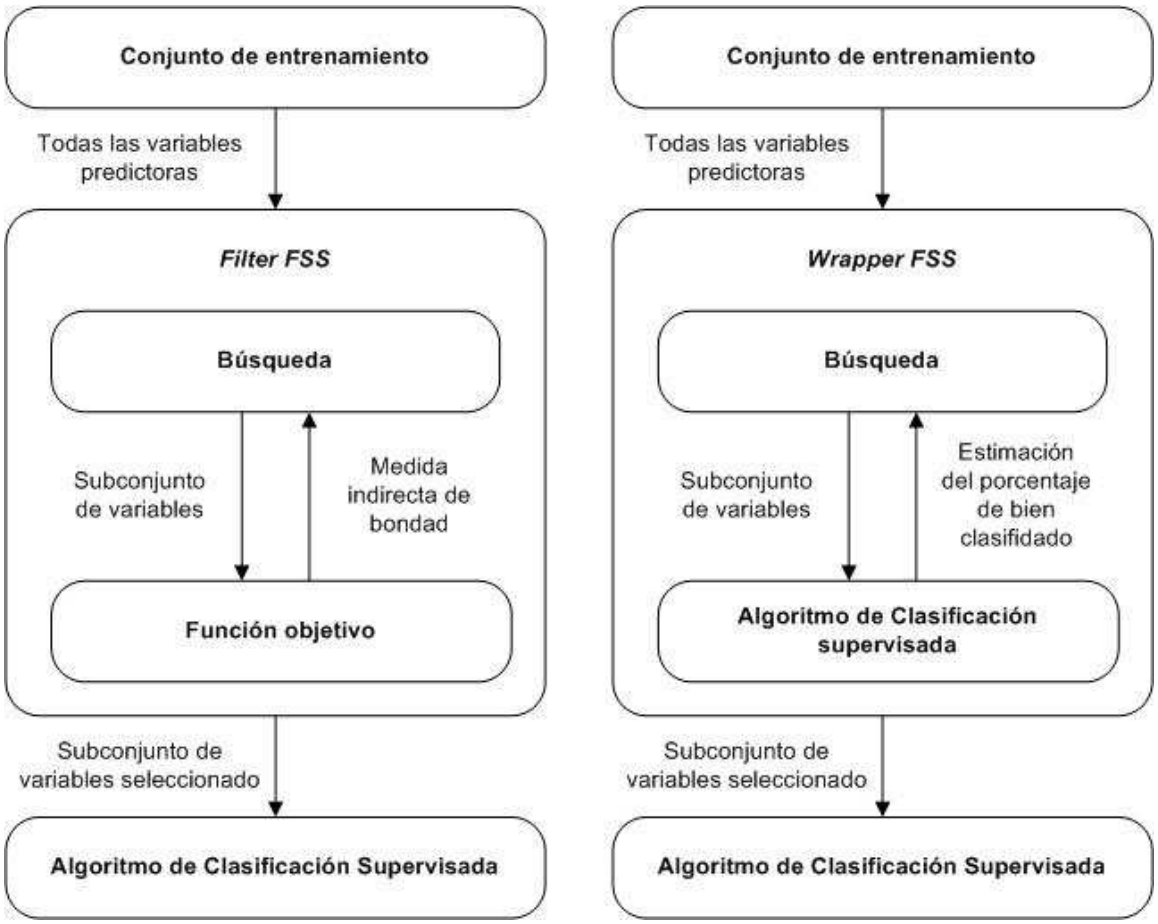

Figura 2: Diferencias entre las aproximaciones *filter* y *wrapper* a la clasificación supervisada

La aproximación indirecta establece –véase Figura 2– una medida indirecta de bondad de la selección de las variables, habitualmente un ranking entre las variables predictoras teniendo en cuenta un criterio –previamente fijado– de relevancia entre una variable predictora y la variable clase. A partir de dicha medida de relevancia las variables predictoras quedan ordenadas –supongamos de mayor a menor relevancia respecto de la variable clase– seleccionándose las k, $k < n$, primeras para inducir con ellas el modelo clasificador. Nótese que la medida de relevancia con la que hemos ordenado las variables predictivas no tiene en cuenta el paradigma (Naive–Bayes, Árbol de Clasificación, etc) con el que se va a inducir el modelo clasificatorio. Es por ello por lo que este tipo de aproximación al problema puede verse como una selección indirecta.

Ejemplo 12.1:
Supongamos que las variables predictoras, $X_1, \ldots, X_n$ sean discretas. Podemos construir una tabla de contingencia cruzando cada variable predictora X_i con la variable a predecir C, para a continuación calcular el valor del estadístico chi–cuadrado. Cuanto menor sea la significatividad relacionada con el estadístico, mayor es la relevancia de la variable en cuestión con respecto de la variable C. Es decir que el estadístico de la chi–cuadrado lo podemos utilizar como valor para ordenar las variables predictoras y posteriormente seleccionar las k primeras. Otra medida de relevancia usada habitualmente es la cantidad de información mutua entre cada variable predictora X_i y la

variable clase C.

Ejemplo 12.2:
En caso de que las variables sean continuas, podemos estudiar –véase por ejemplo la Figura 3– las funciones de densidad de cada variable predictora condicionada a los distintos valores de la variable clase, C. Definiendo una distancia entre dichas funciones de densidad condicionadas, ordenaríamos las variables predictoras. La primera de dichas variables predictoras en la ordenación anterior, sería aquella para la cual la distancia entre las funciones de densidad condicionadas sea mayor.

De manera análoga la variable predictora que ocupase el último lugar en la ordenación

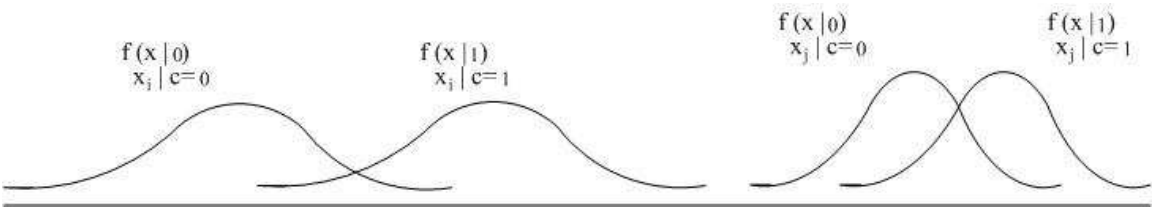

Figura 3: Funciones de densidad de las variables X_i y X_j condicionadas a los dos valores de la variable clase, C.

sería aquella para la cual la distancia entre las funciones de densidad condicionadas a los valores de la clase es menor entre las n.

En el supuesto representando en la Figura 3, la variable X_i se ordena antes que la variable X_j, y por tanto en caso de no escoger la variable X_i tampoco se seleccionará la X_j.

12.3.2 Aproximación directa o *wrapper*

En la aproximación directa o *wrapper* al FSS, cada posible subconjunto de variables candidato es evaluado por medio del modelo clasificatorio inducido –con el paradigma utilizado– a partir del subfichero conteniendo exclusivamente las variables seleccionadas junto con la variable clase.

Con esta aproximación *wrapper*, el problema FSS puede ser visto como un problema de búsqueda en un espacio de cardinalidad 2^n. Este es el motivo por el que los heurísticos estocásticos de optimización estudiados en la parte I de la asignatura –búsqueda voraz, algoritmos genéticos y algoritmos de estimación de distribuciones– se han venido aplicando con éxito al problema FSS desde un perspectiva de aproximación directa.

Referencias

1. M. Ben–Bassat (1982). Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation. *Handbook of Statistics*, Vol. 2, 773–791.

2. J. Doak (1992). *An evaluation of feature selectionmethods and their application to computer security*. Technical Report CSE-92-18. University of California at Davis.

3. I. Inza, P. Larrañaga, R. Etxeberria, B. Sierra (2000). Feature Subset Selection by Bayesian networks-based optimization. *Artificial Intelligence*, 123: 157–184.

4. R. Kohavi, G. John (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97, (1–2): 273–324.

5. H. Lui, H. Motoda (1998). *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers.

6. J. Yang, V. Honavar (1998). Feature subset selection using a genetic algorithm. *IEEE Inteligent Systems*, 13(2): 44–49.