

# **RESOLUCIÓN DEL PROBLEMA DE SELECCION DE VARIABLES CUANTITATIVAS MEDIANTE GRASP. APLICACIÓN A RATIOS FINANCIEROS**

Olga Gómez<sup>(2)</sup>, Silvia Casado<sup>(1)</sup>, Laura Núñez<sup>(2)</sup> y Joaquín Pacheco<sup>(1)</sup>

(1) Dpto. Economía Aplicada. Universidad de Burgos.  
Fac. C Económicas y Empresariales. Plaza Infanta Elena s/n Burgos 09001, España  
(2) Instituto de Empresa, María de Molina 11, Madrid 28006, España

## **RESUMEN**

En este trabajo se propone un algoritmo tipo GRASP para el problema de Selección de Variables en el ámbito de la clasificación, en el caso concreto en el que las variables explicativas son todas cuantitativas. El problema consiste en dado un conjunto de variables usadas en la clasificación seleccionar el subconjunto de estas que lleve a cabo la tarea de forma óptima. Reducir la dimensionalidad conlleva diversas ventajas (Inza et al. 2000) como la reducción del coste en la adquisición de datos, mejora en la comprensión del modelo final de clasificador, incremento de la eficiencia del clasificador y mejora en la eficacia del clasificador. La búsqueda del subconjunto de variables es un problema NP-duro (Kohavi 1995), de modo que es recomendable el uso de estrategias metaheurísticas para obtener soluciones razonablemente buenas sin explorar todo el espacio de soluciones. No existe ningún trabajo previo que aborde este problema de clasificación en el caso concreto donde todas las variables son cuantitativas. El algoritmo propuesto está diseñado “ad-hoc” para este tipo de variables con objeto de aumentar su eficacia. Se va a aplicar a un problema de selección de ratios financieros para predecir la situación de insolvencia empresarial en España.

## 1.- INTRODUCCIÓN

El problema de selección de variables consiste en determinar la clase a la que pertenecen un conjunto de instancias caracterizadas por atributos o variables. En el aprendizaje supervisado, se cuenta con un conjunto de ejemplos caracterizados por los mismos atributos que los de las instancias y con otro atributo adicional correspondiente a la clase a la que pertenecen. A través de este conjunto de ejemplos se pretende crear y generalizar una regla o conjunto de reglas que permitan clasificar, con la mayor precisión posible, el conjunto de instancias.

Métodos para clasificar, (o crear reglas para clasificar), existen muchos en la literatura:

- *Instance Based Learning* (clasificación basada en la proximidad a los ejemplos)
- Análisis Discriminante clásico
- Análisis Probit-Logit
- Redes Neuronales: Perceptrón Multicapa

Sin embargo, independientemente del método usado, en muchos casos se observa que no todas las variables o características aportan la misma calidad de información, llegando incluso a degradar la creación de reglas de clasificación.

El problema de la selección de variables consiste en encontrar un subconjunto de variables con las que se pueda llevar a cabo la tarea de clasificar de forma óptima. Reducir la dimensión conlleva diversas ventajas (Inza et al. 2000) como la de reducción del coste en la adquisición de datos, mejora en la comprensión del modelo final obtenido, incremento de la eficiencia del clasificador y mejora en la eficacia del clasificador.

La búsqueda del subconjunto de variables es un problema NP-duro (Kohavi 1995), de modo que el uso de metaheurísticas permite obtener soluciones razonablemente buenas sin explorar todo el espacio de soluciones. Los Algoritmos Genéticos (AG) han sido aplicados al problema de selección de variables (Bala et al. 1996, Jourdan et al. 2001) y más recientemente la Búsqueda Dispersa (BD) (García et al. 2003) alcanzándose, en ambos casos, resultados prometedores.

No es por tanto un problema muy estudiado en la literatura, y como se ve las pocas aportaciones que hay son relativamente recientes. En nuestro caso, a diferencia de las

anteriores referencias, se va a tratar este problema en el caso especial en el que las variables que se usan para clasificar son todas cuantitativas. El uso de todas las variables cuantitativas permite una mejor medición y comparación de su capacidad clasificatoria o discriminante. Esto hace que se puedan desarrollar métodos de selección variables, adaptados a este tipo de variables, en los que se incorpore estas medidas discriminantes, y por tanto más eficaces que los anteriores (para este caso).

Más concretamente, se va a desarrollar un algoritmo basado en la estrategia metaheurística GRASP (*Greedy Randomized Adaptive Search Procedure*) para este problema. En cada iteración se genera una solución mediante un método ávido-aleatorio, que se mejora mediante búsqueda local. En nuestro caso el método ávido aleatorio usa como guía la información resultante de la descomposición de la varianza. Las soluciones obtenidas por nuestro método GRASP son mejores que las obtenidas por otros métodos recientes.

En este trabajo se van a aplicar los algoritmos desarrollados en el mismo al problema de selección de ratios financieros para predecir la quiebra y/o solvencia empresarial. Como se verá mas adelante, en el campo financiero contable existen varios trabajos en los que se usan dichos ratios para analizar los fenómenos de quiebra y suspensión de pagos. Los ratios financieros son todos variables cuantitativas. Por tanto, al menos en el campo financiero-contable, existen aplicaciones que justifican el análisis del problema de selección de variables cuantitativas para clasificación.

El trabajo se estructura de la forma siguiente: en la sección siguiente se formula y modeliza el problema; en la sección 3 se describe el algoritmo GRASP, incluyendo el constructivo ávido-aleatorio y la búsqueda local; en la sección 4 se describen una serie de experiencias computacionales con datos ficticios, que sirven para analizar la eficacia de nuestra estrategia; en la sección 5 se muestran las conclusiones y finalmente la sección 6 se dedica a las referencias bibliográficas.

## **2.- MODELIZACIÓN Y FORMULACIÓN**

Considérese un conjunto de  $n$  casos o instancias  $A = \{ a_1, a_2, \dots, a_n \}$ , y un conjunto de  $m$  variables  $V = \{ v_1, v_2, \dots, v_m \}$ ; (para simplificar también se identificará

indistintamente  $V$  con los coeficientes, es decir,  $V = \{1, 2, \dots, m\}$ ). Cada instancia  $a_i$  viene definida de la forma

$$a_i = (a_{i1}, a_{i2}, \dots, a_{im} | c_i),$$

es decir por el valor que toman las variables y la clase a la que pertenece.

Sea un valor predefinido  $p \in N$  se trata de encontrar un subconjunto  $S \subset V$ , de tamaño  $p$  de mayor capacidad clasificatoria.

Para medir la capacidad clasificatoria de los diferentes subconjuntos  $S$  considérese  $k$  particiones previamente definidas del conjunto  $A$ . En cada una de ellas hay 2 subconjuntos,  $A_1$  (conjunto de entrenamiento) y  $A_2$  (conjunto de validación). Es decir  $A = A_1 \cup A_2$ , donde  $A_1$  y  $A_2$  tienen aproximadamente la misma proporción de elementos de cada clase que  $A$ . El cardinal de todos los subconjuntos  $A_1$  es el mismo (y por tanto el de los  $A_2$ ). Para cada subconjunto de variables  $S$ , y para cada par de instancias  $a_i$  y  $a_t$  se define la siguiente distancia

$$d(a_i, a_t) = \sum_{j \in S} d_j^2(a_i, a_t)$$

donde

$$d_j(a_i, a_t) = \frac{|a_{ij} - a_{tj}|}{\max_j - \min_j}$$

con  $\max_j$  y  $\min_j$  los valores máximos y mínimos de la variable  $v_j$  observados en el conjunto de entrenamiento.

Para determinar el valor de la bondad  $f(S)$  de cada subconjunto de variables  $S$  se actúa como se indica a continuación. Para cada una de las particiones consideradas se realiza el siguiente proceso: Para cada instancia  $a_i$  del conjunto de validación  $A_2$  se determina la instancia más cercana del conjunto de entrenamiento  $A_1$ ,  $a_{i^*}$  y se asigna a  $a_i$  a la clase a la que pertenece  $a_{i^*}$ . El porcentaje de aciertos total es la medida de la bondad  $f(S)$  de cada subconjunto  $S$ .

### 3.- DESCRIPCIÓN DE UN ALGORITMO GRASP

Nuestro método está basado en constructivos GRASP. GRASP, o Greedy Randomized Adaptive Search Procedure, es una estrategia metaheurística que construye soluciones usando una aleatoriedad controlada mediante una función voraz. La mayoría de las implementaciones GRASP, como la que se propone en este trabajo, además incluyen una búsqueda local que se usa para mejorar las soluciones generadas con el método ávido-aleatorio. **GRASP fue propuesto originalmente para el set covering problem (Feo y Resende, 1989)**. Detalles de esta metodología y aplicaciones recientes se pueden encontrar en Feo y Resende (1995) y Pitsoulis y Resende (2002).

Un esquema del funcionamiento de nuestro algoritmo GRASP es el siguiente

*Repetir*

*Construir una solución por el método Ávido-aleatorio*

*Mejorar dicha solución mediante búsqueda local*

*Actualizar la mejor solución obtenida hasta ese momento*

*hasta alcanzar un criterio de parada*

El criterio de parada se alcanza cuando transcurren un predeterminado número de iteraciones (*max\_iter*) sin mejora. A continuación se describen los dos procedimientos principales: el método ávido aleatorio y la búsqueda local.

#### 3.1.- Procedimiento ávido-aleatorio

La función ávida que guía la entrada de variables en la solución esta basada en conocidísimos resultados sobre la descomposición de la varianza. Más concretamente, sea  $\mathbf{x}$  una variable cualquiera definida sobre los  $n$  casos considerados, es decir,  $\mathbf{x}^T = (x_1, x_2, x_3, \dots, x_n)$ ,  $ng$  el número de clases y  $nm_i$  el número de casos del grupo  $i$ ,  $i = 1 \dots ng$ . Sea así mismo

$\bar{x}$  : media de la variable  $\mathbf{x}$  en el conjunto de los  $n$  casos;

$\bar{x}_i$  : media de la variable  $\mathbf{x}$  en los casos de la clase  $i$ ;  $i = 1.. ng$ ;

$cl(j)$  : clase a la que pertenece el individuo  $j$ .

Se definen

$$VT(\mathbf{x}) = \sum_{j=1}^n (x_j - \bar{x})^2 \quad (\text{Variabilidad Total})$$

$$VE(\mathbf{x}) = \sum_{i=1}^{ng} n_i (x_i - \bar{x})^2 \quad (\text{Variabilidad entregrupos})$$

$$VI(\mathbf{x}) = \sum_{j=1}^n (x_j - \bar{x}_{cl(j)})^2 \quad (\text{Variabilidad intragrupos})$$

y 
$$F(\mathbf{x}) = \frac{VE(\mathbf{x})}{VI(\mathbf{x})}.$$

Se sabe que  $VT(\mathbf{x}) = VE(\mathbf{x}) + VI(\mathbf{x})$ ; también es conocido que la función  $F(\mathbf{x})$  es un buen medidor de la capacidad discriminatoria de cada variable.

Sea  $S$  la solución que se va a construir, el procedimiento ávido-aleatorio se describe de la forma siguiente

1. *Iniciar: Hacer  $S = \emptyset$*
2. *Calcular  $F_j = F(v_j), j = 1.. m$*
3. *Determinar  $F_{max} = \max \{F_j / j = 1..m\}$  y  $F_{min} = \min \{F_j / j = 1..m\}$*
4. *Construir  $L = \{j / F_j \geq \alpha \cdot F_{max} + (1-\alpha) \cdot F_{min}\}$*
5. *Elegir  $j^* \in L$  aleatoriamente y hacer  $S = \{j^*\}$*
6. *Mientras  $|S| < p$  hacer:*
  - a. *Sea  $S = \{j_1, j_2, \dots, j_t\}$  (las variables que ya están en la solución)*

$\forall j \notin S$ : - *Determinar los valores de la variable  $r_j$  en el siguiente modelo lineal por mínimos cuadrados ordinarios*

$$v_j = \alpha + \beta_1 \cdot v_{j_1} + \beta_2 \cdot v_{j_2} + \dots + \beta_t \cdot v_{j_t} + r_j$$

- *Calcular  $F_j = F(r_j)$*
  - b. *Determinar  $F_{max} = \max \{F_j / j \notin S\}$  y  $F_{min} = \min \{F_j / j \notin S\}$*
  - c. *Construir  $L = \{j / F_j \geq \alpha \cdot F_{max} + (1-\alpha) \cdot F_{min}\}$*
  - d.  *$j^* \in L$  aleatoriamente y hacer  $S = S \cup \{j^*\}$*

Como se observa la función  $F$ , antes definida, es la que guía el procedimiento de selección de variables. Sin embargo en cada paso no se elige necesariamente la variable correspondiente al mayor valor de  $F$ ,  $F_{max}$ . En este caso se construye un conjunto  $L$  (denominado “lista de candidatos”), formado por los de mayor valor, y se elige aleatoriamente uno de esa lista.

Inicialmente la función guía es el valor de la función  $F$  en las variables originales. Posteriormente se usa el valor de la  $F$  pero no en las variable originales candidatas a entrar, sino en los residuos que se obtienen al quitar en dichas variables la información que nos suministran las variables que ya están en la solución  $S$ . Esta idea la usan algunos programas estadísticos conocidos como BMDP y SPSS en sus procedimientos para seleccionar variables previos a la ejecución de los métodos de discriminación propiamente dichos. La diferencia con nuestro método GRASP, es que la selección de variables que usan estos programas es determinística, es decir, siempre se selecciona la variable correspondiente a  $F_{max}$ .

Precisamente, una de las ideas para el uso de estos procedimientos ávido aleatorios, es que la mejor solución a la que se llega repitiendo varias veces la ejecución de este procedimiento, suele ser mejor que la obtenida con la selección determinística. Como se verá en los apartados siguientes esto también ocurre en este caso.

El parámetro  $\alpha$  sirve para controlar el grado de aleatoriedad del procedimiento. A mayor valor de  $\alpha$  menor aleatoriedad. Si  $\alpha = 0$ , el procedimiento es totalmente aleatorio, ya que  $L$  o “Lista de candidatos”, estaría formada por todos los variables que no están en la solución; Si  $\alpha = 1$   $L$  estaría únicamente formado por la variable correspondiente a  $F_{max}$ . En adelante, en este trabajo denominaremos *constructivo determinístico* al método propuesto cuando  $\alpha = 1$ .

### **3.2.- Procedimiento de búsqueda local**

A cada solución completa  $S$  generada por el procedimiento ávido aleatorio se la mejora posteriormente por un sencillo procedimiento de búsqueda local. En este caso, cada paso de la búsqueda local va a consistir en intercambiar una variable que esté dentro de la solución por otra que este fuera. Más concretamente, sea  $S$  una solución se define

$$N(S) = \{ S' / S' = S \cup \{j'\} - \{j\}, \forall j \in S, j' \notin S \}$$

El procedimiento de búsqueda local se puede describir como sigue

*Leer Solución Inicial  $S$*

*Repetir*

*Hacer  $valor\_ant = f(S)$*

*Buscar  $f(S^*) = \max \{ f(S') / S' \in N(S) \}$*

*Si  $f(S^*) > f(S)$  entonces hacer  $S = S^*$*

*hasta  $f(S^*) \leq valor\_ant$*

Como se observa el procedimiento finaliza cuando ningún intercambio produce mejora.

#### **4.- EXPERIENCIAS COMPUTACIONALES**

Para contrastar la eficacia de nuestro algoritmo GRASP y la de sus componentes, se han realizado una serie de pruebas. Para ello se usa la tabla de 141 ratios financieros para un total de 198 empresas (los datos se encuentran disponibles para aquellos lectores que estén interesados). De esta tabla se consideran tablas de menores dimensiones, con un número  $m$  de ratios financieros para las 198 empresas. Así se consideran los siguientes valores de  $m$ ,  $m = 40$  (correspondientes a los primeros 40 ratios financieros), 65, 90, 105 y 120.

El número de casos (empresas) que se consideran es 198, divididos en 2 clases (solventes y no solventes), con 131 y 67 elementos respectivamente. Se considera una partición, obtenida aleatoriamente  $A = A_1 \cup A_2$ , donde  $A_1$  tiene 100 elementos (66 solventes y 34 no) y  $A_2$  98 (65 y 33).

En la tabla 1 se muestra el resultado (proporción de aciertos) obtenido, por el constructivo determinístico, 20 ejecuciones del método ávido aleatorio ( $\alpha = 0.85$ ), y nuestro GRASP ( $\alpha = 0.85$  y  $max\_iter = 20$ )

$m$	$p$	Constructivo determinístico	Avido-aleatorio	GRASP
40	4	0,67346939	<b>0,70408163</b>	0,7755102
	5	0,69387755	0,69387755	0,7755102
	6	0,68367347	0,68367347	0,7755102
	7	0,68367347	0,68367347	0,79591837
	8	0,71428571	0,71428571	0,80612245
65	6	0,67346939	<b>0,71428571</b>	0,80612245
	7	0,69387755	0,69387755	0,81632653
	8	0,70408163	0,70408163	0,82653061
	9	0,70408163	0,70408163	0,84693878
	10	0,73469388	0,68367347	0,85714286
90	8	0,65306122	<b>0,75510204</b>	0,85714286
	9	0,68367347	<b>0,75510204</b>	0,85714286
	10	0,69387755	<b>0,74489796</b>	0,86734694
	11	0,69387755	<b>0,70408163</b>	0,86734694
	12	0,67346939	<b>0,75510204</b>	0,86734694
105	10	0,64285714	<b>0,74489796</b>	0,87755102
	11	0,60204082	<b>0,75510204</b>	0,87755102
	12	0,60204082	<b>0,71428571</b>	0,87755102
	13	0,60204082	<b>0,7244898</b>	0,8877551
	14	0,59183673	<b>0,69387755</b>	0,87755102
120	12	0,66326531	<b>0,78571429</b>	0,90816327
	13	0,65306122	<b>0,78571429</b>	0,8877551
	14	0,68367347	<b>0,7244898</b>	0,90816327
	15	0,70408163	<b>0,73469388</b>	0,8877551
	16	0,68367347	<b>0,7244898</b>	0,8877551

**Tabla 1.- Resultados de las pruebas previas**

Como se observa en la tabla 1, la repetición del método ávido-aleatorio da mejores resultados que el método constructivo determinístico: 17 casos mejor (en negrita), 7 casos empate y sólo 1 peor. El método GRASP (integrando el método ávido-aleatorio y la búsqueda local) mejora notablemente los resultados del método ávido aleatorio aislado. Por tanto la búsqueda local, resulta eficaz para mejorar calidad de las soluciones obtenidas por los diferentes constructivos.

## 5.- CONCLUSIONES

En este trabajo se trata el problema de selección de variables en el ámbito de la clasificación aplicado al mundo financiero, concretamente a la predicción de la

insolvencia financiera en las empresas. Los métodos de solución que se proponen son tres: un algoritmo constructivo determinístico (en el que se inspiran algunos paquetes estadísticos), un algoritmo constructivo ávido-aleatorio y un procedimiento GRASP (ávido-aleatorio + búsqueda local). Tras la realización de las experiencias computacionales se observa como el procedimiento GRASP es el que nos permite obtener los mejores resultados.

## **6.- REFERENCIAS BIBLIOGRÁFICAS**

BALA J., DEJONG K., HUANG J., VAFAIE H. y WECHSLER H. (1996). "Using Learning to Facilitate the Evolution of Features for Recognizing Visual Concepts". *Evolutionary Computation*, 4, 3, 297-311.

FEO T.A. y RESENDE M.G.C. (1989). "A Probabilistic heuristic for a computationally difficult Set Covering Problem". *Oper Res Lett*, 8, 67-71.

FEO T.A. y RESENDE M.G.C. (1995). "Greedy Randomized Adaptive Search Procedures". *Journal of Global Optimization*, vol. 2, pp 1-27.

GARCÍA F.C., GARCÍA-TORRES M., MORENO PÉREZ J.M.y MORENO-VEGA J.M. (2003). "Búsqueda Dispersa para el Problema de la Selección de Variables". CAEPIA-2003.

INZA I., LARRAÑAGA P., ETXEBERRIA R. y SIERRA B. (2000) "Feature Subset Selection by Bayesian networks based optimization". *Artificial Intelligence*, 123, 157-184.

JOURDAN L., DHAENENS C. y TALBI E. (2001). "A Genetic Algorithm for Feature Subset Selection in Data-Mining for Genetics". *MIC 2001 Proceedings*, 4<sup>th</sup> Metaheuristics International Conference, 29-34.

KOHAVI R. (1995). "Wrappers for Performance Enhancement and Oblivious Decision Graphs". Stanford University, Computer Science Department.

PITSOULIS L.S. y RESENDE M.G.C. (2002). "Greedy Randomized Adaptive Search Procedures" in Handbook of Applied Optimization, P. M. Pardalos and M. G. C. Resende (Eds.), Oxford University Press, pp. 168-182.