Selección del Sistema de Conjuntos de Apoyo para ALVOT usando Búsqueda Secuencial Flotante

Erika Danaé López Espinoza, Jesús Ariel Carrasco Ochoa, José Francisco Martínez Trinidad
Departamento de Ciencias Computacionales
Instituto Nacional de Astrofísica, Óptica y Electrónica
Luis Enrique Erro No. 1 Sta. María Tonantzintla Puebla, CP 72840, México
Tel. [+52] (222) 266-3100 Fax [+52] (222) 266-3152
e-mail {danae, ariel, fmartine} @inaoep.mx.

RESUMEN

En varias aplicaciones de clasificación supervisada, los objetos que se desean clasificar se encuentran descritos mediante n-uplos de gran dimensionalidad en los cuales las comparaciones globales no aportan mucha información a los especialistas, por lo cual es necesario aplicar métodos basados en precedencia parcial, donde las comparaciones entre los objetos son hechas entre subdescripciones previamente seleccionadas. Por tal motivo es necesario contar con métodos de selección de variables que permitan elegir no sólo un subconjunto de características relevantes sino un conjunto de subconjuntos de variables, llamados sistema de conjuntos de apoyo, necesarios para los métodos basados en precedencia parcial. En este trabajo se presenta un algoritmo para el cálculo de los p mejores subconjuntos empleando Búsqueda Secuencial Flotante y el clasificador ALVOT. Este clasificador basado en precedencia parcial permite trabajar directamente bases de datos con descripciones en términos de variables cualitativas y cuantitativas. Los experimentos fueron realizados sobre algunas bases de datos, variando los parámetros de búsqueda. Algunas pruebas fueron realizadas empleando los Testores Típicos como el sistema de conjuntos de apoyo. Los resultados de ambos métodos de selección fueron comparados.

Palabras clave: ALVOT, búsqueda secuencial flotante, selección de variables.

I. INTRODUCCIÓN

Uno de los problemas del Reconocimiento de Patrones es la selección de variables, la cual se realiza con al menos uno de dos objetivos: para obtener una mejor representación; o para lograr una mejor clasificación, aumentando con esta última, en algunos casos, la velocidad de procesamiento.

La selección de variables consiste tradicionalmente en buscar entre un conjunto de características un subconjunto menor que mejore o mantenga la eficiencia del clasificador, pero para los métodos de clasificación basados en precedencia parcial lo que se busca es un sistema de conjuntos de apoyo, el cual es un conjunto de subconjuntos de características que son relevantes para hacer comparaciones parciales.

En varios estudios comparativos entre los diferentes métodos de selección de variables los algoritmos Búsqueda Secuencial Flotante (BSF) han mostrado ser superiores [1-3]. En estos comparativos función de evaluación es monótona, lo que implica que cuando se agrega una nueva variable al conjunto actual, el desempeño de la función mejora o mantiene. En otros estudios donde se considera criterio no monótono los resultados han sido también satisfactorios para los métodos flotantes [4,5]. embargo, en todos los trabajos anteriormente mencionados el método empleado para medir eficiencia de los subconjuntos seleccionados solo trabaja con descripciones cuantitativas, teniendo emplear etiquetas numéricas para manipular información no numérica. Entre estos métodos podemos mencionar a la distancia Mahalanobis, clasificador KNN (K Nearest Neighbor) o clasificador Gaussiano.

En Reconocimiento Lógico Combinatorio de Patrones existen clasificadores supervisados que permiten trabajar con información tanto cualitativa como cuantitativa. Además una de las ventajas de estos clasificadores es poder trabajar incluso existiendo ausencia de información [6]. ALVOT (algoritmos votación) es uno de los clasificadores desarrollados bajo este enfoque. Este clasificador se basa en principio de precedencia parcial, y por tal motivo requiere de un sistema de conjuntos de apoyo indique las subdescripciones que se tomarán cuenta para comparar y clasificar objetos.

En este trabajo se expone un algoritmo para seleccionar los p mejores subconjuntos de variables empleando BSF y ALVOT para medir el desempeño de los subconjuntos seleccionados. Se realizaron varios experimentos con diferentes bases de datos. Otra prueba fue considerar a los Testores Típicos para encontrar el sistema de conjuntos de apoyo. resultados de ambos métodos fueron comparados.

J. Díaz de León, G. González, J. Figueroa (Eds.): Avances en Ciencias de la Computación, pp. 152-156, 2003.

II. FUNDAMENTOS

BÚSQUEDA SECUENCIAL FLOTANTE

Los algoritmos BSF forman parte de los métodos que usan la estrategia Wrapper de selección de variables. En esta estrategia la selección se realiza usando un clasificador, el cual evalúa los subconjuntos de características en base a su eficiencia de clasificación J.

Existen dos métodos para realizar una búsqueda flotante, BSF Hacia Adelante y BSF Hacia Atrás.

La idea de la BSF Hacia Adelante es iniciar con un conjunto vacío de variables y después de realizar la mejor inclusión al conjunto, de la variable que maximizó la eficiencia del clasificador. Posteriormente se realizan exclusiones de variables mientras el subconjunto resultante mejore la eficiencia de clasificación, en comparación con la eficiencia del subconjunto del paso anterior. El algoritmo termina hasta alcanzar un subconjunto de variables de la cardinalidad buscada.

En la BSF Hacia Atrás la idea es la misma sólo que en lugar de iniciar con el conjunto vacío se inicia con el conjunto completo de variables y se realiza exclusión seguida de inclusiones.

A continuación se muestra el algoritmo de BSF Hacia Adelante [7].

1. Iniciar con el conjunto vacío, $Y=\phi$, $k=0$

   Terminar cuando k sea igual al número de características requeridas.

   (en la práctica se puede iniciar con $k=2$, aplicando dos inclusiones)

2. *Inclusión:* Seleccionar la característica mas significativa.

$$x^{+} = \arg\max_{x \subset X - Y_k} \left[J(Y_k + x) \right]$$

$$Y_{k+1} = Y_k + x^{+}; \quad k = k + 1$$

3. *Exclusión:* Seleccionar la característica menos significativa

$$x^{-} = \arg\max_{x \subset Y_k} \left[J(Y_k - x) \right]$$

Si $J(Y_k - x^{-}) > J(Y_k)$ entonces

$$Y_{k-1} = Y_k - x^{-}; \quad k = k - 1$$

ir al paso 3

de lo contrario

ir al paso 2.

El mejor desempeño del método de BSF Hacia Adelante se da cuando el subconjunto de variables que se está buscando no es muy grande, a diferencia de esto, la BSF Hacia Atrás tiene un mejor desempeño cuando el subconjunto de variables que se busca es grande o próximo al conjunto total.

En estos métodos de selección de variables como en todos los métodos secuenciales, el conjunto encontrado de características es sólo el de la cardinalidad que se haya requerido. En el algoritmo presentado en este trabajo proponemos encontrar los p subconjuntos con mejor comportamiento, para algún p especificado.

ALVOT

Como ya fue mencionado ALVOT es uno de los algoritmos desarrollados, dentro del Reconocimiento Lógico Combinatorio de Patrones, para realizar clasificación supervisada basándose en el principio de precedencia parcial. Bajo este principio la clasificación de un objeto se hace por medio de comparaciones entre subdescripciones previamente seleccionadas.

ALVOT realiza la clasificación mediante las siguientes seis etapas:

1. Definición del sistema de conjuntos de apoyo.
2. Definición de la función de semejanza.
3. Evaluación por fila dado un conjunto de apoyo fijo.
4. Evaluación por clase dado un conjunto de apoyo fijo.
5. Evaluación por clase para todo el sistema de conjuntos de apoyo.
6. Aplicación de la regla de solución.

En la primera etapa se define el sistema de conjuntos de apoyo. Entendiéndose por éste, cualquier conjunto de subconjuntos de atributos, que indican qué subdescripciones se toman en cuenta para realizar las comparaciones entre los objetos. A cada uno de estos subconjuntos se les llama un conjunto de apoyo.

En la segunda etapa se define la función de semejanza. Esta función establece en qué forma son comparadas las subdescripciones y debería reflejar cómo es que se realizan las comparaciones entre los objetos en la vida real.

En la tercera etapa, evaluación por fila dado un conjunto de apoyo fijo, se inicia el proceso de votación evaluando las semejanzas entre las diferentes subdescripciones de los objetos ya clasificados y los que se desean clasificar.

Posteriormente, en la evaluación por clase, dado un conjunto de apoyo fijo se totalizan las evaluaciones obtenidas para cada uno de los objetos ya clasificados respecto a los objetos que se están clasificando.

En la evaluación por clase para todo el sistema de conjuntos de apoyo se totalizan las evaluaciones obtenidas para cada una de las clases para todo el sistema de conjuntos de apoyo.

Por último se aplica la regla de solución, la cual establece el criterio que se tomará para decidir en qué clase se clasificarán los objetos considerando las votaciones obtenidas en la etapa anterior.

Testores Típicos

En el Reconocimiento Lógico Combinatorio de Patrones la selección de variables se realiza mediante la teoría de testores. Los Testores Típicos o un subconjunto de ellos puede ser empleados como el sistema de conjuntos de apoyo para clasificador ALVOT.

Un subconjunto de variables T es un Testor si considerando sólo las variables de T no existen subdescripciones iguales de objetos de clases diferentes, es decir, los objetos de la clase i no se confunden con algún otro objeto de la clase j siendo $i \neq j$.

Un Testor T es un Testor Típico (irreducible) si al eliminar cualquiera de las variables de T resulta que T deja de ser Testor. Esto significa que no existe algún otro Testor T' tal que $T' \subset T$.

Los Testores Típicos son combinaciones irreducibles de características que permiten diferenciar objetos de clases diferentes. Es natural pensar que si una variable aparece en muchas combinaciones irreducibles o Testores Típicos, resulta más difícil prescindir de ella. Sobre la base de esta idea, Zhuravlev formula su definición de peso informacional de una variable como la frecuencia relativa de aparición de esa variable en la familia de todos los Testores Típicos [6].

Sea τ el número de Testores Típicos que tiene una muestra y sea $\tau(j)$ el número de aquellos Testores Típicos en los que aparece la variable correspondiente a la característica x_j. Diremos que el *peso informacional* (*relevancia*) de x_j es:

$$P(x_j) = \frac{\tau(j)}{\tau} \qquad (1)$$

para $j = 1,...,n$.

III. ALGORITMO PROPUESTO

Como se ha visto, ALVOT requiere de un sistema de conjuntos de apoyo para realizar las comparaciones entre los objetos que se van a clasificar.

El algoritmo propuesto encuentra los p mejores subconjuntos para un p especificado, formando con éstos el sistema de conjuntos de apoyo para ALVOT.

Para encontrar los p mejores subconjuntos se realizó una extensión al algoritmo de BSF Hacia Adelante descrito en la sección II. Lo mismo aplica para la BSF Hacia Atrás.

1. Iniciar con el conjunto vacío $Y = \phi$; k=0;
   total_variables = total de características que describen a los objetos de la muestra.
   p = número de subconjuntos buscados.
2. Para i = 1 a *total_variables*
   2.1. Mientras Y no tenga la cardinalidad i
   Inclusión:

$$x^+ = \arg\max_{x \in X - Y_k} [J(Y_k + x)]$$

$$Y_{k+1} = Y_k + x^+; \quad k = k + 1$$

Exclusión:

$$x^- = \arg\max_{x \in Y_k} [J(Y_k - x)]$$

Si $J(Y_k - x^-) > J(Y_k)$ entonces

$$Y_{k-1} = Y_k - x^-; \quad k = k - 1$$

ir al paso Exclusión
de lo contrario
ir al paso Inclusión

2.2. Si ($i > p$)
agregar el nuevo Y_i y $J(Y_i)$ a los subconjuntos anteriores.
buscar y guardar los p mejores subconjuntos entre los i examinados.
de lo contrario
guardar Y_i y $J(Y_i)$

La modificación consistió en aumentar un ciclo (paso 2) con el objetivo de encontrar cada unos de los mejores subconjuntos de cardinalidad i empleando BSF Hacia Adelante (paso 2.1). Durante este ciclo van almacenando los p primeros subconjuntos independientemente de su eficiencia de clasificación (paso 2.2 si $i \leq p$). Una vez que se hayan alcanzado $p+1$ subconjuntos, se van guardando sólo los mejores.

El algoritmo termina hasta que son revisados todos los subconjuntos de cardinalidad i, donde i=*1,...,total_variables*.

IV. EXPERIMENTOS

En esta sección se presentan los resultados obtenidos con el algoritmo desarrollado. Las pruebas fueron realizadas sobre algunas bases de datos tomadas de [8].

La eficiencia de cada conjunto fue medida en base los objetos clasificados correctamente con clasificador ALVOT. Para los experimentos muestra de aprendizaje y la muestra de prueba fueron todos los objetos de la base de datos.

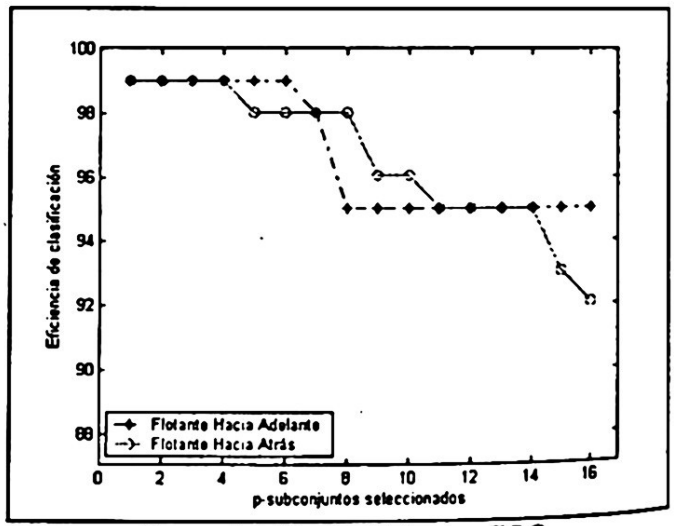

Figura 1. Métodos flotantes ZOO.

Los Testores Típicos que se consideraron para las pruebas se seleccionaron en base a su relevancia, esto se hizo considerando el peso informacional de los rasgos en base a la Ec.(1). La relevancia del Testor Típico se calculó con la siguiente formula:

$$R(\tau_j) = \frac{\sum_{x_{ji} \in \tau_j} P(x_{ji})}{|\tau_j|} \quad (2)$$

donde τ_j es el Testor en cuestión, $|\tau_j|$ es la cardinalidad del Testor y $P(x_{ji})$ es el peso informacional Ec.(1) de la característica x_{ji} del Testor j en el rasgo i.

El primer experimento se realizó con la base de datos Zoo. Esta base tiene 101 animales distribuidos en 7 clases. Cada uno se encuentra descrito con 16 atributos de los cuales 15 son Booleanos y 1 es nominal. La mayor eficiencia de clasificación para el método de BSF Hacia Adelante se alcanzó con 1 y hasta con 6 subconjuntos, mientras que para BSF Hacia Atrás la mayor eficiencia fue con 1 a 4 subconjuntos (fig. 1). Por otro lado en esta muestra se encontraron 32 Testores Típicos siendo la mayor eficiencia a partir de un conjunto formado por cinco subconjuntos. El comportamiento de ambos métodos se puede observar en la fig. 2.

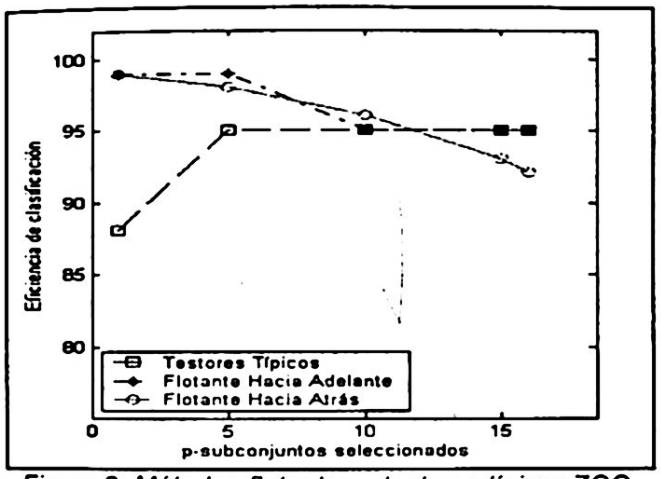

Figura 2. Métodos flotantes y testores típicos ZOO.

El segundo experimento se realizó sobre la base de datos HEPATITIS. Ésta contiene 115 objetos agrupados en 2 clases con 19 atributos de los cuales 6 son numéricos y 13 Booleanos. En esta base de datos existe ausencia de información. La mejor eficiencia para BSF Hacia Adelante se alcanzó con uno y 14 subconjuntos. Para BSF Hacia Atrás con uno y 12 subconjuntos (fig. 3). En esta base se encontraron 35 testores típicos alcanzándose una mayor eficiencia con diez subconjuntos (fig. 4).

El tercer experimento se hizo sobre la base SPECTF heart data. Esta base tiene 80 objetos en 2 clases con 22 atributos todos Booleanos. Se encontraron 26 Testores Típicos dando la mejor eficiencia con un solo subconjunto (fig. 6). Para BSF Hacia Adelante se obtuvo la mayor eficiencia con 1 a 5 y 14 a 18 subconjuntos, mientras que para BSF Hacia Atrás con 1 a 8 subconjuntos (fig. 5 y 6).

EL último experimento se realizó sobre la base de datos FLAGS. Tiene 193 objetos distribuidos en 8 clases y descritos con 28 atributos. Existen 15 atributos nominales, 2 numéricos y 11 Booleanos. Se encontraron 1469 Testores Típicos obteniéndose una mayor eficiencia con un conjunto de cinco testores típicos (fig. 8). Para BSF Hacia Adelante la mayor eficiencia se obtuvo con 1, 2 y 6 subconjuntos. Para BSF Hacia Atrás con 23 a 25 subconjuntos (fig. 7).

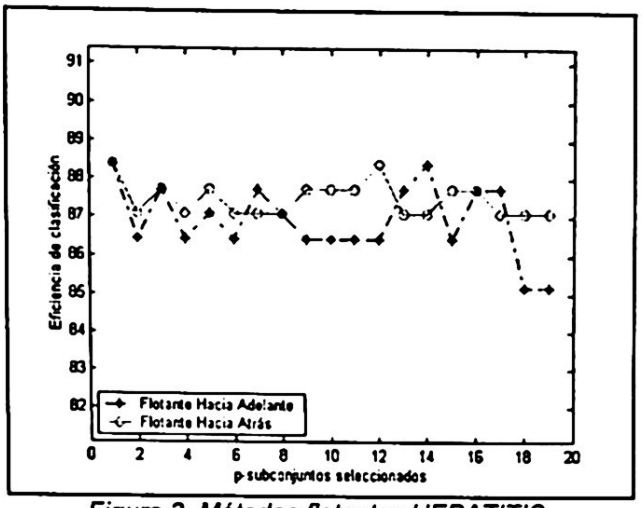

Figura 3. Métodos flotantes HEPATITIS.

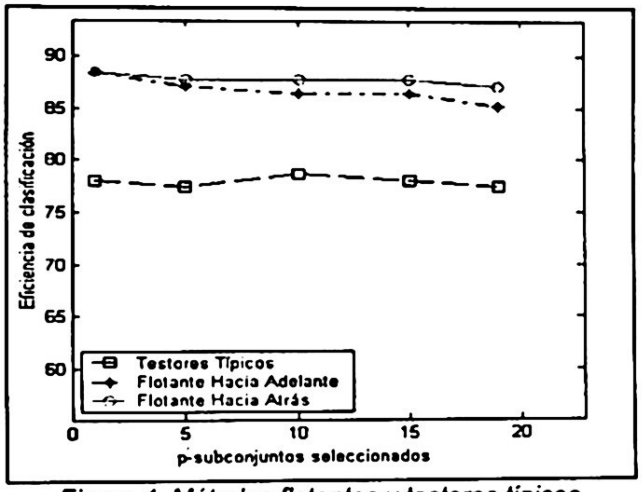

Figura 4. Métodos flotantes y testores típicos HEPATITIS.

IV. CONCLUSIONES

La selección de variables es una tarea muy útil cuando los objetos con los que se está trabajando se encuentran descritos mediante n-uplos de gran dimensionalidad. Con ésta se puede conseguir un subconjunto de características mucho menor al original y reducir el tiempo de procesamiento y/o mejorar la calidad de la clasificación.

Los métodos de selección de variables bajo la estrategia Wrapper han empleado comúnmente clasificadores con los que sólo se trabaja información numérica. Además sólo se busca un único conjunto de variables de cardinalidad k. En este trabajo se presentó un algoritmo de selección de variables

encontrando los p mejores subconjuntos empleando BSF y ALVOT para medir el desempeño de los subconjuntos seleccionados. Emplear este clasificador permitió trabajar directamente con objetos descritos mediante variables cualitativas y cuantitativas. Debido a la flexibilidad de ALVOT se pudo realizar un experimento sobre una base de datos en la cual existía ausencia de información.

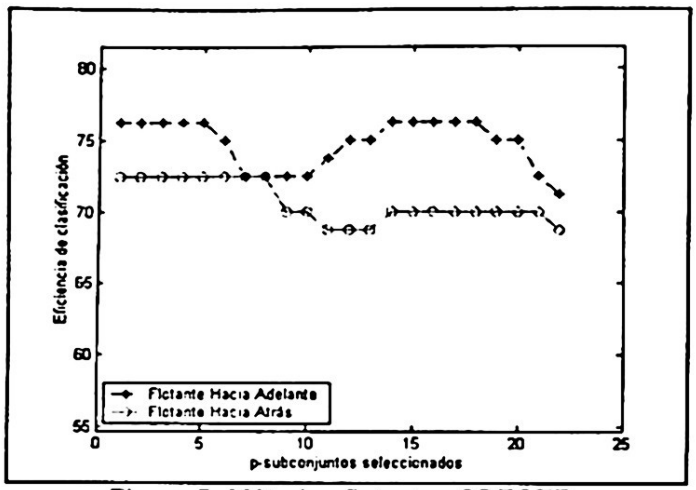

Figura 5. Métodos flotantes SPECTF.

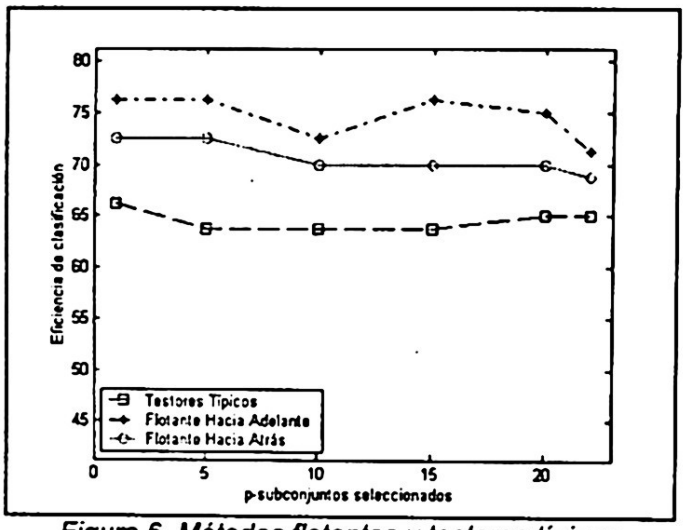

Figura 6. Métodos flotantes y testores típicos SPECTF.

En los experimentos realizados el mejor desempeño lo tuvieron los métodos flotantes, sin embargo al usar Testores Típicos la eficiencia de clasificación puede ser mejorada aumentando el número de subconjuntos.

El algoritmo expuesto es una alternativa para obtener el sistema de conjuntos de apoyo para ALVOT empleando BSF. Sin embargo, existen diferentes maneras de encontrar el sistema de conjuntos de apoyo empleando esta técnica, mismas que estamos explorando.

Agradecimiento. Este trabajo fue financiado por CONACYT México bajo los proyectos I38436-A y J38707-A.

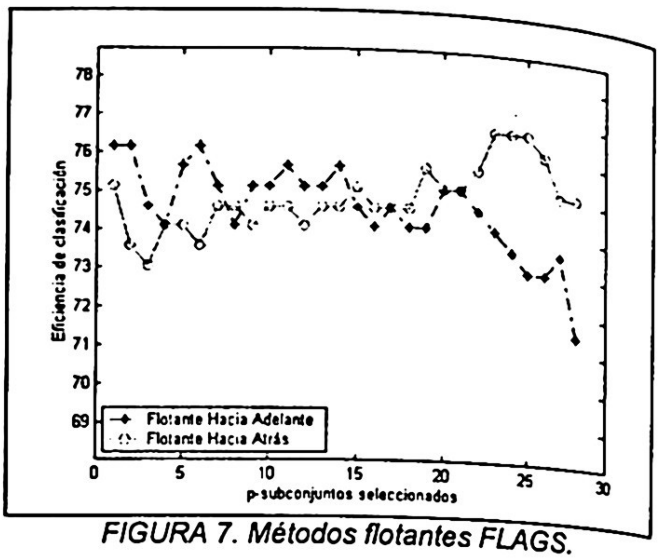

FIGURA 7. Métodos flotantes FLAGS.

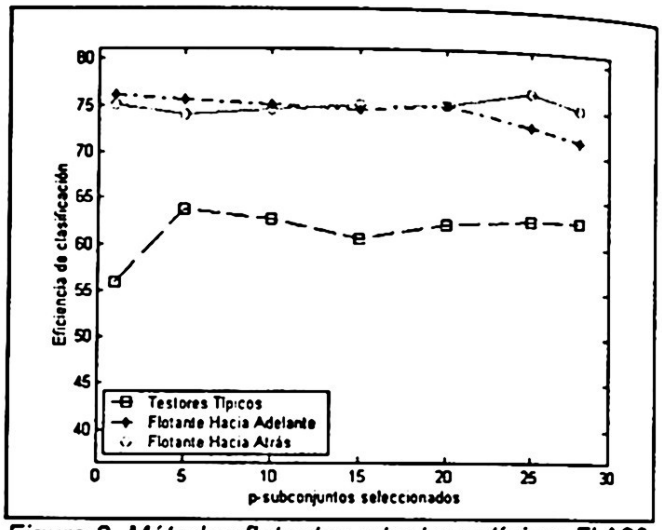

Figura 8. Métodos flotantes y testores típicos FLAGS.

REFERENCIAS

[1] Zongker D., Jain A. "Algorithms for feature selection: evaluation", In 13th International Conference on Pattern Recognition, 1996, pp. 18-22.

[2] Jain, A., Zongker, D., "Feature selection: Evaluation, application and small sample performance", IEEE Transactions on PAMI 19, 1997, pp.153-158.

[3] J. J. Ferri, P. Pudil, M. Hatef, y J. Kittler, "Comparative study of techniques for large-scale feature selection," in Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems, eds. E. S. Gelsema y L. N. Kanal, Elsevier Science B.V., 1994, pp. 403-413.

[4] P. Pudil, F.J. Ferri, J. Novovičová, and J. Kittler. "Floating search methods for feature selection with non monotonic criterion functions", In Proceedings of the 14th International Conference on Pattern Recognition (ICPR'97) 1994, pp. 279-283.

[5] M.Kudo and J. Sklansky "A comparative evaluation of medium- and large-scale feature selectors for pattern classifiers", In 1st International Workshop on Statistical Techniques in Pattern Recognition (STIPR'97), Prague, 1997, pp. 91-96.

[6] J. Ruiz Shulcloper, Adolfo Guzmán Arenas y J. Fco. Martínez Trinidad, "Enfoque Lógico Combinatorio al Reconocimiento de Patrones", Editorial Politécnica, ISBN: 970-18-2384-1, 1999.

[7] Ricardo Gutierrez Osuna, "Feature selection: sequential", Introduction to Pattern Recognition, Wright State University http://faculty.cs.tamu.edu/rgutier/courses/cs790_wi02/

[8] Machine Learning Databases, University of California, Irvine, Department of Information & Computer Science http://ftp.ics.uci.edu/pub/machine-learning-databases/