Un nuevo algoritmo de selección de rasgos basado en la teoría de los conjuntos aproximados

A new algorithm for feature selection based on rough sets theory

*Yailé Caballero[a, *], Delia Álvarez[a], Analay Baltá[a], Rafael Bello[b], María García[b],*

[a]Departamento de Computación. Universidad de Camagüey. Carr. Circunv. Norte km 5 ½. CP: 74650. Camagüey, Camagüey, Cuba

[b]Departamento de Ciencia de la Computación. Universidad Central de Las Villas. Carretera de Camajuaní km 5 ½. CP 54830. Santa Clara, Villa Clara, Cuba.

(Recibido el 30 de junio de 2006. Aceptado el 12 de abril de 2007)

Resumen

La teoría de los conjuntos aproximados ha abierto nuevas tendencias en el desarrollo de las técnicas de análisis de datos. Dentro de estas es significativo el concepto de reducto, cuya obtención en un sistema de decisión es un proceso computacionalmente costoso aunque importante en análisis de datos y nuevo conocimiento. Debido a esto, se ha hecho necesario desarrollar diferentes variantes para calcular reductos. El presente trabajo investiga la utilidad que ofrece el modelo de los conjuntos aproximados en selección de rasgos y se presenta un nuevo método con el propósito de calcular un buen reducto. Este nuevo método consiste en un algoritmo glotón que usa heurísticas para encontrar un buen reducto en tiempos aceptables. Se presentan, además, los resultados experimentales obtenidos usando diferentes conjuntos de datos.

---------- *Palabras clave:* selección de rasgos, reducto, conjuntos aproximados

Abstract

Rough Sets Theory has opened new trends for the development of data analysis techniques. In this theory, the notion of *reduct* is very significant, but obtaining a *reduct* in a decision system is an expensive computing process although very important in data analysis and new discoveries. Because of this, it has been necessary

* Autor de correspondencia: teléfono: 53+ +3 +228 13 63, fax: +53 +3 +226 15 87, correo electrónico: yailec@yahoo.com (Y. Caballero).

to develop different variants to calculate *reducts*. The present work looks into the utility that offers Rough Sets in feature selection and a new method is presented with the purpose of calculate a good *reduct*. This new method consists of a greedy algorithm that uses heuristics to work out a good *reduct* in acceptable times. Experimental results obtained by using different data sets are presented.

---------- *Key words:* Feature selection, *reduct*, rough sets.

Introducción

Con el crecimiento actual de los volúmenes de información en bases de datos tanto científicas como corporativas, la necesidad de determinar qué información es realmente importante se convierte en un reto para los desarrolladores con fines de facilitar las tareas para la minería de datos *(data mining)* y el aprendizaje automatizado *(machine learning).*

La selección de rasgos en un conjunto de datos es un problema en cuya solución se han utilizado diversas variantes dentro de la inteligencia artificial, debido a su utilidad en gran cantidad de áreas de la informática y la ciencia de la computación que tienen como denominador común reducir la dimensionalidad de los problemas. De esta forma, grandes volúmenes de datos se pueden manipular rápidamente al extraer de ellos sólo la información necesaria que los describa, sin perder la calidad del sistema, y obteniendo conocimiento sobre ellos. Aplicaciones recientes de la selección de rasgos se pueden encontrar en clasificación de contenido web, procesamiento de textos, diagnóstico médico, entre otras.

Los conjuntos de datos consisten en un sistema de información descrito por un conjunto de atributos y los objetos que representan con combinaciones de valores válidos dentro del dominio de cada atributo. De esta forma, la selección de rasgos en un sistema de información de este tipo, consiste en obtener un subconjunto de atributos tal que describa el sistema como si se tratara del conjunto completo. Esto quiere decir que el proceso se centra en encontrar los atributos más importantes dentro de los que se han utilizado para representar los datos y eliminar aquellos que se consideran irrelevantes y hacen más difícil el proceso de descubrimiento de conocimiento dentro de una base de datos. Dicho de otro modo, la selección de rasgos representa el problema de encontrar un subconjunto óptimo de características (rasgos o atributos) en una base de datos y según cierto criterio, tales que se pueda generar un clasificador con la mayor calidad posible a través de un algoritmo inductivo que corra sobre los datos, pero sólo tomando en cuenta el subconjunto de atributos obtenido [1].

El proceso de selección de rasgos consta de dos componentes principales: una función de evaluación y un método de búsqueda. La función de evaluación permite calcular la calidad de un subconjunto de rasgos; mientras que el método de búsqueda, por lo general heurística, es el encargo de generar los subconjuntos de rasgos; dados N rasgos se tienen 2^N - 1 posibles subconjuntos. Seleccionar los rasgos relevantes de un conjunto de datos es una tarea necesaria en el aprendizaje automatizado *(machine learning),* dada su importancia en el descubrimiento de reglas y relaciones en grandes volúmenes de datos entre otras aplicaciones. Por esta razón, la selección de características relevantes de un conjunto de datos con tiempos y costo de cómputo aceptables, ha sido en los últimos años tema de investigación de muchos autores en diferentes variantes [2, 3, 4, 5, 6, 7].

Una herramienta matemática muy potente para la selección es la teoría de conjuntos aproximados (Rough Sets Theory), propuesta por el profesor polaco Z. Pawlak y su equipo en 1982 [8]. La filosofía de conjuntos aproximados (Rough Sets) se basa en asumir que existe información asociada con cada objeto del universo de discurso [9]. Un conjunto de entrenamiento se representa por una tabla donde cada fila representa un objeto y cada fila un atributo, a este conjunto se le llama sistema de información, más formalmente, es un par S= (U, A) donde U es un conjunto no vacío y finito de objetos llamado universo y A es un conjunto no vacío y finito de atributos. Un sistema de decisión es cualquier sistema de información de la forma SD=(U, $A \cup \{d\}$) donde $d \notin A$ es un atributo de decisión, o sea, un atributo que indica a qué grupo pertenece el objeto.

El modelo de los conjuntos aproximados posee importantes ventajas en el análisis de datos. La principal se basa únicamente en los datos originales y no requiere de información externa para obtener conocimiento sobre el sistema, de forma que no es necesario hacer suposiciones sobre

este; la otra ventaja importante consiste en que esta herramienta permite analizar atributos tanto cuantitativos como cualitativos.

Seguidamente se introduce la teoría de conjuntos aproximados (epígrafe 2) y posteriormente se presenta un algoritmo para la selección de rasgos, en el que se estudian tres alternativas de función de evaluación, las cuales se usan como función de evaluación heurística en un algoritmo glotón *(greedy)* (epígrafe 3), y finalmente se desarrolla un estudio experimental del mismo con su correspondiente procesamiento estadístico (epígrafe 4).

La teoría de conjuntos aproximados

Con frecuencia se almacenan grandes volúmenes de información en bases de datos con diferentes objetivos; estos pueden ser adquiridos de mediciones obtenidas por expertos humanos o de representaciones de hechos específicos de problemas de la vida cotidiana.

Una base de datos puede contener cierta cantidad de atributos que son redundantes u objetos que se encuentran repetidos en distintos niveles de esta, pero sobre todo sucede que contiene información insuficiente o incompleta. La teoría de conjuntos aproximados emerge desde el contexto del aprendizaje supervisado, donde los conjuntos de datos se refieren a un universo de objetos descritos por un conjunto de atributos y cada objeto pertenece a una clase predefinida por uno de los atributos, llamado atributo de decisión. Para una aproximación inicial, considérese que cada conjunto de datos está representado por una tabla, donde cada fila constituye un caso, un evento, un paciente o simplemente un objeto; y cada columna, un atributo que puede ser una variable, una observación, una columna, una propiedad, etc., tal que posee un valor específico para cada objeto. A esta tabla se le llama sistema de información, una definición formal del mismo según [10], es:

$I = \left\{ U , A , V_q , f_q \right\} \quad q \in A$ donde:

U es un conjunto finito y no vacío de objetos llamado universo.

A es un conjunto finito y no vacío de atributos.

Vq es el conjunto de valores posibles para cada atributo de A de modo que $q : U \to V_q$ para todo $q \in A$.

fq es una función de información tal que $f_q : U \to V_q$.

Dado un sistema de información como el descrito anteriormente, existe una relación de equivalencia:

$$IND_1(B) = \{(x, y) \in U^2 / \forall a \in B\, a(x) = a(y)\}$$

En la que IND_1 es la relación de inseparabilidad basada en el subconjunto de atributos B. Si $(x, y) \in IND_I(B)$ entonces X y Y son inseparables con respecto a B.

Una relación de inseparabilidad induce una partición del universo. Estas particiones pueden ser usadas para construir nuevos subconjuntos del universo. Los subconjuntos de mayor interés son aquellos que tienen incluso el mismo valor para el atributo de decisión, sin embargo, puede suceder que los atributos condicionales de algunos objetos sean inseparables, mientras que el de decisión sea diferente; la solución a problemas de este tipo se encuentra en el uso de los conjuntos aproximados, los cuales se definen a través de sus aproximaciones superior e inferior.

Dado un sistema de información

$I = \left\{ U, A, V_q, f_q \right\}_{q \in A}$, $B \subseteq A$ y $X \subseteq U$

entonces se puede aproximar X usando sólo la información contenida en B construyendo las aproximaciones inferior y superior de X, denotadas B^* y B_* respectivamente, y definidas [11] de la siguiente forma:

$$B_* = \{x / [x]_B \subseteq X\} \text{ y } B^* = \{x / [x]_B \cap X \neq \phi\}$$

En [12] se definen de la siguiente forma: La aproximación inferior de un conjunto (con respecto a un conjunto dado de atributos) se define como la colección de casos cuyas clases de equivalencia están contenidas completamente en el

conjunto; mientras que la aproximación superior se define como la colección de casos cuyas clases de equivalencia están al menos parcialmente contenidas en el conjunto.

A partir de las aproximaciones inferior y superior, se pueden determinar medidas para inferir conocimiento sobre las bases de casos a través de una serie de medidas que se definen sobre estos conceptos, una de ellas es la *precisión de la clasificación,* que se especifica según la cantidad de valores de decisión que posee el sistema, la medida, también llamada *vaguedad* del concepto o conjunto X con respecto a la relación B, se puede caracterizar matemáticamente por el coeficiente:

$$\alpha_B(X) = \frac{|B_*(X)|}{|B^*(X)|} \qquad (1)$$

Donde $|X|$ denota la cardinalidad del conjunto X, y $0 \leq \alpha_B(X) \leq 1$. Si $\alpha_B(X) = 1$, el conjunto X será *duro* o *exacto* con respecto a la relación de equivalencia B, mientras que si $\alpha_B(X) < 1$, el conjunto X es *aproximado* o *vago* con respecto a B. Esta magnitud mide el grado de perfección o integridad del conocimiento sobre el conjunto X considerando los atributos incluidos en la relación de equivalencia. Un punto importante dentro del análisis de datos es el descubrimiento de dependencias entre estos. Intuitivamente, un conjunto de atributos D depende totalmente de un conjunto de atributos C, denotado $C \rightarrow D$, si todos los valores de los atributos de D están únicamente determinados por valores de los atributos de C. En otras palabras, D depende totalmente de C si existen dependencias funcionales entre los atributos de C y D [9]. Formalmente la dependencia puede ser definida de la siguiente forma: si C y D son subconjuntos de A entonces se dice que D depende de C en grado k $(0 \leq k \leq 1)$ si:

$$k = \gamma(C, D) = \frac{|POS_C(D)|}{|U|} \qquad (2)$$

Donde POS es la región positiva de la partición U/D con respecto a C, o sea, el conjunto de todos los elementos de U que pueden ser únicamente clasificados en bloques de la partición U/D por medio de C.

Si k = 1, se dice que D depende totalmente de C y si k < 1, se dice que D depende parcialmente, en grado k, de C.

Si D depende totalmente de C entonces $IND(C) \subseteq IND(D)$, lo que significa que la partición generada por C es mucho mejor o más fina que la generada por D. A la dependencia de los atributos también se le llama calidad de la clasificación.

Los conjuntos aproximados se aplican eficientemente en la reducción de atributos o selección de rasgos sobre la base del concepto de reducto. Dado un sistema de información $I = (U, A)$, un reducto es un conjunto mínimo de atributos $B \subseteq A$ tal que $IND_I(B) = IND_I(A)$. En otras palabras, un reducto es un conjunto mínimo de atributos de A que preservan la partición del universo y de esta forma la habilidad de ejecutar clasificaciones como si se tratara del mismo conjunto A. El uso de reductos en selección de rasgos ha sido estudiado por varios autores [3, 10, 12, 13, 14, 15].

Sin embargo, esta beneficiosa alternativa se encuentra limitada por el hecho de que encontrar esos conjuntos mínimos de atributos es un problema NP-*hard*. Esto constituye un cuello de botella de la teoría de conjuntos aproximados [9], dado que computar todos los reductos no es una tarea con alto costo computacional. Diversos autores han propuesto métodos para el cálculo de reductos a través de los conjuntos aproximados [1, 5, 11, 16].

RSReduct: un nuevo método de selección de rasgos

RSReduct es un método que trata de encontrar un reducto de manera que éste sea lo suficientemente bueno para el análisis de datos en tiempos aceptables. Para ello se utiliza la búsqueda heurística como estrategia de búsqueda, debido a que si se tomara en cuenta la variante

exhaustiva o completa, el consumo de tiempo y recursos de cómputo sería bastante grande y la no determinística dificulta saber cuándo aparece un subconjunto mínimo. El método consiste en un algoritmo glotón que comienza por un conjunto vacío de atributos, y a través de heurísticas va formando un reducto mediante la selección de los atributos uno a uno de una lista, hasta que se cumple la condición de parada; en la lista de atributos; estos se encuentran ordenados según el valor arrojado por la función de evaluación heurística para cada uno de estos. Para la construcción de las funciones de evaluación heurística se siguen criterios del método ID3 con respecto a la entropía y la ganancia de los atributos, dependencia entre atributos mediante los conjuntos aproximados, así como la opción de otorgar costos a los atributos, es decir, manipulación de atributos con costos diferentes.

En este algoritmo se utilizan las medidas R(A) y H(A) que se proponen en [17].

$$R(A) = \sum_{i=1}^{k} \frac{|S_i|}{|S|} e^{(1-C_i)} \quad (3)$$

En la expresión *k* es el número de valores diferentes del rasgo A. C_i es el número de clases diferentes presentes en los objetos que tienen el valor i para el rasgo A y $\frac{|S_i|}{|S|}$ es la frecuencia relativa del valor i en S (cantidad de objetos con el valor i en el rasgo A sobre la cantidad de objetos de toda la muestra). La principal idea de esta medida es maximizar la heterogeneidad entre objetos que pertenecen a clases diferentes y minimizar la homogeneidad entre aquellos que son de la misma clase, además $0 \le R(A) \le 1$.

H(A) también se obtiene a través del siguiente algoritmo:

I. Se calcula el vector R (T)= (R(A1),...). Para todos los atributos del problema se calcula su R(A) y así con todos los valores se forma el vector R (T).

II. Se determinan los *n* mejores atributos por los cálculos del paso anterior. El valor de *n* se puede seleccionar por el usuario. Como resultado de este paso se obtiene el vector RM = (R(Ai), R(Aj),...) con n = $|RM|$

III. Se determinan las combinaciones de *n* en *p* (valores seleccionados por el usuario) desde los atributos seleccionados en el paso II. Se obtiene el vector de combinaciones $Comb = (\{a_i, a_j, a_k\},..., \{a_i, a_t, a_p\})$. Por ejemplo: si *n = 4* y *p = 3* y los atributos seleccionados en II son: (a_1, a_3, a_5, a_8) el vector de combinaciones $C_p^n = \frac{n!}{p!(n-p)!}$ tendría cuatro componentes que serían: $Comb = (\{a_1, a_3, a_5\}, \{a_1, a_3, a_8\}, \{a_3, a_5, a_8\}, \{a_1, a_5, a_8\})$.

IV. Se calcula el grado de dependencia de las clases con respecto a cada una de las combinaciones obtenidas en el paso anterior. Como resultado de este paso se obtiene el vector de dependencias: $DEP = (k(comb_1, d),..., k(comb_r, d))$, donde k representa la medida para el grado de dependencia entre atributos de los conjuntos aproximados con respecto a los valores de decisión *d.*

V. Para cada atributo A se calcula H(A) según la siguiente ecuación:

$$H(A) = \sum_{\forall i / A \in comb_i} k(comb_i, d)$$

La ganancia radial ***G(A)*** o ganancia de Quinlan es una medida alternativa para seleccionar atributos [18]:

$$G(A) = \frac{Ganancia(S, A)}{SplitInformation(S, A)} \quad (4)$$

Donde:

Ganancia(S, A) = entropía(s)

$$- \sum_{v \in valores(A)} \frac{|S_v|}{|S|} Entropía(S_v)$$

Donde, *valores (A)* es el conjunto de valores posibles por el atributo A y S_v es el subconjunto de *S* para el cual A tiene el valor *v*, es decir, $S_v = \{s \in S | A(S) = v\}$.

$$Entropía(S) = \sum_{i=1}^{c} - P_i \log_2 P_i$$

Es la clásica medida del valor de entropía de un sistema de información propuesta por Shannon y donde P_i es la proporción de *S* perteneciente a la clase *i*. La $Entropía = 0$ si todos los elementos en S pertenecen a la misma clase y la $Entropía = 1$ si todas las clases tienen igual número de ejemplos.

$$SplitInformation(S, A) = -\sum_{i=1}^{C} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Donde *C* son los valores del atributo A. Esta medida es la entropía de *S* con respecto al atributo A. Para calcular *C(A)* existen dos variantes, ambas aparecen en [18]:

Medida de Schlimmer y Tan:

$$C(A) = \frac{Ganancia^2(S, A)}{Cost(A)} \quad (4)$$

Donde *Cost(A)* es un parámetro entrado por el usuario que representa el costo del atributo A.

Medida de Núñez:

$$C(A) = \frac{2^{Ganancia(S,A)} - 1}{(Cost(A) + 1)^w} \quad (5)$$

Donde *Cost(A)* es un parámetro entrado por el usuario que representa el costo del atributo A y *w* es un valor constante entre 0 y 1 que determina la importancia relativa del costo contra la información de la ganancia.

Sobre la base de las medidas ya planteadas, se proponen tres funciones de evaluación heurística para el algoritmo RSReduct, estas permiten obtener una medida de la relevancia de los atributos dentro del conjunto de datos

Heurística 1. RG(A) = R(A) + H(A) (6)

Heurística 2. RG(A) = H(A) + G(A) (7)

Heurística 3. RG(A) = H(A) + C(A) (8)

A continuación, se muestran los pasos del algoritmo RSReduct:

P1. Formar la tabla de distinción.

Sea B matriz binaria $(M^2 - M)/2$ x $(N+1)$. Cada fila corresponde a un par de objetos diferentes. Cada columna de esta matriz corresponde a un atributo, la última columna corresponde a la decisión (tratada como un atributo).

Sea *b((k, n), i)* un elemento de B correspondiente al par O_k, O_n y al atributo *i*, para $i \in \{1,..N\}$:

$$b((k,n),i) = \begin{Bmatrix} 1, sia_i(O_k) \neg \Re a_i(O_n) \\ 0, sia_i(O_k) \Re a_i(O_n) \end{Bmatrix} parai \in \{1,...N\}$$

$$b((k,n),N+1) = \begin{Bmatrix} 0, sid_i(O_k) \neq d_i(O_n) \\ 1, sid_i(O_k) = d_i(O_n) \end{Bmatrix}$$

Donde $\Re$ es una relación de similaridad en dependencia del tipo del atributo a_i.

P2. Para cada atributo "A" se calcula el valor de RG(A) por cualquiera de las tres heurísticas. Se forma una lista ordenada de atributos comenzando por el atributo más relevante (el que maximice RG(A)).

Heurística1. RG(A) = R(A) + H(A)

Heurística2. RG(A) = H(A) + G(A)

Heurística3. RG(A) = H(A) + C(A)

P3. Se tiene $i = 1$, R = conjunto vacío y se tiene $A_1, A_2, ..., A_n$ lista ordenada de atributos según el paso 2, si $i \leq n$ entonces $R = R \cup A_i$, $i = i + 1$.

P4. Si *R* satisface la Condición I parar (que significa terminar).

$$\forall k,n \quad \forall a_i \in R \quad a_i(o_k) \Re a_i(o_n) \Rightarrow d(o_k) = d(o_n) \quad \text{(condición I).}$$

Encontrar un reducto significa *encontrar un conjunto mínimo de atributos* que cubran a B, es decir, un subconjunto que satisfaga:

$$\forall(k,n)\exists i \in R : b((k,n),i) = 1 Ub((k,n),N+1) = 1$$

Esto significa que para cada par de objetos que se tomen, o son de la misma clase o tienen algún atributo de *R* para el cual los objetos son "similares", es decir, cumplen la relación $\Re$.

P5. En otro caso repetir desde el paso 3.

Resultados experimentales

El algoritmo RSReduct fue probado con diferentes conjuntos de datos disponibles en el sitio ftp de la Universidad de California [19]. Algunas de estas bases de datos pertenecen a datos del mundo real *como Vote, Iris, Breast Cancer, Heart y Credit;* las otras representan resultados obtenidos en laboratorio como *Ballons-a, Hayes-Roth, LED, M-of-N, Lung Cancer y Mushroom.*

En un primer experimento se tomó la base de casos *Breast Cancer,* la cual representa un estudio del cáncer de mama realizado en julio de 1988 en el Instituto de Oncología perteneciente al Centro Médico Universitario de Ljubljana, Yugoslavia. Dicho conjunto de datos contiene 286 ejemplos de los cuales 201 pertenecen a una clase y 85 a otra; los ejemplos son descritos por 9 atributos algunos de los cuales son lineales y otros nominales, su descripción aparece en la tabla 1.

Tabla 1 Descripción de la base de casos *Breast Cancer*

Nombre del atributo	***Valores de su dominio***
Edad	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
Menopausia	lt40 *(antes de los 40)*, ge40 *(después o a los 40)*, premeno *(premenopausia)*
Tamaño del tumor	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59
Nodos invertidos	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
Capas de nodos	Yes, No
Grado de malignidad	1, 2, 3
Mama	Izquierda, Derecha
Cuadrante de mama	left-up, left-low, right-up, right-low, central
Irradiaciones	Yes, No
Clase	no-recurrence-events, recurrence-events

Luego de haber probado esta base de casos con el método RSReduct a través de las tres funciones de evaluación heurística definidas para este se obtuvieron los subconjuntos de rasgos que se muestran en la tabla 2.

En la tabla 3 se muestra un estudio cuantitativo de los resultados obtenidos usando las tres funciones de evaluación heurísticas definidas para RSReduct para diferentes bases de casos; se compilaron la longitud del reducto y el tiempo de ejecución, en segundos, para cada base de caso.

Tabla 2 Descripción de los reductos obtenidos para la base de casos *Breast Cancer*

Heurística utilizada	***Reductos obtenidos***
Heurística 1	Tamaño del tumor, grado de malignidad, edad, menopausia
Heurística 2	Tamaño del tumor, grado de malignidad, capas de nodos, nodos invertidos
Heurística 3 (en sus dos variantes)	Tamaño del tumor, grado de malignidad, edad, nodos invertidos, irradiaciones

Tabla 3 Resultados experimentales de las tres funciones de evaluación heurística de RSReduct para las diferentes bases de casos

Nombre de la base de casos (cantidad de casos, cantidad de atributos)	***Heurística 1***		***Heurística 2***		***Heurística 3***	
	Tiempo (segundos)	***Longitud del reducto***	***Tiempo (segundos)***	***Longitud del reducto***	***Tiempo (segundos)***	***Longitud del reducto***
Ballons-a (20,4)	5,31	2	3,12	2	16,34	2
Iris (150,4)	40,15	3	30,79	3	34,73	3
Hayes-Roth (133,4)	36,00	3	32,30	3	39,00	3
Bupa (345,6)	74,20	6	89,00	6	89,00	6
E. Coli (336,7)	57,00	5	41,15	5	46,60	5
Heart (270,13)	30,89	9	16,75	9	54,78	10
Pima (768,8)	110,00	8	110,00	8	110,00	8
Breast- Cancer (683,9)	39,62	4	31,15	4	32,56	5
Yeast (1484,8)	82,00	6	78,00	6	85,70	6
Dermatology (358,34)	148,70	8	125,90	8	190,00	9
Lung-Cancer (27,56)	25,46	7	18,59	7	31,50	8
LED (226,25)	78,10	9	185.00	8	185,00	9
M-of-N (1000,14)	230,26	6	162.50	6	79,40	6
Exactly (780,13)	230,00	11	215.00	11	230,00	11
Mushroom (3954,22)	86,20	8	64.10	8	67,20	8
Credit (876,20)	91,20	14	86.01	14	90,20	15
Vote (435,16)	37,93	12	21.25	11	26,90	12

Para ilustrar un poco mejor cuán buena fue la reducción del método propuesto, la figura 1 muestra el tamaño original de las bases de casos y el tamaño al que fueron reducidas al usar las tres heurísticas de RSReduct.

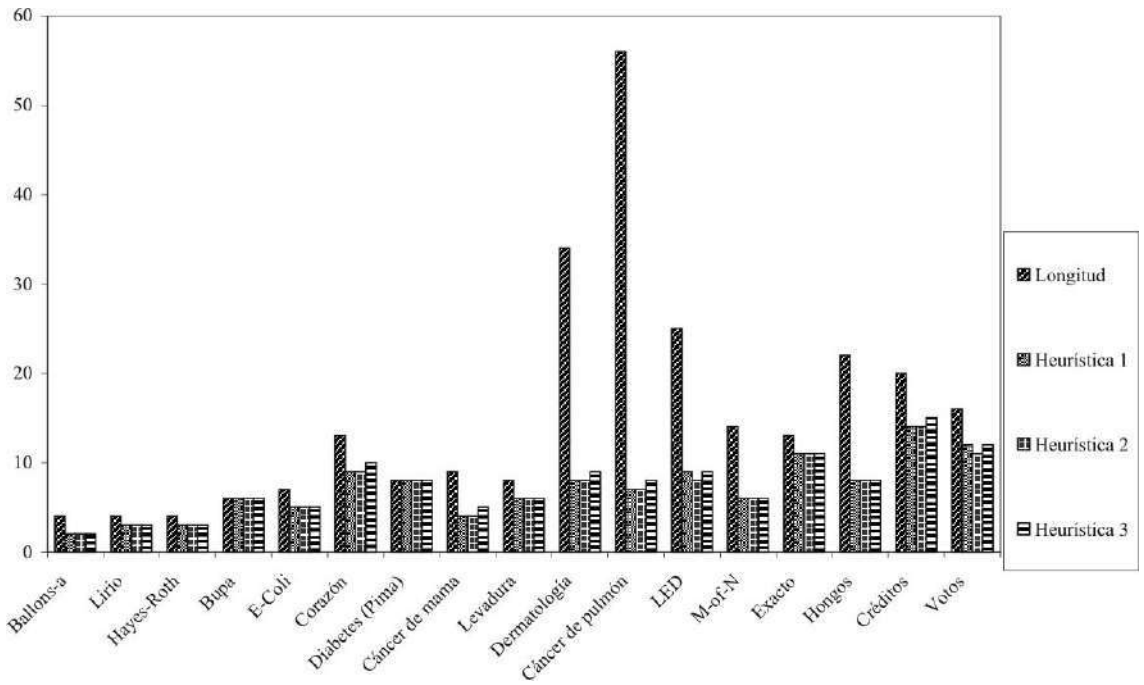

Figura 1 Representación gráfica de los resultados experimentales de RSReduct para las diferentes bases de casos

Estos resultados experimentales se compararon estadísticamente en aras de buscar diferencias con otros métodos de selección de rasgos implementados con técnicas de reconocimiento de patrones [20], algoritmos de estimación de distribuciones de tipo UMDA [20], y algoritmos genéticos [15]. Inicialmente se probaron cada uno de los algoritmos mencionados anteriormente en el mismo procesador y para las mismas bases de casos que RSReduct, con las medias de los resultados obtenidos en [20, 15], aunque en la tabla 4 se muestran las longitudes promedio de los reductos obtenidos para cinco bases de casos que se consideran representativas por la calidad de la reducción obtenida para las mismas y su heterogeneidad en cuanto a cantidad de objetos y atributos, que puede verificarse en la tabla 3; aun con estas características, tal como aparece en la tabla 4, hubo bases de casos para las cuales algunas de estas técnicas no pudieron obtener un reducto, las longitudes promedio obtenidas son bastante similares a las de RSReduct, especialmente con respecto a la segunda función de evaluación heurística del mismo.

Para hacer las comparaciones estadísticas, se usó la prueba de Kruskal-Wallis, esta es una prueba no paramétrica basada en suma de rangos que compara más de dos grupos relacionados entre sí de una vez con el objetivo de descubrir diferencias entre ellos. En la tabla 5 aparecen los resultados de la prueba de Kruskal-Wallis entre los algoritmos descritos que fueron comparados en cuanto a tiempo de ejecución

Tabla 4 Resultados experimentales para otros métodos de selección de rasgos

Bases representativas	*Promedio de la longitud del reducto obtenido por técnicas de reconocimiento de patrones*	*Promedio de la longitud del reducto obtenido con EDA en la variante UMDA*	*Promedio de la longitud del reducto obtenido con el algoritmo genético SAVGeneticReducer*
Breast Cancer	4,50	7,12	4,60
Lung Cancer	6,50	20,33	9,49
Mushroom	No pudo calcular	No pudo calcular	12,00
Heart	4,35	10,62	9,50
Dermatology	10,50	31,58	18,54

de los mismos; como se puede observar, para todos los casos los resultados fueron menores de 0,05 lo que indica un 95% de significancia estadística, en otras palabras, existen diferencias significativas entre los métodos a favor de RSReduct.

Tabla 5 Significancia de la prueba de Kruskal-Wallis entre las tres funciones heurísticas de RSReduct con otros métodos de selección de rasgos

Bases de casos representativas	*Significancia de Heurística 1 vs. otros métodos*	*Significancia de Heurística 2 vs. otros métodos*	*Significancia de Heurística 3 vs. otros métodos*
Breast Cancer	0,0039	0,0039	0,0039
Lung Cancer	0,0020	0,0020	0,0020
Mushroom	0,0034	0,0034	0,0034
Heart	0,0039	0,0039	0,0039
Dermatology	0,0265	0,0265	0,0265

La conclusión de este análisis es que en un tiempo más pequeño se puede obtener un reducto lo suficientemente bueno en relación con longitud y diferenciación entre clases, de modo que el nuevo método RSReduct disminuye el costo computacional de problemas de clasificación.

Para comprobar aún más la eficiencia del método, se calcularon la calidad y la precisión por clase de la clasificación para el conjunto de datos completo y el conjunto de datos reducido, para las bases de casos. Ver tabla 6.

Al realizar una simple inspección de la tabla 5, es posible observar que no hay diferencias en calidad y precisión de la clasificación entre la base de casos completas y el reducto que se obtuvo a través de la segunda función heurística de RSReduct, esto nos lleva a concluir, sobre la base de los conjuntos aproximados, que no existen

Tabla 6 Medidas de inferencia basadas en los conjuntos aproximados para algunas bases de casos y el reducto obtenido por RSReduct para las mismas

Bases de casos representativas	*Calidad de la clasificación para la base completa*	*Calidad de la clasificación para el reducto*	*Precisión de la clasificación por clase para la base completa*		*Precisión de la clasificación por clase para el reducto*	
Breast Cancer	1,0	1,0	Clase 0	1,0	Clase 0	1,0
			Clase 1	1,0	Clase 1	1,0
Lung Cancer	0,8519	0,8519	Clase 1	1,0	Clase 1	1,0
			Clase 2	0,6667	Clase 2	0,6667
			Clase 3	0,6364	Clase 3	0,6364
Mushroom	1,0	1,0	Clase 1	1,0	Clase 1	1,0
			Clase 2	1,0	Clase 2	1,0
Heart	1,0	1,0	Clase 0	1,0	Clase 0	1,0
			Clase 1	1,0	Clase 1	1,0
Dermatology	1,0	0,9944	Clase 1	1,0	Clase 1	1,0
			Clase 2	1,0	Clase 2	0,9815
			Clase 3	1,0	Clase 3	0,9728
			Clase 4	1,0	Clase 4	0,9667
			Clase 5	1,0	Clase 5	1,0

pérdidas de información, o al menos estas son lo suficientemente pequeñas, al emplear el nuevo método de selección de rasgos propuesto.

Conclusiones

El nuevo método de selección de rasgos a través del concepto de reducto que se propone: RSReduct se analizó en cuanto a su eficiencia con diferentes bases de casos, la mayoría de las cuales redujo en gran medida; también se realizaron comparaciones con otros métodos, obteniendo un 95% de significancia estadística que indica importantes diferencias entre RSReduct y los otros métodos, favorables para el primero en cuanto a que este es capaz de encontrar un reducto lo suficientemente bueno en tiempos aceptables. Para probar, además, que no existían pérdidas de información sobre el sistema en el reducto obtenido, se realizaron pruebas con las medidas de inferencia de calidad y precisión de los conjuntos aproximados sobre dicho reducto y sobre la base completa, para obtener los mismos valores.

Referencias

1. N. Zhong, J. Dong, S. Ohsuga "Using Rough sets with heuristics for feature selection" *Journal of Intelligent Information Systems*. Vol. 16. 2001. pp. 199-214.
2. J. Wroblewski. "Finding minimal reducts using genetic algorithms". En: Wang, P.P. (Ed). *Proceedings of the International Workshop on Rough Sets Soft Computing at Second Annual Joint Conference on Information Sciences*, North Carolina, USA. 1995. pp. 186-189.
3. J.S Deogun. "Feature selection and effective classifiers". *Journal of ASIS.* Vol 49. 1998. pp. 423-434.
4. S.K. Choubey. "A comparison of feature selection algorithms in the context of rough classifiers". *Proceedings*

of Fifth IEEE International Conference on Fuzzy Systems. Vol. 2. 1996. pp. 1122-1128.

5. R. Kohavi, B. Frasca. "Useful feature subsets and Rough set Reducts". *Proceedings of the Third International Workshop on Rough Sets and Soft Computing*. San José, California. 1994. pp. 310-317.
6. H. Liu, H. Motoda. "Feature Selection Boston, MA : Kluwer academic Publishers". En: http://citeseer.ist.psu.edu/321378.html 1998. Consultado el 7 de mayo de 2006.
7. R. Jensen, S. Qiang. "Finding rough sets reducts with Ant colony optimization". http://www.inf.ed.ac.uk/publications/online/0201.pdf. 2003. Consultado el 5 de noviembre de 2005.
8. Z. Pawlak. "Rough sets". *International Journal of Information & Computer Sciences*. Vol. 11. 1982. pp. 341-356.
9. J. Komorowski, Z. Pawlak, Z. "Rough Sets: A tutorial. In Pal, S.K. and Skowron, A. (Eds.) Rough Fuzzy Hybridization: A new trend in decision-making." *Spring*er. 1999. pp. 3-98.
10. I. Dunstsh, Ivo, G. Gunter. "Rough set data analysis". En: http://citeseer.nj.nec.com/dntsch00rough.html. 2000. Consultado el 20 de mayo de 2006.
11. Z. Pawlak, "Rough SetsTheoretical Aspects of Reasoning About Data. Kluwer Academic Publishing, Dordrecht". 1991. En: http://citeseer.ist.psu.edu/context/36378.html. Consultado el 7 de abril de 2006.
12. R. Bello, Z. Valdivia, M.M. García, L. Reynoso. *Aplicaciones de la inteligencia artificial. México.* 1ª. ed. Guadalajara. Universidad de Guadalajara. 2002. pp. 62-64
13. R. Bello, A. Nowe, A. Puris. "Two Step Ant Colony System to Solve the Feature Selection Problem". *Lectures Notes on Computer Sciences*. Springer-Verlag. 2006. pp. 588-596.
14. B.S. Ahn. "The integrated methodology of rough set theory and artifi cial neural networks for business failure predictions". *Expert Systems with Applications*. Vol. 18. 2000. pp. 65-74.
15. A. Ohrn. "Rosetta Technical Reference Manual. Department of Computer and Information Science" Norwegian University of Science and Technology. Noruega, 2002.
16. Y. S. Wong, C. J. Butz. "Methodologies for Knowledge Discovery and Data Mining". Zhong y Zhou (Eds.) *On Information-Theoretic Measures of attribute importance*. pp. 231-238. En: http://citeseer.ist.psu.edu/yao99informationtheoretic.html. Consultado el 15 de febrero de 2006.
17. P. Piñero, L. Arco, M.M. García, Y. Caballero."Two New Metrics for Feature Selection in Pattern Recognition", *Lectures Notes in Computer Science* (LNCS 2905) Springer Verlag, Berlin Heidelberg. 2003. pp. 488-497
18. T. Mitchell, M. Hill. "Machine Learning". 1997. http://www.cs.cmu.edu/~tom/mlbook.html. Consultado el 14 de enero de 2006.
19. C. L Blake, C. J. Merz. "UCI Repository of machine learning databases". University of California, Irvine, 1998 http://www.ics.uci.edu/~mlearn/. Consultado el 10 de mayo de 2005.
20. D. Álvarez. Feature selection for data analysis using Rough Sets Theory. Master Thesis of Computer Science Engineering. University of Camagüey. Cuba. 2005. pp. 38-66.