Un algoritmo para el cálculo de la relevancia entrópica multivariada y su uso en la selección de variables

Félix F. González Navarro Lluís A. Belanche Muñoz
Departamento de Lenguajes y Sistemas Informáticos
Universitat Politècnica de Catalunya
fgonzalez@lsi.upc.edu belanche@lsi.upc.edu

Resumen

La reducción de la dimensionalidad mediante la selección de variables es uno de los pasos fundamentales del preprocesado de datos, como fase previa al análisis de información y descubrimiento de conocimiento. De entre las diversos algoritmos de reducción de la dimensionalidad, muchos se basan en dos enfoques, el *wrapper* que utiliza un inductor (e.g. un clasificador) como medida de desempeño del proceso, y el *filter*, el cual analiza los datos sin recurrir a un inductor. La desventaja del primero es el alto coste computacional que implica evaluar un clasificador cientos o miles de veces. En este artículo se presenta un algoritmo *filter* muy eficiente para la reducción de la dimensionalidad mediante el cálculo de la relevancia entrópica multivariada. Se presentan resultados comparativos experimentales en trece problemas de *benchmarking*. El algoritmo propuesto es comparado con un algoritmo de búsqueda hacia adelante tipo *wrapper*. Los resultados indican que el algoritmo propuesto encuentra soluciones iguales o mejores con un coste inferior.

1. Introducción

La reducción de la dimensionalidad mediante la *selección de variables* es uno de los pasos fundamentales del preprocesado de información [10]. La selección de variables es un proceso en el cual se selecciona un subconjunto de variables optimizando lo mejor posible un criterio de evaluación dado. Esta reducción es esencial por varias razones: reduce la complejidad de la construcción de un clasificador, elimina conocimiento irrelevante o redundante y mejora la capacidad de generalización [4].

Es común plantear un problema de tipo combinatorio, en el que el espacio de búsqueda es analizado buscando soluciones subóptimas, que poseean al máximo las características de la totalidad del modelo estudiado. Siendo $\{X_1, \ldots, X_n\}$ el conjunto total de variables que representan la medición de un sistema o modelo, se selecciona un subconjunto $\{X_{i_1}, ...X_{i_m}\}$ con $m < n$ e $i_j \in \{1, \ldots, n\}$, en el que es posible encontrar más de una solución igualmente satisfactoria.

Los métodos de selección de variables son divididos en dos grandes categorías: 1) el enfoque *wrapper*, el cual depende de un inductor (e.g. un clasificador) para determinar la habilidad discriminante (con la desventaja del alto coste computacional) y 2) el enfoque *filter*, donde el proceso está basado en los datos de manera independiente de un inductor [5]. Bajo este punto de vista, el proceso de selección de variables debe contener dos elementos principales: 1) la función de evaluación, que permitirá juzgar si un subconjunto es mejor (más *relevante*) que otro y 2) la estrategia de búsqueda, que decide cómo se siguen explorando nuevas soluciones.

En el espectro computacional de técnicas y algoritmos existen diversas funciones de evaluación y estrategias de búsqueda:

1. Funciones de evaluación: descripción de conceptos mínimos [1], información mutua, conteo de inconsistencias [12], sepa-

rabilidad interclase [6], entre otras.

2. Estrategias de búsqueda: Método *Branch & bound* [8], búsqueda secuencial hacia delante y hacia atrás [10], búsqueda flotante [17], entre otras.

Este artículo centra su objetivo en la Teoría de la Información como marco conceptual para la determinación de la relevancia de un subconjunto de variables. En primer lugar, se presenta un método eficiente para el cálculo de la relevancia entrópica multivariada. En segundo lugar, se presenta un algoritmo de búsqueda simple que la maximiza. Se muestra a continuación una serie de experimentos en trece problemas de *benchmarking*. El algoritmo propuesto es comparado con un algoritmo de búsqueda hacia adelante tipo *wrapper*. Los resultados indican que el algoritmo propuesto encuentra soluciones iguales o mejores con un coste computacional inferior.

El resto del trabajo está organizado como sigue. En la segunda y tercera secciones se da un panorama de los conceptos de entropía, entropía condicional multivariada e información mutua multivariada y su uso como función de evaluación en la selección de variables. Se revisan asimismo definiciones computacionales para calcular la información mutua multivariada. En la cuarta sección se ofrece un método de cálculo de relevancia multivariada basada en los datos y se detalla un algoritmo para dicho cálculo, además de un algoritmo de búsqueda bidireccional de subconjuntos de variables, el cual usa al primero como medida de evaluación de subconjuntos. Finalmente, en la quinta sección se ofrecen resultados experimentales con diversos clasificadores y problemas para probar la bondad de los subconjuntos generados por la combinación de ambos algoritmos y un estudio comparativo con un algoritmo de busqueda hacia adelante tipo *wrapper*, con el propósito de contrastar el método propuesto.

2. Entropía e información mutua

Se revisan en esta sección conceptos fundamentales de la teoría de la información. Claude E. Shannon, en su trabajo pionero [18], sentó las bases de lo que hoy conocemos como la *teoría de la información*. En éste introdujo dos conceptos fundamentales sobre la información: *incertidumbre* y *entropía*. La entropía se puede ver como un promedio de la incertidumbre de una variable aleatoria. Si X es una variable aleatoria discreta con masa de probabilidad p, su entropía es definida por:

$$H(X) = -\sum_x p(x)\, log p(x) = -E_X[log\, p(X)] \tag{1}$$

siendo $E[]$ el operador esperanza. Si una variable (X) es conocida y otra (Y) no, la *entropía condicional* de Y respecto X es la entropía media con respecto a la distribución condicional:

$$H(Y|X) = -\sum_x \sum_y p(x,y) log\, p(y|x) \tag{2}$$

es decir, la esperanza de la distribución condicional de Y respecto X. A partir de estas dos definiciones se construye el concepto de *información mutua* (IM), que puede interpretarse como una medida de información que una variable aleatoria contiene acerca de otra:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= E_{X,Y}[log\, \tfrac{p(x,y)}{p(x)p(y)}] \end{aligned} \tag{3}$$

Nótese que $I(X;X) = H(X)$ dado que $H(X|X) = 0$ y que $I(X;Y) = I(Y;X)$. En la Figura 1 se da una impresión gráfica entre estos conceptos.

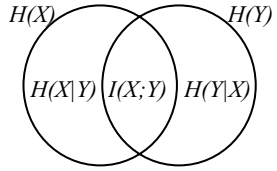

Figura 1. Relación de conceptos de entropía.

Se puede apreciar que al aumentar el área de $I(X;Y)$, la de $H(Y|X)$ decrece, es decir,

hay una reducción en el desconocimiento de la variable Y debido a la acción de la variable X.

El cálculo de la IM se puede extender al caso multivariado. Los primeros trabajos se al respecto se remontan a [13] y [7]. Se define la IM entre un número arbitrario n de variables y otra variable Y como una extensión del caso bivariado (3):

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, \ldots, X_{i-1})$$

donde se define la *IM condicional* de la manera natural, es decir, condicionando en (3):

$$I(X; Y|Z) = H(Y|Z) - H(Y|X, Z) \qquad (4)$$

El cálculo práctico de la IM multivariada plantea dificultades computacionales. Por ese motivo se han propuesto en la literatura diferentes alternativas que la capturen de manera aproximada.

Una alternativa existente es la *información de interacción* (descrita por ejemplo en [9]). Para tres variables X, Y, Z, la información de interacción $I(X; Y; Z)$ se define:

$$I(X; Y; Z) = I(X; Y|Z) - I(X; Y) \qquad (5)$$

La extensión a conjuntos de n variables $I(X_1, \ldots, X_n)$ es en términos de las entropías marginales y viene dada por

$$I(X_1, \ldots, X_n) = -\sum_{\tau \subseteq \{X_1, \ldots, X_n\}} (-1)^{n-|\tau|} H(\tau)$$

que tampoco es inmediata de calcular dado su carácter combinatorio.

3. Trabajo relacionado

El uso de la *Información Mutua* (IM) como base para la evaluación de suconjuntos de variables es amplio (por ejemplo, [19, 21, 16, 2]. La correcta identificación de variables que mayormente *caracterizan* datos observados es crítica para la minimización del error de clasificación. En [16] se plantea un esquema de *maximización de relevancia* del subconjunto S de variables con respecto a la variable de clasificación Y y *minimización de redundancia* de S, basado en la IM:

$$D(S, Y) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i; Y)$$

$$R(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \in S} I(X_i; X_j)$$

a partir de los cuales se construye la función $D(S, Y) - R(S)$ que debe maximizarse respecto a S. Similar enfoque se utiliza en [2], donde se plantea un algoritmo voraz para la selección de variables basado en IM. Se van seleccionando variables X_i iterativamente mediante la maximización de:

$$I(Y, X_i) - \beta \sum_{X_j \in S} I(X_i; X_j)$$

donde S son las variables ya seleccionadas y β es un factor que penaliza valores altos de redundancia.

En cuanto a aplicaciones prácticas, [19] reporta un algoritmo para la selección de variables en estudios de clasificacion de señales electrocardiográficas utilizando la IM. Se tratan los variables individualmente y se seleccionan los mejores variables como aquellas que simplemente presenten mayor IM con respecto a la variable de clasificación. También se ha utilizado la IM como función de evaluación en problemas de clasificación de datos genéticos (*microarray data*), caracterizados por una muy alta dimensionalidad [21]. Por otro lado, en [20] se propone un *Indice de Relevancia* (R) dado por la siguiente relación:

$$R(X_i; Y) = \frac{I(X_i; Y)}{H(Y)}, \; i = 1, \ldots, n \qquad (6)$$

donde $R(X_i; Y)$ toma valores entre cero (nula relevancia) y uno (máxima relevancia). Tomando Y como la variable de clase, se pueden asignar valores a los variables como relevancia numérica para tareas de clasificación.

4. Los Algoritmos ARM y ABBM

Se propone una forma heurística del cálculo de la IM multivariada a partir de las instancias presentes en los datos, considerando cada combinación de instancias como un posible valor de una variable nueva que representa la unión de todas las variables. Sea $X = \{X_1, ..., X_n\}$ el conjunto de todas los variables y V_i el conjunto de valores que puede tomar la variable X_i. Para un conjunto $\tau \subseteq X$, definimos el operador $\biguplus$ como la *concatenación para variables* y su correspondiente $\uplus$ para los valores de las variables. Entonces definimos a $\mathcal{V}_\tau$ como:

$$\mathcal{V}_\tau = \biguplus_{X_i \in \tau} X_i \tag{7}$$

Dado τ, se obtiene una única variable $\mathcal{V}_\tau$ cuyos posibles valores son todas las concatenaciones de cada posible valor de las variables de τ. Se pueden entonces calcular la entropía condicional entre $\mathcal{V}\tau$ y la variable de clase Y como:

$$H(Y|\mathcal{V}_\tau) = -\sum_{v \in \mathcal{V}_\tau} \sum_{y \in Y} p(v,y) log \frac{p(v,y)}{p(y)} \tag{8}$$

De esta forma, la IM es determinada como un caso bivariado simple:

$$I(\mathcal{V}_\tau; Y) = H(Y) - H(Y|\mathcal{V}_\tau) \tag{9}$$

Inspirándose en [20] se propone un *Indice de Relevancia* R de la variable $X_i \in X$ con respecto al subconjunto $\tau \subset X$ dado por la siguiente relación:

$$\begin{aligned} R(X_i; Y|\tau) &= \frac{I(X_i; Y|\mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)} \\ &= \frac{H(Y|\mathcal{V}_\tau) - H(Y|X_i; \mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)} \end{aligned} \tag{10}$$

Este método de cálculo de relevancia fue utilizado como función de evaluación en la selección de variables, en un algoritmo tipo *filter* y con estrategia de búsqueda bidireccional. A continuación se detallan los algoritmos para el cálculo de R, al que llamaremos *Algoritmo de Relevancia Multivariada* (ARM), y el de búsqueda *Algoritmo de Búsqueda Bidireccional Modificado* (ABBM).

4.1. Algoritmo ARM

Consideramos una matriz $D_{p\times(n+1)}$ de p datos descrita por n variables $X = \{X_1 \dots X_n\}$ (más la variable de clase Y, que supondremos ocupa la columna $n+1$). Previamente al inicio del proceso de selección, se ordena D por filas, usando como criterio el orden lexicográfico, lo cual permite acelerar los futuros cálculos de relevancia. Para la determinación de $R(X_i; Y|\tau)$, se contemplan las variables como una sola columna concatenada, y por tanto una nueva matriz $T_{p\times 2}$ formada por la nueva variable compuesta ($\mathcal{V}_\tau$) y la clase Y.

En los Algoritmos 1 y 2 se muestra el pseudocódigo del cálculo.

Algoritmo AH $(Y, \tau' \subset X)$
$V^- \leftarrow \uplus \{d_{1,j} \mid X_j \in \tau\}$
$Y^- \leftarrow d_{1,n+1}$
$cv \leftarrow 1$
$cy \leftarrow 1$
$H \leftarrow 0$
for $i \leftarrow 2$ **to** p **do**
  $V \leftarrow \uplus \{d_{i,j} \mid X_j \in \tau\}$
  $Y \leftarrow d_{i,n+1}$
  if $V^- = V$ **and** $Y^- = Y$ **then**
    $cv \leftarrow cv + 1$
    $cy \leftarrow cy + 1$
  else if $V^- \neq V$ **then**
    $H \leftarrow H + \frac{cy}{p} \log \frac{cy}{cv}$
    $cv \leftarrow 1$
    $cy \leftarrow 1$
  else
    // nº de apariciones de V en T
    $t \leftarrow |\{v \mid v = T_{i,1}, i = 1, \dots, p\}|$
    $H \leftarrow H + \frac{cy}{p} \log \frac{cy}{t}$
    $cv \leftarrow cv + 1$
    $cy \leftarrow 1$
  end
  $V^- \leftarrow V$
  $Y^- \leftarrow Y$
end
returns $-(H + \frac{cy}{p} \log \frac{cy}{cv})$

Algoritmo 1: AH para entropía multivariada.

Algoritmo ARM $(X_i \in X \setminus \tau, Y, \tau \subseteq X)$
returns $\frac{\mathbf{AH}_{(Y,\tau)} - \mathbf{AH}_{(Y,\tau\cup\{X_i\})}}{\mathbf{AH}_{(Y,\tau)}}$

Algoritmo 2: ARM para relevancia multivariada.

El primer **if** se encarga del caso en que haya dos filas iguales de la misma clase, el segundo dos filas distintas (sin importar la clase) y el tercero dos filas iguales de distinta clase (esto puede suceder al trabajar con un subconjunto de las variables). El algoritmo ARM puede implementarse haciendo un sólo recorrido a lo largo de los datos, detectando variaciones y calculando la entropía en las instancias de $\mathcal{V}_\tau$ e Y, obteniendo incrementalmente la entropía condicional multivariada. Como resultado de ello el cálculo de la relevancia no depende tanto del número de variables n como de la longitud del conjunto de datos p.

4.2. Algoritmo ABBM

Para la obtención de los mejores subconjuntos de variables (MSV, en adelante), se propone a continuación una *búsqueda bidireccional* de variables, que opera con dos conjuntos simultáneamente: Φ que contiene en todo momento el mejor subconjunto de variables de búsqueda hacia adelante; y β, que contiene el mejor subconjunto de variables de búsqueda hacia atrás.

Partiendo de $\Phi = \emptyset$ y $\beta = \{X_1, \ldots, X_n\}$, incrementalmente se van agregando (a Φ) y quitando (de β) variables, con la premisa de no agregar a Φ candidatos no presentes en β; y no quitar de β variables que hayan sido previamente seleccionadas por Φ. El algoritmo será detenido cuando $R(\Phi; Y) = 1$ o cuando ambos conjuntos coincidan. En cualquier caso, se ejecuta un proceso de mejora condicional sobre Φ, que acaba siendo la solución final. Este algoritmo –esquematizado en pseudocódigo en el Algoritmo 3- es una mejora sobre el puramente adelante usado en [3]. El algoritmo ABBM usa ARM y AH como subrutinas de la siguiente manera general:

$R(x; Y, Z)$ se resuelve con $\mathbf{ARM}\,(x, Y, Z)$

$R(Y, Z)$ se resuelve con $\mathbf{AH}\,(Y, Z)$

para $x \in X \setminus Z$ y $Z \subseteq X$, formando la combinación ARM-ABBM.

$\Phi \leftarrow \emptyset$ MSV hacia delante
$\beta \leftarrow \{X_1, \ldots, X_n\}$ MSV hacia atrás
repeat
  $x' \leftarrow \underset{x\notin\Phi,\ x\in\beta}{\operatorname{argmax}} \{R(x; Y, \Phi)\}$
  $\Phi \leftarrow \Phi \cup \{x'\}$
  $x'' \leftarrow \underset{x\notin\Phi,\ x\in\beta}{\operatorname{argmin}} \{R(x; Y, \beta \setminus \{x\})\}$
  $\beta \leftarrow \beta \setminus \{x''\}$
until $R(Y, \Phi) = 1$ **or** $\Phi = \beta$;
repeat
  $x' \leftarrow \underset{x\in\Phi}{\operatorname{argmin}} \{R(x; Y, \phi \setminus \{x\})\}$
  if $R(\Phi \setminus \{x'\}; Y) \geq R(\Phi; Y)$
  then $\Phi \leftarrow \Phi \setminus \{x'\}$
until $R(\Phi \setminus \{x'\}; Y) < R(\Phi; Y)$;

Algoritmo 3: ABBM de búsqueda bidireccional.

Para prevenir el caso de empates en los valores de $R(x; Y, \Phi), R(x; Y, \beta \setminus \{x\})$ ó $R(x; Y, \phi \setminus \{x\})$ para alguna x, se analizaron tres posibles criterios de desempate: Aleatorio (AL), donde se selecciona de forma aleatoria la variable x de entre las que han empatado; Relevancia Individual (RI), donde la variable a seleccionarse es aquella que presente mejor relevancia entrópica por sí misma (es decir, con mayor $R(x; Y, \emptyset)$) y Recursivo (REC), donde se ejecuta un subproceso de selección de variables con subespacio de búsqueda igual al conjunto de variables causantes del empate.

5. Experimentos y resultados

En esta sección se presentan los experimentos realizados para probar el funcionamiento de la combinación ARM-ABBM. Los experimentos constan de tres etapas. En la primera de ellas se aplicó ARM-ABBM a diversos conjuntos de datos, a fin de obtener los mejores subconjuntos de variables de acuerdo a este algoritmo; en la segunda etapa se hicieron pruebas con tres clasificadores utilizando los mejores subconjuntos de variables (MSV) resultantes y se compararon con los resultados sin reducción de

la dimensionalidad. En la tercera se usa un algoritmo *wrapper* con los mismos clasificadores con el propósito de comparar sus rendimientos con los de la parte segunda.

5.1. Selección del MSV

Se analizaron 13 conjuntos de datos, a los que se aplicó ARM-ABBM con las tres variaciones en el criterio de desempate de variables. Tres de ellos son generados artificialmente y los otros diez son del repositorio de datos de aprendizaje UCI [15]. Las características de estos conjuntos[1]. se muestran en la Tabla 1. Se distinguen tres tipos de problemas: generados artificialmente y discretos (AD), de procesos reales y discretos (RD) ó de procesos reales y continuos (RC). Los datos de tipo RC fueron discretizados utilizando el algoritmo de maximización de la interdependencia clase-variable (CAIM) [11], el cual posee dos características importantes: 1) está diseñado para trabajar en un ambiente supervisado, y 2) no requiere de un número de intervalos predefinido. Posteriormente, los conjuntos de datos fueron procesados con ARM-ABBM, obteniéndose diferentes configuraciones de MSV para los criterios de desempate. En la Tabla 1 se muestran los resultados para cada uno de ellos: el tamaño del mejor subconjuntos de variables y su valor de relevancia R.

En todos los casos se alcanza una notable reducción en el tamaño del MSV. En la mayoría de ellos es por lo menos un 50 % de los variables y con una relevancia R muy alta, sobresaliendo *Mammogram* con $R = 1$ (la máxima) y reducción del 90,76 %; similarmente *Sonar signals* y *Spectf* alcanzan una $R = 1$ y una reducción del 88,33 %. Los tipos de datos AD arrojan $R = 1$ y reducciones altas en el tamaño, excepto en un caso. Para los tipos RD, sólo en un caso se llega a $R = 1$ aunque casi siempre se llega a R superiores a 0.9 y las reducciones están alrededor del 50 %. La excepción es el problema Spect, que tan sólo obtiene $R = 0{,}64$. En lo referente a los criterios de desempate, AL logra los mejores resultados, con ocho soluciones de $R = 1$ y una reducción de más del 50 %; los criterios REC y RI logran esta condición en siete y seis conjuntos respectivamente.

Estos resultados muestran que la combinación ARM-ABBM es en general muy efectiva encontrando soluciones que maximicen R y minimicen el tamaño del MSV. Un asunto distinto es que estas soluciones ofrecidas por ARM-ABBM sean realmente útiles para inducir clasificadores.

5.2. Pruebas de clasificación

Para verificar la bondad de los MSV generados, se efectuaron experimentos con tres clasificadores mediante validación cruzada (*10-Fold Cross Validation*): *1-nearest neighbour* (1NN), *C4.5* y *Naïve Bayes* (NB), utilizándose el entorno de cómputo YALE [14]. En éstos se aplicaron los 39 subconjuntos de datos generados a partir de los 13 MSV obtenidos por cada uno de los tres criterio de desempate. En la Tabla 2 se muestra la exactitud en porcentaje de los clasificadores para los tres criterios, así como la exactitud obtenida por los mismos clasificadores *sin reducción de variables* (SR).

En la gran mayoría de los conjuntos de datos, se observa que la exactitud es numéricamente similar con o sin reducción de variables, y en varios de ellos es superior; tal es el caso de los conjunto de datos *Gmonks*, *Zoo Database*, *Breast Cancer*, *Mammogram* y *Sonar signals*. En particular destaca el problema *Spect*, que muestra buen desempeño en todos los clasificadores y en los tres criterios, despuntando *Naïve Bayes* con el criterio AL, obteniendo una exactitud de 96.67 %, comparado con un 82.37 % sin reducción, y *Mammogram*, con todos sus valores por encima de SR, especialmente el criterio AL en el clasificador 1NN, con 87.50 % de exactitud. Se aprecia también que la pérdida de poder clasificador es en todos los casos mínima. Es interesante observar que, comparando los clasificadores entre ellos, con *Naïve Bayes* se obtienen los mejores resultados promedio, seguido de *1NN* (Tabla 2, última fila). Pensamos que *C4.5* ofrece peores resultados ya que realiza su propia selección de

[1] En algunos de ellos se eliminaron las filas que contenían valores perdidos.

Conjunto	Tipo	p	n	AL $\|MSV\|$	AL R	AL t	RI $\|MSV\|$	RI R	RI t	REC $\|MSV\|$	REC R	REC t
Corral	AD	150	18	7	1	1.95	9	0.92	1.95	8	1	3.11
Gmonks	AD	100	21	5	1	2.15	5	1	2.25	5	1	3.42
Majority	AD	90	25	8	1	3.95	9	1	3.85	8	1	5.94
Hepatitis	RD	129	17	9	0.89	1.64	9	0.89	1.61	9	0.89	1.67
Horse Colic	RD	114	14	7	0.93	0.84	7	0.95	0.84	7	0.94	1.11
Lymphography	RD	148	18	6	1	2.15	9	0.98	2.14	9	0.98	3.07
Spect	RD	187	22	11	0.64	1.17	11	0.64	1.16	11	0.64	1.23
Zoo Database	RD	101	16	8	0.98	0.97	8	0.98	0.98	8	0.98	1.03
Breast Cancer	RC	569	30	7	1	32.4	7	1	32.38	7	1	52.45
Ionosphere	RC	351	33	17	0.98	25.64	17	0.97	25.45	17	0.98	35.24
Mammogram	RC	86	65	6	1	5.30	6	1	5.31	6	1	9.60
Sonar signals	RC	208	60	7	1	10.35	7	1	10.36	7	1	18.77
Spectf	RC	269	44	9	1	6.9	9	1	6.9	9	1	11.29

Tabla 1. Descripción de los conjuntos de datos y resultados de ARM-ABBM. La columna *Tipo* indica si los datos son artificiales y discretos (AD), reales y discretos (RD), ó reales y continuos (RC); p el número de instancias y n el número de variables; AL el criterio aleatorio de desempate de subconjuntos, RI el de relevancia individual, y REC el criterio recursivo; $|MSV|$ el tamaño del mejor subconjunto de variables y R su relevancia entrópica.

Conjunto	1NN SR	1NN AL	1NN RI	1NN REC	C4.5 SR	C4.5 AL	C4.5 RI	C4.5 REC	Naïve Bayes SR	Naïve Bayes AL	Naïve Bayes RI	Naïve Bayes REC
Corral	92.67	98.00	92.67	52.00	95.33	97.33	92.67	61.33	95.33	96.00	92.67	50.67
Gmonks	81.00	82.00	83.00	83.00	79.00	83.00	83.00	83.00	72.00	84.00	85.00	85.00
Majority	76.67	80.00	72.22	72.22	80.00	81.11	76.67	82.22	86.67	82.22	74.44	86.67
Hepatitis	85.96	82.88	83.72	82.88	85.26	79.87	86.79	79.87	84.49	83.65	89.04	83.65
Horse Colic	70.23	70.30	66.74	66.74	70.91	69.09	75.98	75.98	75.30	72.80	75.68	75.68
Lymphography	81.05	82.38	74.95	79.67	77.62	77.62	81.10	74.95	85.14	81.76	81.19	82.48
Spect	88.83	83.89	92.05	92.05	92.05	77.08	92.05	92.05	82.37	96.67	88.80	88.80
Zoo Database	93.18	83.27	89.27	77.27	92.09	93.18	89.18	78.36	94.09	96.00	95.09	73.27
Breast Cancer	96.31	95.78	95.08	95.78	94.20	94.20	94.55	94.20	96.49	93.85	93.50	93.85
Ionosphere	88.86	87.73	89.16	88.02	92.04	91.18	90.33	91.75	92.58	91.44	92.29	92.59
Mammogram	75.69	87.50	83.89	86.39	71.39	81.67	77.08	79.31	87.50	93.33	96.67	92.22
Sonar signals	86.05	81.76	81.26	81.76	78.43	79.74	86.07	79.74	91.31	81.29	82.24	81.29
Spectf	85.53	86.25	82.55	84.77	77.34	82.56	83.28	82.52	85.50	84.76	80.68	83.66
MEDIA	87.70	87.18	86.39	84.07	86.19	85.98	87.78	84.53	89.21	88.89	87.99	85.04

Tabla 2. Exactitud de los clasificadores (en %) por criterio de desempate. SR: sin reducción de variables. Última fila: media de la exactitud, ponderada por los respectivos números de filas p.

variables, que se suma a la ya efectuada, y esto provoca demasiada pérdida de información.

5.3. Pruebas de selección de variables y clasificación con un algoritmo tipo *wrapper*

Con el objetivo de comparar los rendimientos obtenidos mediante ARM-ABBM, se muestran a continuación pruebas de selección de variables mediante un algoritmo *wrapper* con *1-nearest neighbour* (1NN) y *Naïve Bayes* (NB) en los mismos conjuntos de datos (esta vez *C4.5* se dejó de lado por el comportamiento observado con anterioridad). Se muestran resultados con búsqueda hacia delante (*Forward Selection*). Este algoritmo es el más sencillo dentro de los que realizan una búsqueda por el espacio de posibles combinaciones y tiene el menor coste posible. Cuando se usa en modo *wrapper* y *10-Fold Cross Validation*, su coste es $\theta(n^2 J_{p,n})$, donde $J_{p,n}$ es el coste de una llamada al clasificador. El coste de ARM-ABBM es $\theta(p \cdot \max(n^2, \log p))$. Por tanto, siendo $J_{p,n} \geq pn$ necesariamente, el coste de ARM-ABBM es inferior.

Los resultados, que se muestran en la Tabla 3, indican tamaños de subconjuntos en gener-

al inferiores a los logrados por ARM-ABBM. Sin embargo, éste logra exactitudes de clasificación superiores en varios casos. Es muy reseñable el rendimiento en función del tipo de problema. Para los problemas de tipo AD y RC, y comparando con los resultados usando el criterio de desempate AL –aparentemente el mejor según la Tabla 2–, la exactitud obtenida con los clasificadores usando las soluciones ofrecidas por ARM-ABBM es superior en 13 de los 16 casos, un resultado muy notable. Sin embargo, para los problemas de tipo RD, la situación se invierte y es superior sólo en 2 de los 10 casos. Se conjetura que esta situación puede ser debida a que el criterio de discretización usado influye en el resultado, siendo que una mejor discretización favorece una mejor selección entrópica de variables.

Conjunto	1NN	NB
Corral	97.33 (4)	93.33 (4)
Gmonks	80.00 (5)	85.00 (8)
Majority	85.56 (5)	81.11 (9)
Hepatitis	86.09 (4)	87.63 (4)
Horse Colic	78.26 (4)	81.59 (6)
Lymphography	83.76 (4)	88.25 (5)
Spect	92.05 (14)	88.25 (8)
Zoo Database	85.00 (6)	94.00 (6)
Breast Cancer	94.74 (4)	94.73 (4)
Ionosphere	86.89 (4)	90.30 (5)
Mammogram	79.31 (8)	81.81 (4)
Sonar signals	76.46 (4)	77.88 (7)
Spectf	82.15 (6)	81.81 (5)

Tabla 3. Pruebas de clasificacion con *forward selection* en modo *wrapper* y utilizando 1NN y NB como clasificadores. Entre paréntesis el MSV encontrado.

6. Conclusiones y trabajo futuro

En este trabajo se ha propuesto una forma de cálculo de relevancia de variables, basado en un cálculo heurístico de la información mutua multivariada, mediante el algoritmo ARM. Se ha propuesto un algoritmo de búsqueda bidireccional para selección de variables (ABBM), con tres criterios de desempate, en el que se utilizó como función de evaluación la relevancia calculada por ARM, obteniéndose así el algoritmo ARM-ABBM. Este algoritmo se aplicó a 13 conjuntos de datos de clasificación, tres artificiales y diez reales. Se han contrastado además los resultados del ARM-ABBM (un algoritmo tipo *filter*) con uno tipo *wrapper*, computacionalmente más costoso. De los experimentos se puede concluir lo siguiente:

- La reducción de variables fue significativamente alta, rondando el 50 % en todos los casos y un 95 % en el mejor de los casos.
- Las soluciones obtenidas mediante ARM-ABBM arrojan relevancias razonablemente similares y, en su gran mayoría, cercanas a la máxima, sin importar el origen de los datos.
- Los experimentos con tres clasificadores usando las soluciones obtenidas mediante ARM-ABBM ofrecen exactitudes en media muy similares a las obtenidas por los mismos clasificadores sin reducción de variables.
- Pruebas experimentales adicionales contra un algoritmo adelante tipo *wrapper* indican que el ARM-ABBM es una opción aceptable para la selección de variables, a un coste computacional inferior.

Como trabajo futuro se desarrollarán los siguientes aspectos:

- Ampliar el espectro de problemas a aquellos con mucha mayor dimensionalidad (del orden de cientos de variables).
- Estudiar la relación entre el tipo de discretización y el rendimiento final.
- Estudiar la naturaleza del error cometido en la aproximación de la información mutua multivariada y su influencia en el rendimiento final.

Agradecimientos

Los autores desean agradecer al CICyT No. de proyecto CGL2004-04702-C02-02, CONACyT y a la UABC por el apoyo a esta investigación.

Referencias

[1] Almuallim, H., and Dietterich, T. Learning with many irrelevant features. In *Procs. of AAAI-91)*, vol. 2, pp. 547–552.

[2] Battiti, R. Using mutual information for selecting features in supervises neural net learning. *IEEE Trans. on Neural Networks 5*, 4 (1994), 537–590.

[3] Bell, D. A., and Wang, H. A formalism for relevance and its application in feature subset selection. *Machine Learning 41*, 2 (2000), 175–195.

[4] Cios, K. J., Pedrycz, W., and Swiniarski, R. W. *Data Mining Methods for Knowledge Discovery.* Kluwer, 1998.

[5] Dash, M., and Liu, H. Feature selection for classification. *Intelligent Data Analysis 1*, 3 (1997), 131–156.

[6] Duda, R., and Hart, P. *Pattern Recognition and Scene Analysis.* Wiley, 2001.

[7] Fano, R. M. *Transmission Of Information: A Statistical Theory of Communication.* MIT Press, 1961.

[8] Fukunaga, K. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[9] Jakulin, A., and Bratko, I. Quantifying and visualizing attribute interactions: An aproach based on entropy. In *http://arxiv.org/abs/cs.AI/0308002v3* (2004).

[10] Kittler, J. Feature selection and extraction, *Handbook of Pattern Recognition and Image Processing.* pp. 59–83.

[11] Kurgan, L. A., and Cios, K. J. Caim discretization algorithm. *IEEE Trans. Knowl. Data Eng 16*, 2 (2004), 145–153.

[12] Liu, H., and Setiono, R. A probabilistic approach to feature selection: a filter solution. In *Int. Conf. on Machine Learning* (1996), pp. 319–327.

[13] McGill, W. Multivariate information transmission. *IEEE Trans. Information Theory 4*, 4 (1954), 93–111.

[14] Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., and Euler, T. Yale: Rapid prototyping for complex data mining tasks. In *Proc. of the 12th ACM SIGKDD 2006.*

[15] Newman, D.J., Hettich, S., Blake, C.L, Merz, C.J. (1998). UCI Repository of machine learning databases (directory `mlearn/MLRepository.html` at `www.ics.uci.edu`). Irvine, CA.

[16] Peng, H., Long, F., and Ding, C. H. Q. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 8 (2005), 1226–1238.

[17] P. Pudil, J. Novovicová, J. Kittler. Floating search methods in feature selection *Pattern Recognition Letters*, 15(11), 1994.

[18] Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal. 27* (1948), 379–423.

[19] Sheikholeslami, N., and Stashuk, D. Supervised mutual-information based feature selection for motor unit action potential classification. *Medical and Biological Eng. and Comp. 35*, 6 (1997), 661–670.

[20] Wang, H. *Towards a unified framework of relevance.* PhD thesis, U. of Ulster, 1996.

[21] Zhou, W., Zhou, C., Liu, G., and Zhu, H. Feature selection for microarray data analysis using mutual information and rough set theory. In *Artificial Intelligence Applications and Innovations* , IFIP International Federation for Information Processing vol. 204, pp. 492–499, 2006.