# SELECCIÓN DE CARACTERÍSTICAS RELEVANTES USANDO INFORMACIÓN MUTUA

# SELECTION OF EXCELLENT CHARACTERISTICS USING MUTUAL INFORMATION

CESAR AUGUSTO CARDONA M.
*Escuela de Sistemas, Facultad de Minas, Universidad Nacional de Colombia*

JUAN DAVID VELÁSQUEZ H.
*Escuela de Sistemas, Facultad de Minas, Universidad Nacional de Colombia, jdvelasq@unal.edu.co*

Recibido para revisión 24 de Noviembre de 2004, aceptado 29 de Agosto de 2005, versión final recibida 7 de Abril de 2004

**RESUMEN**: Un nuevo método para la selección de atributos relevantes basado en Información Mutua es presentado. Este se basa en el concepto de probabilidad de relevancia de cada atributo, el cual es medido a través de una prueba de permutación, y permite descartar variables irrelevantes así como ordenar por importancia aquellas relevantes. La metodología propuesta es probada usando tres problemas de clasificación bien conocidos. Igualmente, se realiza una investigación con miras a esclarecer su robustez cuando las variables relevantes están contaminadas con ruido, o existen variables aleatorias artificiales irrelevantes. Los resultados indican las bondades de la metodología propuesta, por lo que se sugiere que ella debe ser una parte integral de las herramientas usadas en la selección de características relevantes.

**PALABRAS CLAVE**: Información Mutua, Clasificadores, Selección de Atributos.

**ABSTRACT**: A new method for relevant attribute selection based on Mutual Information is presented. It is based on the concept of relevance probability of each attribute, which is measured using a permutation test, and it allows to discard irrelevant variables thus as to order by importance the relevant variables. The proposed methodology is tested using three well know classification problems. Also, a research is conducted to clarify its performance when the relevant variables are contained with noise, or there are irrelevant artificial random variables. The results show the success of the proposed methodology, by which it is recommended its use as integral part of the tools used in relevant attribute selection

**KEYWORDS**: Mutual Information, Classifiers, Attribute Selection.

## 1 INTRODUCCIÓN

El desarrollo de metodologías para la solución del problema general de clasificación ha tenido un progreso muy importante en las últimas décadas (Duda et al., 2001). Este se encuentra íntimamente ligado al problema de reconocimiento de patrones por lo que a veces es difícil diferenciarlos.

Mientras las metodologías usadas para la solución del primer problema son conocidas de forma genérica como clasificadores, las usadas para la solución del segundo son llamadas técnicas de agrupamiento o clustering. Dada la importancia de ambos problemas, muchos esfuerzos han sido invertidos en el desarrollo de metodologías para su solución [véase Duda *et al*. (2001) y Jain et al. (2000)].

El problema general de clasificación está comprendido por dos subproblemas (Kasavob, 1997) que corresponden: primero, a determinar la clase a la que pertenece un

objeto a partir de un conjunto de ejemplos de objetos cuya clasificación es conocida; en el segundo, se tiene un conjunto de objetos sin ninguna agrupación y se desea determinar tanto el número de clases existente como a la que pertenece cada objeto.

Entre los métodos más comúnmente usados en la construcción de clasificadores se encuentran los métodos estadísticos, sistemas expertos, árboles de decisión, sistemas difusos, redes neuronales artificiales o los sistemas híbridos, entre otros.

Una etapa previa a la construcción del clasificador, consiste en determinar que características o atributos de los objetos permitirán diferenciar las clases a las que pertenecen. Ello puede ser determinado directamente a través de medidas de relación entre cada atributo y la clase a la que el objeto pertenece; o indirectamente, a través de una medida del desempeño del clasificador usando el atributo evaluado. No obstante, la selección de atributos basada en el desempeño del clasificador depende directamente de la metodología usada para su especificación y estimación de parámetros, de tal forma que clasificadores diferentes puedan llevar a seleccionar distintos grupos de atributos, e inclusive para un mismo clasificador construido con diferentes metodologías.

De esta forma, los procesos de selección de atributos relevantes basados en la medición directa de sus relaciones con las clases a las que pertenecen los objetos, tienen claras ventajas, ya que se independiza este proceso de la construcción del clasificador, haciendose más fácil la preparación de este último.

Diversas medidas de la relación existente entre cada atributo y la clase pueden ser usadas. Una de las más comunes es la correlación serial, pero es inadecuada ante la presencia de relaciones no lineales; una medida más general es la Información Mutua (IM), la cual permite detectar tanto relaciones lineales como no lineales. Sin embargo, uno de sus principales inconvenientes está relacionado con la decisión de cuándo dicha medida tiene un valor significativamente diferente de cero, debido a que la presencia de relaciones espurias entre los valores de los atributos, y las aproximaciones realizadas en su proceso de cálculo, pueden producir valores diferentes de cero para atributos irrelevantes. Otro problema está relacionado con que no existe garantía de que el clasificador pueda aprender las relaciones entre los atributos finalmente seleccionados y la clase a la que pertenece cada objeto. En consecuencia, este trabajo se centra en explorar el uso de las pruebas estadísticas de permutación como una herramienta para determinar si la IM es significativamente diferente de cero, y en cómo un algoritmo secuencial de selección de atributos dentro del proceso de construcción del clasificador permite realizar el afinamiento final del conjunto de atributos relevantes realizados.

Para ello, el resto de este artículo está organizado de la siguiente forma: En la próxima sección, se define formalmente el problema de selección de características relevantes; posteriormente, se exponen los principios básicos de la IM y los problemas que debe resolver; se prosigue con una discusión del problema de la estimación de la función de densidad de probabilidad (FDP) y su problemática; después, se propone formalmente una metodología para la selección de características relevantes basada en la estimación de la IM y un test de permutación; posteriormente, se procede a la validación de la metodología usando tres problemas de clasificación ampliamente estudiados en la literatura; y finalmente en el sexto se discuten las conclusiones pertinentes.

## 2 EL PROBLEMA DE SELECCIÓN DE CARACTERÍSTICAS RELEVANTES

Un sistema clasificador se define formalmente como un algoritmo que ubica nuevos objetos en un grupo o clase perteneciente a un conjunto finito previamente definido, con base en la

observación de particularidades de los objetos llamados atributos o características (Koller y Sahami, 1996). Las categorías a las que pertenecen los objetos pueden estar definidas de antemano, de tal forma que existen etiquetas lingüísticas asociadas a cada clase representando conceptos; en este caso, el aprendizaje realizado por el clasificador se considera como supervisado. No obstante, las clases existentes pueden ser desconocidas, por lo que el clasificador deberá determinarlas, realizándose en este caso un aprendizaje no supervisado.

La selección de características es un paso básico en el proceso de construcción de un sistema clasificador. Stoppiglia et al. (2003) indican que los objetivos de este proceso son: satisfacer la meta general de maximizar el desempeño del clasificador mientras se minimiza los costos de medida asociados; mejorar la exactitud del clasificador reduciendo características irrelevantes y redundante; reducir la complejidad y los costos computacionales asociados; reducir la cantidad de datos necesarios para el entrenamiento, hacer que el modelo obtenido sea entendible y práctico.

La definición previa asume que se conocen de antemano, los atributos que permiten determinar a cual clase pertenece un determinado objeto. No obstante, dichos atributos no son conocidos en todos los casos, y deben ser determinados a partir del universo de características que poseen los objetos.

La selección de características relevantes independiente del clasificador, busca determinar el subconjunto óptimo de atributos, $x_{opt}$, obtenido del universo de atributos $X$, que maximiza la medida de relación $G$, con la clase $C$, tal que:

$$x_{opt} = \arg\max_{x \subset X} G(x, C) \qquad (1)$$

Cuando la selección es dependiente del clasificador $f$, sus parámetros $W$ son optimizados para cada posible subconjunto de atributos $x$, de tal forma, que se minimice la medida de error $M$, entre los resultados del clasificador y las clases $C$, tal que:

$$x_{opt} = \arg\min_{x \subset X} \left\{ \arg\min_{W} M\left[ f(W, x), C \right] \right\} \qquad (2)$$

En este caso, los procesos de selección del modelo, selección de variables y estimación de parámetros se hacen de forma conjunta, por lo que puede dificultar enormemente el proceso de construcción del modelo.

El uso de (1) y (2) implican la necesidad de hacer un recorrido de todos los posibles subconjuntos de $X$. Jain *et al*. (2000) indica que los principales métodos son: búsqueda exhaustiva, algoritmos de ramificar y acotar, mejores variables individuales, selección secuencial hacia adelante, eliminación secuencial hacia atrás, step-wise, Búsqueda Tabú y Algoritmos Genéticos

## 3 INFORMACIÓN MUTUA

Una forma de definir $G$ en (2), es a través de la Información Mutua (IM); Su formulación original se basa en el trabajo de Shannon (1949). Para definir la IM, se hace necesario definir primero la entropía $H(X)$:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \qquad (3)$$

Donde $p(x)$ representa la función de densidad de probabilidad marginal de la variable aleatoria $X$ y el logaritmo puede ser en base 2, 10 ó logaritmo natural, produciendo correspondientemente unidades de bits, Hartleys o nats.

La entropía $H(X)$, puede entenderse como una medida de sorpresa o incertidumbre; mientras más grande su valor, más incertidumbre se tiene acerca del valor que tomará en algún momento.

Análogamente a (3), la entropía conjunta de dos variables $X$ y $Y$, cuyas probabilidades

están definidas sobre el mismo espacio de probabilidad, se expresan como:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) \quad (4)$$

La entropía condicional representa la incertidumbre que se tiene sobre Y cuando se conoce el valor de *X*:

$$\begin{aligned} H(Y/X) &= \sum_{x\in X} p(x)H(Y/X=x) \\ &= -\sum_{x\in X} p(x)\sum_{y\in Y} p(y/x)\log p(y/x) \end{aligned} \quad (5)$$

Si se sabe que:

$$I(X;Y) = H(X) + H(Y) - H(X/Y) \quad (6)$$

Entonces la IM es:

$$I(X;Y) = \sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (7)$$

A partir de la definición para el caso discreto, es fácil entender que la definición de IM para el caso continuo es:

$$I(X;Y) = \int_x\int_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \quad (8)$$

La IM mide indirectamente la relación entre el conjunto de variables supuestamente explicativas *X* y la variable de salida *Y*, a través de la medida de la distancia entre la distribución conjunta actual de los datos $p(x_i,y)$ y la distribución que ellos tendrían si $x_i$ y *y* fueran independientes, esto es $p(x_i,y) = p(x_i)p(y)$ en el caso de independencia lineal. Dicha medida es conocida como la distancia de Kullback-Leiber.

El cálculo o estimación de la IM debe salvar un serio obstáculo como es la estimación de las funciones de densidad $p(x_i)$, $p(y)$ y $p(x_i,y)$ las cuales son desconocidas a priori y deben ser estimadas de los datos mismos. Para superar este problema, los valores deben discretizarse o aproximar sus densidades con métodos paramétricos o no paramétricos. En el caso de variables discretas, este problema tiene fácil solución debido a que se trabaja con sumatorias, donde las probabilidades son estimadas desde el conteo de frecuencia en los datos.

Algunas propuestas que usan IM de alguna manera para la selección de características relevantes pueden ser encontradas en Bonnlander (1996), Battiti (1994), Kwak y Choi (2002) y Hall (1999).

## 4 ESTIMACIÓN DE LA FUNCIÓN DE DENSIDAD DE PROBABILIDAD (FDP)

La FDP es un modelo matemático que describe el comportamiento probabilístico de una variable aleatoria. Las funciones utilizadas como estimadores (Wilfahrt, 2002), deben cumplir las propiedades de verdadera densidad, consistencia y adicionalmente deben ser insesgados.

Existen dos tipos de acercamientos para llevar a cabo la estimación de la FDP, ellos son los métodos paramétricos y los no paramétricos. Los primeros consideran que la FDP que se desea estimar pertenece a una determinada clase de funciones paramétricas como la distribución normal, exponencial, poisson, etc. Bajo esta suposición, la estimación se reduce a determinar el valor de los parámetros del modelo a partir de la muestra. Este método solo es útil cuando la distribución subyacente es conocida de antemano e impone una estructura a la función que puede conducir a inferencias y predicciones incorrectas.

Los métodos no paramétricos, a diferencia de los paramétricos, no exigen a priori la especificación de una forma de la FDP, en su lugar tratan de estimarla dejando que los datos “hablen” por sí mismos. Es por lo anterior, que son más consistentes y flexibles bajo una menor cantidad de suposiciones restrictivas, cuando son comparados con las funciones obtenidas mediante los métodos paramétricos (Bonnlander, 1996).

Se ha demostrado la imposibilidad de existencia de un estimador que cumpla la propiedad de verdadera densidad y que sea insesgado para todas las propiedades continuas, esto ha hecho que se centre la atención en secuencias de estimadores no paramétricos que sean asintóticamente insesgados, es decir, que la esperanza de la estimación tienda a la función verdadera cuando el tamaño de la muestra tiende a infinito.

Algunos de los métodos no parámetricos utilizados son el histograma, estimador Naive, estimador de núcleo, estimador del vecino más cercano, estimador de núcleo variable, estimador de máxima verosimilitud penalizada y estimador de series ortogonales. El estimador de núcleo (kernel) es el método de estimación no paramétrico más ampliamente usado y analizado. Su FDP para algún punto $x$, es estimada como:

$$\hat{f}(x) = \left(\frac{1}{nh}\right)\sum K\left(\left(x_i - x\right)/h\right) \qquad (9)$$

Donde $K$ es la función núcleo o kernel, $n$ es el número de observaciones, $h$ es el ancho de ventana escogido para la función núcleo, $x_i$ es el elemento número $i$ en el grupo de datos.

El estimador de núcleo puede interpretarse como una suma de protuberancias situadas en cada una de las observaciones. Esto puede verificarse en la Figura 1. La función núcleo $K$ determina la forma de las protuberancias, mientras el parámetro $h$ determina el ancho y nivel de suavizamiento de la función obtenida como estimativo. La función núcleo más comúnmente utilizada es la Epanechnikov, aunque se utilizan otras como Biweight, Triweight, Triangular y Gauss.

Cuando se utiliza el estimador de núcleo se obtienen estimaciones de FDP continuas y diferenciables, aunque se tiene un nuevo reto como es la elección del ancho de ventana $h$ que se utilizará, una tarea compleja que requiere de un juicioso análisis.

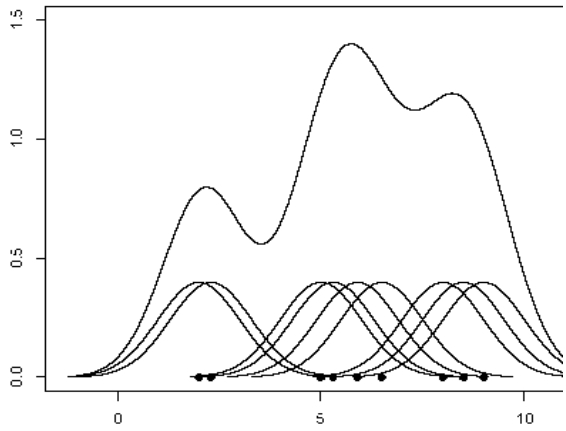



**Figura 1.** Función de densidad a partir del núcleo de Gauss. Tomado de Wand y Jones (1995).
**Figure 1.** Density Function using the Gauss's kernel. From Wand and Jones (1995).

Si cada punto de datos $x$ tiene $d$ elementos, la función de Epanechnikov se expresa como:

$$K(x) = \begin{cases} \prod_{i=1}^{d} \frac{3}{4}(1-x^2) & para\ |x|<1 \\ 0 & de\ otra\ forma \end{cases} \qquad (10)$$

En la búsqueda del mejor tamaño de ventana se trata de minimizar el error absoluto o minimizar el error cuadrático de la estimación. Con base en la minimización de alguno de estos errores, se han propuesto diversas metodologías que tratan de determinar el tamaño óptimo de ventana, algunas de ellas son validación cruzada de mínimos cuadrados, validación de máxima verosimilitud, distribución estándar, métodos plug-in y bootstrapping.
A pesar de que los dos últimos métodos presentan los mejores resultados, sus altos costos computacionales los hacen prohibitivos, por lo cual uno de los métodos más comúnmente usados es el que utiliza el criterio de verosimilitud. La idea básica en este método es escoger un $h$ que maximice el valor de verosimilitud:

$$\log L = \sum_{i=1}^{n} \log f\left(x_i\right) \quad (11)$$

Un estimativo para log *L* es:

$$\log L = \sum_{i=1}^{n} \log \hat{f}\left(x_i\right) = \log L(h) \quad (12)$$

Donde $\hat{f}\left(x_i\right)$ es un estimador de densidad de *f* y depende de *h*.

Cuando se maximiza con respecto a *h*, se produce un máximo trivial en *h* = 0. Para sobrellevar este problema, se adopta el principio de validación cruzada en el cual $\hat{f}\left(x_i\right)$ es reemplazado por $\hat{f}_{-i}(x)$

## 5 PROPUESTA METODOLÓGICA

La metodología de selección de características relevantes esta enfocada en los aspectos que a continuación se presentan.

### 5.1 Estimación de la Información Mutua

Para la estimación de la IM se propone utilizar la expresión:

$$\hat{I}(X;Y) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i, y_i)}{p(x_i)\,p(y_i)} \quad (13)$$

Dado que esta estimación es insesgada y empíricamente converge al valor verdadero, en función del tamaño de la muestra, como es demostrado por Bonnlander (1996).

A partir de la estimación de IM, clasificar o hacer un ranking de las variables de entrada, según el valor de esta métrica.

### 5.2 Estimación de la Función de Densidad de Probabilidad

Se propone estimar la FDP desde los datos de la muestra, utilizando funciones de núcleo continuo tipo Epanechnikov. En la determinación de este estimativo se tiene en cuenta cuando se trabaja con variables continuas (variables de entrada) y discretas (variables o clases de salida). Esto sería:

$$I(X;Y) = \sum_{y \in C_1}^{C_n} \int_x p(x, y/y = C_i) \log \frac{p(x, y/y = C_i)}{p(x)\,p(y/y = C_i)} dx \quad (14)$$

De igual forma, en el proceso de estimación de la FDP, se propone utilizar el método de histograma para encontrar soluciones aproximadas de las integrales que deben resolverse.

### 5.3 Tamaño de la Ventana

Se propone utilizar la función de máxima verosimilitud. Dado que el espacio de búsqueda del tamaño de ventana o *h* óptimo es muy grande y se ocasionan altos costos computacionales, se propone un algoritmo que permita encontrar valores de *h* óptimos o cercanos a él.

La idea básica del algoritmo propuesto, consiste en reducir inicialmente el espacio de búsqueda del *h* óptimo, dado que él debe estar entre la menor y mayor distancia entre dos puntos cualquiera de la variable que se desee analizar. El intervalo que tiene como extremos esos dos puntos se llama, para efectos explicativos, intervalo *Z* y constituye el espacio de búsqueda.

El intervalo *Z* se divide en una cantidad de subintervalos iguales, mientras más divisiones presente, más exhaustiva será la búsqueda. Luego se inicia un proceso recursivo en el que inicialmente se toma la primera partición de *Z* y se calcula el valor de verosimilitud tomando como valor de *h* los puntos extremos y medio de él. Se verifica en cual de los tres puntos se obtuvo el máximo valor de verosimilitud.
Si el máximo valor se obtuvo en el primer punto, se repite el mismo proceso tomando como punto inicial del intervalo este mismo valor y como punto final el punto medio antes obtenido. Ese punto medio se tiene en cuenta como una nueva división del intervalo *Z*.

Si el máximo valor se obtuvo en el punto medio, se repite el mismo proceso tomando como punto inicial del intervalo este mismo

valor y como punto extremo, el punto extremo antes obtenido.

Si el máximo valor se obtuvo en el último punto, se repite el proceso tomando como punto inicial del intervalo este mismo valor y como punto extremo, el valor de la próxima división del intervalo *Z*.

A lo largo de todo el proceso se guarda el valor de *h* que produjo la máxima verosimilitud.

Cabe anotar que el valor de *h* obtenido de esta forma se usa para determinar $p(x,y)$, $p(x)$ y $p(y)$.

### 5.4 Prueba de Permutación

Se propone aplicar una prueba estadística de permutación. Esta prueba consiste en calcular la IM entre cada una de las variables de entrada y la variable de salida en al menos 1000 oportunidades diferentes, permutando en cada cálculo el ordenamiento de la variable de salida. El objetivo es determinar una muestra de valores de la IM cuando no existe una relación entre *x* y *y*. Si realmente existe una relación de dependencia entre ambas variables, el valor de la IM debe ser superior a cuando no existe dicha relación; estos últimos estimados de la IM se obtienen al permutar el orden de la variable de salida tal como ya se indicó.

### 5.5 Prueba de Umbral

Se propone una segunda prueba estadística de umbral de selección, que permite determinar hasta que nivel, en el ranking de variables, se pueden considerar que ellas tienen alguna relevancia.
Esta prueba consiste en crear una nueva variable de entrada a partir de valores aleatorios y medir la IM que comparte ella con la variable de salida. Este valor se clasifica en el ranking de variables hecho inicialmente. Desde su posición en el ranking hacia abajo, se considera que las variables se comportan como variables aleatorias, ya que su valor de IM no es significativamente diferente de la variable ficticia aleatoria agregada a los patrones de ejemplo, por lo que estos atributos pueden ser descartados.

A partir de las características seleccionadas se propone llevar a cabo la clasificación mediante un perceptrón multicapa (PMC).

## 6 CASOS DE ESTUDIO

Se tomaron tres conjuntos de datos: Ionosphere (Sigillito et al., 1989), Iris Plant (Fisher, 1936) y Sonar (Gorman y Sejnowski, 1988) de la base de datos dedicada a la investigación sobre aprendizaje de máquina UCI Repostory (http://www.ics.uci.edu/mlearn/MLRepostory.html). Sobre estos conjuntos fue aplicada la metodología propuesta sobre selección de características relevantes. El resumen de las propiedades presentes en estos conjuntos de datos puede ser observado en la Tabla 1.

Como método alterno de selección de características relevantes y clasificación se utilizó MARS (Friedman, 1991), el cual es un método adaptativo de regresión multivariable basado en splines. Este puede entenderse como un modelo de la forma:

$$f(x) = a_0 + \sum_{K_m=1} f_i(x_i) + \sum_{K_m=2} f_{ij}(x_i, x_j) + \sum_{K_m=3} f_{ijk}(x_i, x_j, x_k) + \ldots \quad (15)$$

Donde la aproximación es construida como la suma de una constante, más la suma de las funciones base de una variable, más la suma de todas las funciones base de dos variables que representan todas las interacciones entre dos variables y así sucesivamente. Este método realiza un particionamiento recursivo del espacio, formado por las variables de entrada para construir un modelo final, como la suma de los modelos de una sola variable.

Los clasificadores construidos usando las variables seleccionadas, de acuerdo con el criterio de la IM, tuvieron un desempeño algunas veces inferior, cuando son comparados con los obtenidos por MARS, como puede ser verificado en la Tabla 2. No

obstante, estos resultados pueden obedecer a la dificultad intrínseca del entrenamiento y selección de modelos de RNA. En términos generales, MARS presenta una mayor eficiencia cuando se tiene en cuenta su elevada tasa de clasificación correcta, utilizando solo una mínima cantidad de variables

Una de las situaciones observada en los resultados, es la marcada influencia que puede ejercer la relación entre el tamaño del conjunto de entrenamiento y el número de atributos potencialmente relevantes; cuando existen muchas variables y pocos ejemplos, se hace mucho más difícil la selección de características. Esta puede ser la explicación a las bajas tasas de desempeño presentada en el conjunto Sonar, cuando es comparada con los otros dos conjuntos de datos, siendo esto cierto tanto para el método de IM como para MARS.

**Tabla 1.** Grupos de datos usados en la investigación y sus propiedades
**Table 1.** Datasets used in the research and its properties.

| Base de Datos | Número de clases | Número de variables | Conjunto de entrenamiento | Conjunto de prueba |
|---|---|---|---|---|
| Ionosphere | 2 | 34 | 200 | 151 |
| Iris Plant | 3 | 4 | 75 | 75 |
| Sonar | 2 | 60 | 104 | 104 |

Después de calcular la IM, se estimó la importancia relativa de cada una de las variables de entrada con respecto a aquella variable que aportaba un mayor valor de IM, obteniéndose los resultados presentados en la Figura 2. De igual forma, en la Figura 3 puede observarse los resultados arrojados por MARS. Los resultados de la prueba de permutación pueden examinarse en la Figura 4.

Una vez realizada la selección de variables relevantes usando IM, se entrenó un clasificador consistente en un perceptrón multicapa (PMC) con cinco neuronas ocultas, 100 épocas, tasa de aprendizaje igual a 0.5 y se usó Regularización Bayesiana. Luego del entrenamiento se procedió a realizar la clasificación del conjunto de prueba y se obtuvo los resultados de desempeño que son presentados en la Tabla 2.

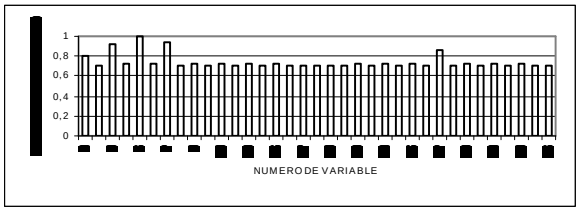
Al observar los resultados sobre la importancia relativa de las variables de entrada y de la prueba de permutación de la variable de salida, presentados en las Figuras 2 y 4 respectivamente, puede notarse en forma clara como algunas variables de entrada pueden ser descartadas por carecer de importancia dado su incipiente aporte de IM. Esto es bastante evidente en el caso de los conjuntos de datos Ionosphere y Sonar, donde el comportamiento de gran parte de sus variables de entrada, exhibe un desempeño similar y en algunas oportunidades muy inferior al presentado por la variable aleatoria introducida en cada uno de los conjuntos de variables (la cual corresponde a la variable 35 en el conjunto Ionosphere, 5 en Iris Plant y 61 en Sonar). Esto mismo no puede decirse del conjunto Iris Plant, ya que el comportamiento de sus variables de entrada es muy superior al de su correspondiente variable aleatoria.

Los resultados de selección de características relevantes mediante el uso de la IM, contrastan fuertemente con los obtenidos por MARS, como puede ser observado en la Figura 3. La reducción de variables irrelevantes es drástica, pues en el conjunto Ionosphere consideró que sólo eran relevantes dos variables (5 y 27), en Iris Plant sólo una variable (4) y en Sonar tres variables (49, 10 y 36); estos resultados están en concordancia con los resultados obtenidos con la IM, ya que estas mismas variables aparecen entre las de mayor importancia relativa cuando ella se usó, sin embargo, el conjunto de atributos relevantes finalmente seleccionado difiere en ambas metodologías.

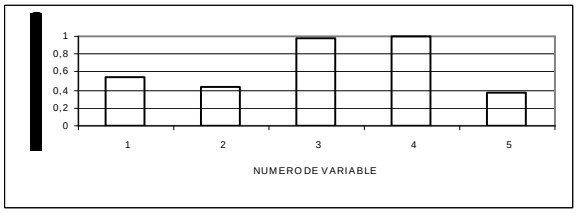
El proceso de clasificación, usando IM, se realizó con las variables que aparecían como las más relevantes en cada uno de los casos. Para el grupo Ionosphere las variables más relevantes son la 1, 3, 5, 7 y 27. En el grupo Iris Plant ninguna variable puede ser descartada, por lo cual todas son consideradas

relevantes; y en el grupo Sonar pueden considerarse relevantes cerca de veinte variables, pero solo se tuvieron en cuenta las primeras doce. Para tomar la decisión de cuales variables eran más relevantes, se tuvo en cuenta los resultados de la prueba de permutación en unión con los resultados obtenidos en la prueba de umbral.

(a) Ionosphere




(b) Iris Plant




(c) Sonar

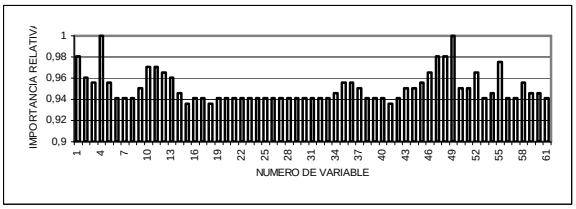



**Figura 2.** Importancia relativa de las variables de entrada respecto a las más importantes desde el punto de vista de la IM.
**Figure 2.** Relative importance for the input variables related to the most important using MI.

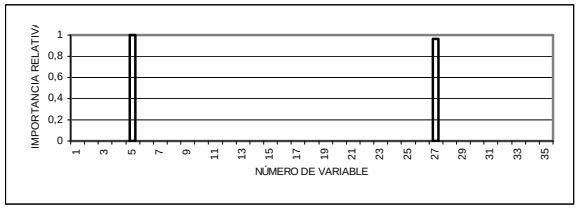
En la prueba de permutación, Figura 4, puede observarse que algunas variables tienen un desempeño óptimo (100% de eficiencia o cercano a él), mientras que para otras no sucedió así (0%), por lo cual las ultimas se consideran totalmente irrelevantes. Dado que se presentaban muchas variables con un óptimo desempeño, se tuvo en cuenta aquellas variables que presentaban los mayores valores de IM, es aquí en este punto, donde la prueba de umbral prestó una valiosa colaboración ayudando a determinar el límite entre variables relevantes y no relevantes.

Cuando se comparan los resultados de clasificación obtenidos por medio de la IM con los reportados por Kwak y Choi (2003), Tabla 2, usando el método Taguchi con IM, puede notarse como en términos generales, se presenta un mejor desempeño del método probado por estos últimos, sin embargo debe tenerse presente que el clasificador utilizado por ellos fue un perceptrón multicapa con 3 neuronas ocultas, 300 épocas y 0.5 de tasa de aprendizaje, lo cual hace que no se pueda tener una base de comparación equitativa.

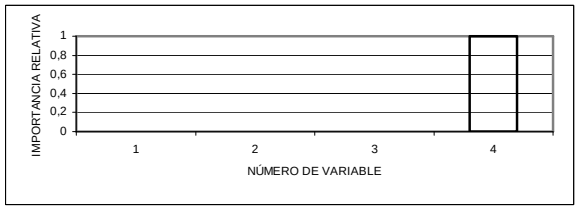
Los resultados obtenidos en este trabajo, así como el de Kwak y Choi (2003), muestran que aumentar el número de atributos relevantes para el clasificador, no necesariamente redundan en un mejor desempeño de este; por el contrario, puede llegarse a clasificadores con desempeños mucho más pobres. Lo anterior es congruente con Koller y Sahami (1996), cuando afirman que características redundantes e irrelevantes pueden causar problemas a los algoritmos de aprendizaje, dado que pueden eclipsar el pequeño subconjunto de variables relevantes que poseen la información más valiosa, haciendo más difícil el proceso de entrenamiento y selección del modelo.

Se llevó a cabo una prueba complementaria que permitiera medir la capacidad de la IM para encontrar relación entre las variables de entrada y la variable de salida de funciones no lineales con diferentes grados de complejidad, cuando ellas presentaban y carecían de señales de ruido. Con este fin se utilizaron las funciones Aditiva, Harmónica, Interacción Simple, Interacción Complicada y Radial, propuestas por Hwang et al. (1994).

(a) Ionosphere




(b) Iris Plant




(c) Sonar

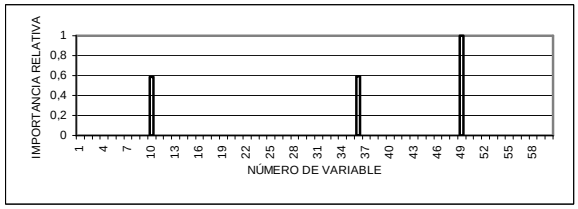



**Figura 3**. Importancia relativa de las variables de entrada respecto a la más importante cuando se usa MARS
**Figure 3.** Relative importance for the input variables selected by MARS

Para iniciar la prueba, cada una de las cinco funciones fue evaluada en los mismos 225 puntos aleatorios del plano en el intervalo [0 , 1]X[0 , 1] . Se adicionó a cada una de las variables un ruido aleatorio modelado como una distribución normal de varianzas iguales a la unidad, 0.5 y 0.05. Ello implica que si la varianza de las variables independientes es unitaria, se están adicionando ruidos equivalentes al 100%, al 50% y 5% de su varianza. De esta forma, se obtuvieron ocho variables para las que se calculó la IM con respecto al valor evaluado de su correspondiente función. Para finalizar, se realizó una prueba de permutación de la variable de salida de la misma forma que se hizo en la selección de características relevantes. Los resultados obtenidos cuando se aplicó esta prueba son presentados en la Tabla 3. Ellos permiten llegar a las siguientes conclusiones:

La capacidad de la IM para detectar la relevancia de variables es robusta ante la complejidad de las relaciones entre los atributos. Esto puede corroborarse, al observar que para todas las variables sin ruido, la prueba de permutación indica relevancias cercanas al 100%, independientemente de la complejidad de la función.

A medida que aumenta el ruido adicionado a las variables, la prueba de permutación indica un aumento de la probabilidad de encontrar combinaciones aleatorias entre las variables independientes y la variable dependiente que pueden tener valores de la IM superiores a los estimados para las variables originales contaminadas con ruido. No obstante, la IM es tolerante al ruido e indica relevancia con una probabilidad mayor al 50% para todos los casos

Ante variables contaminadas con ruido, la IM es más robusta en la medida de que la función generadora de la variable dependiente sea menos lineal. Es así como para la función Aditiva, ante niveles de ruido del 100% la probabilidad de relevancia es muy cercana a la unidad; mientras que para la función Radial, dicha probabilidad se reduce a un valor cercano al 50%.

## 7 CONCLUSIONES

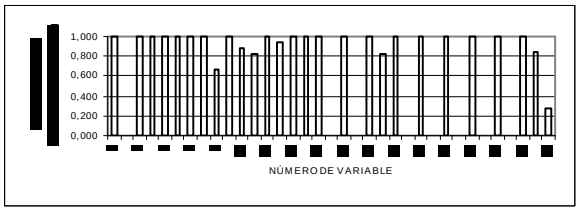
El problema de la selección de características relevantes depende del clasificador y de las características. Cuando se intenta realizar la selección de características conjuntamente con la clasificación, el problema aumenta en complejidad debido a que es más difícil comprobar mayor cantidad de variables y estimar los parámetros del clasificador al mismo tiempo. Es así como los métodos destinados a la eliminación de variables no relevantes y clasificación por contenido de información, permiten romper este problema en dos más simples, tal que se pueda

eliminar aquellas variables que no aportan conocimientos para la solución del problema y posteriormente, suponiéndolas como fijas, diseñar el clasificador como tal. Es de esta manera como la IM evita el problema en ese sentido, con los correspondientes beneficios interpretativos, no sólo de los datos, sino también computacionales

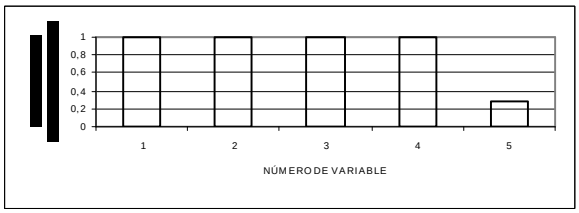
El uso de la IM permite establecer una medida de relevancia entre los atributos y las clases, de tal forma que se puede probar objetivamente la existencia de la relación, independizando la construcción del clasificador del problema de selección de características relevantes, más aún, el establecer un ordenamiento por orden de importancia, permite establecer cuales variables contribuyen de una manera más fuerte a la clasificación de los datos, con los consecuentes beneficios para el entendimiento del problema.

Pueden existir variables colineales con otras, tal que, cada una independientemente aporta información relevante para la clasificación, pero en su conjunto, no hay ganancia en el desempeño del clasificador ya que la información de una está contenida en la otra. Este hecho puede ser detectado al observar los valores de la IM ya que variables colineales deberán tener valores similares en este parámetro. Ante la sospecha de colinealidad entre dos o más atributos, puede realizarse un refinamiento de los atributos seleccionados, construyendo clasificadores en los cuales el primero contiene la variable más relevante, el segundo las dos más relevantes y así sucesivamente; y posteriormente, revisando el desempeño de los clasificadores construidos, buscando detectar aquellas variables que al ser incluidas no mejoran sustancialmente dicho desempeño. Lo anterior implica que ya no es necesario utilizar en la construcción del clasificador esquemas elaborados de recorrido del espacio de combinaciones de atributos relevantes, tales como: búsqueda exhaustiva, algoritmos de ramificar y acotar, etc.

(a) Ionosphere




(b) Iris Plant




(c) Sonar

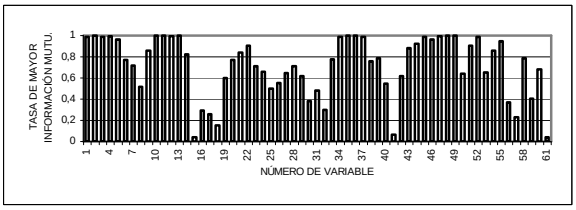



**Figura 4.** Tasa en que la IM de las variables de entrada es mayor que ese mismo valor en 1000 ensayos de permutación de la variable de salida
**Figure 4.** Frequency of higher values for the IM of the input variables, calculated in a simulation of 1000 tests of permutation

El uso de la prueba de permutación propuesto en este trabajo, como un método estadístico para determinar cuando la IM es significativamente diferente de cero, permite eliminar criterios subjetivos de clasificación, de tal forma que se provee al investigador de una herramienta objetiva de análisis. Es de anotar que en la revisión bibliográfica realizada, este problema no es resuelto, de tal forma que se hace un aporte importante a la aplicabilidad de esta metodología.

La inclusión de variables aleatorias es una herramienta que permite medir la incidencia de otros tipos de errores, presentes en el algoritmo de estimación de la IM y que están asociados a aspectos como los problemas relacionados con la estimación de la FDP.

**Tabla 2.** Tasas de clasificación correcta usando diferentes métodos de selección de características relevantes. (SIMU: IM Uniforme como selector de características, TSIM: Método de Taguchi con IM como selector de características, TSIMU: Método Taguchi con IM Uniforme como selector de características).
**Table 2.** Rate of corrected clasification using several methods for feature selection. (SIMU: Uniform using IM as method of selection, TSIM: Taguchi method using IM for feature selection, TSIUM: Taguchi method using uniform IM).

| CONJUNTO DE DATOS | VARIABLES INCLUIDAS EN PMC | CANTIDAD DE VARIABLES INCLUIDAS | PMC USANDO IM | MARS | * SIMU * TSIMU | |
|---|---|---|---|---|---|---|
| IONOSPHERE | 5 7 3 | 3 | 0,87 | | 0,91 | 0,93 |
| | 5 7 3 27 1 | 5 | 0,92 | | 0,91 | 0,93 |
| | 5 7 3 27 1 9 15 23 33 21 | 10 | 0,91 | | 0,92 | 0,92 |
| | 5 27 | 2 | | 0,93 | | |
| IRIS PLANT | 4 | 1 | 0,95 | | | |
| | 3 4 | 2 | 0,95 | | | |
| | 1 3 4 | 3 | 0,96 | | | |
| | 1 2 3 4 | 4 | 0,97 | | | |
| | 4 | 1 | | 0,95 | | |
| SONAR | 49 4 47 | 3 | 0,66 | | 0,65 | 0,65 |
| | 49 4 47 1 48 55 | 6 | 0,62 | | 0,77 | 0,71 |
| | 49 4 47 1 48 55 10 11 46 | 9 | 0,77 | | 0,78 | 0,79 |
| | 49 4 47 1 48 55 10 11 46 52 12 13 | 12 | 0,77 | | 0,89 | 0,82 |
| | 10 36 49 | 3 | | 0,75 | | |

* Resultados obtenidos por Kwak y Choi (2002).

**Tabla 3.** Tasa en que la IM calculada inicialmente es mayor que la IM obtenida en 1000 permutaciones de la variable de salida. (X = Abscisa, Y = Ordenada, R1 = Ruido con distribución N(0,1), R2 = Ruido con distribución N(0,0.5), R3 = Ruido con distribución N(0,0.05)).
**Table 3.** Frequency in that the IM calculated from the original dataset is higher than the IM estimated in 1000 permutations of the dependent variable. (X = Abscise, Y = Dependent variable, R1 = random noise with distribution N(0,1), R2 = random noise with distribution N(0,0.5), R3 = random noise with distribution N(0, 0.005)).

| FUNCIÓN | X | X+R1 | X+R2 | X+R3 | Y | Y+R1 | Y+R2 | Y+R3 |
|---|---|---|---|---|---|---|---|---|
| Aditiva | 1,000 | 0,900 | 0,995 | 0,999 | 0,979 | 0,829 | 0,831 | 0,949 |
| Harmónica | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 0,975 | 0,998 | 1,000 |
| Interacción-C | 1,000 | 0,834 | 0,995 | 1,000 | 1,000 | 0,324 | 0,915 | 0,999 |
| Interacción-S | 1,000 | 0,621 | 0,991 | 1,000 | 1,000 | 0,997 | 1,000 | 1,000 |
| Radial | 1,000 | 0,532 | 0,926 | 0,998 | 1,000 | 0,802 | 0,952 | 0,997 |

Dentro del problema de la estimación del parámetro *h*, de las funciones núcleo de la FDP, se ha hallado que en el proceso de minimización existen múltiples puntos de mínimo local, por lo que se hace necesario refinar el algoritmo de búsqueda. Esta problemática no es analizada dentro de la literatura consultada, por lo que no hay una advertencia clara para el investigador cuando desee estimar dicho parámetro.

Respecto al modelo MARS, con el cual fueron clasificados los conjuntos de datos y que sirvió como metodología alterna de validación de la que se propone en este trabajo, se encontró que usando esta última se lograba detectar más variables con relaciones no lineales que las obtenidas cuando se utilizó MARS. Esto conlleva a que la simplificación del problema basada en la eliminación de variables irrelevantes, unida a modelos no lineales como RNA, las cuales fueron entrenados con potentes algoritmos de aprendizaje (regularización bayesiana), puede llevar a clasificadores claramente superiores.

Finalmente, logró mostrarse como la IM es una medida capaz de detectar relaciones de dependencia entre variables, cuando esas relaciones son no lineales y complejas. De igual manera se puso en evidencia la robustez de la IM ante el ruido presente en dichas variables.

Al aplicar la metodología propuesta, quedan abiertos temas de estudio como:

- La validación de las dos pruebas estadísticas propuestas para la selección de características relevantes, la prueba de permutación y de umbral, a través de un mayor número de ensayos.

- La calidad de los valores encontrados para el parámetro utilizando el algoritmo propuesto, comparado con los valores obtenidos cuando se hace una búsqueda exhaustiva.

- La calidad de los resultados que se producen cuando para el cálculo de la IM se emplea los valores de $p(x,y)$, $p(x)$ y $p(y)$ obtenidos a partir de un mismo valor del parámetro *h*, comparada con el valor que se obtendría si esas densidades fueran calculadas a partir de valores de *h* óptimos para cada una de las variables.

## BIBLIOGRAFÍA

[1] BATTITI R. 1994. Using mutual information for selecting features in supervised neural net learning. En: IEEE

transactions Neural Networks. Vol. 5, No.4; p.537-550.

[2] BERLINET A. & DEVROYE L. 1994. A comparison of kernel density estimates. Institut de statistique de la Université de Paris, Vol.38, No.3, p. 3-59.

[3] BLUM A. & LANGLEY P. 1997. Selection of relevant features and examples in machine learning. En: Artificial intelligence. Vol. 97, No.1-2; p. 245-271.

[4] BONNLANDER B. 1996. Nonparametric selection input variables for connectionist learning. PhD thesis, University of Colorado.

[5] CARUANA R. & DE SA V. 2003. Benefitting from the variables that variable selection discards. En: Journal of Machine Learning Research, Vol. 3; p. 1245-1264.

[6] DEVROYE L. 1997. Universal smoothing factor selection in density estimation: Theory and Practice. En: Test. Vol. 6; p. 223-320.

[7] DINARDO J. & TOBIAS J. 2001. Nonparametric and Regression Estimation. En: Journal of Economic Perspectives. Vol. 15, No. 4; p. 11-28.

[8] DUDA R.O., HART P.E. & STORK D.G. 2001. Pattern Classification. New York: John Wiley & Sons. 2da edición.

[9] FADDA D., SLEZAK E. & BIJAOUI A. 1998. Density estimation with non-parametric methods. En: Astronomy and Astrophysics supplement series. Vol. 127, No.1; p. 335-352.

[10] FISHER R. 1936. The use of multiple measurements in taxonomic problems. En: Annual Eugenics. Vol. 7, No.2; p. 179-188.

[11] FRIEDMAN J.H. 1991. Multivariate adaptive regression splines. En: Annals of Statistics, Vol 19, No.1.

[12] GOLDBERG D.E. & HOLLAND J.H. 1988. Genetic Algorithms and Machine Learning. En: Machine Learning. Vol.3.

[13] GORMAN R. & SEJNOWSKI T. 1988. Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets. En: Neural Networks, Vol. 1; p. 75-89.

[14] GUYON I. & ELISSEEFF A. 2003. An introduction to variable and feature selection. En: Journal of Machine Learning Research. Vol 3; p. 1157-1182.

[15] HALL M. 1999. Correlation-Based Feature Selection for Machine Learning. PhD thesis, University of Waikato, New Zeland.

[16] HWANG J., LAY S., MAECHLER M., MARTIN D. & SCHIMERT J. 1994. Regression Modeling in Back-Propagation and Projection Pursuit Learning. Seminar für Statistik, ETH, Zurich, Switzerland.

[17] JAIN A. K., FELLOW, IEEE, DUIN R. & MAO J. 2000. Statistical Pattern Recognition: A Review. En: IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol. 22, No.1; p. 4-37.

[18] JOHN G., KOHAVI R. & PFLEGER K. 1994. Irrelevant Features and the Subset Selection Problem. En: Proceedings of the 11th International Conference on Machine Learning. San Francisco, California; p. 121-129.

[19] KASAVOB, N. 1997. Foundations of Neural Networks, Fuzzy Systems and Knowledge Engineering, Second Edn. Massachussets Intitute of Technology.

[20] KOHAVI R. & JOHN G. 1997. Wrappers for feature subset selection. En: Artificial Intelligence. Vol 97, No.(1-2); p.273-324.

[21] KOLLER D. & SAHAMI M. 1996. Toward optimal feature selection. En: 13th International Conference on Machine Learning. S.L; p. 284-292.

[22] KWAK N. & CHOI C. 2002. Input Feature Selection for Classification Problems. En: IEEE Transactions on Neural Networks, Vol 13, No.1; p. 143-159.

[23] MACKAY D.J. 1993. Bayesian non-linear modeling for the energy prediction competition.

[24] Salford Systems 2001. MARS USER GUIDE.

[25] SHANNON C.E. & WEAVER W. 1949. The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press.

[26] SIGILLITO V., WING S., HUTTON L. & BAKER K. 1989. Classification of radar returns from the ionosphere using neural networks. En: Johns Hopkins APL Technical Digest. Vol 10; p. 262-266.

[27] SILVERMAN B. W. 1986. Density estimation for statistics and data analisys. London: Chapman and Hall.

[28] STOPPIGLIA H., DREYFUS G., DEBOIS R. & OUSSAR Y. 2003. Ranking a random feature for variable and feature selection. En: Journal of Machine Learning Research. Vol 3; p. 1399-1414.

[29] TAKADA T. 2001. Nonparametric density estimation: A comparative study. En: Economics Bulletin. Vol. 3, No.16. p.1-10.

[30] TORKKOLA K. 2003. Feature extraction by non-parametric mutual information maximization. En: Journal of Machine Learning Research. Vol 3; p. 1415-1438.

[31] TORKKOLA K. & CAMPBELL W. 2000. Mutual information in learning feature transformations. En: Proceedings of the 17th International Conference on Machine Learning, Stanford, California. p. 1015-1022.

[32] WAND M. & JONES M. 1995. Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

[33] WILFAHRT R. 2002. An Analysis of Feature Subset Selection Methods. En: Computer Science Seminar Fall, University of Minnesota.