UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

SELECCIÓN E IDENTIFICACIÓN DE GRUPOS COMPLEMENTARIOS DE CARACTERÍSTICAS UTILIZANDO INFORMACIÓN MUTUA

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

JORGE RAMÓN VERGARA QUEZADA

PROFESOR GUÍA:
PABLO ESTÉVEZ VALENCIA

MIEMBROS DE LA COMISIÓN:
GONZALO RUZ HEREDIA
JORGE SILVA SÁNCHEZ
PABLO ZEGERS FERNÁNDEZ

SANTIAGO DE CHILE
2015

RESUMEN DE TESIS PARA OPTAR AL GRADO DE: Doctor en Ingeniería Eléctrica
POR: Jorge Ramón Vergara Quezada
FECHA: 20/03/2015
PROFESOR GUÍA: Dr. Pablo A. Estévez Valencia

SELECCIÓN E IDENTIFICACIÓN DE GRUPOS COMPLEMENTARIOS DE CARACTERÍSTICAS UTILIZANDO INFORMACIÓN MUTUA

El constante crecimiento del volumen, la tasa de generación y la dimensionalidad de datos en todas las áreas hace cada vez más necesario el análisis automático de éstos. Bajo este contexto, el pre-procesamiento de datos aparece como un área de investigación de gran interés, principalmente porque hace inteligible los datos para un posterior análisis. Una de las técnicas de preprocesamiento es la selección de características, la cual estudia cómo seleccionar un subconjunto mínimo de características para construir un modelo que genera los datos. La selección de características tiene como propósito: (i) mejorar el desempeño de predicción de los predictores, (ii) proveer mayor rapidez y efectividad en la predicción, y (iii) posibilitar una comprensión más clara del proceso que genera los datos. Tradicionalmente la literatura ofrece métodos focalizados en mejorar los dos primeros puntos, destacando aquellos métodos con las siguientes cualidades: independientes del clasificador, reflejándose en un bajo costo computacional; y criterios de selección de características basados en información mutua que tienen la capacidad de detectar relaciones no-lineales entre las características. La importancia del tercer punto es que no solo ayuda a la comprensión del proceso, sino que además permite identificar las interacciones existentes entre características (grupos de características). Por otro lado, su principal dificultad consiste en la poca precisión (y consecuentemente difícil cuantificación) de los conceptos de relevancia e interacción. La presente tesis extiende los actuales criterios de selección de características basados en información mutua a través del diseño de un criterio de selección e identificación de grupos complementarios de características. En este contexto, un grupo complementario se define como aquellas características que al actuar conjuntamente entregan mayor información de la variable de salida en relación a la suma de información que cada una de éstas por separado tiene de la variable de salida. Este trabajo se inicia con una detallada revisión de los actuales métodos de selección de características, permitiendo conocer la necesidad de incluir los términos de interacción, específicamente el término de complementariedad o sinergia. Posteriormente se realiza una formalización teórica e identificación de los conceptos de relevancia única, redundancia y complementariedad existentes entre características y la variable de salida. Además se formula una estrategia para determinar el límite inferior de información que cada característica tiene de la variable de salida. Finalmente se realiza el diseño y formulación de un nuevo criterio de selección e identificación de grupos complementarios. Las principales contribuciones de esta tesis son desarrollar una mejora del criterio CMIM (*Conditional Mutual Information Maximization*) y proponer un nuevo criterio para la identificación y selección de grupos complementarios de características. Los métodos propuestos se comparan con los criterios de selección de características más usados de la literatura. Los resultados muestran que los aportes de la propuesta van más allá de identificar grupos complementarios de características, sino que además permite mejorar el desempeño de los clasificadores en la etapa posterior al preprocesamiento. También el criterio propuesto permite visualizar los grados de interacción que existen entre las características y variable de salida. La complejidad computacional se incrementa levemente, siendo todavía comparable a los criterios más eficientes existentes en el área. Los alcances de esta propuesta permiten establecer en términos generales, una mayor y mejor comprensión de los procesos que generan los datos, y en términos específicos, la información necesaria para una futura detección del tamaño óptimo de características.

La vida debe ser una superación voluntaria de las dificultades, las que nos ocurren y las que creamos voluntariamente; de otra manera, es solo un juego de azar.
George I. Gurdjieff

Agradecimientos

Esta tesis está limitada a documentar los resultados de un dedicado y arduo trabajo de investigación, sin embargo, la contribución humana de profesores, familia, compañer@s y amig@s, es un pilar fundamental de este trabajo que generalmente es omitido. En esta sección quisiera rectificar esta omisión.

La persona más involucrada en este trabajo es mi tutor Pablo Estévez. Pablo es una persona de gran paciencia, perseverancia, buen humor y enorme intelecto. A menudo me ha recordado que cualquier dificultad es también una oportunidad para hacer una contribución. Junto con sus otras cualidades, Pablo también es un investigador formidable, con una asombrosa habilidad para detectar nuevas y fructíferas líneas de investigación, y con una capacidad de escribir fantásticos documentos técnicos. Pablo ha dedicado innumerables horas a trabajar conmigo, y también me ha compartido algunos de sus notables conocimientos sobre diversos temas. Ha sido un placer y un privilegio trabajar con Pablo durante mi carrera de posgrado.

También quisiera gradecer a los miembros de la comisión Pablo Zegers, Jorge Silva y Gonzalo Ruz por la detallada lectura de este trabajo, y sus valiosos comentarios que me permitieron mejorarlo. Me gustaría agradecer especialmente a Pablo Zegers (Pablo2, para diferenciar a los muchos Pablos que conocí en mi paso por el doctorado) su importante e inestimable apoyo desde el inicio de mi doctorado, ¡Gracias Pablo2!

Agradezco a mis compañeros de laboratorio de la Universidad de Chile: Vera (Pablo3), Daniel, Rafa, Carloncho, Shulz, Alonso, Cament, Chubo, Lucho, Juan, Cata, David, Huijse (Pablo4), Carlos, su valiosa compañía, y además por hacer del laboratorio un espacio de trabajo divertido y agradable para compartir. Ellos representaron mi nueva familia en esta aventura académica.

A cientos de kilómetros mi familia siempre ha sido una presencia reconfortante para mí durante toda mi carrera de posgrado. Ellos siempre han confiado y apoyado en mis decisiones, entregándome consejos en muchos puntos críticos en mi vida. Este trabajo no hubiera sido posible sin su apoyo.

Para finalizar y muy especialmente quiero expresar mi agradecimiento a la persona de la cual recibí el mayor de apoyo durante toda mi carrera de posgrado, mi compañera de viaje Angélica. Desde que nos conocimos, ella siempre me ha visto preocupado de la tesis y el doctorado. Con su enorme sabiduría, amor, fuerza interior, paciencia y nobleza, logró alentarme en los momentos difíciles y ha disfrutado conmigo aquellos momentos de éxitos y alegrías. ¡Gracias Angélica!

Esta tesis ha sido parcialmente financiada por el proyecto FONDECYT 1140816.

Tabla de contenido

Capítulo 1

Introducción

1.1. Motivación y fundamentación teórica

La presente tesis se refiere al diseño de un criterio para la selección e identificación de grupos complementarios de características basado en información mutua. La selección de características se define como el proceso de búsqueda de aquellas características o grupos de éstas que tienen la mayor información de la variable de salida, usualmente la clase. Por otro lado, un grupo complementario se define como aquellas características que al actuar conjuntamente entregan mayor información de la variable de salida en relación a la suma de la información que cada una de éstas tiene por separado de la variable de salida. El grado de complementariedad se establece de acuerdo a la cantidad de información "extra" que posee el grupo de características versus la suma de información que se puede obtener de cualquier combinación de éstas. En el caso extremo los grupos altamente complementarios se caracterizan por entregar información de la variable de salida solamente cuando estas características actúan conjuntamente.

Dentro de las particularidades del criterio propuesto, se puede mencionar el cálculo del límite inferior de información que cada característica tiene de la variable de salida. Esto permite realizar una comparación justa en el proceso de selección de características con el fin de detectar eficientemente aquellas características relevantes y altamente complementarias. Además, el criterio propuesto cuantifica los diferentes niveles de interacción que cada característica tiene con la variable de salida y con el resto de las características. Esta cuantificación permite identificar los grupos de características complementarias y además entregar las magnitudes de información de relevancia, redundancia y complementariedad.

Una motivación para el desarrollo de esta tesis es que el preprocesamiento de datos representa uno de los primeros pasos para un efectivo análisis de datos. Dentro de las técnicas de preprocesamiento está la selección de características, la cual estudia cómo seleccionar un subconjunto de atributos, variables o características que son usadas para construir un modelo, que en el caso supervisado, relaciona la clase con los datos. La selección de características es comúnmente utilizada en espacios de alta dimensionalidad y tiene como propósito: (i) mejorar el desempeño de predicción de los predictores, (ii) proveer mayor rapidez y efectividad en

la predicción, y (iii) tener una comprensión más clara del proceso que genera los datos [1, 2]. Aunque la literatura ofrece una variada gama de métodos de selección de características, en la mayoría de los casos se trata de aproximaciones a soluciones teóricamente óptimas. Entre otras razones, la búsqueda del subconjunto óptimo de características se realiza mediante criterios que sólo manejan un conocimiento parcial de la información total del sistema (relevante o redundante), despreciando las interacciones entre características. La cantidad de subconjuntos que se pueden formar es de 2^m, siendo m el número total de características. Esta explosión combinatorial hace que el problema no se pueda resolver en forma exhaustiva para $m \geq 20$ y por lo tanto se prefiera un enfoque heurístico.

De acuerdo a lo anterior, se requiere un criterio que cuantifique de manera eficiente la relevancia de cada característica y de grupos de éstas. Los métodos basados en teoría de la información, específicamente información mutua y entropía, han demostrado ser efectivos en el problema de selección de características [3, 4]. La principal virtud de la información mutua radica en que es una métrica independiente del clasificador o inductor, requisito fundamental en preprocesamiento de datos [5]. Además, la información mutua permite capturar relaciones no-lineales entre las características, permitiendo definir y cuantificar conceptos tales como relevancia [6, 7], redundancia [8, 9], y complementariedad, también conocida como sinergia [7, 10, 11]; las cuales representan las nociones esenciales en selección de características.

Esta tesis pretende, además de buscar un subconjunto mínimo de características relevantes, identificar cuáles son las interacciones entre características que aportan en el conocimiento de la variable de salida. Tradicionalmente, una característica relevante se define como aquella que individualmente entrega una alta información de la variable de salida [6, 12]. La definición formal de relevancia indica que la máxima información de la variable de salida se obtiene a través de la propia característica o cuando ésta interactúa con algún subconjunto de características [8, 13, 14], sin embargo, la dificultad para encontrar cuál es el subconjunto de características relevantes es un problema combinatorio de tipo NP [1, 15–17]. Por otro lado, es importante mencionar que algunas áreas de investigación, tales como neurociencia [18–22], clasificación de texto [23–25], bioinformática [14, 26–28], por mencionar algunos, tienen como principal foco determinar cuáles características, y cómo éstas permiten predecir la variable de salida, y no solo se limitan a determinar si una característica es individualmente relevante o no.

El primer problema en selección de características es la selección de la primera característica [5, 7]. Comúnmente, los métodos de selección de características utilizan una estrategia de búsqueda *greedy* que consiste en una búsqueda secuencial de características [5, 12, 29, 30]. Esta búsqueda se caracteriza por la selección de una característica a la vez, con el fin de evitar evaluar todos los posibles subconjuntos que se pueden generar. A pesar de reducir la cantidad de cálculos, tiene la desventaja de que una vez seleccionada una característica ya no se puede eliminar o cambiar su posición asignada en la selección. Este inconveniente se produce porque el valor de información de la característica seleccionada puede variar a medida que ingresan nuevas características al conjunto de variables seleccionadas. Existen criterios de selección de características que intentan evitar este problema re-evaluando la posición de la característica sobre pequeños subconjuntos de características o realizando una búsqueda combinada hacia adelante y hacia atrás [31], sin embargo, estas estrategias tienen la desventaja de aumentar la cantidad de cálculos y tiempos de selección de características [32, 33].

La gran mayoría de los criterios de selección de características basados en información mutua fueron desarrollados en forma heurística sin un fundamento teórico sólido, con lo cual se tiene poca claridad de las ventajas de un criterio u otro. Recientes trabajos proponen un marco conceptual basado en la maximización de la verosimilitud [34], desde el cual se deriva una gran cantidad de criterios. A pesar de que la propuesta de maximización de la verosimilitud permite comparar diferentes criterios de selección de características basados en información mutua existentes en la literatura, no establece claramente cuáles son las ventajas o limitaciones de éstos.

Uno de los aspectos más importantes realizados en esta tesis es una detallada revisión bibliográfica de los criterios más utilizados en selección de características basados en información mutua existentes en la literatura, con el fin de conocer las ventajas y limitaciones de éstos. Esta revisión, y su posterior análisis, permitió detectar la importancia de la complementariedad entre características para la predicción de la variable de salida, desarrollando primero una mejora del criterio CMIM (*Conditional Mutual Information Maximization*) [3] al incluirle el concepto de complementariedad. Posteriormente, dado que la complementariedad está íntimamente relacionada con la redundancia, se desarrolló una metodología para cuantificar, en componentes de información no traslapadas (independientes), los distintos grados de interacción que una característica tiene con la variable de salida o con otras características. Con esto se establecen las bases para el desarrollo de una metodología de identificación de grupos complementarios de características. Finalmente, a partir de la variación de información que una característica tiene de la variable de salida al interactuar con otras características, se diseña una metodología que establece límites inferiores de información para cada característica.

1.2. Objetivos

1.2.1. Objetivo general

- Diseñar un criterio para identificar y seleccionar características y grupos complementarios de características basado en información mutua.

1.2.2. Objetivos específicos

- Determinar las limitaciones de los actuales métodos de selección de características basados en información mutua mediante el diseño de un marco teórico basada en información mutua.
- Detectar las interacciones de redundancia y complementariedad entre características y cuantificar el efecto de estas interacciones en el conocimiento de la variable de salida.
- Establecer límites de información que cada característica tiene de la variable de salida de acuerdo a su interacción con otras características.
- Diseñar una metodología para la identificación de grupos complementarios de características relevantes utilizando teoría de la información.

1.3. Contribuciones de esta tesis

Los principales aportes desarrollados en esta tesis son expuestos a continuación:

- Desarrollo de un nuevo marco teórico basado en principios de teoría de la información que permite deducir una gran cantidad de criterios de selección de características basados en información mutua existentes en la literatura.
- Presentar las ventajas y desventajas de los actuales criterios de selección de características basados en información mutua para aproximarse al subconjunto óptimo de características.
- Desarrollo de una versión mejorada del criterio de selección de características CMIM [3], llamado CMIM2, al incluir el concepto de complementariedad.
- Desarrollo de una metodología para descomponer la información mutua entre las características y la variable de salida en componentes no traslapadas de información que cuantifican los conceptos básicos en selección de características de: relevancia única, redundancia y complementariedad.
- Establecer rangos de relevancia o información que una característica posee de la variable de salida cuando ésta interactúa con otras características.
- Desarrollo de un nuevo criterio de selección e identificación de características y grupos complementarios de éstas basado en información mutua. La ventaja de este criterio es que permite realizar un ordenamiento de características o grupos complementarios de características como una etapa previa a cualquier proceso o máquina de aprendizaje.
- Desarrollo de un *toolbox* implementado en MATLAB y C++ para realizar selección e identificación de características y grupos complementarios de características utilizando información mutua. El *toolbox* implementa varios de los tradicionales criterios de selección existentes en la literatura, además del criterio propuesto.

1.4. Publicaciones

Las siguientes publicaciones fueron resultado de este trabajo de tesis.

- **VERGARA, J. R.**; ESTÉVEZ, P. A. *A review of feature selection methods based on mutual information.* Neural Computing and Applications, 2014, vol. 24, no 1, p. 175-186.
- **VERGARA, J. R.**.; ESTÉVEZ, P. A. *CMIM2: an enhanced conditional mutual information maximization criterion for feature selection.* Journal of Applied Computer Science Methods, 2010, vol. 2, no 1, p. 5-20.

Otras publicaciones que presentan el nuevo criterio de rangos de relevancia de características e identificación de grupos complementarios de características están en proceso de redacción.

1.5. Definiciones básicas

Sea F un conjunto de características compuesto de m características y n muestras o ejemplos, y C un vector de salida que representa las clases de un proceso. Se asume que F es la realización de un muestreo aleatorio de una distribución desconocida, siendo f_i la i-ésima característica de F y $f_i(j)$ la j-ésima muestra de la característica f_i. Igualmente, c_i es la i-ésima componente de C y $c_i(j)$ es la j-ésima muestra de la clase c_i. Las letras mayúsculas denotan conjuntos de variables aleatorias (características), y letras minúsculas denotan características individuales de este conjunto.

Otras notaciones y terminologías usadas en este trabajo son las siguientes.

S	Subconjunto de características seleccionadas hasta el momento (en un proceso secuencial).
f_i	Característica candidata que será agregada (o eliminada) al (del) subconjunto de características seleccionadas S.
$\{f_i, f_j\}$	Subconjunto compuesto de las características f_i y f_j.
$\neg f_i$	Subconjunto de todas las características en F excepto la característica f_i. $\neg f_i = F \backslash f_i$.
$\{f_i, S\}$	Subconjunto compuesto de la característica f_i y el subconjunto S.
$\neg\{f_i, S\}$	Subconjunto de todas las características en F excepto el subconjunto $\{f_i, S\}$. $\neg\{f_i, S\} = F \backslash \{f_i, S\}$
$p(f_i, C)$	Probabilidad de masa conjunta entre la característica f_i y la variable de salida C.
$\lvert\cdot\rvert$	Valor absoluto / cardinalidad de un conjunto.

Los conjuntos, subconjuntos y características mencionados anteriormente están relacionados de la siguiente forma: $F = f_i \cup S \cup \neg\{f_i, S\}$, $\emptyset = f_i \cap S \cap \neg\{f_i, S\}$.

1.6. Estructura de la tesis

A continuación se presenta un resumen de los temas abordados en cada capítulo de esta tesis.

Capítulo 1: Se presenta una introducción al problema a resolver, cuáles son sus inconvenientes y la forma en que serán trabajadas en esta investigación. Además se presentan algunas definiciones básicas y notaciones utilizadas en este trabajo. Finalmente se presentan el aporte de esta tesis en el área de selección de características y las contribuciones (publicaciones) logradas desde esta investigación.

Capítulo 2: Se presentan los elementos teóricos fundamentales utilizados en este trabajo, los cuales son: teoría de la información, selección de características y algoritmos de aprendizaje supervisado. En la sección 2.1 se definen los elementos básicos de teoría de la información necesarios para comprender y diseñar un criterio que cuantifique la cantidad de información que contienen las características analizadas de la variable de salida. En la sección 2.2 se presenta una revisión de la teoría desarrollada en selección de características. Las preguntas que intenta responder esta sección son: ¿cuál es el objetivo de la selección de características?, ¿cómo se clasifican las características?, ¿qué estrategias existen para la búsqueda de características?, entre otras preguntas. En la sección 2.3 se presenta el funcionamiento de dos algoritmos de aprendizaje supervisado que serán utilizados para evaluar y comparar el desempeño del criterio propuesto.

Capítulo 3: : Se presenta el estado del arte de los criterios de selección de características basados en información mutua existentes en la literatura. Este recorrido comienza con la reciente propuesta de un marco unificador que permite explicar gran parte de los métodos de selección de características basados en información mutua. Posteriormente se detallan algunos de los criterios más utilizados.

Capítulo 4: Se presentan los dos temas desarrollados en esta tesis. En la sección 4.1 se presenta la mejora del criterio heurístico de selección de característica CMIM. La nueva versión de CMIM, llamado CMIM2, nace como resultado de un detallado análisis para determinar cuáles son las ventajas de CMIM por sobre otros criterios de su clase, detectando la importancia de incluirle el término de complementariedad. Además, el análisis desarrollado permitió identificar qué información relevante se pierde en varios de los criterios existentes en la literatura como resultado de la aproximación que hacen. En la sección 4.2 se presenta un criterio para la detección e identificación de grupos de características complementarias basado en información mutua. La novedad de este criterio se halla en que, además de ordenar las características de acuerdo a su conocimiento de la variable de salida, mide el grado de interacción de grupos de características en el conocimiento de la variable de salida. Además, se presenta una propuesta para dividir y cuantificar la información que se tiene de la variable de salida en componentes no traslapadas de relevancia, redundancia y complementariedad.

Capítulo 5: Se presenta las principales conclusiones obtenidas en el desarrollo de esta tesis. Los resultados empíricos muestran la superioridad del ordenamiento entregado por el criterio propuesto en relación a los criterios de selección de características basados en información mutua más utilizados en la literatura. También, se muestra que la selección de grupos complementarios permite tener una mejor comprensión del proceso que genera los datos. Finalmente se exponen nuevas líneas de investigación a desarrollar a futuro.

Capítulo 2

Fundamentos de selección de características utilizando información mutua

En este capítulo se exponen los fundamentos de teoría de la información y de selección de características que sirven de base para el desarrollo de esta tesis. En la sección 2.1 se presentan las unidades centrales de la teoría de la información utilizados para el desarrollo del funcional propuesto. Entre los conceptos abordados en esta sección se encuentran: entropía, información mutua, información mutua multivariable, así como un estimador para calcular entropía e información mutua desde los datos. En la sección 2.2 se presentan los fundamentos de selección de características donde se revisan: objetivos de la selección de características, estrategias de búsqueda en el proceso de selección de características, clasificación de las características, y definición del subconjunto óptimo de características. Finalmente, en la sección 2.3 se presenta resumidamente el funcionamiento de dos algoritmos de aprendizaje utilizados en esta tesis para comparar y validar los resultados entregados por el funcional propuesto en el capítulo 4.

2.1. Elementos de teoría de la información

La entropía, divergencia e información mutua son conceptos básicos definidos en teoría de la información [35]. En su origen, la teoría de la información se utilizó dentro del contexto de la teoría de la comunicación, para encontrar respuestas acerca de la compresión de datos y la velocidad de transmisión [36]. Desde entonces, los principios de la teoría de la información se han incorporado en gran medida en el aprendizaje automático y específicamente en la selección de características, véase por ejemplo Principe *et al.* [37].

2.1.1. Entropía

Entropía (H) es una medida que determina el valor medio ponderado de incertidumbre de los eventos ocurridos en una variable aleatoria. La incertidumbre está relacionada con

la probabilidad de ocurrencia de un evento. En términos simples, una alta entropía indica que cada evento de la variable aleatoria tiene aproximadamente la misma probabilidad de ocurrencia, mientras que una baja entropía significa que cada evento tiene una probabilidad diferente de ocurrencia. Formalmente, la entropía de una variable aleatoria discreta x, con probabilidad de masa $p(x(i)) = Pr\{x = x(i)\}$, $x(i) \in x$ se define como:

$$H(x) = -\sum_{i=1}^{n_i} p(x(i)) \log_2(p(x(i))), \tag{2.1}$$

donde n_i representa el número de muestras de x. La ecuación (2.1) también se interpreta como el valor esperado del negativo del logaritmo de la probabilidad de masa de x, es decir, $\mathbb{E}_{p(x)}\left[-\log\left(p(x)\right)\right]$.

La entropía conjunta de las variables aleatorias discretas x e y con probabilidad de masa conjunta $p(x, y)$, corresponde a la suma ponderada de la incertidumbre contenida por estas dos características. Formalmente, la entropía conjunta se define como:

$$H(\{x, y\}) = -\sum_{i=1}^{n_i}\sum_{j=1}^{n_j} p(x(i), y(j)) \cdot \log_2(p(x(i), y(j))), \tag{2.2}$$

donde n_i y n_j representan el número de muestras de x e y respectivamente. El rango de valores que puede tomar la entropía conjunta corresponde a:

$$\text{máx}\left(H(x), H(y)\right) \leq H(\{x, y\}) \leq H(x) + H(y), \tag{2.3}$$

donde la desigualdad (2.3) alcanza su valor máximo cuando x y y son completamente independientes, mientras que el mínimo valor ocurre cuando x es completamente dependiente de y.

Para conocer la cantidad de incertidumbre restante de la variable aleatoria x cuando se conoce la variable aleatoria y, se utiliza la entropía condicional. Formalmente, la entropía condicional se define como:

$$H(x|y) = \sum_{j=1}^{n_j} p(y(j)) \cdot H(x|y = y(j)), \tag{2.4}$$

donde $H(x|y = y(j))$ es la entropía de x asociada al evento (clase) $y(j)$ de la variable y. Otra forma de representar la entropía condicional es:

$$H(x|y) = H(\{x, y\}) - H(y). \tag{2.5}$$

El rango de valores de la entropía condicional es:

$$0 < H(x|y) < H(x), \tag{2.6}$$

donde el valor cero en la ecuación 2.6 ocurre cuando x es estadísticamente dependiente de y, es decir, no queda incertidumbre en x si conocemos y. Por otro lado, el valor máximo de la entropía condicional ocurre cuando x e y son estadísticamente independientes, es decir, la variable y no añade información para reducir la incertidumbre de x.

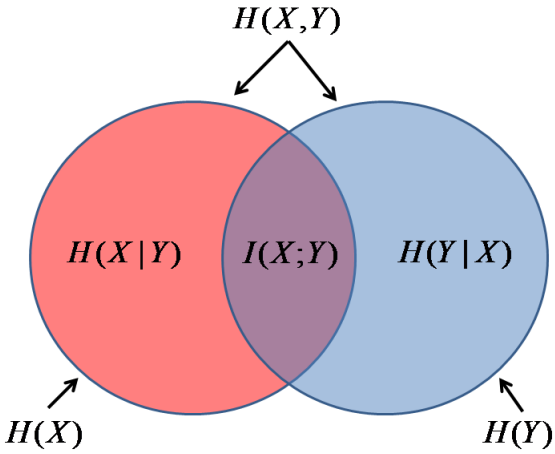


Figura 2.1: Diagrama de Venn mostrando las relaciones entre la información mutua y entropías de las variables aleatorias x e y.

2.1.2. Información mutua

La información mutua (I) mide la cantidad de información que una variable aleatoria tiene sobre otra variable [35]. Esta definición es útil en el contexto de la selección de características, ya que permite cuantificar la información que entrega un subconjunto de características con respecto a la variable de salida (vector de clase C). Formalmente, la información mutua entre dos variables aleatorias x e y se define como:

$$I(x;y) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} p(x(i), y(j)) \cdot \log \left(\frac{p(x(i), y(j))}{p(x(i)) \cdot p(y(j))} \right), \tag{2.7}$$

donde la ecuación (2.7) entrega un valor cero cuando x e y son estadísticamente independientes, es decir, $p(x(i), y(j)) = p(x(i)) \cdot p(y(j))$. La información mutua está relacionada linealmente con la entropía de las variables a través de las siguientes ecuaciones:

$$I(x;y) = \begin{cases} H(x) - H(x|y) \\ H(y) - H(y|x) \\ H(x) + H(y) - H(x,y). \end{cases} \tag{2.8}$$

La Figura 2.1 muestra el diagrama de Venn con las relaciones descritas en (2.8).

La información mutua puede ser condicionada sobre una tercera variable obteniéndose de esta forma la información mutua condicional. La información mutua condicional se define como:

$$I(x;y|z) = \sum_{k=1}^{n_k} p(z(k)) I\left(x;y|z=z(k)\right), \tag{2.9}$$

donde z es una nueva variable aleatoria discreta de n_k muestras, e $I\left(x;y|z=z(k)\right)$ corresponde a la información mutua entre x e y en el contexto del evento $z(k)$ de la variable z. La información mutua condicional permite medir la información mutua entre dos variables en el contexto de una tercera, pero no mide la información mutua contenida en las tres variables. Los valores de la información mutua condicional $I(x;y|z)$ pueden ser mayores o menores que la información mutua $I(x;y)$.

2.1.3. Información mutua multivariable

El primer intento por medir la información mutua entre más de dos variables fue propuesto por McGill [38] para el caso de tres variables. La multi-información o información de interacción mutua fue definida por McGill como:

$$I(x;y;z) = \begin{cases} I(x;y|z) - I(x;y) \\ I(y;z|x) - I(y;z) \\ I(x;z|y) - I(x;z), \end{cases} \tag{2.10}$$

donde las tres expresiones del lado derecho de la ecuación (2.10) son equivalentes. También la multi-información puede ser definida como:

$$I(x;y;z) = I(\{x,y\};z) - I(x;z) - I(y;z). \tag{2.11}$$

Posteriormente, Fano [39] extendió la definición de información mutua al caso general (sobre tres variables), la cual fue reformulada sobre un marco teórico por Han [40]. A pesar que Fano y Han consideraron diferentes aproximaciones, sus resultados conducen a lo mismo.

Fano [39] calcula la información mutua entre un número arbitrario de variables (mayor a dos) como una extensión del caso de información mutua de dos variables. Para comprender esta generalización de la información mutua, es conveniente comenzar del caso básico de dos variables aleatorias x_1 y x_2 definido como:

$$\begin{aligned} I(x_1;x_2) &= H(x_1) - H(x_1|x_2) \\ &= -I(x_1) + I(x_1|x_2). \end{aligned} \tag{2.12}$$

En la segunda igualdad de la ecuación (2.12), Fano define la auto-información de una variable como el negativo de la entropía de la variable, es decir, $I(x_1) = -H(x_1)$. Extendiendo la información mutua para tres variables se tiene:

$$I(x_1; x_2; x_3) = I(x_1; x_2|x_3) - I(x_1; x_2), \tag{2.13}$$

Para el caso de m variables, la multi-información entre éstas se define como:

$$I(x_1; x_2; ...; x_m) = I(x_1; x_2; ...; x_{m-1}|x_m) - I(x_1; x_2; ...; x_{m-1}). \tag{2.14}$$

Por otro lado, la información mutua básica que existe entre un conjunto de m variables y la clase C, puede ser representada como la suma de las multi-informaciones de todas las posibles combinaciones de las m variables y C. Matemáticamente la representación de la información mutua básica de un conjunto de m variables y C en términos de multi-información se define como:

$$I(\{x_1, x_2, ..., x_m\}; C) = \sum_{k=1}^{m} \sum_{\substack{\forall S \subseteq \{x_1, ..., x_m\} \\ |S| = k}} I([S \cup C]), \tag{2.15}$$

donde $I([S \cup C]) = I(s_1; s_2; \cdots; s_k; C)$. Nótese que la suma del lado derecho de la ecuación (2.15) es realizada sobre todos los subconjuntos S de tamaño k obtenidos desde el conjunto $\{x_1, ..., x_m\}$[1].

Algunas propiedades de la información mutua multivariable (multi-información) definidas en [41] son:

- Una interpretación intuitiva de la información mutua multivariable corresponde a la ganancia (o pérdida) en la información entre un grupo de variables debido al conocimiento de una nueva variable.
- A diferencia de la información mutua bivariable que siempre es positiva, la información mutua multivariable puede tomar valores positivos y negativos. Esto es posible ya que el efecto de una variable puede incrementar o disminuir la dependencia de las otras. La ecuación (2.10) indica que si la dependencia entre dos variables es mayor que la dependencia de estas dos variables bajo el contexto de una tercera variable, entonces la información mutua multivariable será negativa.
- La información mutua multivariable es simétrica, por lo que cualquier combinación en el orden de las variables utilizadas en el cálculo de la información mutua entre ellas entregará el mismo resultado.
- La información mutua multivariable posee la propiedad de semi-independencia, esto significa, que un valor cero de información mutua multivariable no significa necesariamente que las variables sean independientes entre sí.
- Las propiedades de recursividad y regla de la cadena de la información mutua bivariable son extensibles a la información mutua multivariable [42].

[1]Para evitar confusiones futuras en casos como: $I(x_1, x_2; C)$ y $I(x_1; x_2; C)$, donde $I(x_1, x_2; C)$ corresponde a la información mutua bivariable $I(\{x_1, x_2\}; C)$, mientras que $I(x_1; x_2; C)$ corresponde a la información mutual multivariable, se utilizará la notación $II(:)$ para indicar la información mutua entre más de dos variables ($II(x_1; x_2; C) \equiv I(x_1; x_2; C)$).

2.1.4. Estimación de la entropía

La estimación de la entropía de Shannon [36] a partir de una distribución de probabilidad dado un conjunto de muestras ha sido ampliamente estudiada [43–52]. Los estimadores de entropía pueden ser divididos en dos categorías: "*plug-in*" y "*non plug-in*". Los métodos *plug-in* [47, 51, 53–55] estiman la densidad de probabilidad mediante la creación de histogramas o métodos basados en las ventanas de Parzen. Por otro lado, los métodos *non plug-in* [44, 45, 48, 56–58] estiman la entropía directamente desde las muestras, al derivar desde éstas la distribución de densidad de probabilidad.

Los métodos *plug-in* tienen un gran problema cuando el número de dimensiones es grande. Debido a la *maldición de las dimensiones,*[2] no es posible obtener una buena estimación de la distribución de probabilidad real que genera las muestras desde el conjunto de las muestras cuando existen más de dos o tres dimensiones. Además, el número de muestras necesarias para estimar correctamente la densidad de probabilidad crece exponencialmente al aumentar las dimensiones [47,59], por lo tanto, la limitante para estos métodos es que no sólo dependen del número de dimensiones, sino que también del número de datos.

La aproximación basada en las ventanas de Parzen [60, 61] es ampliamente usada para estimar funciones de probabilidad desde un conjunto finito de muestras. Esta aproximación utiliza un kernel (usualmente Gaussiano) que tiene un parámetro para ajustar el ancho del kernel usado. Un inadecuado valor de este parámetro afecta el resultado de la estimación [47].

Entre los métodos *non plug-in*, los más comunes son aquellos métodos basados en grafos entrópicos o grafos de vecinos más cercanos. Estos métodos son capaces de estimar la entropía aun para un alto número de dimensiones en el espacio. Los grafos entrópicos o de vecindad están basados en la distancia entre las muestras. Al calcular las distancias, sin importar el gran número de dimensiones, el problema es simplificado a medidas de distancias. De esta forma, la complejidad computacional no depende de la dimensionalidad, sino sólo del número de muestras.

Por otro lado, una ventaja de los métodos *non plug-in* por sobre los métodos *plug-in* se observa al comparar las ventanas de Parzen y los gráficos entrópicos. En la aproximación de las ventanas de Parzen existe problema con la esparcidad de los datos. El kernel Gaussiano no considera las muestras que se encuentran a una distancia mayor que tres desviaciones estándar, ya que el peso asociado a estas muestras es cercano a cero. Este efecto es menos notorio cuando se usa los grafos entrópicos para estimar la entropía (ver Figura 2.2).

2.1.4.1. Estimación de la entropía de Kozachenko-Leonenko

El estimador de Kozachenko-Leonenko [45, 63] es un método *non plug-in* para estimar la entropía de Shannon basado en la distancia de las muestras con sus vecinos más cercanos. Como se mencionó anteriormente, la ventaja de este tipo de estimador es que sus cálculos están basados en las distancias de las muestras, sin importar la dimensión en que éstas se encuentren. Esta última cualidad es una de las principales razones por las cuales se eligió

[2]Definida posteriormente en la sección 2.2.4.1.

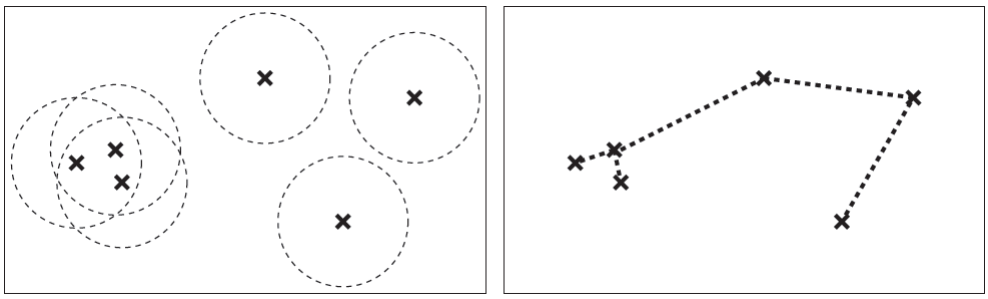


Figura 2.2: Efecto de la distribución de datos en la estimación de la entropía. En el gráfico izquierdo (método *plug-in*) se muestra la utilización de ventanas de Parzen para estimar la distribución de probabilidad de los datos. Una inadecuada selección en el ancho del kernel hará que algunas muestras no interactúen con el resto en la estimación de la distribución de datos. Por otro lado, en el gráfico de la derecha (método *non plug-in*) se muestra que los métodos basados en distancias (vecinos más cercanos) si consideran aquellas muestras en la estimación de la distribución los datos. Fuente: [62].

este estimador para el desarrollo de esta tesis. Algunas pruebas donde se muestra la ventaja de este estimador en relación a otros estimadores pueden ser encontradas en [43, 64, 65].

El objetivo del estimador de Kozachenko-Leonenko consiste en encontrar una estimación rápida y eficiente de la entropía de Shannon. Para lograr esto, se comienza con la definición de entropía de Shannon para el caso continuo, la cual se define como [35]:

$$H(X) = -\int f(x) \log f(x) dx, \tag{2.16}$$

siendo $f(x)$ una función de densidad de probabilidad. La primera aproximación utilizada por Kozachenko-Leonenko consiste en considerar la integral de la ecuación (2.16) como el promedio de $\log f(x)$ $\left(\widehat{\log f(x)}\right)$. Por lo tanto, si se pudiera obtener un estimador insesgado[3] para $\widehat{\log f(x)}$, entonces se podría calcular un estimador insesgado para la ecuación (2.16). De esta forma se tiene que la aproximación realizada por Kozachenko-Leonenko es:

$$\widehat{H}(X) = -n^{-1} \sum_{i=1}^{n} \widehat{\log f(x(i))}, \tag{2.17}$$

donde n es el número de muestras obtenidas de la función de densidad de probabilidad $f(x)$.

Para obtener $\widehat{\log f(x)}$, se utiliza la distribución de probabilidad $P_k(\epsilon)$ de la distancia entre la muestra $x(i)$ y su k-vecino más cercano. La distribución $P_k(\epsilon)$ se obtiene considerando una bola de diámetro ϵ centrada en $x(i)$, y donde existe una muestra a una distancia $\epsilon/2$, entonces existen $k-1$ muestras más cercanas a $x(i)$ (muestras dentro de la bola) y $n-k-1$ muestras más alejadas de x_i (muestras fuera de la bola). La probabilidad de que esto suceda es [45]:

[3]Un estimador es insesgado si el valor esperado del mismo es igual al parámetro de la población que estima.

$$P_k(\epsilon) = k \cdot \binom{n-1}{k} \cdot \frac{dp_i(\epsilon)}{d\epsilon} \cdot p_i^{k-1} \cdot (1-p_i)^{n-k-1}, \tag{2.18}$$

siendo p_i la probabilidad de masa de la ϵ-bola definida como:

$$p_i(\epsilon) = \int_{\|\xi - x(i)\| < \epsilon/2} f(\xi) d\xi. \tag{2.19}$$

Por lo tanto, la esperanza de $\log p_i(\epsilon)$ es:

$$\mathbb{E}\left[\log p_i(\epsilon)\right] = \int_0^\infty P_k(\epsilon) \log p_i(\epsilon) d\epsilon \tag{2.20}$$

$$= k \cdot \binom{n-1}{k} \cdot \int_0^1 p^{k-1} \cdot (1-p)^{N-k-1} \cdot \log p \cdot dp \tag{2.21}$$

$$= \psi(k) - \psi(n), \tag{2.22}$$

donde $\psi(\cdot)$ es la función digamma. Si se asume que $f(x)$ es constante en toda la ϵ-bola, entonces es posible aproximar

$$p_i(\epsilon) \approx \frac{V_d}{2^d} \cdot \epsilon^d \cdot f(x(i)), \tag{2.23}$$

siendo d la dimensión de $x(i)$ y V_d el volumen de la bola unitaria $\mathcal{B}(0,1)$. Para la norma máximo se tiene que $V_d = 1$, mientras que para norma euclidiana, V_d se define como:

$$V_d = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}. \tag{2.24}$$

Despejando $f(x(i))$ de la ecuación (2.23) y luego aplicando el logaritmo y operador esperanza, es posible obtener $\widehat{\log f(x)}$ como:

$$\mathbb{E}\left[\log f_i(\epsilon)\right] \approx \mathbb{E}\left[\log p_i(\epsilon)\right] - \mathbb{E}\left[\log \epsilon^d\right] - \mathbb{E}\left[\log \frac{V_d}{2^d}\right]. \tag{2.25}$$

Reemplazando la ecuación (2.22) en (2.25) y utilizando las propiedades de la esperanza, la ecuación (2.25) queda como:

$$\mathbb{E}\left[\log f_i(\epsilon)\right] \approx \psi(k) - \psi(n) - d \cdot \mathbb{E}\left[\log \epsilon\right] - \log \frac{V_d}{2^d}, \tag{2.26}$$

donde la entropía de Shannon estimada por Kozachenko-Leonenko se obtiene reemplazando la ecuación (2.26) en (2.17), quedando como

$$\widehat{H}(X) = -\psi(k) + \psi(n) + \log \frac{V_d}{2^d} + \frac{d}{n} \sum_{i=1}^{n} \log \epsilon(i). \tag{2.27}$$

El estimador de Kozachenko-Leonenko puede ser usado también para estimar la información mutua [66]. Para ello es conveniente recordar que en problemas de clasificación la información mutua puede escribirse en términos de la entropía como:

$$\hat{I}(X;C) = \hat{H}(X) - \sum_{c \in C} \hat{p}_C(c) \hat{H}(X|C=c). \tag{2.28}$$

Reemplazando la entropía en la ecuación (2.28) por el estimador definido en la ecuación (2.27) se obtiene

$$\hat{I}(X;C) = \psi(n) - \frac{1}{n} \sum_{c \in C} n_c \psi(n_c) + \frac{d}{n} \left[\sum_{i=1}^{n} \log \epsilon_k(i) - \sum_{c \in C} \sum_{i|c_i=c} \log \epsilon_k(i|c) \right], \tag{2.29}$$

donde n_c es el número de muestras que pertenecen a la clase c, $\epsilon_k(i)$ es diámetro de la hiperesfera centrada en $x(i)$ que contiene los $k-1$ vecinos más cercanos de $x(i)$, y $\epsilon_k(i|c)$ es diámetro de la hiperesfera centrada en $x(i)$ que contiene los $k-1$ vecinos más cercanos de $x(i)$ de la clase c. La distancia $\epsilon_k(i)$ está dada por

$$\epsilon_k(i) = 2 \cdot \|x(i) - x(j)\|, \tag{2.30}$$

siendo $x(j)$ el k-vecino más cercano de $x(i)$, y $\|\bullet\|$ la norma euclideana.

2.2. Selección de características

2.2.1. Objetivo de la selección de características

La selección de características es una técnica de preprocesamiento de datos que es utilizada para reducir el número de características en aplicaciones con decenas o miles de características[4]. En términos generales la selección de características estudia cómo seleccionar un subconjunto de características relevantes del total de características existentes, con el fin de construir modelos que describan el vector de clases C. Este propósito incluye remover características irrelevantes y/o redundantes para reducir la cantidad de características necesarias en el proceso de aprendizaje [67] (reducir la dimensionalidad), mejorar la precisión de predicción de los algoritmos [1], y mejorar la comprensión de los modelos construidos [68].

2.2.2. Clasificación de los métodos de selección de características

En el contexto de aprendizaje supervisado (clasificación), las técnicas de selección de características han sido agrupadas en tres categorías dependiendo de cómo éstas realizan la búsqueda de características con respecto a la construcción del modelo del clasificador. Estas categorías son: métodos de filtros, métodos envolventes y métodos embebidos.

[4] A través de este trabajo se usará indistintamente el término "variable" o "característica" para indicar los atributos que describen un dato, muestra o instancia.

2.2.2.1. Métodos de Filtro

Los métodos tipo filtro evalúan la relevancia de las características solamente analizando las propiedades intrínsecas de los datos y su relación con la clase, sin ocupar un algoritmo de inducción (clasificador). En la mayoría de los casos, estos métodos calculan una puntuación de relevancia de las características, descartando aquellas características que presentan valores inferiores en relación a un umbral establecido. Una vez seleccionado el subconjunto por el filtro éste es presentado como entrada a un clasificador.

Las ventajas de los métodos tipo filtros es que son fácilmente aplicables a datos de alta dimensionalidad, son computacionalmente rápidos y simples, y son independientes del clasificador. Como resultado, la selección de características es realizada sólo una vez, y diferentes clasificadores pueden evaluar el mismo subconjunto. La desventaja de estos métodos es que ignoran la interacción con el clasificador (la búsqueda en el subespacio de características está separada del espacio de hipótesis). Además la mayoría de los métodos propuestos en la literatura son univariables, es decir, cada característica es considerada individualmente, ignorando posibles dependencias entre ellas, lo cual puede producir un pobre desempeño de clasificación comparado con otras técnicas de selección de características. Para resolver en parte esto, actualmente las técnicas multivariables incluyen en sus funcionales medidas para incorporar grados de dependencia entre características. Ejemplos de algoritmos tipo filtros se pueden encontrar en los siguientes trabajos [69–75].

2.2.2.2. Métodos envolventes

Los métodos envolventes incluyen la búsqueda del modelo de hipótesis dentro de la búsqueda del subconjunto de características. En este esquema, el procedimiento de búsqueda de subconjuntos de características consiste en generar varios subconjuntos obtenidos del conjunto original de características, los que son evaluados utilizando un clasificador. En el espacio de todos los subconjuntos de características, la búsqueda del subconjunto óptimo de características es envuelta *wrapped* en torno al clasificador. La dificultad de estos métodos consiste en que el número de subconjuntos crece exponencialmente al aumentar el número de características, siendo necesario utilizar métodos de búsqueda heurísticos para guiar la búsqueda del subconjunto óptimo de características. Las ventajas de los métodos envolventes es que éstos si incluyen la interacción entre el subconjunto de características y el clasificador, además de tomar en cuenta la dependencia entre características. Un inconveniente común es que estas técnicas son propensas al sobre-entrenamiento y tienen un gran costo computacional, especialmente al construir el clasificador. Aplicaciones con estos métodos en el áreas de selección de características se encuentran en los siguientes trabajos: [13, 76, 77].

2.2.2.3. Métodos embebidos

Los métodos embebidos construyen un subconjunto óptimo de características conjuntamente con el aprendizaje del clasificador. Esto puede ser visto como una búsqueda en el espacio combinado de subconjuntos y de hipótesis. Al igual que los métodos envolventes, la

Tabla 2.1: Resumen de los métodos de selección de características según Guyon [1].

	Filtros	Envolvente	Embebido
Criterio	Medida sobre subconjunto característica/característica: *Relevancia*	Medida sobre subconjunto de características: *Utilidad*	Medida sobre subconjunto de características: *Utilidad*
Búsqueda	Usualmente ordena las características (*ranking*)	Busca en el espacio de todas las características	Búsqueda guiada por el proceso de aprendizaje
Validación	Usa prueba estadística	Validación cruzada	Validación cruzada
Cualidades	Son (relativamente) robustos ante el sobreajuste, pero puede fallar en la selección de la característica más relevante	Pueden en principio encontrar la característica más relevante, pero es propenso al sobreajuste	Similar a los *wrapper*, pero es menos costoso computacionalmente y menos propenso al sobreajuste

selección de las características depende fuertemente del algoritmo de aprendizaje utilizado. Los métodos embebidos tienen la ventaja de incluir la interacción con el clasificador, siendo al mismo tiempo menos costosos computacionalmente que los métodos envolventes. Ejemplos de estos métodos pueden ser encontrados utilizando *Support Vector Machines* [78–80] y árboles de decisión [81–83], por mencionar algunos.

La Tabla 2.1 presenta un resumen de las características de los tres tipos de métodos revisados anteriormente.

2.2.3. Clasificación de las características

En términos generales, el proceso de selección de características busca detectar el subconjunto mínimo de características (S_{opt}) que conserve la información contenida en el conjunto original de características (F) con respecto a la variable deseada (C). La clasificación de una característica (f_i) dependerá de la cantidad de información que ésta tenga de la variable de salida C o de un subconjunto de $F \backslash f_i$. Tradicionalmente, una característica (f_i) se clasifica como relevante cuando ésta tiene información de la variable de salida (C), por lo que mientras mayor sea la información de C, mayor es su relevancia. Si no presenta información de C, entonces será clasificada como irrelevante. Por otro lado, una característica (f_i) será clasificada como redundante cuando la información que ésta tenga sea de un subconjunto de $F \backslash f_i$. Recientes estudios [14, 19, 84–92] han propuesto la complementariedad o sinergia como una nueva taxonomía de las características. Esta sinergia ocurre cuando f_i interactúa con algún subconjunto de $F \backslash f_i$, lográndose una mayor cantidad de información de C en relación a la suma de informaciones que f_i y el subconjunto generan individualmente de C.

A continuación se revisan en detalle las diferentes clasificaciones de características mencionadas.

2.2.3.1. Relevancia

Una característica f_i se dice relevante cuando ésta, individualmente o junto con otras características, proporciona información acerca de C. En la literatura existen varias definiciones de interés en las cuales incluyen diferentes niveles de relevancia [1, 4, 6, 8, 12, 13, 15, 34, 69, 93]. Kohavi y John [13] utilizaron un marco probabilístico para definir tres niveles de relevancia: (i) características fuertemente relevantes, (ii) características débilmente relevantes, y (iii) características irrelevantes. Posteriormente Meyer [14] redefinió estos niveles de relevancia en términos de información mutua.

Una característica f_i se define como fuertemente relevante cuando ésta proporciona información única de C, es decir, la información de f_i no pueden ser reemplazadas por el resto de características pertenecientes a $\neg fi$. Utilizando la información mutua es posible determinar una característica relevante mediante la evaluación de $I(fi; C|\neg f_i)$. Si esta medida es mayor que cero (o superior a un umbral definido previamente), f_i se clasifica como característica fuertemente relevante.

Una característica f_i se define como débilmente relevante cuando ésta proporciona información de C, sin embargo, existe en $\neg f_i$ alguna característica (o grupo de características) que tiene la misma información que f_i tiene de C. La evaluación de $I(fi; C|\neg f_i)$ es cero dado que en $\neg f_i$ ya existe la información que f_i tiene de C. Por lo tanto, para clasificar una características como débilmente relevante se tiene que encontrar el subconjunto S de f_i donde no estén incluidas las características que entregan la misma información que f_i tiene de C. De esta forma se obtendrá $I(fi; C|S)$ mayor que cero.

Finalmente, las características irrelevantes no proporcionan información de C, y pueden ser eliminadas sin perder información de C. Para definir si f_i es irrelevantes, es necesario que $I(fi; C|S)$ sea igual a cero para todos los subconjuntos $S \subseteq \neg f_i$.

La Tabla 2.2 muestra los enfoques probabilístico y de información mutua para definir los niveles de relevancia, junto con las condición para definir cada una de los tipos de características relevantes. Es importante mencionar que el enfoque probabilístico presenta dos inconvenientes para determinar las diferentes clases de relevancia: 1) es necesario probar independencia condicional de todos los posibles subconjuntos de características y, 2) es necesario estimar las funciones de densidad de probabilidad (pdfs) [94].

Una definición alternativa de relevancia se da en el marco de la información mutua [3, 6–8, 71, 95–97]. Una ventaja de este enfoque es que existen buenos métodos para la estimación de la información mutua. La última columna de la Tabla 2.2 muestra cómo se definen los tres niveles de relevancia en términos de información mutua.

Las definiciones mostradas en la Tabla 2.2 presentan algunos inconvenientes que se resumen de la siguiente manera:

1. Para clasificar una característica f_i como irrelevante, es necesario evaluar todos los posibles subconjuntos de $\neg f_i$. Por tanto, este procedimiento está sujeto a la maldición de la dimensionalidad [98, 99].

2. La definición de características fuertemente relevantes es muy restrictiva. Si dos características proporcionan información acerca de la clase, pero son redundantes, entonces ambas características serán descartadas por este criterio. Por ejemplo, sea $\{x_1, x_2, x_3\}$ un conjunto de 3 características, donde $x_1 = x_2$ y x_3 es ruido, y la clase de salida se define como $C = x_1$. Considerando el criterio de relevancia fuerte tenemos $I(x_1, C|\{x_2, x_3\}) = I(x_2; C|\{x_1, x_3\}) = I(x_3; C|\{x_1, x_2\}) = 0$.

3. La definición de relevancia débil no es suficiente para decidir cuál característica debe quedarse en el subconjunto óptimo de características, ya que no se discrimina cuáles características son redundantes entre sí.

Tabla 2.2: Niveles de relevancia de la característica candidata f_i, de acuerdo al esquema probabilístico y al esquema basado en información mutua.

Nivel de relevancia	Condición	Aproximación probabilística [13]	Aproximación de información mutua [14]
Relevancia Fuerte	$\nexists$	$p(C\|f_i, \neg f_i) \neq p(C\|\neg f_i)$	$I(f_i; C\|\neg f_i) > 0$
Relevancia Débil	$\exists\, S \subset \neg f_i$	$p(C\|f_i, \neg f_i) = p(C\|\neg f_i)$ $\wedge$ $p(C\|f_i, S) \neq p(C\|S)$	$I(f_i; C\|\neg f_i) = 0$ $\wedge$ $I(f_i; C\|S) > 0$
Irrelevancia	$\forall\, S \subseteq \neg f_i$	$p(C\|f_i, S) = p(C\|S)$	$I(f_i; C\|S) = 0$

2.2.3.2. Redundancia

La ambigüedad de si una característica débilmente relevante debe ser descartada o no en la búsqueda del subconjunto óptimo, fue aclarada por Yu y Liu [8], quienes propusieron una subdivisión de las características débilmente relevantes: en redundantes y no-redundantes. Además, los autores definen el conjunto óptimo de características como aquel compuesto por características fuertemente relevantes más aquellas características débilmente relevantes pero no redundantes.

Recordando que la redundancia se asocia con el nivel de dependencia entre dos o más características, es posible medir la dependencia de una determinada característica f_i con respecto al subconjunto de características $S \subseteq \neg f_i$, simplemente usando la información mutua, $I(f_i, S)$. Esta medida de información teórica de redundancia satisface las siguientes propiedades: es simétrica, no lineal, no negativa, y no disminuye su valor al añadir nuevas características [14].

Un inconveniente de esta medida (la información mutua) es que no es posible determinar concretamente cuál o cuáles características de S son redundantes con f_i. Esto requiere de

criterios más elaborados de redundancia, tales como *Markov Blanket* [8, 71], y la correlación total [100].

Markov Blanket es una fuerte condición de independencia condicional, y se define como:

Definición 2.2.1 ***(Markov Blanket)*** *Dada la característica f_i, el subconjunto $M \subseteq \neg f_i$ es un Markov Blanket de f_i ssi [8, 71]:*

$$p(\{F \backslash \{f_i\,, M\}, C\} \mid \{f_i\,, M\}) = p(\{F \backslash \{f_i\,, M\}, C\} \mid M). \tag{2.31}$$

Esta condición requiere que M no solo contenga toda la información que f_i tiene en relación a C, sino que además, M debe contener toda la información que f_i tiene con las otras características de $F \backslash \{f_i, M\}$. Se ha demostrado que las características fuertemente relevantes no tienen un *Markov Blanket* [8].

La condición de *Markov Blanket* dada por la ecuación (2.31) puede ser reescrita en términos de la teoría de la información como [14]:

$$I(f_i; \{C, \neg\{f_i, M\} \mid M) = 0. \tag{2.32}$$

Una medida alternativa de redundancia es la correlación total o correlación multivariable [100]. Dado un conjunto de características $F = \{f_1, ..., f_m\}$, la correlación total se define como:

$$TC(f_1, ..., f_m) = \sum_{i=1}^{m} H(f_i) - H(f_1, ..., f_m). \tag{2.33}$$

La correlación total mide la información común (redundancia) en cualquier subconjunto de características de F. Si se desea medir la redundancia entre la característica f_i y cualquier subconjunto $S \subseteq \neg f_i$, entonces podemos utilizar la correlación total como:

$$TC(f_i, S) = H(f_i) + H(S) - H(f_i, S), \tag{2.34}$$

sin embargo, esto corresponde a la definición clásica de información mutua, es decir, $TC(f_i; S) = I(f_i; S)$.

2.2.3.3. Sinergia o complementariedad

El concepto de sinergia o complementariedad ha sido recientemente utilizado en el contexto de selección de características [14, 19, 84–92]. En términos generales, la sinergia busca detectar aquellas características que, actuando conjuntamente, generan mayor información de C con respecto a la suma de las informaciones que cada una de ellas por separado tienen de

C. Distintas áreas de investigación han utilizado este concepto, entre las cuales podemos mencionar: sistemas físicos [101–103], sistemas biológicos [14,86,104,105], neurociencia [18,22,87], transmisión de datos/codificación [106] , minería de datos [85], entre otros.

En el contexto de selección de características, el concepto de sinergia no tiene una definición clara, ya que resulta difícil medir directamente el aporte "sinérgico" que presenta un grupo de características. Para entender este concepto consideremos la descomposición de la información mutua de dos variables aleatorias $\{x_1, x_2\}$ y la clase C como:

$$I(\{x_1, x_2\}; C) = I(x_1; C) + I(x_2; C) + II(x_1; x_2; C). \tag{2.35}$$

El tercer término del lado derecho de la ecuación (2.35), es la información de interacción ($II(x_1; x_2; C)$). Esta cumple un rol fundamental para indicar la interacción de las dos variables y la clase. Cuando la interacción toma un valor negativo, la información que entrega el conjunto de dos variables será menor que la suma de la información que cada una de las variables tiene de C, lo que bajo el contexto de selección de características se relaciona al concepto de redundancia. Además en la literatura, la noción de redundancia se representa por el término $I(x_1; x_2)$, el cual puede ser derivado de $II(x_1; x_2; C)$ de acuerdo a la ecuación (2.10). Siguiendo este razonamiento, un valor positivo de interacción significa que la combinación de las características generan "nueva información" de C que no está en la información que cada una de las características tiene por separado de C. Esta nueva información se denomina sinergia.

A continuación se presentan algunas técnicas basadas en teoría de la información que han intentado explicar y cuantificar la sinergia, donde algunas de ellas sutilmente han vinculado su origen con el concepto de redundancia. Al final de esta revisión se presenta una descomposición no negativa de la información mutua, la cual fundamenta claramente el vínculo existente entre los conceptos de redundancia y sinergia.

Información de interacción (II) Esta métrica fue propuesta por McGill [38] para medir la información mutua entre tres o más variables. Para el caso de tres variables $\{x_1, x_2, x_3\}$, la información de interacción se define como:

$$II(x_1; x_2; x_3) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} p(x_1^i, x_2^j, x_3^k) \cdot \log \left(\frac{p(x_1^i, x_2^j, x_3^k) \cdot p(x_1^i) \cdot p(x_2^j) \cdot p(x_3^k)}{p(x_1^i, x_2^j) \cdot p(x_1^i, x_3^k) \cdot p(x_2^j, x_3^k)} \right), \tag{2.36}$$

donde $p(\bullet)$ corresponde a la probabilidad de masa, $x_1^i = x_1(i)$, $x_2^j = x_2(j)$, $x_3^k = x_3(k)$, y n_i, n_j, n_k corresponden al número de muestras de x_1, x_2, x_3, respectivamente. En la subsección 2,1,3 se presentó más detalles de esta métrica.

La información de interacción ha sido ampliamente utilizada en la literatura y a menudo ha sido llamada como sinergia [10, 22, 86, 87, 104], índice de redundancia-sinergia [91, 107] o co-información [14, 103, 108].

Correlación total (TC) Esta métrica fue Introducida por Watanabe [100], y busca cuantificar la redundancia o dependencia entre un conjunto de variables aleatorias. La correlación total se define como la divergencia de Kullback-Liebler [109, 110] entre la distribución conjunta y el modelo independiente de las variables aleatorias analizadas. Para el caso de tres variables $\{x_1, x_2, x_3\}$, la correlación total se define como:

$$TC(x_1, x_2, x_3) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} p(x_1^i, x_2^j, x_3^k) \cdot \log \left(\frac{p(x_1^i, x_2^j, x_3^k)}{p(x_1^i) \cdot p(x_2^j) \cdot p(x_3^k)} \right). \tag{2.37}$$

La correlación total también puede ser escrita en términos de entropías (ecuación (2.38)) o en términos de información mutua (ecuación (2.39)). Para el conjunto X de m variables, se tiene:

$$TC(X) = \left(\sum_{x_i \in X} H(x_i) \right) - H(X), \tag{2.38}$$

$$= I(x_1; x_2) + I(\{x_1, x_2\}; x_3) + \cdots + I(\{x_1, ..., x_{m-1}\}; x_m). \tag{2.39}$$

Correlación total dual (DTC) Introducida por Han [111, 112], esta métrica posee una estructura semejante a la correlación total y se define para el caso de tres variables $\{x_1, x_2, x_3\}$ como

$$DTC(X) = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} \sum_{k=1}^{n_k} p(x_1^i, x_2^j, x_3^k) \cdot \log \left(\frac{p(x_1^i | x_2^j, x_3^k) \cdot p(x_2^j | x_1^i, x_3^k) \cdot p(x_3^k | x_1^i, x_2^j)}{p(x_1^i, x_2^j, x_3^k)} \right). \tag{2.40}$$

La correlación total dual calcula el excedente de entropía que existe en X por sobre la suma de las entropías de cada variable de X condicionadas al resto de variables en X [113]. En términos de entropía, la correlación total dual se define como:

$$DTC(X) = \left(\sum_{x_i \in XS} H(X \backslash x_i) \right) - (m-1) \cdot H(X) \tag{2.41}$$

$$= H(X) - \sum_{x_i \in X} H(x_i | X \backslash x_i), \tag{2.42}$$

siendo m el número de variables en X. La correlación total dual a veces ha sido referida como exceso de entropía [114] o información vinculante [113].

Sinergia de Varadan (VS) Esta medida fue introducida por Varadan [115] como sinergia. Formalmente esta métrica se define como:

$$VS(X;C) = I(X;C) - \max_{i} \sum_{S_i} I(S_i;C), \tag{2.43}$$

donde X es un conjunto de variables aleatorias, C es el vector de clase y S_i corresponde a una posible partición del conjunto X, a excepción de la partición $S = X$. El máximo en la ecuación (2.43) busca la partición de X que entregue la máxima suma de información mutua entre los subconjuntos de X y el vector de clase C. Para el caso de tres variables $\{x_1, x_2, x_3\}$, la sinergia de Varadam está dada por

$$VS(X;C) = I(X;C) - max \begin{cases} I\left(x_1;C\right) + I\left(\{x_2,x_3\};C\right) \\ I\left(x_2;C\right) + I\left(\{x_1,x_3\};C\right) \\ I\left(x_3;C\right) + I\left(\{x_1,x_2\};C\right) \\ I\left(x_1;C\right) + I\left(x_2;C\right) + I\left(x_3;C\right) \end{cases}. \tag{2.44}$$

Descomposición de información parcial (PID) Esta descomposición de información fue desarrollada por Williams y Beer [116] y marca el inicio de varias investigaciones recientes [84, 90]. La novedad de este trabajo radica en que la información común entre un conjunto de variables aleatorias X y la variable de salida C se puede descomponer en una suma de términos que no se traslapan. De esta forma, esta métrica permite medir la interacción de una característica o subconjuntos del características de X con la clase en forma independiente, a diferencia de lo que hacen otras métricas que consideran la información de todo el conjunto X. Además, permite identificar claramente los términos de redundancia, relevancia única, sinergia y mezclas de interacciones de estos términos. Otra de las ventajas de esta métrica es que los resultados entregados son positivos, a diferencias de las medidas de interacción mencionadas anteriormente como la información de interacción, sinergia de Varadan, entre otras. Además PID considera a la sinergia y la redundancia como cantidades de información conceptualmente diferentes que pueden ser medidas simultáneamente en una interacción, a diferencia de la información de interacción que entrega un solo valor de estos dos conceptos. El único inconveniente es que los tipos de interacción que se generan entre subconjuntos de variables y las clase crecen exponencialmente a medida que aumenta el número de características, y algunas de las interacciones generadas no tienen una interpretación clara.

Para entender la propuesta de Williams y Beer consideremos el caso más simple donde el conjunto de variables está compuesto por dos elementos, es decir, $X = \{x_1, x_2\}$. De acuerdo a este caso, la información mutua entre X y la clase C se puede descomponer en una suma de términos de información relevante, los cuales son: la información única que cada variable de X provee de C, la información redundante que posee el conjunto X, y la información sinérgica que provee en conjunto X. Escribiendo en ecuaciones lo mencionado anteriormente se tiene:

$$\begin{aligned} I(\{x_1,x_2\};C) \equiv & \acute{U}nica(C;x_1) + \acute{U}nica(C;x_2) + \\ & Redundancia(C;\{x_1,x_2\}) + Sinergia(C;\{x_1,x_2\}) \end{aligned} \tag{2.45}$$

$$I(x_1; C) \equiv \acute{U}nica(C; x_1) + Redundancia(C; \{x_1, x_2\}) \tag{2.46}$$

$$I(x_2; C) \equiv \acute{U}nica(C; x_2) + Redundancia(C; \{x_1, x_2\}) \tag{2.47}$$

Como se observa en las ecuaciones (2.45), (2.46) y (2.47) se tienen más términos (4) que ecuaciones (3), por lo que no se puede obtener una solución única del valor de cada término. Williams y Beer proponen que el término de redundancia es igual a una nueva expresión de información propuesta por ellos mismos llamada la función de mínima información. Ésta se obtiene al comparar la cantidad de información que cada variable de X tiene en cada estado de C. Formalmente la función de mínima información se define para el caso de dos variables como:

$$I_{min}(\{x_1, x_2\}; C) = \sum_{c \in C} p(C = c) \min_i I(x_i; C = c). \tag{2.48}$$

Es conveniente destacar que el concepto de redundancia usado por Williams y Beer en la ecuación (2.48) es distinto a la idea de redundancia en selección de características, ya que esta última considera la información común entre todas las variables de X. La mínima información de Williams y Beer es el promedio ponderado de la información mínima que se obtiene de cualquiera de las variables de X en cada una de las clases del vector C.

La descomposición parcial de información permite explicar el valor negativo de la información de interacción. Esto se puede observar al expresar la información de interacción en términos de información mutua (ecuación (2.11)) y luego expresar los términos de información mutua en función de la descomposición parcial de información dada en las ecuaciones (2.45), (2.46) y (2.47).

$$\begin{aligned} II(x_1; x_2; C) &= I(\{x_1, x_2\}; C) - I(x_1; C) - I(x_2; C) \\ &= Sinergia(C; \{x_1, x_2\}) - Redundancia(C; \{x_1, x_2\}). \end{aligned} \tag{2.49}$$

De esta forma la descomposición de información encuentra que un valor negativo en la información de interacción implica que la interacción de redundancia entre las variables es mayor a la interacción de sinergia.

Con el fin de simplificar la notación de las interacciones entre variables de los posibles subconjuntos de variables, Williams y Beer proponen la notación PI. Utilizando esta notación, los términos obtenidos para el caso analizado de dos variables son: la sinergia o complementariedad que se expresa como $\Pi_R(C; \{x_1, x_2\})$ o PID sinergia, la redundancia que se expresa como $\Pi_R(C; \{x_1\}\{x_2\})$ o PID redundancia, y las informaciones únicas de cada variable que se expresan como $\Pi_R(C; \{x_1\})$ y $\Pi_R(C; \{x_2\})$ o PID información única. Gráficamente, los términos de la descomposición de información para el caso de 2 variables se muestran en la Figura 2.3.

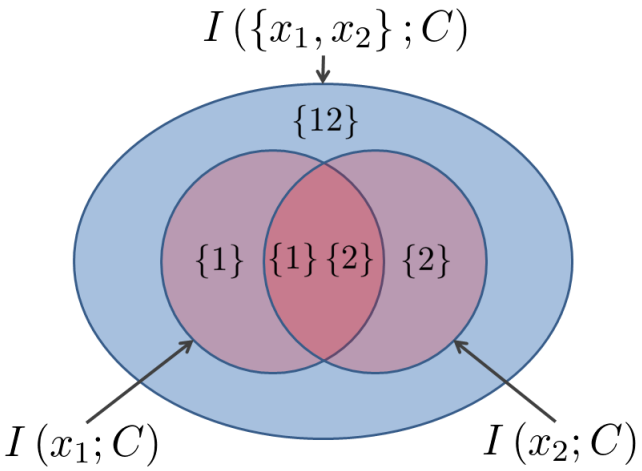


Figura 2.3: Diagrama de información parcial mostrando las relaciones entre complementariedad ({12}), redundancia ({1}{2}) y relevancia ({1} *y* {2}), asumiendo que la multiinformación entre f_i, S y C es positiva.

Si se realiza la descomposición parcial de información a un conjunto de tres variables más la clase la expansión tiene 18 términos. La Figura 2.4 muestra un diagrama con todas las posibles interacciones de las tres variables y la clase.

La propuesta de Williams y Beer no fue desarrollada en el contexto de selección de características por lo que su medida de redundancia mínima no está completamente relacionada con el concepto de redundancia utilizado en selección de características.

2.2.4. Subconjunto óptimo de características

En la literatura existen diferentes definiciones del subconjunto óptimo de características, S_{opt}, así como las estrategias de búsqueda utilizadas. Según [117], en la práctica el problema de selección de características debe incluir un clasificador o un conjunto de clasificadores y una métrica de rendimiento. El subconjunto óptimo de características se define como aquel subconjunto de mínima cardinalidad que maximiza la métrica de desempeño. Yu y Liu [8] definen el subconjunto óptimo de características como aquel subconjunto compuesto por todas las características fuertemente relevantes y débilmente relevantes pero no redundantes.

Las definiciones del subconjunto óptimo de características desde el punto de vista de los métodos de filtro, en particular de los métodos de selección de características basados en información mutua, poseen la noción de independencia condicional, que permite definir el subconjunto suficiente de características como sigue [6, 118]:

Definición 2.2.2 *$S \subseteq F$ es un subconjunto suficiente de características ssi:*

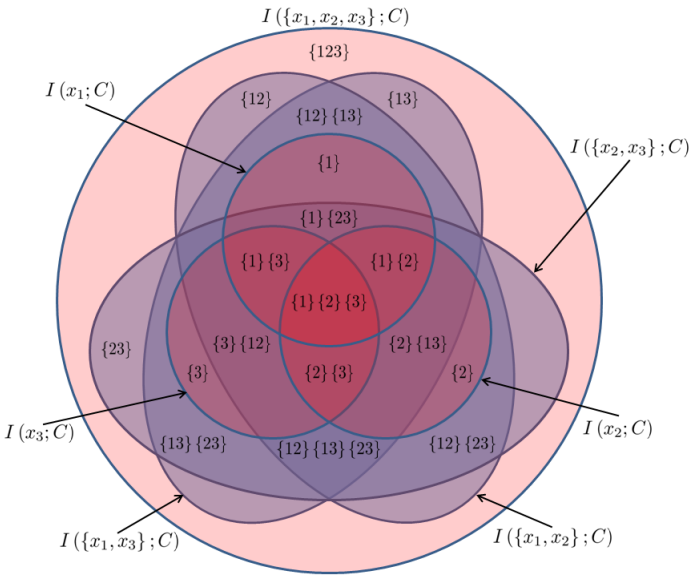


Figura 2.4: Diagrama de información parcial entre las variables $\{x_1, x_2, x_3\}$ y la clase C.

$$p(C|F) = p(C|S). \tag{2.50}$$

Esta definición implica que C y $\neg S$ son condicionalmente independientes, es decir, $\neg S$ no provee de información adicional con respecto a C en el contexto de S. Sin embargo, todavía se requiere una estrategia de búsqueda para seleccionar el subconjunto de características S. Una búsqueda exhaustiva usando este criterio, resulta impráctica debido a la maldición de la dimensionalidad.

En probabilidad, la medida de un subconjunto suficiente de características puede ser expresada como el valor esperado sobre $p(F)$ de la divergencia de Kullback-Liebler entre $p(C|F)$ y $p(C|S)$ [71]. Según Guyon *et al.* [118], esto puede ser expresado en términos de la información mutua como:

$$DMI(S) = I(F; C) - I(S; C). \tag{2.51}$$

Guyon *et al.* [118] propone resolver el siguiente problema de optimización:

$$\min_{S \subseteq F} |S| + \lambda \cdot DMI(S), \tag{2.52}$$

donde $\lambda > 0$ representa un multiplicador de Lagrange. Si S es un subconjunto suficiente de características, entonces $DMI(S) = 0$ y la ecuación(2.52) se reduce a $\min_{S \subseteq F} |S|$. Dado que $I(F; C)$ es constante, la ecuación (2.52) es equivalente a:

$$\min_{S \subseteq F} |S| - \lambda \cdot I(S; C). \tag{2.53}$$

El problema de selección de características corresponde a encontrar el mínimo subconjunto que maximice $I(S;C)$. Dado que el término $\min_{S\subseteq F}|S|$ es discreto, la optimización de (2.53) es difícil [118]. Tishby *et al.* [97] propone reemplazar el término $\min_{S\subseteq F}|S|$ por $I(F;S)$.

Una aproximación alternativa para seleccionar el subconjunto óptimo de características es utilizar el concepto de *Markov Blanket*. El *Markov Blanket*, M, de la variable deseada C corresponde al menor subconjunto de F tal que C sea independiente del resto de las características $F\backslash M$. Koller y Sahami [71] proponen usar los *Markov Blanket* como base para la eliminación de características. Ellos demostraron que las características eliminadas secuencialmente sobre la base de este criterio son innecesarias. Sin embargo, el tiempo requerido para encontrar el *Markov Blanket* crece exponencialmente con el tamaño de este conjunto cuando se consideran dependencias completas. Por lo tanto, muchos algoritmos usan aproximaciones de la definición de *Markov Blanket*, por ejemplo, encontrar el conjunto de k características que están fuertemente correlacionadas con una determinada característica [71]. Los algoritmos para encontrar *Markov Blanket* se han desarrollado para el caso de las distribuciones que son fidedignas a una Red Bayesiana [117, 119]. Sin embargo, estos algoritmos requieren que el subconjunto óptimo de características no contenga características fuertemente complementarias [120]. En la práctica, esto significa que los actuales algoritmos que buscan el *Markov Blanket* no pueden resolver problemas como la función lógica XOR.

Una advertencia importante es que ambos criterios de selección de características, tanto el subconjunto suficiente de características como *Markov Blanket*, están basados en la estimación de la distribución de probabilidad de C dados los datos. La estimación de la probabilidad a posteriori es un problema más difícil que la clasificación, por lo tanto este efecto puede hacer que algunas características contenidas en un subconjunto suficiente características o en el *Markov Blanket* de C sean incorrectamente seleccionadas o eliminadas [117, 118, 121].

2.2.4.1. Maldición de la dimensionalidad

El término de maldición de la dimensionalidad fue propuesto por Richard Bellman en 1961 [98]. En sus palabras, Bellman expresó este concepto como: " *En vista de todo lo que hemos dicho en los apartados anteriores, los muchos obstáculos que al parecer hemos superado, ¿qué arroja la sombra sobre nuestra celebración de la victoria? Es la maldición de la dimensionalidad, una maldición que ha plagado al científico desde los primeros días....*".

La maldición de la dimensionalidad representa el problema causado al aumentar dimensiones (características) extras al espacio. Es posible conocer el efecto de la maldición de la dimensionalidad al comparar un mismo fenómeno en dos espacios de distinta dimensionalidad. Considere dos espacios: uno con 2 dimensiones (2D) y el otro con 10 dimensiones ($10D$), donde ambos describen el mismo fenómeno. Suponiendo que existen 100 muestras distribuidas uniformemente sobre el rango 0 a 1. En el caso del espacio $2D$, si fijamos una característica a un valor fijo, tenemos 10 muestras que describen la otra característica espaciadas 0,1 unidades entre ellas. En el caso del espacio $10D$, si fijamos una característica a un valor fijo, la distancia entre las restantes muestras es de 0,63 unidades a cada eje restante [5]. Esto significa que el efecto del fenómeno no puede ser observado en el espacio $10D$; la

[5]Las distancias son calculadas por $\frac{1}{\sqrt[m]{100}}$, donde m denota el número de dimensiones.

cantidad de muestras necesarias para describir un fenómeno aumenta exponencialmente con el número de dimensiones. Por otra parte, las distancias euclidianas promedio entre dos vecinos más cercanos en estos espacios son de aproximadamente 0,14 y 2,00 unidades[6], lo cual genera duda si los efectos que constituyen el fenómeno observado podrían ser medidos de forma fiable en el espacio $10D$.

2.2.4.2. Relación entre la información mutua y el error de clasificación de Bayes.

Existen algunos interesantes resultados relacionados con la información mutua entre una variable aleatoria f y la variable de salida C, con el mínimo error obtenido de la maximización de la clasificación a posteriori (error de clasificación de Bayes) [35,122–126]. El error de Bayes está limitado superior e inferiormente de acuerdo a la siguiente expresión:

$$1 - \frac{I(f;C) + \log(2)}{\log(|C|)} \leq e_{bayes}(f) \leq \frac{1}{2}\left(H(C) - I(f;C)\right). \tag{2.54}$$

Interesantemente, la ecuación(2.54) muestra que ambos límites son minimizados cuando la información mutua, $I(f;C)$ es maximizada.

2.2.4.3. Estrategias de búsqueda

De acuerdo a Guyon *et al.* [118], un método de selección de características se compone de tres partes: (i) definición del criterio de evaluación, por ejemplo, criterio de relevancia para los métodos de filtro, (ii) evaluación del criterio de estimación, por ejemplo, selección de las características suficientes o *Markov Blanket*, y (iii) estrategia de búsqueda para la generación del subconjunto de características. En esta sección se abordará el punto (iii) en donde se presentan las principales estrategias de búsqueda usadas por los métodos de selección de características basados en la información mutua. Dado un conjunto de características de cardinalidad m, existen 2^m posibles subconjuntos, por lo tanto, una búsqueda exhaustiva es impráctica para bases de datos de alta dimensionalidad. Existen dos estrategias de búsqueda básicas: métodos óptimos y métodos sub-óptimos (métodos *greedy*) [5]. Las estrategias de búsqueda óptima incluyen búsqueda exhaustiva y métodos acelerados basados en la propiedades monotónicas de un criterio de selección de características, tales como *Branch and Bound* [17,83,95,127]. Sin embargo, los métodos óptimos son imposibles de realizar para bases de datos de alta dimensionalidad, por lo que deben usarse estrategias sub-óptimas.

Los métodos más populares de búsqueda sub-óptima son la selección secuencial hacia adelante (*Sequential Forward Selection:* SFS) [128] y la eliminación secuencial hacia atrás (*Sequential Backward Elimination:* SBE) [129]. La selección secuencial hacia adelante comienza con un conjunto vacío ($S = \emptyset$) y agrega a S una característica a la vez. Formalmente, la característica candidata f_i que se agrega a S se obtiene como:

[6] $\simeq \sqrt{m\left(\frac{1}{\sqrt[m]{100}}\right)^2}$

$$S = S \cup \{\arg\operatorname*{máx}_{f_i \in F \setminus S}(I(\{S, f_i\}; C))\}. \tag{2.55}$$

La eliminación secuencial hacia atrás es una búsqueda de características que comienza con el todo el conjunto de características ($S = F$) y elimina una característica a la vez. Formalmente la eliminación de la característica menos informativa (con respecto a la variable de salida C) se define como:

$$S = S \setminus \{\arg\operatorname*{mín}_{f_i \in S}(I(\{S \setminus f_i\}; C)\}. \tag{2.56}$$

Usualmente, la eliminación hacia atrás es computacionalmente más costosa que la selección hacia adelante. Sin embargo, la eliminación hacia atrás puede encontrar mejores subconjuntos de características, ya que la mayoría de los métodos de selección hacia adelante no toman en cuenta la relevancia de las características en el contexto de otras características que no han sido incluidas al subconjunto de las características seleccionadas (características complementarias o sinérgicas) [1]. Ambos tipos de metodologías de búsqueda sufren del efecto anidado, lo que significa que en la selección hacia adelante, una característica no puede ser eliminada una vez que ya ha sido agregada; y en el caso de la eliminación hacia atrás, una característica no puede ser reincorporada una vez que ha sido eliminada. En vez de agregar una única característica a la vez, algunas variantes de selección hacia adelante generalizadas agregan varias características a la vez, tomando en cuenta las relaciones estadísticas entre ellas [5]. Por otro lado, la eliminación hacia atrás generalizada, elimina varias características a la vez. Un mejoramiento de estos dos métodos de búsqueda puede ser obtenido combinando la selección hacia adelante y hacia atrás, evitando el efecto anidado. La estrategia *plus-l-take-away-r* [130] suma a S, l características y remueve las peores r características si $l > r$, o elimina r características y luego suma l características si $r < l$, sin embargo, el costo computacional aumenta.

2.3. Algoritmos de aprendizaje

A continuación se presentan brevemente los principios fundamentales de dos populares clasificadores: k vecinos más cercanos (kNN del inglés *k-Nearest-Neighbour*) y la máquina de soporte vectorial (SVM del inglés *Support Vector Machine*). Una revisión detallada de estos algoritmos puede ser encontrada en [5, 59, 79, 131–135].

2.3.1. k-Nearest-Neighbors (kNN)

En kNN, la clasificación de una muestra x se realiza por la mayoría de votos de la clase a la cual pertenecen los vecinos más cercanos de x. La clase del punto x es asignada a la clase

mayoritaria de los k vecinos de x. Formalmente, dado una salida binaria (dos clases), la regla del kNN [135] está dado por:

$$g(x) = \begin{cases} 1 & Si \sum_{r=1}^{m} \omega_r \delta(C=1) > \sum_{r=1}^{m} \omega_r \delta(C=0) \\ 0 & \text{Otro caso} \end{cases}, \tag{2.57}$$

donde $\delta(\cdot)$ es el indicador de la función y $\omega_r = \frac{1}{k}$ si x_r está entre los k vecinos más cercanos de x y $\omega_r = 0$ en otro caso. Generalmente k es un entero pequeño impar [59].

Para identificar a los vecinos, las muestras se representan por vectores en espacios multidimensionales, donde la distancia euclidiana es la métrica comúnmente usada para medir la distancia hacia los vecinos. La mejor selección de k depende de los datos. Si $k = 1$, entonces la muestra es simplemente clasificada a la clase que tiene su vecino más cercano. La elección de k mayor a 1, reduce los efectos del ruido, pero los límites entre las clases son menos diferenciados. En la práctica, se usa mucho la selección de $k = 3$ [59].

2.3.2. Support Vector Machine (SVM)

Consideremos el problema de clasificación binaria con clases $Y \in \{-1, +1\}$. El hiperplano $y = f(x, \theta) \in \mathbb{R}^m$ en el espacio multidimensional de entradas genera la siguiente función de predicción:

$$\hat{y} = \begin{cases} +1 & Si \quad f(x,\theta) > 0 \\ -1 & Si \quad f(x,\theta) < 0 \end{cases}, \tag{2.58}$$

donde el hiperplano es una combinación lineal de las entradas,

$$f(x, \theta) = \sum_{i=1}^{m} \theta_i x_i + b = \theta \cdot x + b. \tag{2.59}$$

Combinando las ecuaciones (2.58) y (2.59) tenemos la siguiente desigualdad:

$$y(\theta \cdot x + b) - 1 \geq 0. \tag{2.60}$$

Si los datos pertenecientes a cada una de las dos clases son linealmente separables, existe un infinito número de soluciones que separan correctamente los datos. La solución óptima de este problema es el hiperplano que maximiza la distancia entre dos planos paralelos que están lo más alejado posible uno del otro, pero que aún separan los datos. En el proceso de ajuste del SVM, se busca la máxima distancia entre los planos paralelos (también llamado margen) con el fin de obtener el mínimo error de generalización del clasificador. La distancia entre los dos hiperplanos es $2/\|\theta\|$. Como resultado, la maximización del margen es equivalente a la minimización de $\frac{1}{2}\|\theta\|^2$ sujeto a la restricción $y(\theta \cdot x + b) - 1 \geq 0$. El problema de optimización puede ser escrito como [79]:

$$L_p = \frac{1}{2}\|\theta\|^2 - \sum_{r=1}^{n} \lambda_r \left(y_r \left(\theta \cdot x_r + b \right) - 1 \right), \tag{2.61}$$

la cual se llama la forma primal del Lagrangiano, donde λ_r son los multiplicadores de Lagrange. Diferenciando con respecto a θ y b, igualando a cero, y reemplazando en la ecuación (2.61) se obtiene la siguiente ecuación (llamada forma dual del Lagrangiano) [79]

$$L_D = \sum_{r=1}^{n} \lambda_r - \frac{1}{2} \sum_{r=1}^{n} \sum_{s=1}^{n} \lambda_r \lambda_s y_r y_s (x_r \cdot x_s), \tag{2.62}$$

donde $\forall r$, $\lambda_r \geq 0$ y $\sum_{r=1}^{m} \lambda_r y_r = 0$. Los vectores de soporte son aquellos x_r cuyos λ_r son distintos de cero. La ecuación (2.62) se resuelve a través de programación cuadrática convexa.

Para realizar una clasificación no-lineal, es posible reemplazar el producto punto $(x_r \cdot x_s)$ de la ecuación (2.62) con una función de kernel no-lineal $K(x_r, x_s) = (\Phi(x_r), \Phi(x_s))$. Para el caso del kernel Gaussiano se tiene:

$$K(x_r, x_s) = e^{-\|x_r - x_s\|^2 / 2\sigma^2}, \tag{2.63}$$

donde σ es el ancho del kernel Gaussiano. Es conveniente destacar que la función kernel evita trabajar con el espacio de dimensión infinita de la transformación Φ. En vez de ello, el algoritmo internamente maximiza el margen de los hiperplanos en el espacio de dimensión infinita y el resultado de este cálculo es transformado al espacio original, donde esta reducción del tamaño del espacio genera en el espacio original una transformación no lineal de las entradas al clasificador [79].

Capítulo 3

Estado del arte de la selección de características utilizando información mutua

Durante los últimos 20 años se han propuesto variados criterios tipo filtro de selección de características basados en información mutua [3, 4, 6, 12, 34, 62]. En sus inicios, el desarrollo de estos criterios estuvo limitado por el reducido desarrollo computacional de la época para la estimación de la información mutua. Estos aspectos marcaron la pauta para la creación de criterios de búsqueda de características del tipo:

- criterios heurísticos creados con poca rigurosidad teórica y sólo basados en la detección de relevancia y/o redundancia de las características, y
- criterios no-exhaustivos de búsqueda guiada con el fin de evitar la evaluación de una combinatoria exponencial de posibles subconjuntos que aumenta con el número de características del problema. Siguiendo esta lógica, aparecieron las búsquedas de características en forma *greedy*.

A pesar de las desventajas de estos criterios, la mayoría de ellos han sido altamente demandados como etapa de preprocesamiento en *machine learning* y *data mining*, debido a su rapidez e independencia del clasificador [1, 2, 12].

El interés por desarrollar un marco teórico que explique el origen de los actuales criterios de selección de características, está motivado sobre todo para conocer las ventajas y desventajas de los actuales criterios encontrados en la literatura. Recientemente Brown *et al.* [34] presentó un trabajo basado en la Maximización de la Verosimilitud Condicional, donde deriva la mayoría de los criterios de selección de características basados en información mutua existentes en la literatura. Este trabajo divide los actuales criterios de selección de características en aproximaciones lineales y aproximaciones no-lineales, pero no logra explicar la aproximación no-lineal.

El objetivo de este capítulo es presentar una revisión de los actuales criterios de selección de características utilizando información mutua bajo la perspectiva del trabajo de Brown *et al.* [34]. Con esto, se entrega al lector una visión global para entender el origen de los criterios más utilizados en la literatura.

3.1. Marco unificado de teoría de la información para la selección de características

La propuesta de Brown *et al.* [34] asume una función de probabilidad p que genera n muestras *i.i.d.* Esta función de probabilidad p se puede suponer como un proceso $p : F \to C$, el cual se desea modelar. Bajo el contexto de selección de características, el modelado del proceso p se compone de dos etapas: (i) identificar las características del proceso que juegan un rol funcional (características relevantes), y (ii) utilizar estas características relevantes para predecir la variable de salida C. El trabajo de Brown *et al.* se concentra sólo en la primera de estas dos etapas.

Para aproximar el proceso p se usa un modelo predictivo hipotético q con dos capas de parámetros: (i) θ representa las características seleccionadas, y (ii) τ representa los parámetros usados para predecir C. El parámetro θ se define como un vector binario de tamaño m (número de características de F) donde un 1 en su j-ésima posición indica la selección de la j-ésima característica de F, mientras que un 0 en la j-ésima posición, indica que la j-ésima característica de F es descartada. Suponiendo que el proceso p puede ser descrito completamente por un subconjunto óptimo θ^* de características de F, entonces se tiene que $p(C|F) = p(C|F_{\theta^*})$. Sean los datos *i.i.d* generados por p como pares ordenados $\mathcal{D} = \{(F(i), C(i)) \, ; i = 1, ..., n\}$, donde $F(i)$ y $C(i)$ representan la i-ésima muestra de F y C respectivamente. La aproximación de p directamente desde los datos se realiza mediante la función de verosimilitud condicional de las clases dado los parámetros $\{\theta, \tau\}$, definida como:

$$\mathcal{L}\left(\theta, \tau | \mathcal{D}\right) = \prod_{i=1}^{n} q\left(C(i)|F_\theta(i), \tau\right). \tag{3.1}$$

Escalando la ecuación (3.1) por el logaritmo se obtiene la función log-verosimilitud condicional definida como:

$$\ell = \frac{1}{n} \sum_{i=1}^{n} \log\left(q\left(C(i)|F_\theta(i), \tau\right)\right). \tag{3.2}$$

En el contexto de clasificación, la maximización de la verosimilitud condicional equivale a minimizar la divergencia de Kullback-Leibler entre la función de probabilidad a posteriori de la clase verdadera y estimada [59]. Para establecer este nexo, Brown *et al.* introduce en la ecuación (3.2) la distribución de probabilidad verdadera a posteriori de la clase dado las características seleccionadas de F, es decir, $p\left(C|F_\theta\right)$, obteniéndose:

$$\ell = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{q\left(C(i)|F_\theta(i), \tau\right)}{p\left(C(i)|F_\theta(i)\right)}\right) + \frac{1}{n} \sum_{i=1}^{n} \log\left(p\left(C(i)|F_\theta(i)\right)\right). \tag{3.3}$$

El segundo término de la ecuación (3.3) se puede expandir introduciendo la probabilidad $p\left(C|F\right)$, que corresponde a la distribución de probabilidad verdadera a posteriori de la clase dado todas las características de F de la siguiente forma:

$$\ell = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{q\left(C(i)|F_\theta(i),\tau\right)}{p\left(C(i)|F_\theta(i)\right)}\right) + \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p\left(C(i)|F_\theta(i)\right)}{p(C(i)|F(i))}\right) + \frac{1}{n}\sum_{i=1}^{n}\log\left(p(C(i)|F(i))\right). \tag{3.4}$$

Por conveniencia se considera el negativo de la ecuación (3.4) y se aplica el operador esperanza $\mathbb{E}\left\{\bullet\right\}$, obteniéndose:

$$\lim_{n\to\infty} -\ell = \mathbb{E}\left\{\log\left(\frac{p\left(C(i)|F_\theta(i)\right)}{q\left(C(i)|F_\theta(i),\tau\right)}\right)\right\} + I(F\backslash F_\theta; C|F_\theta) + H(C|F). \tag{3.5}$$

Analizando cada uno de los términos de la ecuación (3.5) se obtienen interesantes propiedades que sirven para comprender los fundamentos del proceso de selección de características.

- El primer término de (3.5) corresponde a la razón de verosimilitud entre la distribución de probabilidad a posteriori de la clase verdadera (p) y la estimada (q). El valor de este término depende de cuán bien la distribución estimada se aproxime a la distribución verdadera. Cuando $\theta = \theta^*$ (o un superconjunto de θ^*), puede ser considerado como la divergencia de Kullback-Leibler $p||q$.
- El tercer término corresponde la entropía condicional de la clase con respecto a todas las características de F. Este cuantifica la incertidumbre remanente del vector de clase cuando se conoce la información que entregan todas las características de F. Este término tiene un valor fijo y corresponde al límite inferior alcanzado en un clasificador bayesiano [35, 122–126].
- Finalmente, el segundo término es de especial importancia en el desarrollo del marco propuesto por Brown *et al.*, ya que corresponde a la definición del conjunto óptimo de características introducido por Koller y Sahami [71]. El tamaño de este término depende solamente de la selección de las características, disminuyendo su valor en la medida que el conjunto F_θ posea mayor información de C, llegando a cero en el caso que el remanente de características de F ($F\backslash F_\theta$) no contenga información adicional de C bajo el contexto de F_θ.

De acuerdo a lo establecido por Brown *et al.*, el funcional de un clasificador puede dividirse en dos procesos de optimización: (i) encontrar las características relevantes, y (ii) construir el clasificador utilizando estas características. De esta forma, el segundo término de la ecuación (3.5) puede optimizarse independientemente del primer término de ésta. Este supuesto permite establecer la siguiente igualdad:

$$\arg\max_\theta \mathcal{L}\left(\theta|\mathcal{D}\right) = \arg\min_\theta I(F\backslash F_\theta; C|F_\theta). \tag{3.6}$$

Para evitar los múltiples óptimos globales que pueden existir, se introduce la restricción de mínimo tamaño del subconjunto, redefiniendo el problema de selección de características como sigue:

$$\theta^* = \arg\min_{\theta'} \left\{ |\theta'| : \theta' = \arg\min_{\theta} I(F\backslash F_\theta; C|F_\theta) \right\}. \tag{3.7}$$

La ecuación (3.7) establece un funcional para la selección del subconjunto óptimo de características $S_{opt} = F_{\theta^*}$, sin embargo, tal como mencionó en la subsección 2.2.4, la detección del conjunto óptimo es un problema NP. Para evitar este problema, se realiza una aproximación *greedy* de la ecuación (3.7). Para una búsqueda *greedy* hacia adelante, el vector θ comienza con todos sus valores en cero ($S = \emptyset$) e iterativamente va analizando y agregando nuevas características de $F\backslash S$ que entregan la mayor información de C. Por lo tanto F_θ representa el conjunto de características previamente seleccionadas (S) y la característica candidata (f_i), es decir, $F_\theta = \{S, f_i\}$. Utilizando este hecho, la información mutua condicional de la ecuación (3.7) se puede descomponer como:

$$\begin{aligned} I(F\backslash\{S, f_i\}; C|\{S, f_i\}) &= I(\{F\backslash\{S, f_i\}, f_i\}; C|S) - I(f_i; C|S) \\ &= I(F\backslash S; C|S) - I(f_i; C|S), \end{aligned} \tag{3.8}$$

donde la minimización de $I(F\backslash\{S, f_i\}; C|\{S, f_i\})$ es equivalente a la maximización de $I(f_i; C|S)$ ($\forall f_i \in F\backslash S \Rightarrow I(F\backslash S; C|S) = cte$). La información mutua condicional $I(f_i; C|S)$ corresponde al criterio de relevancia o criterio hacia adelante (Forw) [6, 12, 64] el cual se describe en la siguiente sección.

Por otro lado, Brown *et al.* también deriva desde la ecuación (3.7) la búsqueda *greedy* hacia atrás, al considerar que el vector θ comienza con todos sus valores en uno ($S = F$) e iterativamente va analizando y descartando aquellas características de S que entregan la menor información de C. Por lo tanto F_θ representa el conjunto de características que aún no han sido eliminadas (S) sin considerar la característica candidata (f_i), es decir, $F_\theta = \{S\backslash f_i\}$. Luego, la información mutua condicional de la ecuación (3.7) se puede descomponer como:

$$\begin{aligned} I(\{F\backslash\{S\backslash f_i\}\}; C|\{S\backslash f_i\}) &= I(\{f_i, F\backslash S\}; C|\{S\backslash f_i\}) \\ &= I(f_i; C|S\backslash f_i) + I(F\backslash S; C|S). \end{aligned} \tag{3.9}$$

Luego la minimización de $I(\{F\backslash\{S\backslash f_i\}\}; C|\{S\backslash f_i\})$ es equivalente a la minimización de $I(f_i; C|S\backslash f_i)$ ($\forall f_i \in S \Rightarrow I(F\backslash S; C|S) = cte$).

Hasta el momento se ha derivado un simple criterio *greedy* de optimización que permite evaluar la información que una característica tiene de C. El criterio encontrado en la ecuación (3.8) corresponde a:

$$\mathcal{J}_{CMI}(f_i) = I(f_i; C|S), \tag{3.10}$$

donde el subíndice CMI indica la información mutua condicional (del inglés, *Conditional Mutual Information*). Para Brown *et al.*, este criterio resulta de gran importancia ya que a partir de éste se puede derivar la gran mayoría de los criterios desarrollados en la literatura en los últimos 20 años. El análisis comienza expandiendo la ecuación (3.10) como

$$\mathcal{J}_{CMI}(f_i) = I(f_i;C) - I(f_i;S) + I(f_i;S|C), \tag{3.11}$$

donde se obtienen tres términos de gran familiaridad en los funcionales de selección de características. Estos son la "relevancia" ($I(f_i;C)$), la "redundancia" ($I(f_i;S)$) y la "redundancia condicional" ($I(f_i;S|C)$). A pesar de que conceptualmente la definición de características relevantes o redundantes es fundamental para encontrar el subconjunto óptimo de características, la dificultad en la estimación de funciones de distribución de probabilidad (o probabilidad de masa) en espacios de dimensión mayor a tres, ha hecho recurrir a fuertes supuestos de independencia estadística entre características tales como:

$$p(f_i|S) = \prod_{s_j \in S} p(f_i|s_j) \tag{3.12}$$

$$p(f_i|\{S,C\}) = \prod_{s_j \in S} p(f_i|\{s_j,C\}). \tag{3.13}$$

Utilizando estos supuestos, es posible aproximar la ecuación (3.11) como combinaciones lineales de informaciones mutuas:

$$\mathcal{J}_{CMI}(f_i) \approx \alpha I(f_i;C) - \beta \sum_{s_j \in S} I(f_i;s_j) + \gamma \sum_{s_j \in S} I(f_i;s_j|C). \tag{3.14}$$

Asignando valores a α, β y γ, se puede encontrar una gran cantidad de criterios de selección de características existentes en la literatura. Por otro lado, en la literatura también se encuentran criterios que utilizan operadores no-lineales, tales como máximos o mínimos que no pueden derivarse de la ecuación (3.14). La Tabla 3.1 muestra un resumen de algunos criterios existentes en la literatura.

A continuación se presenta una explicación en detalle de los criterios más importantes desarrollados en la literatura.

3.2. Filtros de selección de características utilizando información mutua

La teoría de la información [35,36] ofrece un sólido marco teórico para diferentes problemas en máquinas de aprendizaje. En el caso de selección de características, algunos métodos de selección de características tipo filtros utilizan la información mutua como métrica para generar el ordenamiento de éstas. La información mutua es considerada como un adecuado criterio para la selección de características [1] principalmente por tres razones: (i) La información mutua es una medida de reducción de la incertidumbre del vector de salida o clase debido al conocimiento de las características [146]. (ii) La maximización de la información mutua entre las características y el vector de clases minimiza el límite inferior del error de

Tabla 3.1: Varios criterios de selección de características basados en información mutua obtenidos desde la literatura.

Sigla	Nombre completo	Autor	Criterio (maximización)
MIM	*Mutual Information Maximisation*	Lewis (1992) [24]	$I(f_i;C)$
MIFS	*Mutual Information Feature Selection*	Battiti (1994) [12]	$I(f_i;C) - \beta \sum_{s_j \in S} I(f_i;s_j)$
JMI	*Joint Mutual Information*	Yang y Moody (1999) [136]	$\sum_{s_j \in S} I(\{f_i, s_j\};C)$
MIFS-U	*MIFS-Uniform*	Kwak y Choi (2002) [137]	$I(f_i;C) - \beta \sum_{s_j \in S} \frac{H(s_j;C)}{H(s_j)} I(f_i;s_j)$
IF	*Informative Fragments*	Vidal-Naquet y Ullman (2003) [138]	$\min_{s_j \in S} [I(\{f_i, s_j\};C) - I(f_i;C)]$
CMIM	*Conditional Mutual Information Maximization*	Fleuret (2004) [3]	$\min_{s_j \in S} [I(f_i;C\|s_j)]$
MRMR	*Max-Relevance Min-Redundance*	Peng et al. (2005) [4]	$I(f_i;C) - \frac{1}{\|S\|} \sum_{s_j \in S} I(f_i;s_j)$
ICAP	*Interaction Capping*	Jakulin (2005) [139]	$I(f_i;C) + \sum_{s_j \in S} \min(0, II(f_i;s_j;C))$
CIFE	*Conditional Infomax Feature Extraction*	Lin y Tang (2006) [140]	$I(f_i;C) - \sum_{s_j \in S} I(f_i;s_p) + \sum_{s_j \in S} I(f_i;s_j\|C)$
DISR	*Doble Input Symetrical Relevance*	Meyer y Bontempi (2006) [141]	$\sum_{s_j \in S} \frac{I(\{f_i, s_j\};C)}{H(\{f_i, s_j, C\})}$
IGFS	*Interaction Gain Feature Selection*	El Akadi et al. (2008) [142]	$I(f_i;C) + \frac{1}{\|S\|} \sum_{s_j \in S} II(f_i;s_j;C)$
SOA	*Second Order Approximation*	Guo y Nixon (2009) [143]	$I(f_i;C) - \sum_{s_j \in S} I(f_i;s_p) + \sum_{s_j \in S} I(f_i;s_j\|C)$
NMIFS	*Normalized MIFS*	Estévez et al. (2009) [144]	$I(f_i;C) - \frac{1}{\|S\|} \sum_{s_j \in S} \frac{I(f_i;s_j)}{\min[H(f_i), H(s_j)]}$
CMIM-2	*CMIM-2*	Vergara y Estévez (2010) [145]	$\frac{1}{\|S\|} \sum_{s_j \in S} I(f_i;C\|s_j)$
CMIFS	*Conditional MIFS*	Cheng et al. (2011) [91]	$I(f_i;C\|s_1) + [I(f_i;s_p\|C) + I(f_i;s_p\|s_1)]$, donde s_1 y s_p son la primera y última característica incluida en S.

clasificación de Bayes [35, 122, 123, 126]. (iii) La información mutua permite capturar relaciones no-lineales entre las características, permitiendo definir y cuantificar conceptos tales como relevancia [6, 7], redundancia [8, 9], y complementariedad (o sinergia) [7, 10, 11]

Algunos criterios propuestos para la selección de características computan la información mutua de cada una de las características del problema con el vector de clase C. La razón principal para utilizar este enfoque *greedy* es evitar la estimación de información mutua es espacios de alta dimensionalidad, sin embargo, se pierde la capacidad de detectar características que interactúen conjuntamente para predecir C. Con el objetivo en detectar estas interacciones entre características en favor de obtener mayor información de C, se han generado una gran cantidad de criterios que calculan interacciones entre pares de características [4,11,12,89], los cuales utilizan una búsqueda *greedy* hacia adelante. Un pseudo código de la búsqueda *greedy* se presenta en el Algoritmo 3.1

Como se mencionó en la sección 2.2.4, el objetivo de todos los criterios de selección de características basados en información mutua es:

> Dado un conjunto de datos F de n muestras, m características, una variable de salida C, y un entero $d < n$, se desea encontrar un subconjunto de características $S \subset F$ de tamaño d que maximice la información mutua $I(S; C)$.

Matemáticamente, encontrar el subconjunto de características S de tamaño d consiste en el problema de optimización combinatorio el cual se define como:

$$S_{opt} = \arg \max_{S \subset F : |S| = d} I(S; C). \tag{3.15}$$

A continuación se presentan detalladamente los criterios de selección de características basados en información mutua más utilizados en la literatura.

Algoritmo 3.1 Pseudo-código de selección de características basado en información mutua utilizando el algoritmo *greedy* con búsqueda hacia adelante.

Entrada: Conjunto de datos F, variable de clase C, número de características a seleccionar d ($d \leq |F|$), y criterio de selección de características $\mathcal{J}(f_i, F, S)$.

Salida: Conjunto S de tamaño d que contiene las características seleccionadas.

1: *Inicialización*: Establecer $S \leftarrow \emptyset$.
2: *Cómputo de la información mutua* entre cada una de las característica de F y C. $\forall f_i \in F$, computar $I(f_i; C)$.
3: *Selección de la primera característica*: Encontrar la característica f_i que maximice $I(f_i; C)$. Establecer $F \leftarrow F \backslash f_i$, $S \leftarrow S \cup f_i$.
4: **Repetir**
5: *Selección de la próxima característica*: Encontrar la característica $f_i \in F$ según el criterio $\mathcal{J}(f_i, F, S)$. Establecer $F \leftarrow F \backslash f_i$, $S \leftarrow S \cup f_i$.
6: **Hasta que** $|S| = d$

3.2.1. RANK

Este criterio *greedy* realiza un ordenamiento de las características de acuerdo a la información que individualmente cada característica tiene de la variable de salida C. Esto significa que dado m características de entrada, este criterio calcula la información mutua $I(f_i; C)$, $i = 1, ..., m$, y posteriormente ordena las características de mayor a menor valor de información mutua [147]. El funcional de este criterio es:

$$\mathcal{J}_{rank}(f_i) = I(f_i; C). \tag{3.16}$$

La principal ventaja de este criterio es su bajo costo computacional, ya que solo requiere m estimaciones bivariables (característica y variable clase). Por otro lado, el principal inconveniente de este criterio es que no considera la interacción que puede existir entre dos o más características. El problema más evidente es la interacción de redundancia, ya que si se tiene dos o más características que tienen la misma información de C, todas ellas obtendrán el mismo valor de información mutua y por lo tanto serán asignadas en posiciones contiguas en el ordenamiento de este criterio, lo cual va en desmedro de la obtención del conjunto mínimo de características. Por la razón mencionada anteriormente, este criterio no permite determinar características complementarias.

3.2.2. MIFS

Battiti [12] propuso el criterio de selección de características basado en información mutua (MIFS, del inglés *Mutual Information-based Feature Selection*), el cual se define como:

$$\mathcal{J}_{MIFS}(f_i) = I(f_i; C) - \beta \sum_{s_j \in S} I(f_i; s_j). \tag{3.17}$$

El criterio MIFS selecciona las d características más relevantes desde un conjunto inicial de m características en forma secuencial según el Algoritmo 3.1. Este criterio busca la característica f_i que presente la mayor información de la clase C ($I(f_i; C)$) pero que además tenga la mínima redundancia de información con las características previamente seleccionadas en el subconjunto S. La selección es regulada por el término proporcional $\beta I(f_i; s_j)$, el cual mide la redundancia entre la característica candidata f_i y cada una de las características del subconjunto S de características seleccionadas previamente. El parámetro β es muy sensible en la selección de las características, por lo cual su ajuste implica un nuevo problema de optimización.

3.2.3. mRMR

El criterio de Mínima Redundancia - Máxima Relevancia (mRMR) [4,148] fue desarrollado en combinación con la estrategia de búsqueda hacia adelante presentada en el Algoritmo 3.1.

Este criterio es equivalente al criterio MIFS, pero con $\beta = {}^{1}/_{|S|}$, donde $|S|$ representa la cardinalidad del conjunto S. Este criterio se define como:

$$\mathcal{J}_{mRMR}(f_i) = I(f_i; C) - \frac{1}{|S|} \sum_{s_j \in S} I(f_i; s_j) \tag{3.18}$$

En cada paso, este método selecciona la característica f_i con el mayor *trade-off* entre relevancia y redundancia.

3.2.4. NMIFS

Estévez *et. al.* [144] desarrollaron una versión mejorada de mRMR, llamada NMIFS (del inglés *Normalized MIFS*), en la cual se normaliza el término de redundancia de mRMR con el fin de lograr un balance de los términos de relevancia y redundancia en mRMR. El criterio propuesto es:

$$\mathcal{J}_{NMIFS}(f_i) = I(f_i; C) - \frac{1}{|S|} \sum_{s_j \in S} \hat{I}(f_i; s_j), \tag{3.19}$$

donde $\hat{I}$ es el término de redundancia normalizado definido como:

$$\hat{I}(f_i; s_j) = \frac{I(f_i; s_j)}{\text{mín}\left\{H(f_i), H(s_j)\right\}}. \tag{3.20}$$

Una variante al NMIFS fue desarrollada por Vinh *et. al.* [149] al normalizar el término de relevancia por $\log_2(|\Omega_C|)$, donde Ω_C es el espacio de muestras del vector de clase C (en el caso de problema de reconocimiento de dos clases, entonces $|\Omega_C| = 2$).

NMIFS realiza la búsqueda de características utilizando la metodología de búsqueda secuencial hacia adelante descrito en el Algoritmo 3.1.

3.2.5. CMIM

El CMIM (del inglés, *Conditional Mutual Information Maximization*) [3] es un criterio que realiza un ordenamiento de las características de acuerdo a la capacidad que tienen éstas para predecir C, utilizando la información mutua condicional. Para evitar estimar la información mutua condicional en espacios de alta dimensionalidad, CMIM calcula la información mutua condicional solo entre la característica candidata, la variable de clase C y cada una de las características del subconjunto S. La diferencia principal de CMIM con respecto a los criterios mencionados anteriormente, consiste en que en este criterio prioriza la minimización de la redundancia entre las características seleccionadas por sobre la información que éstas tienen de C. El objetivo detrás de esta metodología es obtener aquellas características que

entreguen información única de C, sin embargo, esta cualidad de CMIM le impide encontrar características complementarias (esta dificultad fue estudiada y resuelta en esta tesis). El criterio CMIM se define como:

$$\mathcal{J}_{CMIM}(f_i) = \arg\min_{s_j \in S} I(f_i; C|s_j). \tag{3.21}$$

CMIM desarrolla la búsqueda de características utilizando la metodología del búsqueda secuencial hacia adelante descrito en el Algoritmo 3.1.

3.2.6. JMI

Yang y Moody [136] proponen el criterio JMI (del inglés, *Joint Mutual Information*) el cual está basado en la información mutua conjunta. JMI se define como:

$$\mathcal{J}_{JMI}(f_i) = \sum_{s_j \in S} I(\{f_i, s_j\}; C). \tag{3.22}$$

Este criterio selecciona la característica candidata f_i evaluando si ésta aporta mayor información al conocimiento de C al actuar conjuntamente con cada una de las características de S. Este criterio presenta un efectivo rechazo a características redundantes de S y además puede ser considerado una extensión al criterio $mRMR$ al considerar la descomposición:

$$J_{JMI}(f_i) = I(f_i; C) - \sum_{s_j \in S} I(f_i; s_j) + \sum_{s_j \in S} I(f_i; s_j|C). \tag{3.23}$$

Como se observa en la ecuación (3.23), los dos primeros términos del lado derecho corresponden a una aproximación del criterio mRMR (a diferencia del término $1/|S|$de la ecuación (3.18)). La suma del tercer término de la ecuación (3.23) permite tener mayor conocimiento de la interacción que f_i tiene con s_j y C. El criterio JMI, al igual que mRMR, presenta un fuerte supuesto de independencia de pares de características con respecto al resto de características en S.

3.2.7. Selección de características hacia adelante (Forw)

Una forma de maximizar la cantidad $I(S; C)$ en la ecuación (3.15) es utilizando la regla de la cadena de la información mutua y descomponer esta ecuación de la siguiente forma:

$$I(S'; C) = I(S; C) + I(f_i; C|S), \tag{3.24}$$

donde $S' = \{S, f_i\}$ es el conjunto actualizado de características seleccionadas. En vez de maximizar el término del lado izquierdo de la ecuación (3.24), la idea de este criterio consiste

en maximizar secuencialmente el segundo término del lado derecho de la ecuación (3.24) mediante el Algoritmo 3.1. En otras palabras, se desea obtener el subconjunto S agregando una por una las características f_i de máxima relevancia con respecto al conjunto S. A diferencia de los criterios analizados anteriormente, este criterio si trabaja en espacios de alta dimensionalidad al calcular la información mutua condicional entre f_i y C condicionada a todo el subconjunto S obtenido hasta el momento. Notar que el conjunto S va aumentando de cardinalidad en la medida que progrese la selección.

En términos formales, la característica f_i entregada por el criterio en cada paso es:

$$J_{Forw}(f_i) = \arg \max_{f_i \in F \setminus S} I(f_i; C|S). \tag{3.25}$$

Esta criterio previene seleccionar características relevantes con C, pero que a la vez sean redundantes con las características previamente seleccionadas. Además, este criterio permite identificar ciertas características complementarias, sin embargo falla cuando existen características fuertemente complementarias.

Este criterio se conoce también como criterio de relevancia (REL) y ha sido utilizado en los siguientes trabajos [6, 12, 64].

3.2.8. Eliminación de características hacia atrás (Back)

Otra forma para maximizar la cantidad $I(S; C)$ de la ecuación (3.15) es utilizando la regla de la cadena de la información mutua y descomponer esta ecuación de la siguiente forma:

$$I(F; C) = I(S; C) + I(F \setminus S; C|S). \tag{3.26}$$

El criterio de eliminación de características hacia atrás comienza considerando que $S = F$ y, en cada paso, se elimina la característica $f_i \in S$ que tiene menor relevancia con C bajo el contexto del resto de características de S ($S \setminus f_i$). El criterio para seleccionar la característica f_i más irrelevante de S es:

$$J_{Back}(f_i) = \arg \min_{f_i \in S} I(f_i; C|S \setminus f_i). \tag{3.27}$$

Este criterio realiza cálculos de información mutua sobre el espacio de todas las características del subconjunto S, lo cual tiene la ventaja de capturar las interacciones que f_i tiene con el resto de características de S, incluso aquellas posibles interacciones de características altamente complementarias. Por otro lado, la desventaja de este criterio es su alto costo computacional, dado que para obtener las d características más relevantes de F, es necesario buscar los conjuntos de las $|F| - d$ características más irrelevantes. Otra desventaja de este criterio consiste en no entregar un ordenamiento de relevancia de las características del subconjunto S, sino que realiza un ordenamiento de irrelevancia, por lo que una vez obtenido el subconjunto S de tamaño d, no se tiene conocimiento de cuáles son las características más relevantes de S.

El proceso de eliminación es presentado el Algoritmo 3.2.

Algoritmo 3.2 Pseudo-código de selección de características basado en información mutua utilizando el algoritmo *greedy* con búsqueda hacia atrás.

Entrada: Conjunto de datos F, variable de clase C, número de características a seleccionar d ($d \leq |F|$).

Salida: Conjunto S de tamaño d que contiene las características relevantes.

1: *Inicialización*: Establecer $S \leftarrow F$.
2: **Repetir**
3: *Eliminación de la característica menos relevante*: Encontrar la característica $f_i \in S$ utilizando el criterio $J_{Back}(f_i)$. Establecer $S \leftarrow S \backslash f_i$.
4: **Hasta que** $|S| = |F| - d$

3.2.9. *Markov Blanket* de C

Markov Blanket provee un marco teórico para demostrar que la característica f_i pueden ser eliminada completamente de F sin afectar en nada la cantidad de información que se tenía de C, es decir, $I(F;C) = I(F \backslash f_i; C)$. *Markov Blanket* es definido en términos de independencia condicional, ya que el subconjunto de características $M \subset F \backslash f_i$ es un *Markov Blanket* para la característica f_i, si f_i es condicionalmente independiente del resto de las características $F \backslash \{M, f_i\}$ condicionado a M. En términos probabilísticos, el *Markov Blanket* de f_i se define como:

$$p(F \backslash \{M, f_i\} \,|\, \{M, f_i\}) = p(F \backslash \{M, f_i\} \,|M), \tag{3.28}$$

y en términos de información mutua se define como [14]:

$$I(f_i; \{C, F \backslash \{M, f_i\}\} \,|M) = 0. \tag{3.29}$$

Si la ecuación (3.28) ó (3.29) se cumple, es posible eliminar f_i con total certeza de que no se perderá información de C. La eliminación de características utilizando *Markov Blanket* [71] está compuesta de dos pasos: (1) Para cada característica $f_i \in S$, se selecciona un conjunto $M \subseteq S \backslash f_i$ de k características. (2) La característica f_i^{MB} menos relevante (condicionada sobre M) es eliminada, es decir, $S = S \backslash f_i^{MB}$

$$f_i^{MB} = \arg \min_{f_i \in S} I(f_i; C|M). \tag{3.30}$$

Este proceso es repetido hasta que el subconjunto S no contenga más características irrelevantes y redundantes, o cuando se alcance el tamaño deseado del subconjunto de características. En [71], el coeficiente de correlación de Pearson [150] es usado para encontrar las k características más correlacionadas linealmente con f_i. Las k características son consideradas como el *Markov Blanket* M de la característica candidata f_i. En este caso solo se consideran dependencias lineales entre las características. Encontrar *Markov Blanket* que contengan relaciones no-lineales entre las características implica la utilización de métricas más informativas que complejizan el proceso de eliminación. Es importante mencionar que

la búsqueda de un *Markov Blanket* de cada característica f_i es en sí mismo un problema de selección de características.

Una alternativa para evitar determinar el tamaño k del subconjunto *Markov Blanket*, fue propuesta por Tsamardinos *et al.* [117,119,151,152]. El criterio IAMB (del inglés, *Incremental Association Markov Blanket*) es presentado en el Algoritmo 3.3. IAMB genera incrementalmente una búsqueda hacia adelante de características para obtener un *Markov Blanket* global. En una segunda etapa, el algoritmo realiza una búsqueda hacia atrás de características desde el *Markov Blanket* global hasta obtener el mínimo subconjunto de características.

Algoritmo 3.3 Pseudo-código de la selección de características realizada por el algoritmo IAMB. $\perp\!\!\!\perp$ indica independencia entre variables

Entrada: Conjunto de datos F, y variable de clase C.
Salida: Conjunto M ($M \subseteq F$) que contiene el conjunto Markov blanket de C.

$M = \emptyset$
Flag **cierto**

Fase de Crecimiento: Añadir verdaderos positivos a M

mientras Flag=**cierto** **hacer**
  Flag=**falso**
  Encontrar $f_i \in F \backslash M$ que maximice $I(f_i; C|M)$
  si $(f_i \not\perp\!\!\!\perp C \mid M)$
    $M = M \cup f_i$
    Flag=**cierto**
  fin si
fin mientras

Fase de Reducción: Eliminar falsos positivos desde M

para *cada* $f_i \in M$ **hacer**
  si $(f_i \perp\!\!\!\perp C \mid M \backslash f_i)$
    $M = M \backslash f_i$
  fin si
fin para

Capítulo 4

Contribuciones

4.1. Mejoras al CMIM

4.1.1. Criterio de Maximización de Información Mutua Condicional (CMIM) y sus limitaciones

Una estrategia para encontrar el subconjunto óptimo $S \subset F$ sería evaluar todos los posibles subconjuntos en F, sin embargo, esto es imposible en la práctica debido a la explosión combinatoria de las posibles soluciones. Para evitar una búsqueda exhaustiva, la estrategia de selección *greedy* [12] comienza con un conjunto vacío de características seleccionadas ($S = \emptyset$) y sucesivamente añade características de una en una. El algoritmo de selección *greedy* entrega las características más relevantes utilizando el procedimiento descrito en el Algoritmo (4.1).

Algoritmo 4.1 Pseudo-código de selección de características utilizando el algoritmo *greedy* hacia adelante

---

Entrada: Conjunto de datos F, variable de clase C, y número de características a seleccionar d.

Salida: Conjunto S de tamaño d que contiene las características seleccionadas.

1: Inicialización: Establezca $F \leftarrow$"Conjunto inicial de m características", $S \leftarrow$"Conjunto vacío".
2: Cómputo de la información mutua entre cada una de las característica de F y C. $\forall f_i \in F$, computar $I(f_i; C)$.
3: Selección de la primera característica. Encuentre la característica f_i que maximice $I(f_i; C)$. Establezca $F \leftarrow F \backslash f_i$, $S \leftarrow f_i$.
4: **Repetir**
5: Cómputo de la información mutua del conjunto $\{f_i, S\}$ y C: $\forall f_i \in F \backslash S$, calcule $I(\{f_i, S\}; C)$.
6: Seleccione la próxima característica. Encuentre la característica $f_i \in F$ que maximice $I(\{f_i, S\}; C)$. Establezca $F \leftarrow F \backslash f_i$, $S \leftarrow \{S \cup f_i\}$.
7: **Hasta que** $|S| = d$

El algoritmo de selección *greedy* ideal encuentra el subconjunto óptimo de características S al seleccionar un nuevo $f_i \in F \backslash S$ incrementalmente a través de la maximización de $I(\{f_i, S\}; C)$. Usando la propiedad de la regla de la cadena de información mutua, el funcional $I(\{f_i, S\}; C)$ puede reescribirse como:

$$I(\{f_i, S\}, C) = I(S, C) + I(f_i; C|S). \tag{4.1}$$

En la ecuación (4.1), el segundo término del lado derecho, $I(f_i; C|S)$, mide la relevancia de la característica candidata f_i para predecir la salida C bajo la influencia del conjunto S. El primer término $I(S; C)$ es la información sobre C del subconjunto de características previamente seleccionadas S, pero esta información es común a todas las características candidatas (f_i) y por lo tanto este término puede ser desechado. Por lo tanto, el algoritmo de selección *greedy* puede ser modificado para encontrar el subconjunto S que maximiza $I(f_i; C|S)$, que es un criterio de relevancia [6]. En términos analíticos, este enfoque proporciona las características más relevantes de acuerdo a:

$$REL = \arg \max_{f_i \in F \backslash S} I(f_i; C|S). \tag{4.2}$$

El algoritmo de Maximización de la Información Mutua Condicional (CMIM de sus siglas en inglés) [3, 153] en una aproximación del criterio de relevancia (4.2), al considerar la información mutua entre la característica candidata f_i y la clase C condicionada a una de las características del subconjunto S. Esto permite mantener cierto *trade-off* entre el poder de predicción de f_i con respecto a la salida y la independencia de la característica candidata con cada una de las características previamente seleccionadas. CMIM considera que la característica f_i es relevante solo si la información que tiene con respecto a C no está contenida en cualquiera de las características previamente seleccionadas. Formalmente, el esquema iterativo de selección de CMIM es expresado como:

$$CMIM = \begin{cases} \arg\ \max_{f_i \in F} \{I(f_i; C)\} & , \text{para } S = \emptyset \\ \arg\ \max_{f_i \in F \backslash S} \left\{ \min_{f_j \in S} I(f_i; C|f_j) \right\} & , \text{para } S \neq \emptyset \end{cases} \tag{4.3}$$

La justificación del uso de la función mínimo en (4.3) se basa en una aproximación del concepto de *Markov Blanket*. Fleuret [3] considera que el conjunto M en la ecuación (2.32) se compone de una sola característica en S. Por lo tanto la característica f_i pueden ser descartada del proceso de selección si existe una característica $f_j \in S$ tal que f_i y C son condicionalmente independientes dado f_j. Como la información mutua es siempre positiva tenemos que:

$$\min_{f_j \in S} I(f_i; C|f_j) = 0. \tag{4.4}$$

Por otra parte, la característica $f_i \in F \backslash S$ con el valor más alto de $I(f_i; C|f_j)$ es la más relevante, lo que justifica la función máximo en la ecuación (4.3).

Tabla 4.1: Problema XOR

x_1	x_2	x_3	x_4	$C = x_2 \oplus x_4$
0	1	1	1	0
0	1	1	0	1
0	0	1	1	1
0	0	1	0	0

CMIM selecciona características relevantes evitando la redundancia y evitando el cálculo multidimensional de la información mutua. Sin embargo, su capacidad para identificar y seleccionar las características que interactúan como grupos con la salida [10,14] se ve degradada al seleccionar el valor mínimo de la información mutua condicional. El siguiente ejemplo ilustra esta limitación de CMIM.

Ejemplo 4.1.1 *Sean x_1, x_2, x_3 y x_4 características binarias aleatorias, relacionadas por la función lógica $XOR\,(\oplus)$, $C = x_2 \oplus x_4$. Como se puede ver en la Tabla 4.1, ninguna de las características actuando sola entrega información de la clase C, es decir, la relevancia de cada característica es nula, $I(x_1; C) = I(x_2; C) = I(x_3; C) = I(x_4; C) = 0$. Por otro lado, el par de características $\{x_2, x_4\}$ tiene la mayor relevancia, es decir, $I(\{x_2, x_4\}; C) = I(x_2; C|x_4) = I(x_4; C|x_2) = H(C) > 0$.*

Dado que todas las características individuales tienen relevancia nula, la primera característica que será seleccionada utilizando CMIM depende sólo del orden en que se ingresan, por ejemplo, la característica seleccionada por CMIM es x_1. Para la selección de la segunda característica es necesario determinar $I(x_i; C|x_1)$, $i = \{2, 3, 4\}$ de las características candidatas restantes. Sin embargo, x_1 es independiente de las otras características y de la clase C, por lo tanto $I(x_i; C|x_1) = 0$, $i = \{2, 3, 4\}$. Considerando de nuevo el orden en que se introducen las características, CMIM seleccionará la característica x_2.

Después de seleccionar la característica x_2, la próxima característica que tendría que ser seleccionada es x_4 ya que $I(x_4; C|x_2)$ es mayor que $I(x_3; C|x_2)$, sin embargo, CMIM calcula el mínimo entre $I(x_4; C|x_2)$ y $I(x_4; C|x_1)$, que es cero. Así, la característica x_4 se descarta como característica relevante, y la característica x_3 es erróneamente elegida como la tercera característica. Como el criterio CMIM prioriza aquellas características que entreguen el mínimo de la información mutua condicional, éste no encontrará una solución al problema XOR. En general, en los problemas donde las características son fuertemente complementarias para predecir C, el algoritmo CMIM no podrá encontrar dependencias entre las características. El nuevo criterio que se propone a continuación cambia la función mínimo por la función promedio. En tal caso, como $I(x_i; C|x_1) = 0$, $i = \{2, 3, 4\}$, la función máximo se aplicará a $I(x_4; C|x_2)$ y $I(x_3; C|x_2)$. Dado que este último término es cero, la características x_4 será seleccionada correctamente.

4.1.2. Mejora del criterio de Maximización de la Información Mutua Condicional (CMIM-2)

En esta sección se propone una mejora del criterio CMIM. Se mantienen las ventajas del criterio original, pero se resuelve el problema de la selección de características que son relevantes en pares (fuertemente complementarias).

Considerando que queremos evitar el cálculo de la información mutua condicional en espacios de alta dimensionalidad, se busca aproximar el criterio (4.2) mediante el uso de promedios aritméticos de información mutua condicional. Para ello, definimos $S = \cup_{j=1}^{d} f_j$, donde $|S| = d$, y $G_j = S \backslash f_j$, y usando la regla de la cadena en repetidas ocasiones para la información mutua, se obtiene la siguiente expresión:

$$\begin{aligned}
I(f_i;C|S) = &\quad I(f_i;C|f_1) + [I(G_1;C|\{f_1,f_i\}) - I(G_1;C|f_1)] \\
I(f_i;C|S) = &\quad I(f_i;C|f_2) + [I(G_2;C|\{f_2,f_i\}) - I(G_2;C|f_2)] \\
\vdots \quad = &\quad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
+I(f_i;C|S) = &\quad I(f_i;C|f_d) + [I(G_d;C|\{f_d,f_i\}) - I(G_d;C|f_d)] \\
\hline
d\, I(f_i;C|S) = &\sum_{j\in S} I(f_i;C|f_j) + \sum_{j\in S} [I(G_j;C|\{f_j,f_i\}) - I(G_j;C|f_j)]\,. \qquad (4.5)
\end{aligned}$$

Nótese que el primer término del lado derecho de la ecuación (4.5) representa la información mutua entre f_i y C dado f_j, donde f_j es la $j-$ésima característica perteneciente al subconjunto S. El segundo término representa la restante información mutua condicional presente en espacios de alta dimensión. Limitando el área de trabajo a las características f_i, f_j y la clase C, es decir, evitando la estimación de la información mutua en espacios de más de tres características; la ecuación (4.5) se puede simplificar por una aproximación de primer orden como:

$$I(f_i;C|S) \approx \frac{1}{d} \sum_{j\in S} I(f_i;C|f_j) \qquad (4.6)$$

El costo computacional del criterio CMIM-2 depende de la forma de estimar $I(f_i;C|f_j)$, que en nuestro caso es $O(l \times m \times n \log n)$, donde l es el número de clases del vector C, m es el número de características en el conjunto F y n es el número de muestras disponibles. La estimación de la información mutua se realiza a través de tablas de contingencia para características discretas y el algoritmo del Fraser [55] para las características continuas. El algoritmo de Fraser estima la información mutua mediante el uso de histogramas adaptivos. El método elegido para la clasificación de muestras es heapsort [154], que tiene una complejidad de $O(n \log n)$ en el peor de los casos.

Tabla 4.2: Conjunto de datos utilizados en el experimento 2. La columna **n** contiene el número de muestras, la columna **m** contiene el número de características, la columna **c** contiene el número de clases, y la columna tipo contiene el tipo de características de cada conjunto: D: discreto, C=continuo.

Conjunto de datos	**n**	**m**	**c**	**tipo**
Arrhytmia	452	279	16	D,C
Spambase	4601	57	2	D,C

4.1.3. Experimentos

El criterio propuesto se probó en tres experimentos de selección de características. El primer experimento utiliza el conjunto de datos artificiales MONK-1 [155], con el fin de mostrar la importancia de utilizar el promedio de la información mutua condicional en lugar de la función mínimo. El segundo experimento consiste en dos conjuntos de datos que se describen en la Tabla 4.2 que se utilizan comúnmente para comparar las técnicas de selección de características. El tercer experimento replica parcialmente uno de los experimentos descritos por Fleuret [3] sobre el conjunto de datos de Thrombin. En esta sección el rendimiento del criterio propuesto (CMIM-2) se compara con el método original CMIM y el criterio RANK de selección basado en la información mutua más alta entre una característica y la clase C [147].

Experimento 1: MONK-1 MONK-1 es uno de los tres problemas generados por el problema MONK [155], que describe el dominio artificial de un robot con seis atributos.

$$
\begin{array}{lcl}
x_1 : \text{forma\_cabeza} & \in & \text{redondo,cuadrado, octágono} \\
x_2 : \text{forma\_cuerpo} & \in & \text{redondo,cuadrado, octágono} \\
x_3 : \text{está\_sonriendo} & \in & \text{si, no} \\
x_4 : \text{sosteniendo} & \in & \text{espada, balón, bandera} \\
x_5 : \text{color\_chaqueta} & \in & \text{rojo, amarillo, verde, azul} \\
x_6 : \text{tiene\_cola} & \in & \text{si, no}
\end{array}
$$

Cada problema se genera de acuerdo a la tarea de clasificación que debe realizar el robot, donde las salidas se obtienen como operaciones lógicas de las características. Para el problema de MONK-1, la salida se obtiene como:

$$C = (x_1 \equiv x_2) \vee (x_5 \equiv' rojo') \,, \tag{4.7}$$

donde $\equiv$ y $\vee$ son las funciones 'idéntica a' y OR, respectivamente.

De acuerdo a la ecuación (4.7), las características que proporcionan información sobre la clase C son: $\{x_1, x_2, x_5\}$. La Tabla 4.3 muestra la información relevante de cada característica con respecto a la clase ($I(x_i; C)$, $i = 1, ..., 6$) y en la Tabla 4.4 se muestra la información mutua condicional ($I(x_i; C|x_j)$) para todos los pares de características.

Tabla 4.3: Información mutua $I(x_i; C)$ entre cada característica y el vector de clases para el problema MONK-1.

	x_1	x_2	x_3	x_4	x_5	x_6
C	0,0685	0,0073	0,0030	0,0182	**0,2987**	0,0031

Tabla 4.4: Información mutua condicional $I(x_i; C|x_j)$ para cada par de características del problema MONK-1.

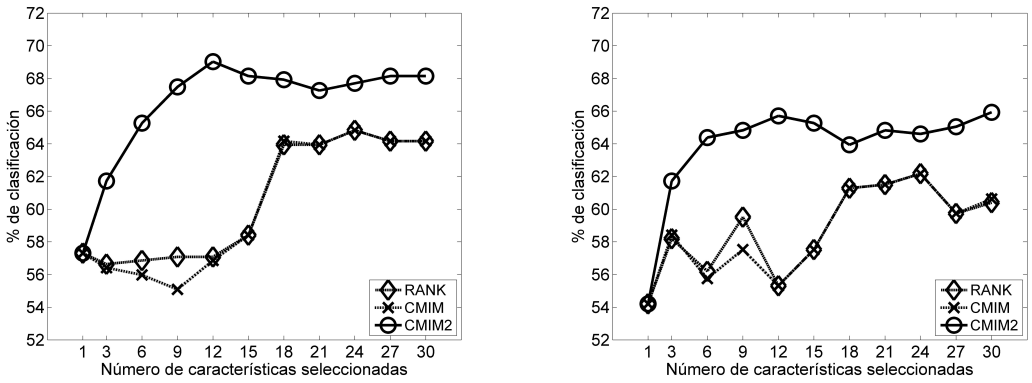
$x_i \backslash x_j$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	0,5096	0,0718	0,0654	**0,0640**	0,0745
x_2	**0,4483**	0	0,0095	0,0080	0,0100	0,0242
x_3	0,062	0,0052	0	0,0116	0,0119	0,0251
x_4	0,0151	0,0189	0,0268	0	0,0331	0,0333
x_5	0,2942	0,3014	0,3076	0,3136	0	0,2993
x_6	0,0091	0,0200	0,0252	0,0182	0,0038	0

Como se muestra en la Tabla 4.3, la característica x_5 tiene la información más alta con respecto a la salida. Este resultado era esperado ya que x_5 no interactúa con ninguna otra característica, excepto con la característica C. Para seleccionar la segunda característica, la información mutua condicional $I(x_i; C|x_5)$ se estima para cada una de las características candidatas $i = (1, 2, 3, 4, 6)$. Los resultados pueden ser observados en la columna x_5 de la Tabla 4.4, donde la característica x_1 tiene el valor más alto de todas las características candidatas. Las características seleccionadas hasta ahora, $S = \{x_5, x_1\}$, son comunes a ambos criterios CMIM y CMIM-2.

La tercera característica que tendría que ser seleccionada es x_2, completando de esta forma la tripleta de características relevantes para el problema MONK-1. Como se observa en la columna x_1 de la Tabla 4.4, la información condicional $I(x_2; C|x_1)$ es máxima, indicando con esto el aporte de información que x_2 tiene de C, sin embargo, CMIM ignorará este hecho (valor máximo en x_2) debido a que su objetivo es encontrar las características que tienen la mayor independencia con respecto a las características previamente seleccionadas (función mínimo en la ecuación 4.3), y que además tengan la mayor información mutua con respecto a la clase C (función máximo en la ecuación 4.3). Realizando el cálculo de CMIM para las restantes características candidatas, es decir, $\max_i \left(\min \left(I(x_i; C|x_1), I(x_i; C|x_5)\right)\right)$, $i = \{2, 3, 4, 6\}$, la tercera característica seleccionada por CMIM es x_4 y no x_2 como era de esperar.

En CMIM-2, los valores de información mutua de la característica candidata x_i, $(i = \{2, 3, 4, 6\})$ con respecto a la clase C condicionada a cada una de las características previamente seleccionadas $(\{x_1, x_5\})$, son promediados en vez de buscar el mínimo, obteniendo el criterio de búsqueda para CMIM-2 como: $\max_i \left(0,5 \cdot \left(I(x_i; C|x_1) + I(x_i; C|x_5)\right)\right)$, $i = \{2, 3, 4, 6\}$. Es de esta forma que CMIM-2 selecciona correctamente la característica x_2.

Experimento 2: Conjunto de datos Arrhythmia y Spambase Con el fin de medir el desempeño del criterio propuesto se utilizan dos bases de datos disponibles en el repositorio UCI [156]. La información básica para cada conjunto de datos está dada en la Tabla 4.2.



(a) (b)

Figura 4.1: Promedio de clasificación de 10 pruebas en la base de datos Arrhythmia vs. el número de características seleccionadas por tres criterios de selección de características: RANK, CMIM y CMIM-2. (a) Usando un clasificador SVM. (b) Usando un clasificador kNN.

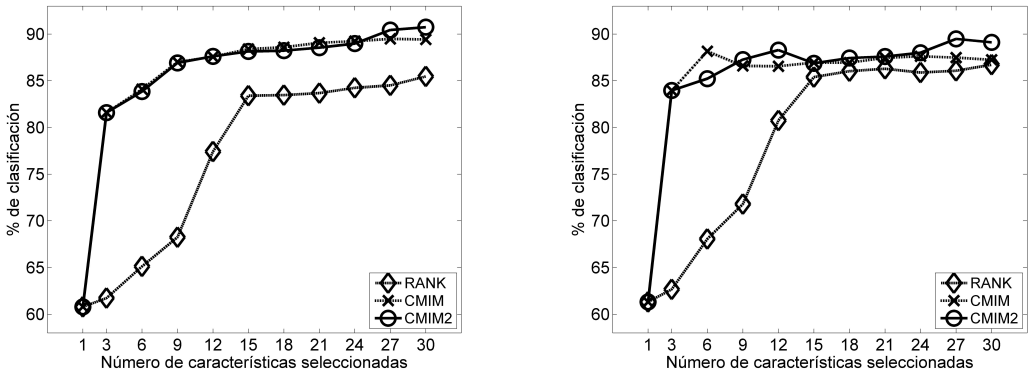
El criterio de selección propuesto (CMIM-2) fue comparado con el método original CMIM [3] y con el método de ordenamiento basado en información mutua (RANK) [147].

Para la validación, se utilizaron dos clasificadores: k-vecinos más cercanos (kNN) y *Support Vector Machine* (SVM) con kernel Gaussiano. La evaluación de subconjuntos de características entregadas por los diferentes métodos se realizó de la siguiente manera: (i) se obtuvieron las primeras 30 características más relevantes entregadas por cada criterio de selección en cada una de las bases de datos (Arrhythmia y Spambase), (ii) se realizaron 10 rondas de validación cruzada en subconjuntos que contienen 1, 3, 6, 9, 12, 15, 18, 21, 24, 27 y 30 características obtenidas desde el ordenamiento entregado por cada criterio de selección de características. El parámetro de vecindad en kNN y el parámetro de tamaño del kernel en SVM fueron seleccionados mediante la optimización del error sobre 10 corridas de validación cruzada. El porcentaje de clasificaciones correctas, es decir, la precisión del clasificador, se presenta en la Figura 4.1 para el conjunto de datos de Arrhythmia.

La Figura 4.1 muestra un aumento significativo en la precisión de clasificador para el criterio CMIM-2 con respecto a los criterios CMIM y RANK en el conjunto de datos de Arrhythmia. Esto significa que existen características en el conjunto de datos de Arrhythmia que son altamente complementarias en la predicción de la variable de salida C [14]. Los resultados entregados por CMIM son comparables a RANK, porque la función mínimo en la ecuación (4.3) busca las características que sean independientes del subconjunto de características previamente seleccionadas S, sin tomar en cuenta la interacción entre la característica candidata y el subconjunto de características seleccionadas de S.

La Figura 4.2 muestra los resultados para el conjunto de datos Spambase, donde se obtuvieron tasas de clasificación muy similares para los métodos CMIM y CMIM-2.

En todas las simulaciones realizadas en estas bases de datos, se observó que el rendimiento CMIM-2 fue superior o igual al rendimiento de CMIM, nunca peor.



(a) (b)

Figura 4.2: Promedio de clasificación de 10 pruebas en la base de datos Spambase vs. el número de características seleccionadas por tres criterios de selección de características: RANK, CMIM y CMIM-2. (a) Usando un clasificador SVM. (b) Usando un clasificador kNN.

Experimento 3: Thrombin El conjunto de datos de Thrombin fue creado para predecir bioactividad molecular en el diseño de fármacos. Esta base de datos contiene 1.909 muestras y 139.351 características (activas o inactivas).

En todos los experimentos se desarrollaron 10 rondas de la validación cruzada para elegir el mejor modelo [157], donde cada partición mantuvo la proporción de ejemplos positivos y negativos. Por otra parte, se realizó la prueba *t-test Student* ($\alpha = 0,05$) de dos colas emparejadas para evaluar si existen diferencias estadísticas significativas entre los promedios de error de validación de CMIM-2 y los de RANK y CMIM.

Debido a que la base de datos Thrombin es muy desbalanceada (42 ejemplos positivos y 1.867 ejemplos negativos), los errores de entrenamiento y validación se midieron mediante el uso de la tasa de error equilibrado (BER), que se define de la siguiente manera:

$$BER = \frac{FP + FN}{2}, \tag{4.8}$$

donde FP es la tasa de falsos positivos y FN es la tasa de falsos negativos.

Los experimentos de validación se realizaron seleccionando las primeras 5 y las primeras 10 características relevantes entregadas por los criterios de selección de características respectivamente. Estas características se introdujeron como entradas de un clasificador SVM con kernel Gaussiano y un clasificador kNN. El parámetro de vecindad de kNN y tamaño del kernel de SVM fueron ajustados minimizando el BER. Debido a los datos desbalanceados, se utilizaron otros dos clasificadores: perceptrón lineal (PCT) y un clasificador bayesiano ingenuo (NB).

Las Tablas 4.5 y 4.6 muestran el error de predicción para diferentes combinaciones de clasificadores y métodos de selección de características utilizando las primeras 5 y primeras 10 características seleccionadas, respectivamente. Además, los valores-p indican que la diferencia

Tabla 4.5: Promedio BER obtenido con diferentes combinaciones de métodos de selección de características y clasificadores para las primeras 5 características seleccionadas en la base Thrombin.

	CMIM-2	CMIM		RANK	
Clasificador	BER	BER	valor-p	BER	valor-p
PCT	**15,17**	18,46	0,32	21,18	0,13
NB	**12,76**	14,22	0,62	18,81	0,05
SVM	**15,85**	26,61	0,03	24,74	0,02
kNN	**18,28**	25,36	0,08	29,41	0,02

Tabla 4.6: Promedio BER obtenido con diferentes combinaciones de métodos de selección de características y clasificadores para las primeras 10 características seleccionadas en la base Thrombin.

	CMIM-2	CMIM		RANK	
Clasificador	BER	BER	valor-p	BER	valor-p
PCT	20.91	26.29	0.18	**18.68**	0.49
NB	**15.21**	16.59	0.55	18.76	0.08
SVM	**17.16**	21.83	0.26	20.91	0.19
kNN	**17.10**	24.33	0.03	22.77	0.14

en el promedio de errores entre CMIM-2 y CMIM y CMIM-2 y RANK, son estadísticamente significativas.

Entre todas las combinaciones, la CMIM-2-NB tiene el menor BER de validación, confirmando la efectividad del método propuesto. En cada clasificador, se encuentra que CMIM-2 obtiene los valores más bajos de BER con respecto a CMIM y RANK, excepto para el PCT con 10 características.

Se debe destacar que los resultados entregados por CMIM-2 para cada clasificador no necesariamente coinciden con los resultados publicados por Fleuret [3], ya en este trabajo se utilizó un conjunto de 139.351 características, mientras que Fleuret utilizó sólo 2.500 características seleccionadas al azar.

4.2. Selección de grupos de características

4.2.1. Criterio propuesto para la selección de grupos complementarios de características utilizando información mutua

4.2.1.1. Detección de características complementarias. Criterio de Máxima Interacción Grupal (MIG)

Como se ha mencionado en capítulos anteriores, la selección de características tiene como objetivo buscar el mínimo subconjunto de características que contengan la información de

C. Considerando que la detección del mínimo subconjuntos es un problema combinatorial, la solución más aceptada en la literatura consiste en identificar y seleccionar una a una las características que entregan información de C (búsqueda *greedy*). De acuerdo a la clasificación de las características realizadas por Yu y Liu [8], el subconjunto óptimo está compuesto por las características fuertemente relevantes y las débilmente relevantes pero no redundantes.

El criterio *greedy* descrito encuentra el subconjunto de variables relevantes, sin embargo el subconjunto de características débilmente relevantes solo puede aproximarse a través de una estrategia eficiente. En este trabajo se está interesado en, por una parte, no se busca una clasificación cualitativa de si una característica es relevante o no, sino que identificar cuál es el grupo complementario para que una característica sea relevante. Además se busca que el criterio propuesto sea rápido y robusto en la búsqueda del mínimo subconjunto de características.

La idea del criterio propuesto es realizar un *ranking* de características, tal como lo hacen los tradicionales métodos existentes en la literatura, pero con la diferencia de agrupar aquellas características que incrementan la relevancia cuando actúan conjuntamente en el conocimiento de C. Al agrupar posible grupos complementarios de características en el *ranking* de características relevantes, es posible detectar en un paso posterior, cuáles son lo grupos que presentan mayor interacción entre ellos.

Para entender esto último, considere un conjunto de características de $A = \{a_1, a_2, a_3, a_4, a_5\}$, donde la característica a_2 es débilmente relevante pero no redundante. El mínimo subconjunto de a_2 para que ésta sea relevante es el subconjunto $\{a_1, a_4\}$, es decir, de acuerdo a la Tabla 2.2 la evaluación de relvancia fuerte de a_2 es $I(a_2; C|\ \{a_1, a_3, a_4, a_5\}) = 0$, mientras que la evaluación de débilmente relevantes es $I(a_2; C|a_1, a_4) > 0$. Por lo tanto, se busca que en el *ranking* de A, la característica a_2 quede de vecina con las características a_1 y a_4.

Para comenzar con el desarrollo de un criterio que identifique grupos complementarios, es necesario establecer límites de relevancia de cada característica f_i. Los límites de relevancia se presentaran dado que, dependiendo del contexto que tenga f_i, ésta tendrá mayor o menor relevancia de C. En el caso que un grupo de características aumente la relevancia de f_i, entonces se habla de características complementarias, mientras que cuando un grupo disminuya la relevancia de f_i, entonces se habla de un grupo redundante.

Es necesario recordar que las características son obtenidas de un proceso o sistema, por lo que el aporte de información que cada característica tiene del proceso no puede ser evaluado en forma individual, ya que las posibles interacciones que cada característica tiene con el resto características para entregar información de C no es conocido a priori. Para conocer el aporte de información que cada característica f_i tiene de C, es necesario evaluar la información que ésta tiene de C bajo el contexto de todo el resto de caracteísticas, es decir, $\neg f_i$. Esta información representa un tipo de información fundamental que cada característica tiene de C. Resulta interesante esta definición porque se relaciona con la definición de características fuertemente relevante, pues cuando la información fundamental es mayor que cero, entonces la característica f_i es clasificada como fuertemente relevante.

Por otro lado, si se considera un subconjunto $S \subset \neg f_i$, $(S \neq \neg f_i)$, el valor de información fundamental de la característica f_i puede ser mayor o menor que la información fundamental,

pero nunca negativa. También resulta interesante observar que esta variación de la información fundamental también ha sido clasificada como característica débilmente relevantes, y esto ocurre cuando la información fundamental de f_i es cero, sin embargo, existe algún subconjunto $S \subset \neg f_i$, $(S \neq \neg f_i)$ donde la información de f_i es mayor que cero.

Tal como se ha mencionado, la información de f_i depende fuertemente de su contexto de características, por lo que es necesario incluir esta información en el proceso de selección de características y en el criterio que se propone. Para ello, se partirá del supuesto que el criterio de selección de características propuesto busca la máxima información que f_i tiene de C. Bajo este supuesto, la información fundamental representa la información mínima que f_i tiene de C, la cual es considerada como información base de f_i. Por lo tanto, las características seleccionadas en el proceso, representan el subconjunto de características que buscan maximizar la información mínima o información fundamental. Considerando este análisis se propone el siguiente criterio:

$$J_{MIG}(f_i) = \begin{cases} \text{máx}\left\{I(f_i;C), I(f_i;C|\neg f_i)\right\} & \text{para } S = \emptyset \\ \text{máx}\left\{I(f_i;C|S), I(f_i;C|\neg f_i)\right\} & \text{para } S \neq \emptyset \end{cases}. \tag{4.9}$$

Las consecuencias de este criterio de considerar la información mínima o fundamental de cada f_i, se traduce en que aquellas características que sean complementarias o fuertemente complementarias, tendrán un valor de información fundamental igual o muy cercano, por lo que en el proceso de selección de características, éstas serán seleccionadas contiguamente. En la próxima sección se describe cómo identificar los grupos complementarios reunidos por el criterio propuesto.

4.2.1.2. Grados de interacción entre características: identificación grupos complementarios

Como se propuso en el criterio MIG, el *ranking* entregado por este criterio deja agrupados contiguamente los posibles grupos complementarios que existen en F. El objetivo en esta etapa consiste en identificar estos grupos complementarios a través de una métrica. Para lograr esto se utiliza la descomposición de la información mutua realizada por William y Beer (revisada en detalle en la Subsección 2.2.3.3), que para el caso de dos variables (característica candidata f_i y el subconjunto de características previamente seleccionadas S) y la clase C se obtiene lo siguiente:

$$\begin{aligned} I(\{f_i, S\}; C) =& \textit{Única}(f_i; C) + \textit{Única}(S; C) + \\ & \textit{Redundancia}(\{f_i, S\}; C) + \textit{Sinergía}(\{f_i, S\}; C). \end{aligned} \tag{4.10}$$

Dado que la información mutua $I(\{f_i, S\}; C)$ nunca es negativa, se asume que cada uno de los términos de la ecuación (4.10) tampoco son negativos. Una de las dificultades del trabajo de Williams y Beer es la determinación de los términos de la descomposición, ya

que en el contexto de selección de características, el cálculo del término de redundancia (ecuación (2.48)) no tiene interpretaciones correctas para algunos problemas. A continuación se presentará en un ejemplo donde se observa este problema.

Ejemplo 4.2.1 *Sean x_1 y x_2 dos características aleatorias binarias independientes entre sí con igual probabilidad (p) de ocurrencia. La función que relaciona estas características esta compuesta por la unión de los valores de cada una de las características, según se muestra en la Tabla 4.7.*

Tabla 4.7: Tabla de resultados del ejemplo 4.2.1.

x_1	x_2	C	p
0	0	00	1/4
0	1	01	1/4
1	0	10	1/4
1	1	11	1/4

Como se observa en este ejemplo, la salida C está construida por cada una de las características x_1 y x_2, por lo tanto, la información que entrega cada una de éstas es única. Por otro lado, dado que las características son independientes entre sí, entonces la información mutua entre ambas es cero ($I(x_1; x_2) = 0$). El valor de redundancia de Williams y Beer es 1, con lo cual, calculando el resto de términos de la descomposición se tiene:

$$Redundancia(\{x_1, x_2\}; C) = 1 \tag{4.11}$$
$$Única(x_1; C) = 0 \tag{4.12}$$
$$Única(x_2; C) = 0 \tag{4.13}$$
$$Sinergia(\{x_1, x_2\}; C) = 1. \tag{4.14}$$

La determinación de la redundancia no es correcta en este problema, ya que se indica que la sinergia y la redundancia tienen valor uno, mientras que la información única de cada característica es cero. En base a esta dificultad, resulta necesario desarrollar, en esta tesis, una nueva metodología para obtener los valores de interacción (redundancia, sinergia, relevancia única), la cual se explica a continuación.

El cálculo de las interacciones parte del supuesto que existen dos tipos de redundancia, una redundancia en la información que ambas características tienen de C, y por otro lado, una redundancia que existe sólo entre las características. A la primera redundancia se le llamará simplemente redundancia, mientras que a la segunda redundancia se le llamará información única entre x_1 y x_2. Conjuntamente al cálculo de interacciones, se parte del supuesto que la complementariedad o sinergia es una información que sólo aparece solo cuando f_i y S interactúan conjuntamente en el conocimiento de C. Considerando lo anterior, se desea cuantificar cada una de estas interacciones en forma individual, con el fin de determinar el grado de complementariedad que existe entre f_i y S. Una representación gráfica de estas interacciones

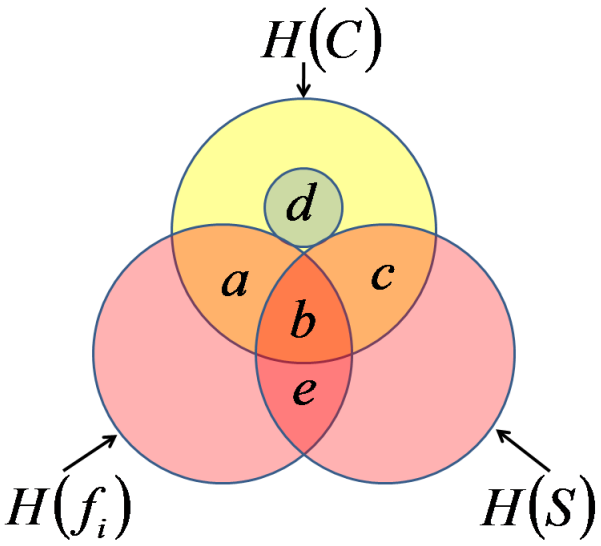


Figura 4.3: Diagrama de Venn que ilustra la organización de la información de f_i ,S y C.

se muestra en la Figura 4.3, la cual ilustra en representaciones de entropía, todas las posibles interacciones de f_i y S en el conocimiento de C.

La representación gráfica de la Figura 4.3 permite visualizar cada uno de los términos propuestos por Williams y Beer, y los cuales están representador por: $\acute{U}nica(f_i;C) = a$, $\acute{U}nica(S;C) = c$, $Redundancia(\{f_i,S\};C) = b$, $Sinergia(\{f_i,S\};C) = d$. A diferencia de la propuesta de William y Beer, en la Figura 4.3 aparece una nueva área e, la cual se puede interpretar como la información única entre f_i y S, es decir, $\acute{U}nica(f_i;S) = e$.

Los círculos grandes en la Figura 4.3, representan la entropía de las variables f_i, S y C, respectivamente. El círculo pequeño corresponde a la entropía que sólo aparece cuando interactúan las variables f_i, S, C y que es representado como d (Revisar anexo A para más detalles de esta propuesta). Formalizando las informaciones mutuas entre éstas tres variables para todas las posibles informaciones mutuas de las combinaciones de éstas tenemos:

$$\begin{array}{lcl,lcl}
1)\, I(f_i;C) & = & a+b & 6)\, I(f_i;C \mid S) & = & a+d \\
2)\, I(S;C) & = & c+b & 7)\, I(S;C \mid f_i) & = & c+d \\
3)\, I(f_i;S) & = & e+b & 8)\, I(f_i;S \mid C) & = & e+d \\
4)\, II(f_i;S;C) & = & d-b & 9)\, I(\{f_i,S\};C) & = & a+b+c+d \\
5)\, I(\{f_i,C\};S) & = & b+c+d+e & 10)\, I(\{S,C\};f_i) & = & a+b+d+e
\end{array} \tag{4.15}$$

Como se observa en las ecuaciones 4.15, las ecuaciones desde la 5 a la 10 son combinaciones lineales de las cuatro primeras ecuaciones. Esto hace que existan cuatro ecuaciones y cinco incógnitas, por lo que no existe una única solución del tamaño de las áreas.

Para determinar el valor de las áreas a, b, c, d y e se considera que una de estas áreas es la variable independiente, dejando al resto de las áreas en términos de esta variable independiente. En esta propuesta se considera como variable independiente el área b, es decir, el

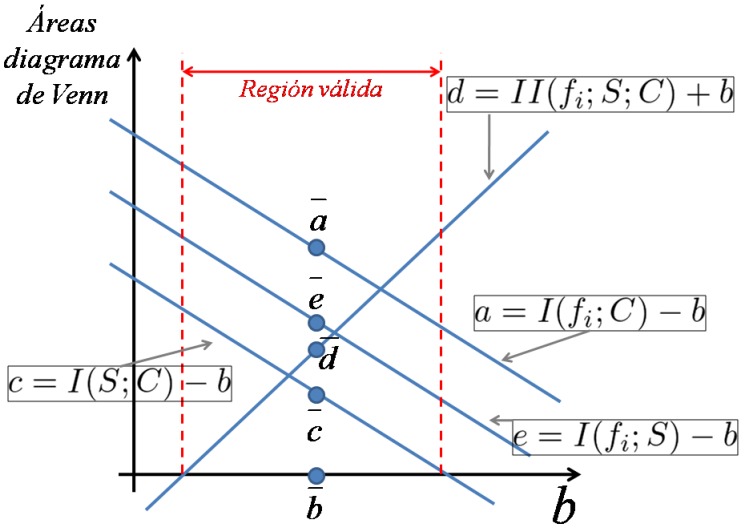


Figura 4.4: Rango válido de variación de los valores de las áreas de la Figura 4.3. Note que la variable independiente es b. Los puntos representan el valor medio de las áreas a, b, c, d, e dentro de la región válida.

término de redundancia. En la Figura 4.4 se observa que al considerar el área b como variable independiente, el resto de variables se determina automáticamente. Ya que los valores que puede tomar b son infinitos, se restringe el rango de b a aquellos valores donde todas las áreas tengan valor positivo, definiendo con ésto un rango válido para todas las áreas. Dado que se busca un valor para cada área y no un rango de valores (región válida presentada en la Figura 4.4) se utiliza, sin pérdida de generalidad, el valor medio de los valores contenidos en la región válida para cada área. En la Figura 4.4 los valores medios para cada área se representan como círculos llamados $\overline{a}$, $\overline{b}$, $\overline{c}$, $\overline{d}$ y $\overline{e}$.

Una vez de determinados el valor de las áreas, es necesario diseñar una métrica que permita cuantificar la interacción entre grupos complementarios. Dentro de las consideraciones de este métrica esta realizar una comparación de la información nueva (complementaria) que se genera, versus la información única que f_i y S entregan de C. De acuerdo a esto, una conveniente métrica de interacción es dividir la complementariedad por sobre suma de la información única que f_i y S tienen de C. Con esto se establece una razón de sinergia la cual se define como:

$$RT_{Sinegia}(f_i, S, C) = \frac{\overline{d}(f_i, S, C)}{\overline{a}(f_i, S, C) + \overline{c}(f_i, S, C)}. \tag{4.16}$$

Finalmente, el último paso consiste en el diseño de una estrategia de búsqueda para la detección de los grupos relevantes. La búsqueda de grupos se realiza utilizando el ordenamiento de características entregado por MIG, criterio que fue diseñado para capturar las posibles interacciones que existen entre las características. La detección de grupos de características complementarias comienza calculando la interacción entre la primera y segunda característica en el ordenamiento. Después, la misma primera características evalúa su interacción con

el grupo de características compuesto por la segunda y tercera características en el ordenamiento. Iterativamente se realiza este proceso de calcular la interacción entre la primera característica y el grupo de características vecinas que aumenta su tamaño hasta llegar al total de las características o hasta un tamaño máximo de características definido por el usuario. Una vez analizada las interacciones de la primera característica se comienza con la segunda características y se realiza el mismo proceso descrito anteriormente, evaluando su interacción con los grupos de características generados con las características vecinas. Este proceso se repite hasta alcanzar la penúltima característica seleccionada donde se calcula la interacción con la última característica del ordenamiento. Al finalizar el proceso, tiene una matriz con los niveles de interacción entre las características vecinas. Esta matriz permite cuantificar los niveles de interacción entre las características.

Para entender el funcionamiento consideremos el siguiente caso hipotético para la seleción e identificación de grupos complementarios de características. Consideremos el conjunto de características $A = \{a_1, a_2, a_3, a_4, a_5\}$. Al ordenar este conjunto de características de acuerdo al criterio MIG, el ordenamiento obtenido es $R = \{a_5, a_3, a_2, a_4, a_1\}$, donde a_5 corresponde a la primera característica seleccionada por MIG (1°), a_3 corresponde a la segunda característica seleccionada por MIG (2°), y así sucesivamente para el resto de características en R.

Considerando que el criterio MIG posiciona contiguamente las características pertenecientes a grupos complementarios, la propuesta para identificar los grupos complementarios es utilizar la siguiente matriz de combinaciones:

$f_i \diagdown S$	$a_5(1°)$	$a_3(2°)$	$a_2(3°)$	$a_4(4°)$	$a_1(5°)$
$a_5(1°)$	$-$	$\{a_5, a_3\}$	$\{a_5, a_3, a_2\}$	$\{a_5, a_3, a_2, a_4\}$	$\{a_5, a_3, a_2, a_4, a_1\}$
$a_3(2°)$		$-$	$\{a_3, a_2\}$	$\{a_3, a_2, a_4\}$	$\{a_3, a_2, a_4, a_1\}$
$a_2(3°)$			$-$	$\{a_2, a_4\}$	$\{a_2, a_4, a_1\}$
$a_4(4°)$				$-$	$\{a_4, a_1\}$
$a_1(5°)$					$-$

Recordando que la interacción entre las características de un grupo complementario debe ser positiva, entonces es necesario evaluar la interacción para cada posible subconjunto de la matriz de combinaciones utilizando la medida de interacción $II(f_i; S; C)$.

Si se toma como ejemplo el subconjunto $\{a_3, a_2, a_4, a_1\}$ de la matriz, la característica $f_i = a_3$, mientras que el subconjunto $S = \{a_2, a_4, a_1\}$. Una vez calculada la medida de interacción para cada subconjunto de la matriz de combinaciones, solo se seleccionan aquellos subconjuntos que presentan interacción mayor a cero. Posteriormente, a éstos últimos subconjuntos se les calcula la tasa de complementariedad utilizando la métrica $RT_{sinergia}$ definida en la ecuación (4.16). Finalmente se visualiza la matriz de combinaciones y se identifican los grupos complementarios.

A continuación se presenta en el Algoritmo 4.2, el pseudo-código del criterio propuesto.

Algoritmo 4.2 Pseudo-código de selección de características e identificación de grupos complementarios de características utilizando el algoritmo propuesto

Entrada: Conjunto de datos F, variable de clase C, y número de características a seleccionar d.

Salida: Conjunto S de tamaño d que contiene las características seleccionadas y matriz de interacciones M entre las características.

Paso 1: Selección de características relevantes y complementarias

1: Inicialización: Establezca $F \leftarrow$"Conjunto inicial de m características", $S \leftarrow$"Conjunto vacío".
2: Cómputo de la información mutua e información mutua condicional de cada una de las característica de F. $\forall f_i \in F$, computar $I(f_i; C)$ y $I(f_i; C|\neg f_i)$.
3: Selección de la primera característica. Encuentre la característica f_i que maximice: $\max\{I(f_i; C), I(f_i; C|\neg f_i)\}$. Establezca $F \leftarrow F \backslash f_i$, $S \leftarrow f_i$.
4: **Repetir**
5: Cómputo de la información mutua condicional entre f_i y C condicionada a S. $\forall f_i \in F \backslash S$, calcule $I(f_i; C|S)$.
6: Seleccione la próxima característica. Encuentre la característica $f_i \in F \backslash S$ que maximice: $\max\{I(f_i; C|S), I(f_i; C|\neg f_i)\}$. Establezca $F \leftarrow F \backslash f_i$, $S \leftarrow \{S \cup f_i\}$.
7: **Hasta que** $|S| = d$

Paso 2: Identificación de grupos complementarios

8: **para** $i = 1$ hasta $(d-1)$ **hacer**
9: **para** $j = i + 1$ hasta d **hacer**
10: $P = \{s_{i+1}, \cdots, s_j\}$ {Generar grupo de características}
11: $idx = II(s_i; P; C)$ {Calcular interacción para identificar si s_i y P son redundantes o complementarios}
12: **si** $idx > 0$
13: $M(i, j) = RT_{sinergia}(s_i; P; C)$ {Complementario}
14: **si no**
15: $M(i, j) = 0$ {Redundante}
16: **fin si**
17: **fin para**
18: **fin para**

Tabla 4.8: Análisis comparativo de las diferentes criterios de selección de características en términos de complejidad computacional.

	Rank	mRMR	CMIM	Forw	Back	MIG
Selecciona relevancia	Si	Si	Si	Si	Si	Si
Elimina redundancia	No	Si	Si	Si	Si	Si
Utiliza interacción de segundo orden entre características	No	No	Si	Si	Si	Si
Utiliza interacción de d-ésimo orden (mayor a 2) entre características	No	No	No	Si	Si	Si
Retorna *ranking* de características	Si	Si	Si	Si	No	Si
Identifica grupos complementarios de características	No	No	No	No	No	Si

4.2.2. Complejidad computacional y comparación teórica con otros métodos

Para medir las ventajas y desventajas del algoritmo propuesto es conveniente realizar comparaciones con los tradicionales métodos de selección de características existentes en la literatura. Los criterios utilizados para realizar la comparación con el criterio propuesto MIG son: Rank, mRMR, CMIM, búsqueda hacia adelante (Forw), búsqueda hacia atrás (Back) (Estos criterios fueron descritos en la sección 3.2). En una primera parte en esta comparación se indicará la presencia o ausencia de ciertas características en el diseño de un filtro de selección de características. La Tabla 4.8 muestra la presencia o ausencia de distintas características deseadas en un filtro de selección de características en cada uno de los criterios analizados.

De acuerdo a la Tabla 4.8 se observa que el criterio propuesto realiza todas las propiedades deseadas en un filtro de selección de características.

Otro etapa importante para evaluar la utilidad del algoritmo propuesto es comparar la complejidad computacional del criterio en la selección de un subconjunto $d \leq m$ de características. La Tabla 4.9 muestra el costo computacional que tiene cada criterio principalmente en tres aspectos: (1) Llamadas al criterio para generar un valor de la relevancia de una característica candidata; (2) Llamados de la información mutua, qiue corresponde a la cantidad de veces que es necesario calcular la informaciòn mutua para generar la nota de relevancia de una característica candidata; (3) Dimensionalidad del espacio de probabilidad donde se realiza el cómputo de información mutua. Finalmente, en la Tabla 4.9 se presenta el orden polinomial del cálculo de cada uno de los criterios analizado.

La última etapa en el proceso de comparación y complejidad computacional está relacionada con la estimación de la información mutua. El estimador utilizado en esta tesis es el propuesto por Kozachenko-Leonenko [45, 63, 158] basado en la distancia de los vecinos más

Tabla 4.9: Costo computacional de distintos criterios de selección de características sin considerar el estimador de entropía.

	Rank	mRMR	CMIM	REL	CMI	MIG
Llamados a la función evaluadora	$\frac{mxd}{2}$	$\frac{m \times d}{2}$	$\frac{m \times d}{2}$	$\frac{m \times d}{2}$	$\frac{m \times (m-d)}{2}$	$\frac{m \times d}{2}$
Llamados de información mutua por evaluación	1	d	$d-1$	1	1	2
Espacio de probabilidad generado	2	2	3	$d+1$	$m-d$	m
Orden del criterio de selección de características	$\mathcal{O}(m \times d)$	$\mathcal{O}(m \times d)$	$\mathcal{O}(m \times d)$	$\mathcal{O}(m \times d)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m \times d)$

cercanos. La dificultad de la estimación de los vecinos más cercanos es que en el peor de los casos, el costo computacional puede ser $\mathcal{O}(n^2)$ en cada dimensión, siendo n el número de muestras de la base de datos. Para evitar este cálculo, en este estudio se utilizó el *toolbox* ANN (*Approximate Nearest Neighbor*), al cual es una librería en C++ para encontrar los vecinos más cercanos de una muestra en forma exacta o aproximada en un espacio p-dimensional. Los n datos en un espacio p-dimensional son procesados dentro de estructuras para una búsqueda eficiente (*kd-trees y box-decomposition*). La complejidad para estimar la entropía depende del número de ejemplos n, el número de vecinos k y la dimensión p del espacio. El costo de cálculo es del orden de $\mathcal{O}(kp \log n)$ [159].

4.2.3. Experimento con base de dato sintética

Los resultados presentados a continuación muestran que el criterio MIG es superior a las dos estrategias de selección características basadas en información mutua: Máxima dependencia con selección hacia adelante (Forw) y máxima dependencia con selección hacia atrás (Back), que son las bases para la deducción de la mayoría de los algoritmos de selección de características existentes en la literatura (ver sección 2.2.4.3). Además, en este experimento se presenta la contribución del algoritmo propuesto en la detección de grupos de características.

4.2.3.1. Descripción de los experimentos

Sea $X = \{x_1, x_2, ..., x_7\}$ un conjunto de características aleatorias binarias obtenidas desde una distribución uniforme. El objetivo de esta prueba es presentar claramente cómo la interacción de las características es crucial para la correcta detección de las características

relevantes. Esta prueba consiste de dos experimentos que relacionan las características de entrada con distintas funciones lógicas.

- El experimento 1 muestra las ventajas de MIG por sobre los tradicionales métodos de selección de características: Máxima Dependencia (Forw) [4] y la Información Mutua Condicional (Back) [34].
- En el experimento 2 se muestra cómo el método propuesto selecciona los grupos de características desde una función matemática donde se pueden generar distintos grupos de características, siendo todos igualmente válidos para predecir el vector de salida. La fácil e intuitiva representación gráfica de las interacciones con el método propuesto, permite representar la mayoría de las combinaciones de características de entrada y además permite asignar grados de confianza de los grupos generados. La estimación de la información mutua se calcula aproximando la función de probabilidad mediante histogramas [3].

4.2.3.2. Experimento 1

En este experimento se utilizó la función $C = x_8 \oplus x_5 \oplus x_6$, donde $x_8 = x_2 \vee x_3 \vee x_4$, y además x_1y x_7 son ruido del sistema. Las funciones $\oplus$ y $\vee$ son las funciones lógicas OR-Exclusivo y OR, respectivamente. La complejidad de este problema es: detectar el conjunto mínimo de características, detectar la complementariedad de las características x_5, x_6 y x_8 para predecir la característica de salida, y además generar los subconjuntos de características relevantes. La prueba consiste en generar el ordenamiento de características en los tres métodos de selección de características: Forw, Back y MIG y repetir esta prueba sobre 1.000 distintas permutaciones en el orden de las características de entrada. Los resultados se presentan en histogramas que muestran la frecuencia que fueron elegidas todas las características de entrada como primera característica más relevante, segunda característica más relevante y hasta la última característica más relevante. Los colores azules de las barras representan las características que generan el conjunto óptimo de características ($\{x_5, x_6, x_8\}$), los colores verdes de las barras representan las características del conjunto redundante de x_8 ($\{x_2, x_3, x_4\}$). Finalmente las barras rojas representan las características irrelevantes del problema ($\{x_1, x_7\}$).

Al comparar las figuras 4.5, 4.6 y 4.7 se observa que el criterio propuesto MIG es superior a Forw y Back.

4.2.3.3. Experimento 2

En este experimento se utilizó la función $C = ((y_1 \vee y_2) \wedge x_5) \vee (x_6 \oplus x_7 \oplus x_8)$, donde $y_1 = (x_2 \vee x_3) \oplus x_1$ y $y_2 = x_2 \oplus x_4$. Las funciones $\oplus$, $\vee$ y $\wedge$ son las funciones lógicas OR-Exclusivo, OR y AND, respectivamente. El objetivo de este experimento es mostrar la utilidad del criterio de selección de grupos de características. La complejidad de este experimento está en la función que relaciona las características, ya que existen grupos de características

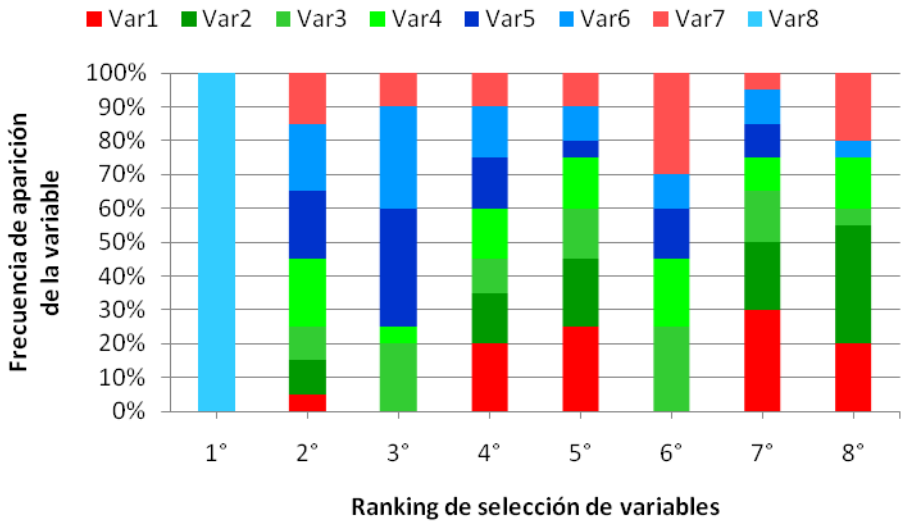


Figura 4.5: Frecuencia con que las características fueron ordenadas por el criterio de selección de características Forw.

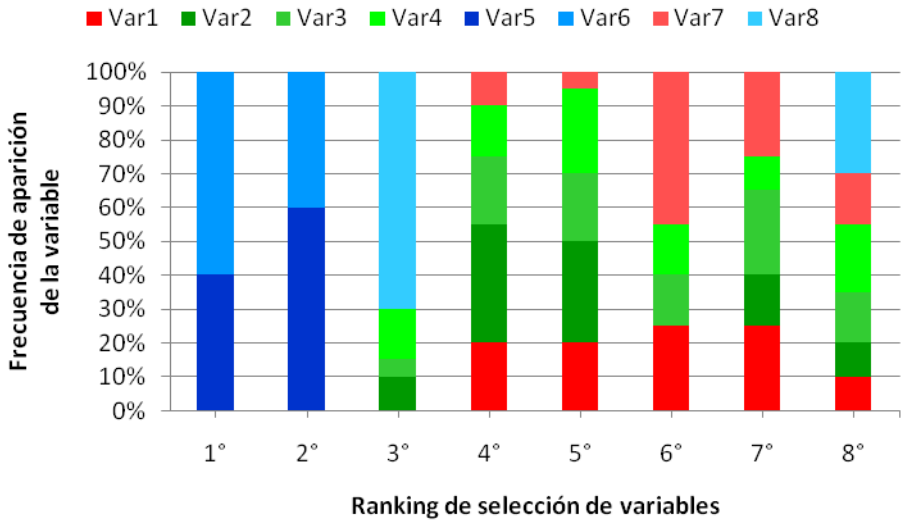


Figura 4.6: Frecuencia con que las características fueron ordenadas por el criterio de selección de características Back.

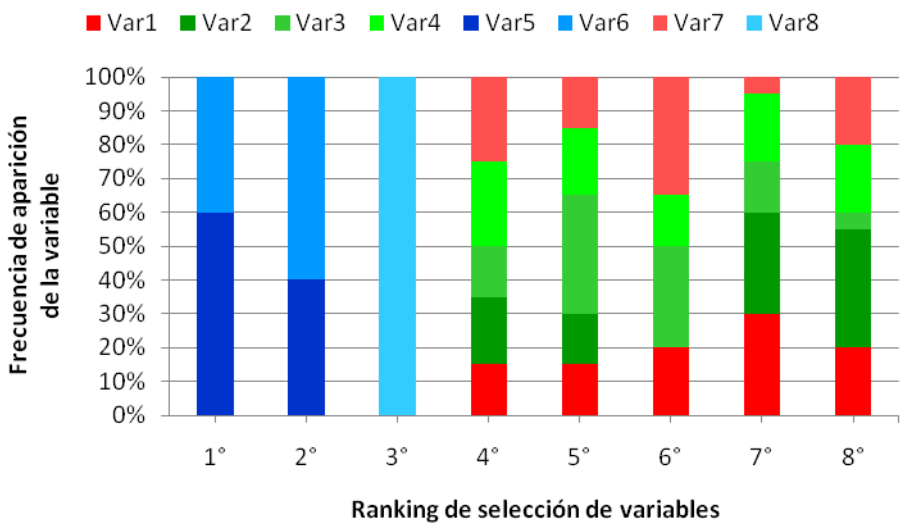


Figura 4.7: Frecuencia con que las características fueron ordenadas por el criterio de selección de características MIG.

complementarias que no son claramente diferenciables del resto de características. El criterio propuesto en la ecuación 4.16, presenta valores elevados de interacción en aquellos grupos de características que son fácilmente identificables como grupos complementarios de características, y entrega valores más pequeños a posibles grupos de características que pueden generarse. El experimento comienza con el ordenamiento entregado por el algoritmo J_{MIG}. El resultado del criterio de selección de características J_{MIG} es $\{x_6, x_7, x_8, x_5, x_1, x_2, x_4, x_3\}$. Para analizar los grupos de características es conveniente reemplazar las características y_1 e y_2 en la ecuación que relaciona las características en este experimento. Reemplazando y reordenando términos la ecuación de este experimento queda de la siguiente forma:

$$C = \begin{cases} ((((x_2 \vee x_3) \oplus x_1) \vee (x_2 \oplus x_4)) \wedge x_5) \vee (x_6 \oplus x_7 \oplus x_8) \\ (((x_1 \oplus x_2) \vee (x_1 \oplus x_3) \vee (x_2 \oplus x_4)) \wedge x_5) \vee (x_6 \oplus x_7 \oplus x_8) \end{cases} \tag{4.17}$$

Observando la ecuación 4.17, es posible apreciar que los grupos fuertemente complementarios de características más importantes son los que están relacionados con la función lógica $\oplus$, y estos son: $\{x_6, x_7, x_8\}$, $\{x_2, x_4\}$. Por otro lado los grupos $\{x_1, x_2\}$ y $\{x_1, x_3\}$ están relacionados por la función $\oplus$, pero la característica x_1 es redundante a ambos grupos, por lo que el grupo de características complementarias puede ser $\{x_1, x_2, x_3\}$, o los grupos $\{x_1, x_2\}$ y $\{x_1, x_3\}$. El análisis de la característica x_5 es aún más complejo, ya que x_5 se relaciona a los grupos $\{x_1, x_2, x_3\}$ y $\{x_2, x_4\}$.

La Tabla 4.10, muestra los valores de interacción obtenidos de acuerdo a la metodología explicada en la subsección 4.2.1.2. Las $\uparrow$ representan valores muy elevados de interacción (tienden a infinito). De acuerdo a la Tabla 4.10, la primera fila muestra que la interacción de la característica x_6 con los subconjuntos $\{x_7\}$ y $\{x_7, x_8\}$ es alta. Esto significa que estas características obligatoriamente deben ser consideradas como un grupo complementario de características relevantes. Considerando que existen dos posibles grupos de características relevantes, es decir, $\{x_6, x_7\}$ y $\{x_6, x_7, x_8\}$ podemos validar si la característica x_8 debe

Tabla 4.10: Matriz de interacción del experimento 2. El signo ↑ indica que el valor de interacción es elevado.

Ranking MIG	***6(1°)***	***7(2°)***	***8(3°)***	***5(4°)***	***1(5°)***	***2(6°)***	***4(7°)***	***3(8°)***
6(1°)	-	↑	↑	4, 58	4, 15	3, 61	2, 59	2, 31
7(2°)		-	↑	0	0	0	0	0
8(3°)			-	0	0	0	0	0
5(4°)				-	0, 08	0, 21	0, 55	0, 67
1(5°)					-	1, 01	0, 54	1, 28
2(6°)						-	↑	↑
4(7°)							-	0
3(8°)								-

pertenecer al grupo de características relevantes $\{x_6, x_7\}$, observando el valor de interacción de las características x_7 y x_8 en la segunda fila de la tabla 4.10. Efectivamente se observa que las características x_7 y x_8 tienen un alto grado de interacción, por lo que se considera que el conjunto de características relevantes está compuesto por $\{x_6, x_7, x_8\}$.

Un caso muy parecido al caso explicado anteriormente, es el que ocurre en la fila 6 de la tabla 4.10. En este caso la característica x_2 presenta un alto grado de interacción con $\{x_4\}$ y $\{x_4, x_3\}$. Solo tenemos certeza que el subconjunto $\{x_2, x_4\}$ es un grupo complementario de características relevantes. Por lo tanto, es necesario corroborar si la característica x_3 pertenece a este grupo complementario de características. Analizando la fila 7 de la tabla 4.10 observamos que la interacción entre x_4 y x_3 es nula, por lo que solo podemos considerar como grupo complementario de características a $\{x_2, x_4\}$.

En lo que sigue del análisis de este experimento, ya no existen grupos complementarios de características que deban ser agrupados por su alto elevado nivel de interacción, por lo tanto es necesario establecer una base para considerar cuáles son grupos complementarios de características y cuáles no. En esta tesis se trabajará con el promedio de las interacciones que no son ↑. El valor de la interacción promedio del sistema es $0, 94$. Valores de interacción por sobre este nivel no serán considerados para asociar grados de interacción fuertes entre características. Observando la fila 4 de la tabla 4.10, vemos que la característica x_5 presenta niveles de interacción menores a la media con los subconjuntos $\{x_1\}$, $\{x_1, x_2\}$, $\{x_1, x_2, x_4\}$ y $\{x_1, x_2, x_4, x_3\}$. Desde la base de la interacción promedio, podemos considerar que las interacciones que tiene x_5 con el resto de las características no es suficientemente fuerte para ser agrupada con otras características, por lo que x_5 es considerada como solo una característica independiente.

Finalmente, analizamos la fila 5 de la tabla 4.10. La característica x_1 interactúa con la característica x_2 con un valor de interacción que es superior a la promedio de interacciones del sistema, sin embargo, cuando x_1 interactúa con $\{x_2, x_4\}$, el valor de interacción es menor a la interacción promedio. Estas disminución de interacción puede ser considerada que el ingreso de la característica x_4 al grupo $\{x_1, x_2\}$, genera interferencia del poder informativo $\{x_1, x_2\}$ tienen de C. Avanzando al siguiente grupo de la fila 5, se observa que la interacción de x_1 y $\{x_2, x_4, x_3\}$ es superior a la media de interacciones, y además es la mayor interacción alcanzada en la fila. Esto nos permite considerar que el grupo que interacciona complementariamente es $\{x_1, x_2, x_4, x_3\}$, o bien los grupos $\{x_1, x_3\}$ y $\{x_1, x_4\}$.

Tabla 4.11: Las bases de datos utilizadas en los experimentos. La columna n indica la cantidad de ejemplos, m es el número de características de la base y $|c|$ es el número de clases de la base.

	Base de datos	n	m	$\|c\|$
1	Madelon	2.600	500	2
2	Arcene	200	10.000	2
3	Splice	3.190	60	3
4	Semeion	1.593	256	10
5	Promoters	106	57	2
6	Lung Cancer	32	56	3

4.2.4. Experimento con bases de datos reales

4.2.4.1. Bases de datos

En esta sección, se compara el desempeño el criterio de selección de características propuesto versus 5 tradicionales métodos de selección de características basados en la información mutua existentes en la literatura. Las 6 bases de datos utilizadas en esta comparación, comprenden un amplio rango de características, en términos de número de muestras, número de características, número de clases, y tipos continuos o discretos de características. El detalle de cada una de estas bases de datos esta dado en la Tabla 4.11.

Todos los criterios de selección de características utilizados en esta comparación utilizan la información mutua en sus funcionales. Para ello, el cálculo de la información mutua se realizó mediante la siguiente descomposición de la información mutua en términos de entropía

$$I(X;C) = H(X) - \sum_{c \in C} p(c) \cdot H(X|C = c). \tag{4.18}$$

La ecuación (4,18) aprovecha de que C es discreta. La estimación de la entropía se realiza con el estimador de Kozachenko-Leonenko (descrito en la sección 2.1.4). Todos los criterios de selección de características utilizados en este experimento fueron implementados en $C++$ utilizando este estimador.

La comparación de MIG se realizó con cuatro aproximaciones de los criterios de selección de características más utilizadas en la literatura. Estos métodos son: Rank, Mrmr, Cmim, MD-forward (Forw). Es necesario destacar que en algunas bases de datos, MIG se comparó con MD-Backward (Back) con el fin de mostrar que MIG entrega resultados comparables a Back, pero con la ventaja que MIG obtiene sus resultados sin el costo computacional de Back.

Para realizar la comparación entre los distintos métodos de selección de características, se utilizo un clasificador de k vecinos más cercanos (kNN). El parámetro k en el clasificador kNN fue de $k = 3$.

El primer paso en esta comparación comienza realizando la selección de características en cada base de datos utilizando cada uno de los criterios considerados en este experimento (Rank, mRMR, CMIM, Forw, MIG y Back). Posteriormente, el ordenamiento de las p características más relevantes (*ranking*) entregado por cada uno de los criterio de selección $R_{crit\_Sel} = \{r_1, ..., r_p\}$, es evaluado para los p subconjuntos de características generados al concatenar las $q \leq p$ primeras características del *ranking*, es decir, el subconjunto de q características se genera como: $S^q_{crit\_sel} = \bigcup_{i=1}^{q} r_i$. La evaluación de cada subconjunto $S^q_{crit\_sel}$ se realiza utilizando 10 rondas de validación cruzada en el clasificador kNN.

Para evaluar la significancia estadística de la diferencia del desempeño del criterio de selección de características propuesto versus los otros métodos de selección de características, se realizó una prueba estadística de t-Student [160] considerando como hipótesis nula que los resultados de clasificación (obtenidos en las 10 rondas de validación cruzada) obtenidos por MIG y otro criterio de selección de características, son resultados independientes de una distribución normal con igual media e igual varianza pero desconocida. La hipótesis alternativa indica que las medias de las distribuciones de los resultados de cada criterio pertenecen a distribuciones con distinta media. Se considera un 5 % de nivel de significancia.

4.2.4.2. Resultados

A continuación se presenta el análisis de cada una de las bases de datos presentadas en la Tabla 4.11. Los resultados son presentados en gráficas de características versus clasificación del clasificador kNN con el cual se comparan los distintos métodos de selección de características.

Madelon Este problema fue utilizado en el desafío NIPS2003 de selección de características [161]. El problema fue diseñado por Isabel Guyon y consiste en un hipercubo de 5 dimensiones donde en cada esquina del hipercubo se asignan cluster de datos. Las 32 esquinas del hipercubo se dividen en dos clases utilizando la función de paridad (generalización de la función lógica XOR). En cada esquina se genera un cluster de 250 muestras aleatoriamente seleccionados desde una distribución gaussiana centrada. En total, el problema consta de 8.000 muestras donde se eligen aleatoriamente (manteniendo la distribución de las clases) 2.000 muestras de entrenamiento y 600 de validación. El problema contiene 500 características que se dividen en: 5 características útiles (ejes del hipercubo), 5 características redundantes, 10 características repetidas y 480 características irrelevantes.

En la Figuras 4.8 se observa los distintos porcentajes de clasificación de las primeras 100 características seleccionadas por cada uno de los métodos de selección de características. Los mejores porcentajes de clasificación ocurren con los criterios Forw y MIG. El *peak* de clasificación ocurre con las 5 primeras características, lo cual coincide correctamente con las 5 características principales (relevantes) que componen el hipercubo. A pesar que no se conocen cuáles son las características útiles, los resultados de Forw y MIG permiten predecir que éstas características son las relevantes para componer el hipercubo. En la Tabla 4.12 se presentan las primeras 5 características seleccionadas por los 5 criterios de selección de características. Como se observa en esta Tabla, la selección de la primera característica de MIG es distinta

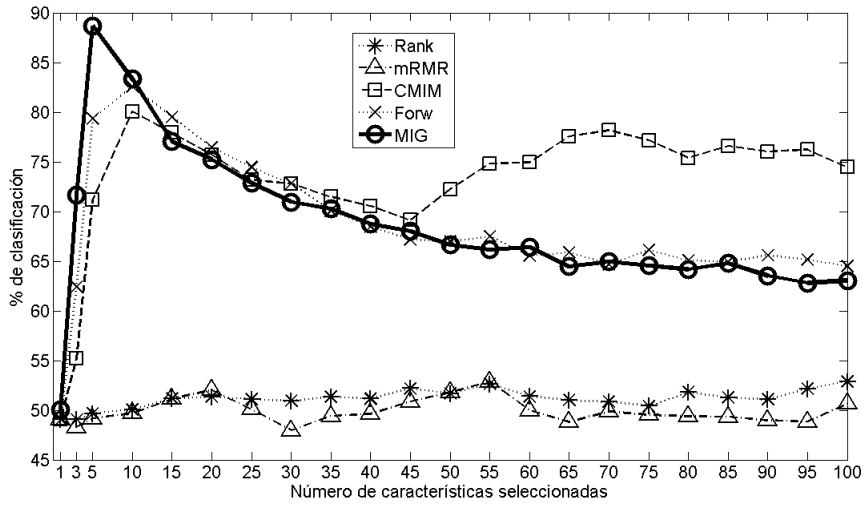


Figura 4.8: Promedio de clasificación de 10 rondas de validación cruzada del clasificador kNN utilizando 5 criterios de selección de características para el conjunto de datos Madelon.

al resto de métodos, permitiendo obtener un mayor porcentaje de clasificación en la primera característica.

Tabla 4.12: *Ranking* de las 5 primeras características entregadas por los criterios de selección de características evaluados en la base de datos Madelon. Los números en las celdas indican la posición real de la característica en la base de datos Madelon.

Ranking	Rank	mRMR	CMIM	Forw	MIG
1º	415	415	415	415	379
2º	460	408	476	476	339
3º	111	208	242	339	476
4º	217	97	339	379	319
5º	380	289	106	319	456

Los grupos de características se muestran en la matriz de interacción de las características presentada en la Figura 4.9, donde se observa la alta interacción de las características: $\{\{379, 339\}, \{339, 476\}, \{476, 319\}, \{319, 456\}\}$ (Ver Tabla 4.12 para ver el orden de las características seleccionadas por MIG). En este ordenamiento, se observa que los 4 pares de características presentan un valor de interacción semejante, por lo que podríamos considerar un grupo de 5 características en vez de 4 grupos de 2 características. Consecuentemente, el grupo de mayor interacción presentado en Madelon son justamente las primeras 5 características que entregan el *peak* de clasificación.

Arcene Esta base de datos fue utilizada en el desafío NIPS2003 de selección de características [161]. El problema consiste en detectar cáncer o no cáncer en los espectros SELDI obtenidos desde el National Center Institute (NCI) Eastern Virginia Medical School (EVMS). Los tipos de cáncer existentes en estas bases son cáncer de ovario y cáncer de prostata. Para

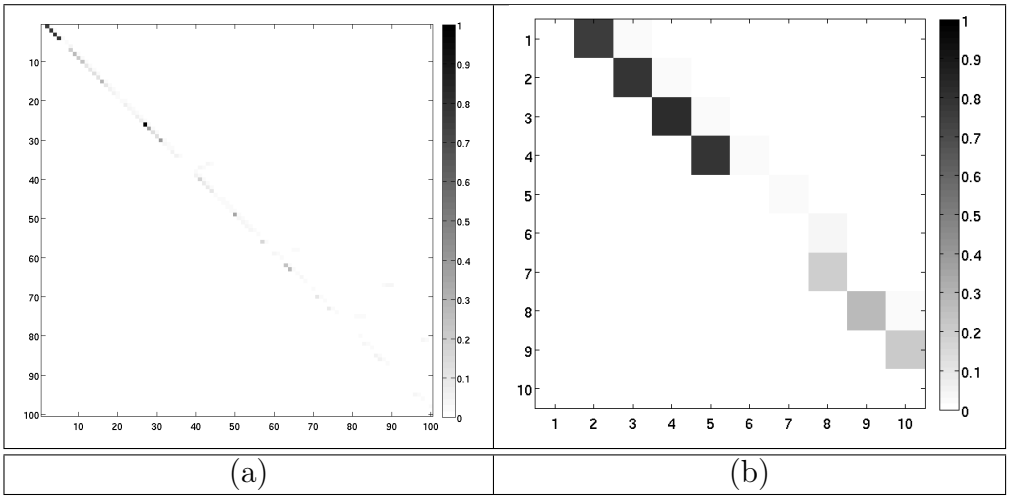


Figura 4.9: Matrices de interacción entre grupos de características para la base de datos Madelon. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de las 100 primeras características seleccionadas por MIG. (b) *Zoom* de la gráfica a, con la matriz de interacción de las 10 primeras características seleccionadas por MIG.

el desafío NIPS2003 se utilizaron 10.000 características de los espectros y la división de las muestras fue la siguiente: 100 muestras para entrenamiento, 100 muestras para validación.

En la Figuras 4.10 se observa que el desempeño de la primera característica seleccionada por MIG es superior a la primera característica del resto de los criterios. Además, se observa que aproximadamente hasta las 12 primeras primeras características seleccionadas, MIG presenta un mejor porcentaje de clasificación que el resto de criterios de selección de características. A pesar de la enorme cantidad de características de este problema (10.000), se demuestra la efectividad del criterio propuesto para detectar y agrupar características complementarias entre sí. En la Figura 4.11 se observa que la primera característica seleccionada por el criterio propuesto interactúan positivamente hasta con la 6^{o} característica, por lo cual se considera que el grupo de las 6 primeras características seleccionadas por MIG como un solo grupo de características relevantes. Otro posible grupo de características ocurre entre la 7^{o} y 9^{o} característica. Esta interacción es relativamente baja en comparación con otros máximos de interacción que se observan en la matriz y que son: $\{\{31^{o}, 32^{o}\}, \{33^{o}, 34^{o}\}, \{34^{o}, 35^{o}\}, \{35^{o}, 36^{o}\}, \{36^{o}, 37^{o}\}\}$. La interacción de estos grupos se relaciona con un mejor resultado de clasificación en las gráficos de desempeño de MIG usando kNN (Figura 4.10).

Lung Cancer Esta base de datos está disponible en el repositorio UCI y en ella se describe 3 tipos de patologías de cáncer al pulmón [156].

Los resultados mostrados en la Figura 4.12 y muestran que el desempeño de clasificación obtenido por MIG es comparable con el criterio Back. Este resultado muestra la ventaja de MIG para detectar características complementarias relevantes en forma secuencial hacia adelante, evitando los elevados costos computacionales que significa obtener características complementarias utilizando la estrategia hacia atrás. En la matriz de interac-

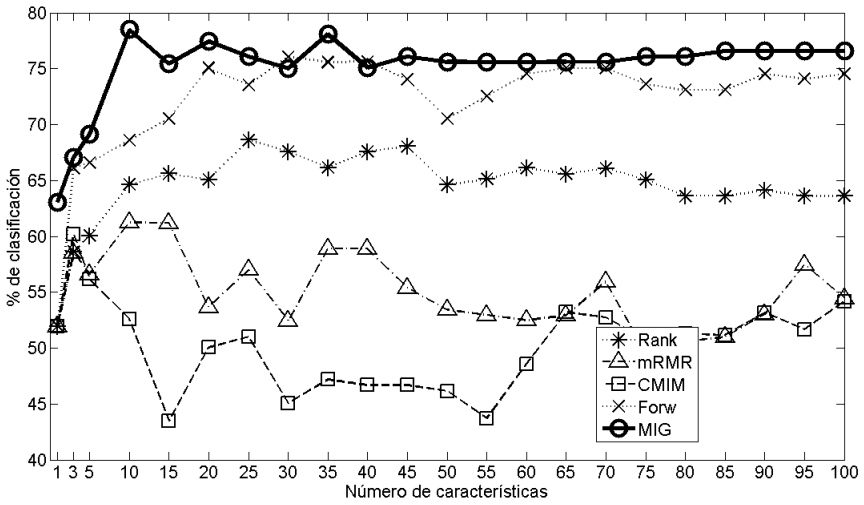


Figura 4.10: Promedio de clasificación de 10 rondas de validación cruzada del clasificador kNN utilizando distintos criterios de selección de características para el conjunto de datos Arcene.

ción de características de la Figura 4.13, se observan varios grupos de características relevantes entre la 4º y 25º característica. Los altos valores de interacción son consistentes con los elevados resultados de clasificación. Algunos de los principales grupos de características son: $\{\{22º, 23°, 24°, 25º\}, \{6º, 7°, 8°, 9°, 10º\}, \{4º, 5°, 6°, 7º\}, \{34º, 35°, 36º\}\}$. Es conveniente destacar que aunque el grupo $\{1°, 2°, 3°, 4°, 5º\}$ no presenta elevados valores de interacción, su aislamiento en relación al resto de las características, permite considerarlo como candidato al grupo de características relevantes.

Promoters Esta base fue obtenida del repositorio UCI [156], y corresponde a una secuencia del gen promoters desarrollada para evaluar un algoritmo híbrido de aprendizaje.

Los resultados mostrados en la Figura 4.14 muestran la importancia de la selección apropiada de la primera característica, como se observa con el resultado de los criterios Forw y MIG. En este base de datos, nuevamente se observa la ventaja del algoritmo propuesto para identificar y seleccionar características relevantes que tienen alta sinergia.

La matriz de sinergia (Figura 4.15) muestra un leve nivel de interacción entre el conjunto de características {1º-3º}. A pesar que el valor de interacción no es alto, es importante mencionar que ésta interacción esta bien aislada del resto de las otras interacciones. Las primeras interacciones fuertes ocurren en los subconjuntos: {{9º,10º}, {9º,10º,11º}, {12º,13º,14º}} se observa cierta interacción que aumenta levemente el porcentaje de clasificación.

Splice Esta base fue obtenida del repositorio UCI [156]. El nombre Splice corresponde a las zonas de unión *(splice junction)*, que son puntos de una secuencia de ADN (Ácido DesoxirriboNucleico) donde se remueve ADN "innecesario" durante el proceso de creación de proteínas en organismos superiores. El problema de esta base de datos consiste en conocer

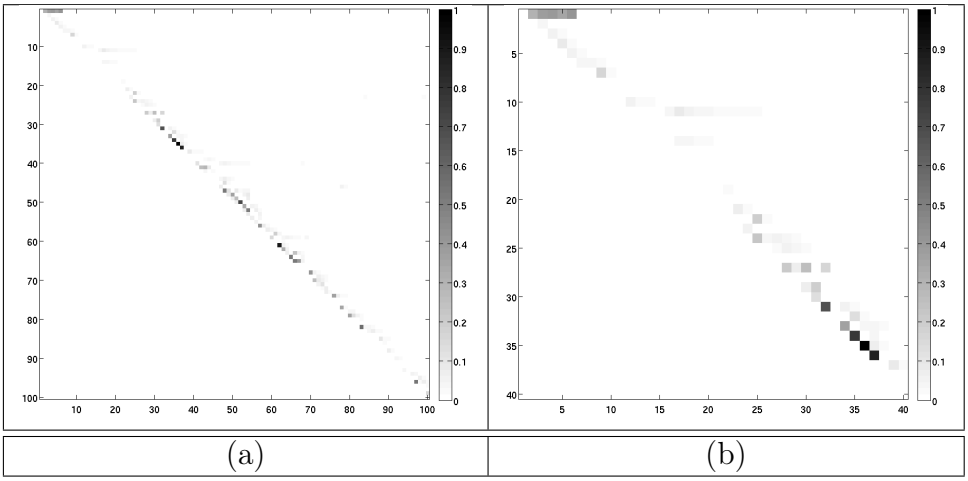


Figura 4.11: Matrices de interacción entre grupos de características para la base de datos Arcene. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de las 100 primeras características seleccionadas por MIG. (b) *Zoom* de la gráfica a, matriz de interacción de las 40 primeras características seleccionadas por MIG.

los límites entre el exón (parte del ADN que no es separada) y el intrón (parte del ADN que es separada). Las tres clases son: (1) Límite entre el Exón/Intrón, (2) límite entre el Intrón/exón, (3) No hay zona de separación.

De acuerdo a la forma en que está diseñado el problema, es altamente probable que varias características interactúen sinérgicamente para identificar los límites entre el exón/intrón (y viceversa). Esta última afirmación es corroborada al observar la matriz de sinergia (Tabla 4.16), en donde se observa una gran cantidad de interacciones fuertes (valores de interacción que existen entre las características).

El efecto de estas interacciones se ve reflejado en la clasificación, ya que en la matriz de sinergia observamos que los *peaks* ocurren entre las características en los subconjuntos $\{4^o, 5^o\}, \{5^o, 6^o\}, \{6^o, 7^o\}$. La Figura 4.17 también entregan *peaks* de clasificaciones en los mismos subconjuntos de características consideradas relevantes.

Otro aspecto importante a mencionar es la selección de la primera característica. Una importante cualidad del criterio propuesto es que permite identificar el verdadero aporte informativo que tiene una característica en relación a la característica de clase, al considerar la cantidad de información mutua que tiene una característica en el contexto del resto de las características. En este problema, donde se busca identificar el límite entre exón/intrón, resulta conveniente ver el desempeño de una característica para identificar el límite en el contexto del resto de las características. La selección de la primera característica del método propuesto MIG presenta un poder de clasificación significativamente superior al resto de los criterios de selección de características analizados.

Semeion Esta base fue obtenida del repositorio UCI [156]. La tarea de esta base es reconocer los números manuscritos del cero al nueve que han sido escritos por 80 personas. Los

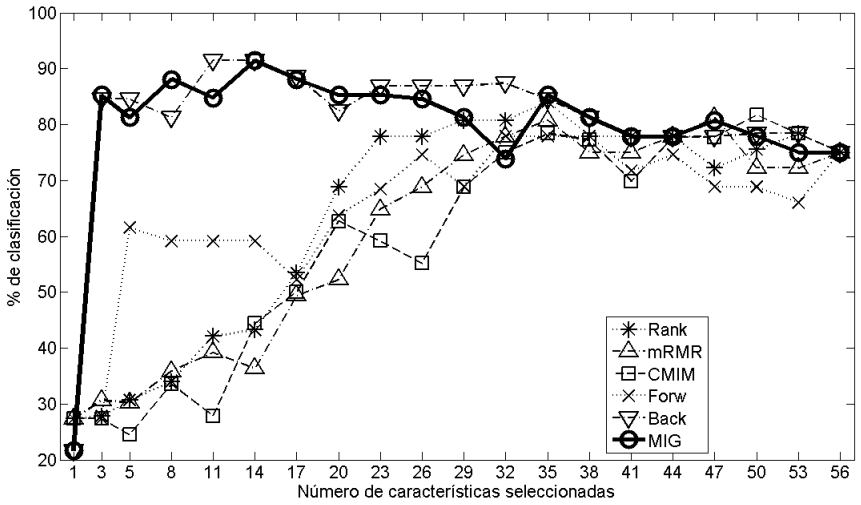


Figura 4.12: Promedio de clasificación de 10 rondas de validación cruzada del clasificador kNN utilizando distintos criterios de selección de características para la base de datos Lung Cancer.

números manuscritos son digitalizados en imágenes de 16×16 pixeles, donde cada pixel es una característica del problema.

Como se observa de la Figura 4.18, los resultados de clasificación de todos los criterios de selección de características son semejantes para este problema. Solo algunas características son relevantes para identificar la la diferencia entre el cero y el nueve (por ejemplo, los pixeles que están en medio de la imagen). Esto último se corrobora al observar que con muy pocas características (pixeles) se alcanza un nivel de clasificación elevado y se mantiene prácticamente constante al aumentar el número de características.

La matriz de sinergia (Figura 4.19) muestra que las primeras características no presentan gran interacción entre ellas. Recién en la característica *rankeada* número 100 se observa pequeños grupos de características que son relevantes, sin embargo, no aportan mayor informaciòn a la clasificaciòn (aproximadamente a partir de la característica 30 se observa que el porcentaje de clasificación se mantiene constante).

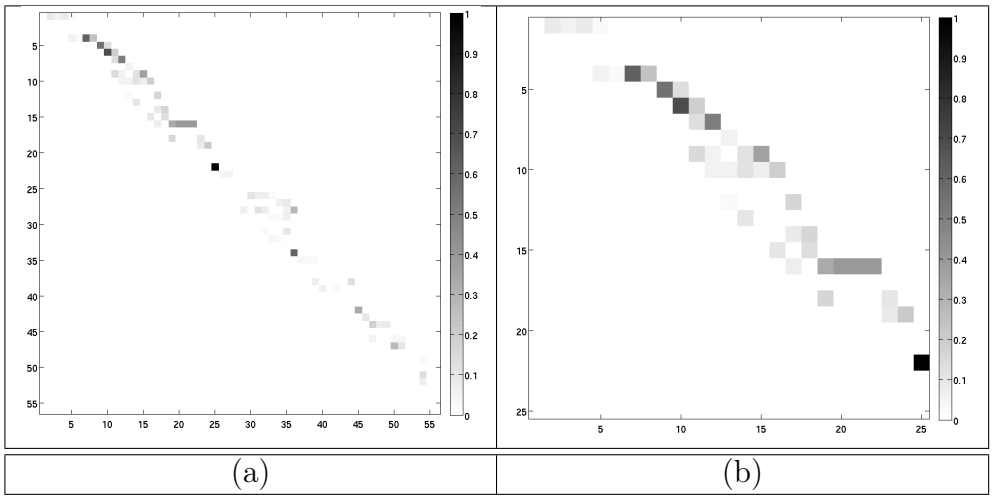


Figura 4.13: Matrices de interacción entre grupos de características para la base de datos Lung Cancer. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de las 56 (todas) características seleccionadas por MIG. (b) *Zoom* de la gráfica a, matriz de interacción de las 25 primeras características seleccionadas por MIG.

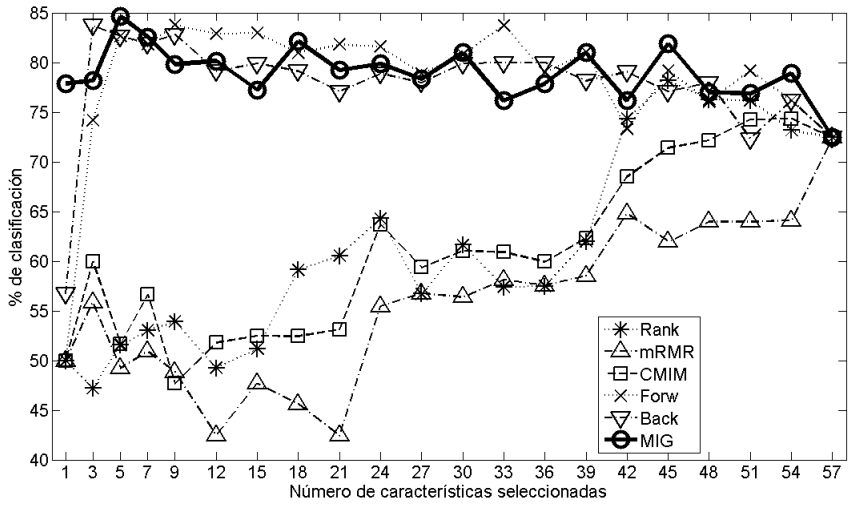


Figura 4.14: Promedio de clasificación de 10 rondas de validación cruzada del clasificador $k - NN$ utilizando distintos criterios de selección de características para la base de datos Promoters.

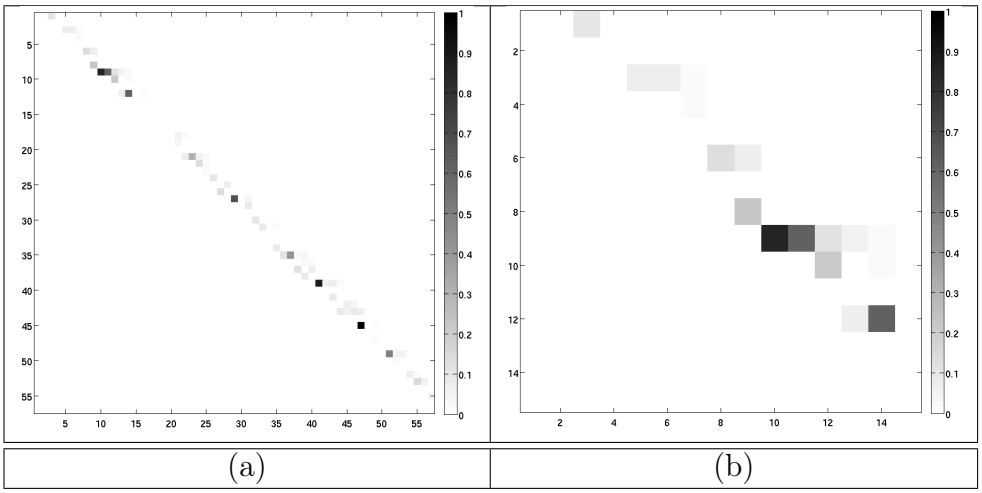


Figura 4.15: Matrices de interacción entre grupos de características para la base de datos Promoters. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de las características seleccionadas por MIG. (b) *Zoom* de las 10 primeras características seleccionadas por MIG.

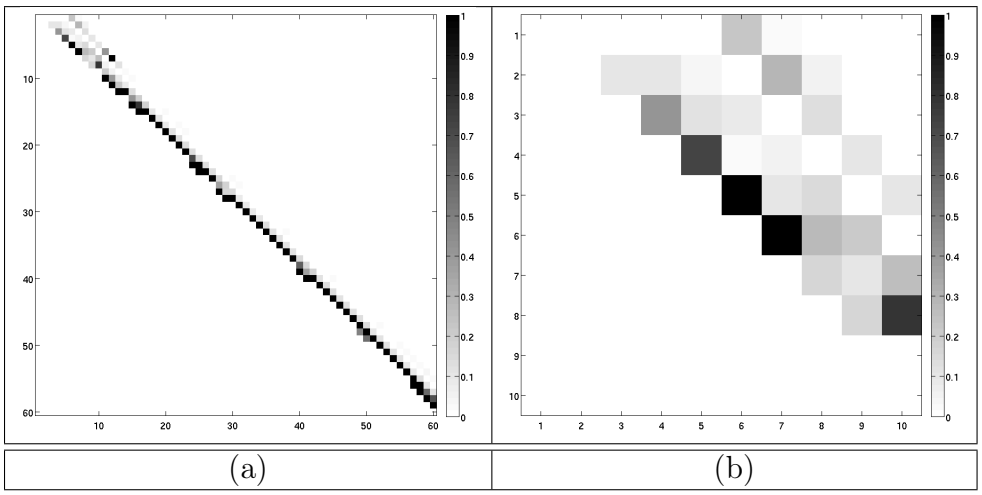


Figura 4.16: Matrices de interacción entre grupos de características para la base de datos Splice. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de las características ordenadas y seleccionadas por MIG. (b) *Zoom* de las 10 primeras características seleccionadas por MIG.

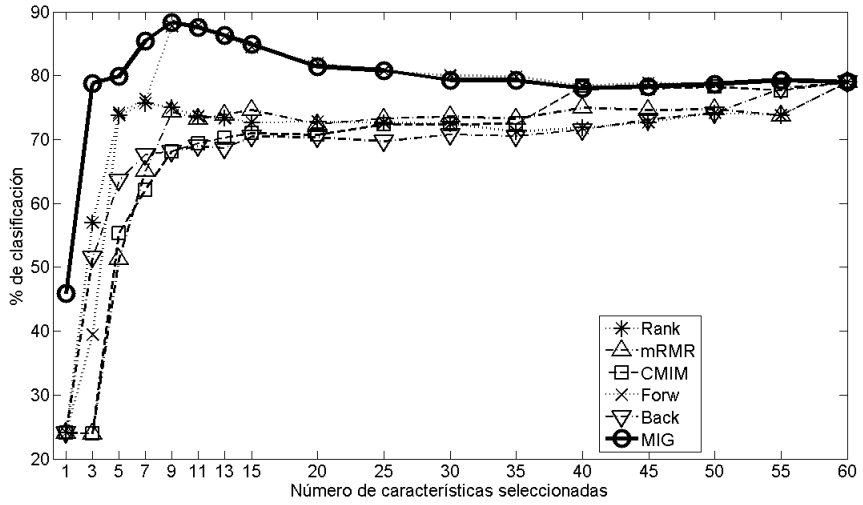


Figura 4.17: Promedio de clasificación de 10 rondas de validación cruzada del clasificador kNN utilizando distintos criterios de selección de características para la base de datos Splice.

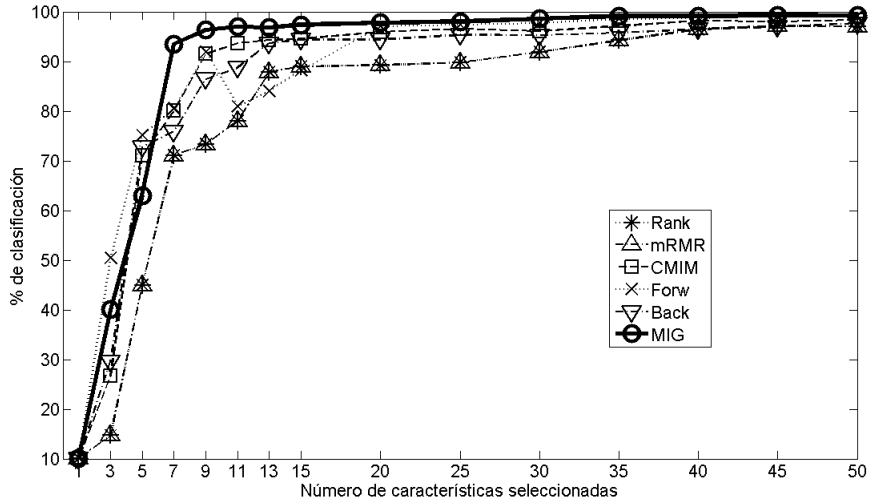


Figura 4.18: Promedio de clasificación de 10 rondas de validación cruzada del clasificador kNN utilizando distintos criterios de selección de características para la base de datos Semeion.

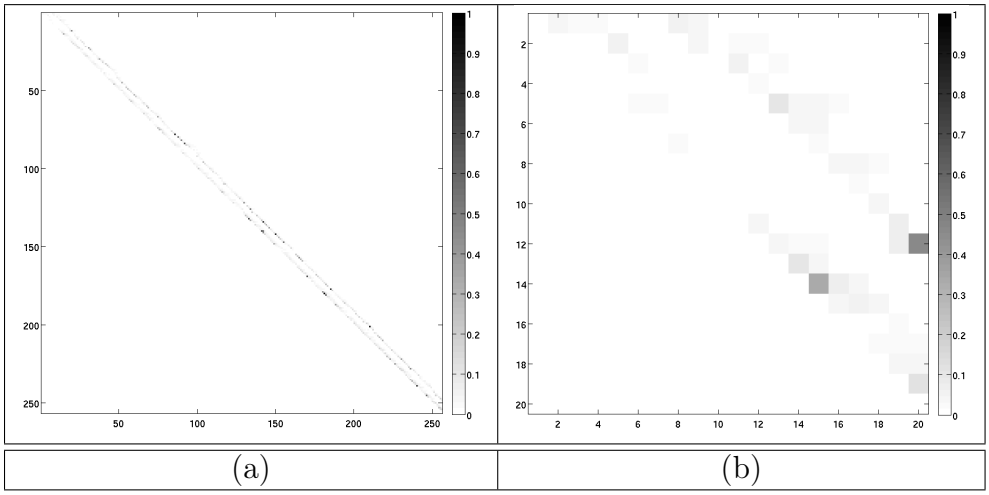


Figura 4.19: Matrices de interacción entre grupos de características para la base de datos Semeion. Los ejes de las matrices indican la posición del *ranking* de las características seleccionadas. Los valores de las interacciones fueron normalizados de 0 a 1. (a) Matriz de interacción de todas las características seleccionadas por MIG. (b) *Zoom* de las 20 primeras características seleccionadas por MIG.

Capítulo 5

Conclusiones

El objetivo de este trabajo fue obtener un criterio que identifique y seleccione grupos complementarios de características utilizando información mutua. La motivación fue obtener la mayor cantidad de información directamente desde los datos en la etapa de preprocesamiento de éstos. Las tradicionales metodologías de selección de características se concentran en determinar cuáles características son relevantes o redundantes, sin tomar en cuenta la interacción de complementariedad. La detección de las características complementarias e identificación de los grupos de éstas, resulta fundamental en la correcta comprensión del proceso que genera los datos.

La búsqueda del criterio propuesto permitió un ordenamiento y estructuración de los actuales métodos de selección de características, detectando con ello la importancia del término de interacción existente entre las características para predecir la variable de salida. Para relevar la importancia de este término, se desarrolló una metodología para identificar cuantitativamente el aporte de información de cada tipo de interacción entre características, encontrando y definiendo los conceptos de: relevancia única, redundancia y complementariedad o sinergia, siendo este último el concepto base para la identificación de grupos complementarios. Otro de los aportes de este trabajo consistió en definir e identificar los límites de información que una característica tiene de C de acuerdo a cómo interactúe con el resto de características. Como parte de esta formulación de la importancia del concepto de complementariedad fue posible identificar la limitación en uno de los criterios más utilizados en selección de características utilizando información mutua, el CMIM, en el cual se propuso una mejora de este criterio llamada CMIM2.

Considerando las enormes potencialidades que se pueden vislumbrar de este esquema de identificación de interacciones entre características, se espera en futuros trabajos mejorar la eficiencia y rapidez en la identificación de grupos relevantes, en algunos aspectos tales como: mejoramiento del cálculo de entropía, optimización para trabajar con bases de datos de miles de características, detección de grupos de características *online*, entre otros trabajos.

Los principales aportes de este trabajo de tesis son:

1. Marco teórico en métodos de selección de características basados en información mutua.

2. Detección de mínima información o información fundamental.

3. Cuantificación de la interacción: Relevancia única, redundancia, complementariedad.

4. Criterio de selección e identificación de grupos complementarios de características utilizando información mutua.

   a) MIG (Maximización de la Interacción Grupal): Criterio de selección de características basado en información mutua que permite seleccionar contiguamente en el *ranking* las características complementarias.

   b) RTsinergia: Criterio para cuantificar e identificar grupos complementarios de características.

A continuación se presenta las discusiones y conclusiones de los aportes realizados por este trabajo de investigación.

5.1. Marco teórico para la clasificación de métodos de selección de características basado en información mutua

En esta propuesta se propone una descomposición de la información mutua que permite derivar la mayoría de los criterios de selección de características existentes en la literatura. Este marco propuesto nace paralelamente a la propuesta de Brown. Sin embargo, la propuesta realizada en esta tesis permite derivar varios criterios que la propuesta de Brown no puede explicar. La ventaja del marco propuesto radica en que es posible utilizar distintos tipos de funciones no-lineales para agrupar las diferentes forma de descomponer la información mutua condicional $I(f_i; C|S)$. Dependiendo del número de características utilizadas en el cálculo de las componentes de información mutua, es posible clasificar los criterios de selección de características de acuerdo al nivel de aproximación que tienen con respecto a $I(f_i; C|S)$. De esta forma, los criterios tales como MIFS, mRMR, Rank, IMIFS son clasificadas como aproximaciones de primer orden de $I(f_i; C|S)$. La característica de estos criterios es que no utilizan el término de complementariedad. Los buenos resultados de estos criterios son explicados si es que las bases datos analizadas por estos criterios poseen características que son independientes entre ellas. Las aproximaciones de segundo orden son aquellos criterios que utilizan dos características en el cálculo de información mutua (aparte de la clase). Entre los criterios de segundo orden se encontraron: CMIM, CMIM2, CMIFS. Estos criterios tiene la capacidad de detectar solamente la complementariedad entre pares de características.

De acuerdo cómo se obtiene la información de las características de la variable estos se pueden clasificar en lineales y no lineales

5.2. Mejoramiento de CMIM

Se ha propuesto un algoritmo de información mutua condicional mejorado con respecto a CMIM para la selección de características. El nuevo algoritmo, llamado CMIM2, es capaz de detectar pares de características relevantes que actúan de forma complementaria en la predicción de la clase. Los resultados experimentales en bases de datos artificiales y UCI muestran que el algoritmo propuesto supera el algoritmo original CMIM.

Una ventaja del enfoque propuesto es que es posible mejorar el criterio CMIM considerando una aproximación de segundo orden de la ecuación (4.5). Sería interesante estudiar cómo influye en la selección de las características relevantes que actúan complementariamente en grupos de tres o más características. Estas ideas pueden ser utilizadas para identificar y formalizar nuevos niveles de interacción entre las características, más allá de las definiciones tradicionales de la relevancia y la redundancia.

5.3. Selección de grupos complementarios de características

El criterio propuesto de selección e identificación de grupos complementarios de característica utilizando información mutua, permite encontrar estos grupos sin la necesidad de evaluar en forma exhaustiva todas las posibles combinaciones de subconjuntos. Las ventajas del criterio propuesto son:

1. Utiliza una estrategia de búsqueda *greedy* hacia adelante, lo cual permite que el método sea muy rápido en la búsqueda de características.
2. El criterio posee las ventajas del criterio *greedy* hacia atrás pero evita el costo computacional de ir descartando características. Esta cualidad está asociada a la capacidad del algoritmo propuesto de capturar las interacciones de las características y a la consideración de la información fundamental.
3. Los resultados obtenidos por el criterio propuesto son superiores o iguales a los métodos de selección de característica existentes en la literatura.
4. La detección de grupos de características presenta ventajas en el desempeño de clasificación.
5. La detección de grupos de características permite comprender mejor el proceso que genera los datos.

La posibilidad que tiene el criterio propuesto para ordenar y seleccionar las características de acuerdo a la capacidad de predecir su características de la clase C, permite, por ejemplo, que la selección de 100 características en la base Madelon se pueda obtener en tiempos razonables (3000 seg.), permitiendo además la detección de grupos complementarios de características relevantes.

5.4. Futuras direcciones de investigación

En esta sección se presentan algunas líneas de investigación, posibles trabajos futuros y desafíos que han sido vislumbrados en esta investigación en el campo de la selección de características, en particular desde el punto de vista de los métodos basados en teoría de la información.

1. *Mejoramiento de la eficacia y la eficiencia de los métodos de selección de características basados en teoría de la información en espacios de alta dimensionalidad.* El tiempo computacional depende de la estrategia y del criterio de evaluación [118]. Actualmente estamos ingresando a la era del *Big Data*, donde se hace urgente el desarrollo de métodos de selección de características más rápidos que sean capaces de trabajar con millones de características y billones de muestras. Un importante desafío es el desarrollo de métodos más eficientes para estimar la información mutua en espacios de alta dimensionalidad. Otro punto de interés es la determinación automática del tamaño óptimo del subconjunto de características, considerando que muchos métodos no tienen un criterio de detención. El desarrollo de nuevas estrategias que superen la optimización *greedy* es otra interesante línea de investigación.

2. *Investigación entre la información mutua y el error de clasificación de Bayes.* Hasta el momento solo se han encontrado los límites superior e inferior del error de clasificación para el caso de una característica aleatoria y la clase. La extensión de estos resultados a la relación entre subconjuntos de características y la clase son de interés para conocer y mejorar las estimaciones de información mutua, como también comparar correctamente el desempeño de subconjunto de características de distinto tamaño.

3. *Efecto del tamaño finito de muestras sobre los criterios estadísticos empleados y la estimación de la información mutua.* Guyon *et al.* [118], argumentan que los subconjuntos de características que no son suficientes pueden tener mejor desempeño que los subconjuntos de características suficientes. Un caso donde se observa este efecto es el dominio de la bioinformática, donde es común tener una enorme cantidad de dimensiones de entrada, y un pequeño tamaño de muestras [162].

4. *Desarrollo de un marco teórico que estudie las relaciones entre la selección de características y el descubrimiento causal.* Guyon *et al.* [163], investiga la selección causal de características. Los autores indican que el conocimiento de las relaciones causales puede beneficiar la selección de características y viceversa. El desafío es el desarrollo de un algoritmo eficiente de *Markov Blanket* de inducción para distribuciones no-fidedignas.

5. *Desarrollo de nuevos criterios de dependencia estadística superiores a la correlación y la información mutua.* Seth y Principe [164] realizaron una revisión de los postulados de medidas de dependencia de acuerdo a Renyi en el contexto de la selección de características. Un importante tópico es la normalización, ya que una medida de dependencia aplicada sobre diferentes tipos de características aleatorias debe ser comparable. No existe una teoría estándar acerca de la normalización de la información mutua [144, 165]. Otro problema es que las medidas de dependencia deben ser robustas cuando se tienen pocas muestras en el sentido de mantener las propiedades deseadas de esta medida. Seth

y Principe [164] y Frénay *et al.* [166] demostraron que esta propiedad no es satisfecha por estimadores de información mutua, porque éstos no alcanzan el máximo valor bajo dependencias estrictas, y no son invariantes a transformaciones uno a uno.

Bibliografía

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[2] A. Araúzo-Azofra and J. Benítez, "Empirical study of feature selection methods in classification," *Hybrid Intelligent Systems, 2008. HIS '08. Eighth International Conference on*, pp. 584–589, Sept. 2008.

[3] F. Fleuret and I. Guyon, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[4] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, pp. 1226–1238, Aug. 2005.

[5] A. R. Webb, *Statistical Pattern Recognition.* John Wiley & Song, Ltd., 2nd ed., 2002.

[6] D. A. Bell and H. Wang, "A formalism for relevance and its application in feature subset selection," *Machine Learning*, vol. 41, no. 2, pp. 175–195, 2000.

[7] I. Kojadinovic, "Relevance measures for subset variable selection in regression problems based on k-additive mutual information," *Computational Statistics and Data Analysis*, vol. 49, pp. 1205–1227, Jun. 2005.

[8] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[9] W. Wienholt and B. Sendhoff, "How to determine the redundancy of noisy chaotic time series," *International Journal of Bifurcation and Chaos*, vol. 6, no. 1, pp. 101–117, 1996.

[10] A. Jakulin and I. Bratko, "Quantifying and visualizing attribute interactions," *CoRR*, vol. cs.AI/0308002, 2003.

[11] Z. Zhao and H. Liu, "Searching for interacting features in subset selection," *Intelligent Data Analysis*, vol. 13, pp. 207–228, Apr. 2009.

[12] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on*, vol. 5, pp. 537–550, Jul. 1994.

[13] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[14] P. Meyer, C. Schretter, and G. Bontempi, "Information-theoretic feature selection in microarray data using variable complementarity," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, pp. 261–274, Jun. 2008.

[15] S. Davies and S. Russell, "Np-completeness of searches for smallest possible feature sets," in *Intelligent Relevance, Association for the Advancement of Artificial Intelligence Symposium on*, pp. 37–39, AAAI Press, 1994.

[16] M. Charikar, V. Guruswami, R. Kumar, S. Rajagopalan, and A. Sahai, "Combinatorial feature selection problems," *Foundations of Computer Science, Annual IEEE Symposium on*, vol. 0, p. 631, 2000.

[17] P. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *Computers, IEEE Transactions on*, vol. C-26, pp. 917–922, Sept. 1977.

[18] T. J. Gawne and B. J. Richmond, "How independent are the messages carried by adjacent inferior temporal cortical neurons?," *The Journal of neuroscience*, vol. 13, no. 7, pp. 2758–2771, 1993.

[19] P. E. Latham and S. Nirenberg, "Synergy, redundancy, and independence in population codes, revisited," *The Journal of neuroscience*, vol. 25, no. 21, pp. 5195–5206, 2005.

[20] S. Nirenberg and P. E. Latham, "Decoding neuronal spike trains: How important are correlations?," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 7348–7353, 2003.

[21] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, "Spatio-temporal correlations and visual signalling in a complete neuronal population," *Nature*, vol. 454, no. 7207, pp. 995–999, 2008.

[22] E. Schneidman, W. Bialek, and M. J. Berry, "Synergy, redundancy, and independence in population codes," *the Journal of Neuroscience*, vol. 23, no. 37, pp. 11539–11553, 2003.

[23] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, March 2003.

[24] D. D. Lewis, "Feature selection and feature extraction for text categorization," in *Proceedings of Speech and Natural Language Workshop*, pp. 212–217, Morgan Kaufmann, 1992.

[25] L.-Q. Qiu, R.-Y. Zhao, G. Zhou, and S.-W. Yi, "An extensive empirical study of feature selection for text categorization," *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on*, pp. 312–315, May 2008.

[26] S.-Y. Kung and M.-W. Mak, "Feature selection for self-supervised classification with applications to microarray and sequence data," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, pp. 297–309, June 2008.

[27] C. Sima and E. R. Dougherty, "What should be expected from feature selection in small-sample settings," *Bioinformatics*, vol. 22, no. 19, pp. 2430–2436, 2006.

[28] X. Xu and A. Zhang, *Boost feature subset selection: A new gene selection algorithm for microarray dataset*, pp. 670–677. 2006.

[29] B. Wang, Y. Jia, and S. Yang, "Forward semi-supervised feature selection based on relevant set correlation," *Computer Science and Software Engineering, 2008 International Conference on*, vol. 4, pp. 210–213, Dec. 2008.

[30] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, pp. 153–158, Feb 1997.

[31] P. Somol, P. Pudil, J. Novovičová, and P. Paclık, "Adaptive floating search methods in feature selection," *Pattern recognition letters*, vol. 20, no. 11, pp. 1157–1163, 1999.

[32] P. Somol, J. Novovičová, and P. Pudil, "Notes on the evolution of feature selection methodology," *Kybernetika*, vol. 43, no. 5, pp. 713–730, 2007.

[33] P. Somol and P. Pudil, "Oscillating search algorithms for feature selection," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 2, pp. 406–409 vol.2, 2000.

[34] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: A unifying framework for information theoretic feature selection," *Journal of Machine Learning Research*, vol. 13, pp. 27–66, Mar. 2012.

[35] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 2006.

[36] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 625–56, Jul., Oct. 1948.

[37] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Publishing Company, Incorporated, 1st ed., 2010.

[38] W. McGill, "Multivariate information transmission," *Psychometrika*, vol. 19, pp. 97–116, Jun. 1954.

[39] R. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA: The MIT Press, 1961.

[40] T. S. Han, "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, no. 1, pp. 26 – 45, 1980.

[41] S. Srinivasa, "A review on multivariate mutual information," *Univ. of Notre Dame, Notre Dame, Indiana*, vol. 2, pp. 1–6, 2005.

[42] A. P. Hekstra and F. M. Willems, "Dependence balance bounds for single-output two-way channels," *Information Theory, IEEE Transactions on*, vol. 35, no. 1, pp. 44–53, 1989.

[43] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[44] I. Hero, A.O., B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *Signal Processing Magazine, IEEE*, vol. 19, pp. 85–95, Sep 2002.

[45] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector.," *Probl. Inf. Transm.*, vol. 23, no. 1-2, pp. 95–101, 1987.

[46] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information.," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 69, June 2004.

[47] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[48] M. P. Wachowiak, R. Smolíková, G. D. Tourassi, and A. S. Elmaghraby, "Estimation of generalized entropies with sample spacing," *Pattern Anal. Appl.*, vol. 8, no. 1, pp. 95–101, 2005.

[49] L. Batina, B. Gierlichs, E. Prouff, M. Rivain, F.-X. Standaert, and N. Veyrat-Charvillon, "Mutual information analysis: a comprehensive study," *Journal of Cryptology*, vol. 24, no. 2, pp. 269–291, 2011.

[50] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–248, 1989.

[51] Y.-I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, p. 2318, 1995.

[52] B. Póczos and J. G. Schneider, "Nonparametric estimation of conditional information and divergences," in *International Conference on Artificial Intelligence and Statistics*, pp. 914–923, 2012.

[53] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 1667–1671, Dec 2002.

[54] H. Liu, L. Wasserman, and J. D. Lafferty, "Exponential concentration for mutual information estimation with application to forests," in *Advances in Neural Information Processing Systems*, pp. 2537–2545, 2012.

[55] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Phys. Rev. A*, vol. 33, pp. 1134–1140, Feb 1986.

[56] A. Hero and O. Michel, "Estimation of renyi information divergence via pruned minimal spanning trees," *Higher-Order Statistics. Proceedings of the IEEE Signal Processing Workshop on*, pp. 264–268, 1999.

[57] M. Sabuncu and P. Ramadge, "Using spanning graphs for efficient image registration," *Image Processing, IEEE Transactions on*, vol. 17, pp. 788–797, May 2008.

[58] J. Costa and A. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *Signal Processing, IEEE Transactions on*, vol. 52, pp. 2210–2221, Aug. 2004.

[59] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* (Chap. 2-4), Wiley-Interscience, 2nd ed., 2001.

[60] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, pp. 1065–1076, 1962.

[61] M. Rosenblatt *et al.*, "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, 1956.

[62] B. Bonev, *Feature selection based on information theory.* PhD thesis, Universidad de Alicante, 2010.

[63] N. Leonenko, L. Pronzato, and V. Savani, "A class of rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36, pp. 2153–2182, 10 2008.

[64] B. Bonev, F. Escolano, and M. Cazorla, "Feature selection, mutual information, and the classification of high-dimensional patterns: Applications to image classification and microarray data analysis," *Pattern Analysis and Applications*, vol. 11, pp. 309–319, Jan. 2008.

[65] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E*, vol. 69, p. 066138, Jun, 2004.

[66] G. Doquire, B. Frénay, M. Verleysen, *et al.*, "Risk estimation and feature selection," in *European Symposium on Artificial Neural Networks (ESANN 2013)*, 2013.

[67] J. De Souza, R. Do Carmo, and G. De Campos, "A novel approach for integrating feature and instance selection," *Machine Learning and Cybernetics, 2008 International Conference on*, vol. 1, pp. 374–379, July 2008.

[68] H. Liu, E. Dougherty, J. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu, and G. Forman, "Evolving feature selection," *Intelligent Systems, IEEE*, vol. 20, pp. 64–76, Nov.-Dec. 2005.

[69] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features," in *Artificial Intelligence, Proceedings of the Ninth National Conference on*, pp. 547–552, AAAI Press, 1991.

[70] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *ML92: Proceedings of the ninth international workshop on Machine learning*, (San Francisco, CA, USA), pp. 249–256, Morgan Kaufmann Publishers Inc., 1992.

[71] D. Koller and M. Sahami, "Toward optimal feature selection," Technical Report 1996-77, Stanford InfoLab, Feb. 1996.

[72] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (Amsterdam, The Netherlands), pp. 407–414, ACM, 2007.

[73] J.-J. Jin-Jie Huang, N. LV, S.-Q. LI, and Y.-Z. CAI, "Feature selection for classificatory analysis based on information-theoretic criteria," *Acta Automatica Sinica*, vol. 34, no. 3, pp. 383 – 392, 2008.

[74] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar, "Ranking a random feature for variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1399–1414, 2003.

[75] H.-L. Wei and S. Billings, "Feature subset selection and ranking for data dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, pp. 162–166, Jan. 2007.

[76] A. Verikas and M. Bacauskiene, "Feature selection with neural networks," *Pattern Recognition Letters*, vol. 23, pp. 1323–1335, Sep 2002.

[77] X. Wang, J. Yang, X. Teng, W. Xia, and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, no. 4, pp. 459 – 471, 2007.

[78] L. Wang, "Feature selection with kernel class separability," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, pp. 1534–1546, Sept. 2008.

[79] B. Schölkopf and A. J. Smola, *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, 2002.

[80] Y. Tang, J. Gao, and G. Cui, "Feature selection based on kernel pattern similarity," *Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on*, pp. 947–954, June 2008.

[81] K. Grabczewski and N. Jankowski, "Feature selection with decision tree criterion," *Hybrid Intelligent Systems, 2005. HIS '05. Fifth International Conference on*, pp. 6 pp.–, Nov. 2005.

[82] M. Zaffalon, "Robust discovery of tree-dependency structures," in *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, (The Netherlands), pp. 394–403, Shaker Publishing, de Cooman, G., Fine, T., 2001.

[83] X. Zhou and T. Dillon, "A statistical-heuristic feature selection criterion for decision tree induction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, pp. 834–841, Aug 1991.

[84] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception* (M. Prokopenko, ed.), vol. 9 of *Emergence, Complexity and Computation*, pp. 159–190, Springer Berlin Heidelberg, 2014.

[85] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Exploiting synergistic and redundant features for multimedia document classification," in *32nd Annual Conference of the German Classification Society-Advances in Data Analysis, Data Handling and Business Intelligence (GfKl 2008), Hamburg, Germany*, 2008.

[86] D. Anastassiou, "Computational analysis of the synergy among multiple interacting genes," *Molecular systems biology*, vol. 3, no. 1, 2007.

[87] N. Brenner, S. Strong, R. Koberle, W. Bialek, and R. Steveninck, "Synergy in a neural code," *Neural Computation*, vol. 12, no. 7, pp. 1531–1552, 2000.

[88] N. Timme, W. Alford, B. Flecker, and J. Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalist's perspective," *Journal of Computational Neuroscience*, vol. 36, no. 2, pp. 119–140, 2014.

[89] S. Yang and J. Gu, "Feature selection based on mutual information and redundancy-synergy coefficient," *Journal of Zhejiang University - Science A*, vol. 5, pp. 1382–1391, Nov 2004.

[90] J. Lizier, B. Flecker, and P. Williams, "Towards a synergy-based approach to measuring information modification," in *Artificial Life (ALIFE), 2013 IEEE Symposium on*, pp. 43–51, April 2013.

[91] H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li, "Conditional mutual information-based feature selection analyzing for synergy and redundancy," *Electronics and Telecommunications Research Institute*, vol. 33, pp. 210–218, Apr. 2011.

[92] G. Brown, "A new perspective for information theoretic feature selection," in *International Conference on Artificial Intelligence and Statistics*, pp. 49–56, 2009.

[93] H. Almuallim and T. G. Dietterich, "Efficient algorithms for identifying relevant features," in *Artificial Intelligence, Proceedings of the Ninth Canadian Conference on*, pp. 38–45, Morgan Kaufmann, 1992.

[94] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, pp. 252–264, Mar. 1991.

[95] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, pp. 900–912, Jul. 2004.

[96] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *Neural Networks, IEEE Transactions on*, vol. 13, pp. 143–159, Jan. 2002.

[97] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *Proc. 37th Annu. Allerton Conf. Communication, Control and Computing*, 1999.

[98] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, first ed., 1961.

[99] G. V. Trunk, "A problem of dimensionality: A simple example," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-1, pp. 306–307, Jul. 1979.

[100] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development*, vol. 4, pp. 66–82, jan. 1960.

[101] L. M. Bettencourt, V. Gintautas, and M. I. Ham, "Identification of functional information subgraphs in complex networks," *Physical review letters*, vol. 100, no. 23, p. 238701, 2008.

[102] N. J. Cerf and C. Adami, "Negative entropy and information in quantum mechanics," *Physical Review Letters*, vol. 79, no. 26, p. 5194, 1997.

[103] H. Matsuda, "Physical nature of higher-order mutual information: Intrinsic correlations and frustration," *Physical Review E*, vol. 62, no. 3, p. 3096, 2000.

[104] I. Gat and N. Tishby, "Synergy and redundancy among brain cells of behaving monkeys," *Advances in neural information processing systems*, pp. 111–117, 1999.

[105] A. A. Margolin, K. Wang, A. Califano, and I. Nemenman, "Multivariate dependence and genetic networks inference," *IET systems biology*, vol. 4, no. 6, pp. 428–440, 2010.

[106] W. Liu, G. Xu, and B. Chen, "The common information of n dependent random variables," in *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pp. 836–843, IEEE, 2010.

[107] G. Chechik and A. G. N. Tishby, "Group redundancy measures reveal redundancy reduction in the auditory pathway," in *Advances in Neural Information Processing Systems 14: Proceedings of the 2001 Neural Information Processing Systems (NIPS) Conference*, vol. 1, p. 173, Bradford Books, 2002.

[108] A. J. Bell, "The co-information lattice," *Analysis*, pp. 921–926, 2003.

[109] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 49–86, 1951.

[110] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1997.

[111] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Information and Control*, vol. 36, no. 2, pp. 133–156, 1978.

[112] T. Sun Han, "Multiple mutual informations and multiple interactions in frequency data," *Information and Control*, vol. 46, no. 1, pp. 26–45, 1980.

[113] S. A. Abdallah and M. D. Plumbley, "A measure of statistical complexity based on predictive information," *arXiv preprint arXiv:1012.1890*, 2010.

[114] E. Olbrich, N. Bertschinger, N. Ay, and J. Jost, "How should complexity scale with system size?," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 63, no. 3, pp. 407–415, 2008.

[115] V. Varadan, D. M. Miller, and D. Anastassiou, "Computational inference of the molecular logic for synaptic connectivity in c. elegans," *Bioinformatics*, vol. 22, no. 14, pp. e497–e506, 2006.

[116] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *CoRR*, vol. abs/1004.2515, 2010.

[117] I. Tsamardinos and C. F. Aliferis, "Towards principled feature selection: Relevancy, filters and wrappers," in *Artificial Intelligence and Statistics, Proceedings of the Ninth International Workshop on*, Morgan Kaufmann Publishers, 2003.

[118] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in *Feature Extraction, Foundations and Applications*, vol. 207 of *Studies in Fuzziness and Soft Computing*, pp. 1–25, Springer Berlin Heidelberg, 2006.

[119] I. Tsamardinos, C. F. Aliferis, and E. Statnikov, "Algorithms for large scale markov blanket discovery," in *The 16th International FLAIRS Conference, St*, pp. 376–380, AAAI Press, 2003.

[120] L. E. Brown and I. Tsamardinos, "Markov blanket-based variable selection in feature space," technical report dsl-08-01, Discovery Systems Laboratory, 2008.

[121] K. Torkkola, "Information-theoretic methods," in *Feature Extraction, Foundations and Applications*, vol. 207 of *Studies in Fuzziness and Soft Computing*, ch. 6, pp. 167–185, Springer Berlin Heidelberg, 2006.

[122] M. Hellman and J. Raviv, "Probability of error, equivocation, and the chernoff bound," *Information Theory, IEEE Transactions on*, vol. 16, pp. 368–372, Jul 1970.

[123] M. Feder and N. Merhav, "Relations between entropy and error probability," *Information Theory, IEEE Transactions on*, vol. 40, no. 1, pp. 259–266, 1994.

[124] S.-W. Ho and S. Verdú, "Conditional entropy and error probability," in *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pp. 1622–1626, July 2008.

[125] B.-G. Hu and H.-J. Xing, "Analytical bounds between entropy and error probability in binary classifications," *CoRR*, vol. abs/1205.6602, 2012.

[126] B.-G. Hu, "What are the differences between bayesian classifiers and mutual-information classifiers?," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, pp. 249–264, Feb 2014.

[127] P. Viswanath, P. Vinay kumar, V. Suresh Babu, and M. Venkateswara Kumar, "Generalized branch and bound algorithm for feature subset selection," *Conference on Computational Intelligence and Multimedia Applications, 2007. International Conference on*, vol. 2, pp. 214–218, Dec. 2007.

[128] A. Whitney, "A direct method of nonparametric measurement selection," *Computers, IEEE Transactions on*, vol. C-20, pp. 1100–1103, Sep. 1971.

[129] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *Information Theory, IEEE Transactions on*, vol. 9, pp. 11–17, Jan. 1963.

[130] S. D. Stearns, "On selecting features for pattern classifiers," in *Pattern Recognition, Proceedings of the 3rd International Conference on*, (Coronado, CA), pp. 71–75, 1976.

[131] M. Murty and V. Devi, "Nearest neighbour based classifiers," in *Pattern Recognition*, vol. 0 of *Undergraduate Topics in Computer Science*, pp. 48–85, Springer London, 2011.

[132] J. Wang and Q. Tao, "Machine learning: The state of the art," *IEEE Intelligent Systems*, vol. 23, no. 6, pp. 49–55, 2008.

[133] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *Advances in Neural Information Processing Systems 13*, pp. 668–674, MIT Press, 2000.

[134] Q. Xu, W. Pei, L. Yang, and Z. He, *Neural Information Processing*, vol. 4232, ch. Support Vector Machine Tree Based on Feature Selection, pp. 856–863. Lecture Notes in Computer Science, 2006.

[135] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[136] H. H. Yang and J. Moody, "Feature selection based on joint mutual information," in *In Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis*, pp. 22–25, 1999.

[137] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on parzen window," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 1667–1671, Dec 2002.

[138] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Computer Vision, 2003. Proceedings Ninth IEEE International Conference on*, vol. 1, pp. 281–288, Oct. 2003.

[139] A. Jakulin, *Machine learning based on attribute interactions*. PhD thesis, University of Ljubljana, Slovenia, 2005.

[140] D. Lin and X. Tang, "Conditional infomax learning: An integrated framework for feature extraction and fusion," in *Computer Vision - ECCV 2006*, vol. 3951 of *Lecture Notes in Computer Science*, pp. 68–82, Springer Berlin Heidelberg, 2006.

[141] P. E. Meyer and G. Bontempi, "On the use of variable complementarity for feature selection in cancer classification," in *Applications of Evolutionary Computing*, pp. 91–102, Springer, 2006.

[142] A. E. Akadi, A. E. Ouardighi, and D. Aboutajdine, "A powerful feature selection approach based on mutual information," *International Journal of Computer Science and Network Security*, vol. 8, pp. 116–121, April 2008.

[143] B. Guo and M. S. Nixon, "Gait feature subset selection by mutual information," *Systems, Man and Cybernetics, Part A, IEEE Transactions on*, vol. 39, pp. 36–46, Jan. 2009.

[144] P. A. Estévez, M. Tesmer, C. A. Pérez, and J. M. Zurada, "Normalized mutual information feature selection," *Neural Networks, IEEE Transactions on*, vol. 20, no. 2, pp. 189–201, 2009.

[145] J. R. Vergara and P. A. Estévez, "Cmim-2: An enhanced conditional mutual information maximization criterion for feature selection," *Applied Computer Science Methods*, vol. 2, no. 1, pp. 5–20, 2010.

[146] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. suppl 2, pp. S231–S240, 2002.

[147] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel, "Feature selection and ranking filter," in *Int. Conf. Artificial Neural Networks (ICANN) and Int. Conf. Neural Information Processing (ICONIP)*, pp. 251–254, Jun. 2003.

[148] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, no. 12, pp. 2394–2402, 2001.

[149] L. Vinh, S. Lee, Y.-T. Park, and B. d'Auriol, "A novel feature selection method based on normalized mutual information," *Applied Intelligence*, vol. 37, no. 1, pp. 100–120, 2012.

[150] M. G. Kendall, A. Stuart, and J. K. Ord, *Kendall's Advanced Theory of Statistics*. New York, NY, USA: Oxford University Press, Inc., 1987.

[151] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of markov blankets and direct causal relations," in *Proceedings of the 9th CAN SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 673–678, 2003.

[152] Y. Zeng, J. Luo, and S. Lin, "Classification using markov blanket for feature selection," in *Granular Computing, 2009, GRC'09. IEEE International Conference on*, pp. 743–747, IEEE, 2009.

[153] G. Wang and F. H. Lochovsky, "Feature selection with conditional mutual information maximin in text categorization," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management, New York, NY, USA*, pp. 342–349, 2004.

[154] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 2nd ed., 1992.

[155] S. Thrun, J. Bala, E. Bloedorn, I. Bratko, B. Cestnik, J. Cheng, K. D. Jong, S. Dzeroski, R. Hamann, K. Kaufman, S. Keller, I. Kononenko, J. Kreuziger, R. Michalski, T. Mitchell, P. Pachowicz, B. Roger, H. Vafaie, W. V. de Velde, W. Wenzel, J. Wnek, and J. Zhang, "The MONK's problems: A performance comparison of different learning algorithms," Tech. Rep. CMU-CS-91-197, Carnegie Mellon University, Computer Science Department, Pittsburgh, PA, 1991.

[156] A. Asuncion and D. Newman, "UCI machine learning repository." http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, School of Information and Computer Sciences, 2007.

[157] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 2008.

[158] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *American Journal of Mathematical and Management Sciences*, vol. 23, no. 3-4, pp. 301–321, 2003.

[159] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *J. ACM*, vol. 45, pp. 891–923, Nov. 1998.

[160] S. E. Edgell and S. M. Noon, "Effect of violation of normality on the t test of the correlation coefficient," *Psychological Bulletin*, vol. 95, no. 3, pp. 576–583, 1984. 00058.

[161] I. Guyon, A. B. Hur, S. Gunn, and G. Dror, "Result analysis of the nips 2003 feature selection challenge," in *Advances in Neural Information Processing Systems 17*, pp. 545–552, MIT Press, 2004.

[162] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[163] I. Guyon, C. Aliferis, and A. Elisseeff, "Causal feature selection," *Chapman and Hall/CRC*, 2007.

[164] S. Seth and J. Príncipe, "Variable selection: A statistical dependence perspective," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*, pp. 931–936, Dec. 2010.

[165] W. Duch, "Filter methods," in *Feature Extraction, Foundations and Applications*, vol. 207 of *Studies in Fuzziness and Soft Computing*, ch. 3, pp. 167–185, Springer Berlin Heidelberg, 2006.

[166] B. Frénay, G. Doquire, and M. Verleysen, "Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification," *Neurocomputing*, vol. 112, no. 0, pp. 64 – 78, 2013.

Apéndice A

Propuesta para incluir complementariedad en diagramas de Venn

Esta propuesta nace como consecuencia de comprender gráficamente por qué dos variables complementarias generan mayor información actuando conjuntamente que la suma de la información de cada una de ellas.

Para ello se parte del supuesto de que la interacción de tres o más variables genera una nueva información. Suponiendo que se tiene tres variables: f_i, S y C, las posibles interacciones e información nueva entre estas variables se representa gráficamente como:

Como se observa en la Figura A.1, las áreas a, b, c y e son partes de informaciones mutuas entre los pares de variables, sin embargo, existe una nueva información que aparece cuando interactúan estas tres variables y que es representada por el círculo d. De esta forma, cualquier cálculo de información mutua que trabaje sobre las tres variables se debe considerar el efecto de esta nueva información. Las posibles informaciones que se pueden obtener utilizando las tres variables a la vez son:

$$I(\{f_i, S\}; C) = a + b + c + d \tag{A.1}$$
$$I(\{f_i, C\}; S) = b + c + d + e \tag{A.2}$$
$$I(\{S, C\}; f_i) = a + b + d + e \tag{A.3}$$
$$I(f_i; C|S) = a + d \tag{A.4}$$
$$I(S; C|f_i) = c + d \tag{A.5}$$
$$I(f_i; S|C) = e + d \tag{A.6}$$
$$II(f_i; S; C) = d - b \tag{A.7}$$

A continuación se muestra gráficamente cómo se representa la entropía de las variables C, $\{f_i, S\}$ y $\{f_i, S, C\}$ cuando se descompone información mutua $I(\{f_i, S\}; C)$ en términos de entropía.

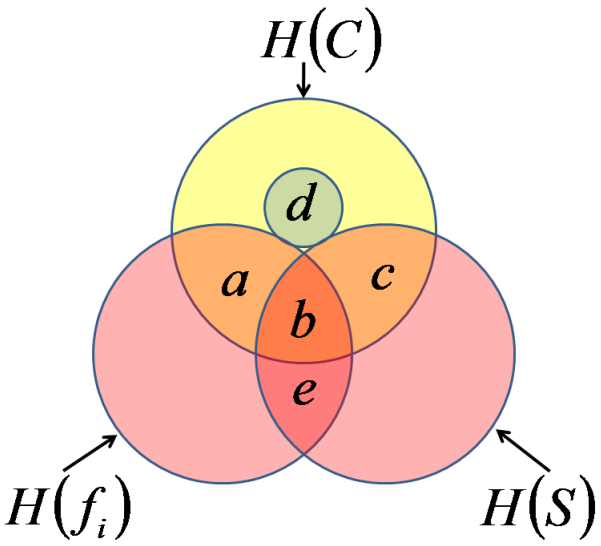


Figura A.1: Diagrama de Venn que ilustra la organización de la información de f_i ,S y C.

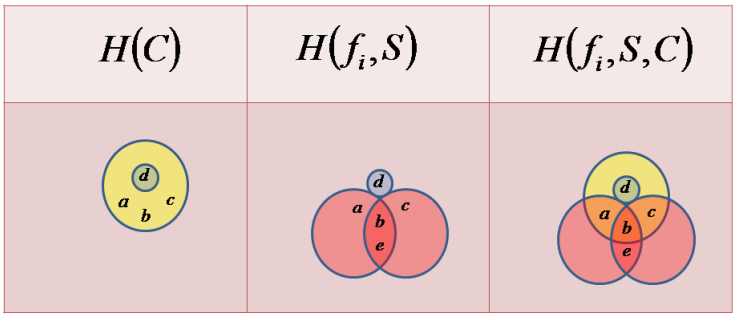


Figura A.2: Representación gráfica de las entropías utilizadas para calcular $I(\{f_i, S\}; C)$.

Como se observa en la Figura A.2, las entropías utilizadas en el cálculo de $I(\{f_i, S\}; C)$ contienen el término de complementariedad (área d). Esta área se mantiene sólo cuando actúan las tres variables a la vez. Por lo tanto, al realizar las operaciones de unión o diferencia de subconjuntos (entropías), el área d considera como un nuevo subconjunto. De esta forma, si se realiza la suma de $H(C)$ y $H(f_i, S)$ se observa que el área d se repite dos veces. Al realizar la diferencia entre $H(C) + H(f_i, S)$ y $H(f_i, S, C)$ el área d solo se mantiene una vez, y esta se suma a las áreas que resultan de esta diferencia y que son: a, b, c, por lo tanto, $I(\{f_i, S\}; C) = a + b + c + d$.