Selección de características en aprendizaje supervisado mediante métricas de teoría de la información

Aníbal Fuentes J., Mario Vicuña A.
Departamento de Ingeniería Eléctrica,
Universidad de Chile,
EL7024 - Teoría de la Información: Fundamentos y aplicaciones. 24 de Agosto de 2018

***Resumen*—El objetivo del proyecto es el estudio e implementación de métodos de selección de características en aprendizaje supervisado, mediante la utilización de métricas de teoría de la información. Se estudian diversos métodos presentes en la literatura basados en heurísticas; además se plantea un marco teórico que de validez y explique la utilización de estos métodos. Se implementan estos métodos y se comparan en términos de desempeño y de tiempo de cómputo al seleccionar subconjuntos de características de distinto tamaño, obteniendo buenos resultados.**

***Palabras Clave*—selección de características, información mutua, aprendizaje supervisado**

I. INTRODUCCIÓN

En la actualidad, en el contexto de aprendizaje de máquinas es común afrontarnos a problemas en que se quiere predecir una variable objetivo en función de otras variables que están asociadas a ella; estos problemas son llamados problemas de aprendizaje supervisado, ya que se aprende de datos ya etiquetados. Se tiene que si la variable a predecir es una variable continua el problema es un problema de "regresión" o "estimación"; mientras que si la variable a predecir es una variable discreta el problema es llamado de "clasificación".

Las variables predictoras en esta clase de problemas serán el "input" o entrada a un clasificador o regresor, también llamados características; mientras que las variables a predecir serán el "output", el cual corresponde a una etiqueta para el dato ingresado. Se recalca que las variables de entrada corresponden a variables aleatorias, las cuales pueden ser tanto continuas como discretas.

Para ejemplificar este tipo de problemas, se presenta un clásico ejemplo de clasificación de dígitos manuscritos de la base de datos MNIST, en la cual los inputs son imágenes de dígitos manuscritos de 28 x 28 pixeles y la etiqueta de cada clase es el número representado en la imagen, es decir, los digitos del 0 al 9; ahora, el problema de clasificación consiste en desarrollar un sistema computacional que a partir de la imagen de entrada, correspondiente a 784 inputs discretos (28 x 28 pixeles) nos entregue la clase correspondiente a esta imagen, y que el sistema sea capaz de generalizar a nuevas imágenes que no hayan sido utilizadas al realizar el entrenamiento. En la figura 1 se presenta un ejemplo de estas imágenes.

El objetivo de nuestro proyecto es, a partir de métricas de teoría de la información y de probabilidades, estudiar y comparar métodos que nos permitan seleccionar un subconjunto de las características de entrada de modo de seleccionar las más importantes al momento de realizar la clasificación. Dicho de otra manera, el problema de selección de características consiste en encontrar el mejor subconjunto de k características que posea un mínimo error de generalización.

Figura 1. Ejemplo de base de datos MNIST

Esto tiene varios propósitos, enumerados a continuación:

- Reducir el costo computacional y la complejidad de entrenar un clasificador: entrenar un clasificador con una cantidad muy grande de features o características es costoso computacionalmente.
- Mejorar el desempeño del predictor: una menor cantidad de características muchas veces va de la mano con un mejor desempeño de los clasificadores, puesto que pueden existir características ruidosas o características que aporten poca información.
- Facilitar la visualización y el entendimiento de los datos, saber que características son las importantes al momento de realizar un clasificador nos ayuda a entender mejor nuestros datos o procesos.
- Afrontar la maldición de la dimensionalidad: esto está asociado al punto 2; la maldición de la dimensionalidad es un concepto que se utiliza ampliamente en aprendizaje de máquinas, minería de datos, estadística. Lo que ocurre es que a medida aumenta la dimensionalidad de los datos aumenta el volumen del espacio exponencialmente, haciendo que los datos se vuelvan dispersos. Esto es negativo ya que métodos que se basan en la organización y búsqueda de datos están basados en la detección de áreas donde los datos forman grupos con propiedades similares, y en datos de alta dimensión, al volverse los datos más dispersos se dificulta encontrar datos con

propiedades similares.

Dentro de los algoritmos de selección de características existen 3 grandes grupos:

- Wrapper: Usa el clasificador para evaluar el subset óptimo, evaluando distintos subset de características. Con muchas dimensiones se vuelve prohibitivo. Además, dentro de este método el clasificador o regresor es visto como una caja negra.
- Embedded: Usa conocimiento específico del algoritmo de aprendizaje; por ejemplo, los árboles de decisión o random forest utilizan la entropía para determinar la clasificación y se puede determinar entonces a partir del clasificador que características son más importantes.
- Filter: asume independencia entre los datos y el algoritmo de clasificación utilizado. Es el tipo de método que se utilizará en este trabajo.

II. MARCO TEÓRICO

A continuación se repasan los principales conceptos necesarios para entender y definir los métodos de selección de características.

II-A. Definiciones básicas de teoría de la información

La teoría de la información nació en el contexto de comunicación, siendo los primeros trabajos presentados por Shannon [1]. Dada una variable aleatoria discreta, se definen las siguientes medidas:

II-A1. Entropía: Se define la **entropía** de una variable aleatoria como:

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

La entropía se interpreta como una medida de incertidumbre de una variable aleatoria, cumple que $H(x) \geq 0$ y $H(x) \leq log_2(n)$, alcanzando la cota máxima para una distribución uniforme.

Dadas dos variables aleatorias discretas x,y; se define la **entropía conjunta** de estas variables aleatorias como:

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=m}^{n} p(x(i),y(j)) log_2(p(x(i),y(j))) \tag{2}$$

Se cumple además que la entropía conjunta de dos variables aleatorias se encuentra acotada por:

$$max H(x), H(y) \leq H(x,y) \leq H(x) + H(y) \tag{3}$$

Alcanzando la cota inferior cuando las variables son estadísticamente dependientes, mientras que la cota superior se alcanza cuando las variables son estadísticamente independientes.

La **entropía condicional** se define como:

$$H(x|y) = \sum_{j=1}^{m} p(y(j)) H(x|y = y(j)) \tag{4}$$

La cual se encuentra acotada por:

$$0 \leq H(x|y) \leq H(x) \tag{5}$$

La entropía condicional se puede interpretar como cuanto se reduce en promedio la entropía de la variable aleatoria x si es conocido el valor de la variable aleatoria y. Esta alcanza su cota inferior cuando x es estadísticamente dependiente de y; mientras que su cota inferior la alcanza cuando x e y son independientes. Otra forma de expresar la entropía condicional es:

$$H(x|y) = H(x,y) - H(y) \tag{6}$$

II-A2. Información mutua: Si se tienen dos variables aleatorias discretas x, y; la **información mutua** de las variables representa la medida de la cantidad de información que tiene una respecto a la otra. Su formulación matemática es:

$$I(x;y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x(i),y(j)) log(\frac{p(x(i),y(j))}{p(x(i))p(y(j))}) \tag{7}$$

Si ambas variables aleatorias son independientes, entonces su información mutua será igual a 0. Además, de acuerdo a su definición matemática, la información mutua se puede escribir de las siguientes maneras:

$$I(x;y) = \begin{cases} H(x) - H(x|y) \\ \\ H(y) - H(y|x) \\ \\ H(x) + H(y) - H(x,y) \end{cases} \tag{8}$$

Si se tienen 3 variables aleatorias x, y, z; la **información mutua condicional** entre x, y dado z se expresa como:

$$I(x;y|z) = \sum_{i=1}^{n} p(z(i)) I(x;y|z = z(i)) \tag{9}$$

Siendo $I(x;y|z = z(i))$ la información mutua entre x e y en el contexto de $z = z(i)$. La información mutua condicional se interpreta como en promedio cuanta información comparten x e y en el contexto de que una tercera variable aleatoria z es conocida.

Por último, un concepto importante en el contexto de selección de características, corresponde a la multi-información, propuesta por McGill [2]. Se define la multi-información entre 3 variables aleatorias x, y, z como:

$$I(x;y;z) = \begin{cases} I(x,y;z) - I(x;z) - I(y;z) \\ \\ I(y;z|x) - H(y;z) \end{cases} \tag{10}$$

La multi-información cumple con que es simétrica, es decir, $I(x;y;z) = I(x;z;y) = I(y;z;x) = ...$, sin embargo no es muy utilizada debido a que su interpretación es difícil, pudiendo tomar valores negativos, entre otras razones. La multi-información se puede interpretar como el monto de información que es común a todas las variables, pero que no está presente en ningún subconjunto de las variables.

II-B. Notación

A continuación se define la notación que será utilizada en el contexto de selección de características. Para simplificar el problema de selección de características se **asumen variables aleatorias discretas**, puesto que en la práctica las variables aleatorias o tienen esta naturaleza, o son continuas y se pueden discretizar. Esta suposición puede no ser la más óptima en términos de precisión, pero ayuda a simplificar los métodos y a la rapidez de cálculo.

- F: Conjunto de características. Contiene m variables.
- C: vector (o matriz) de salida que representa las clases.
- f_i: Característica (variable) i del conjunto F
- S: Conjunto de las variables ya seleccionadas.
- f_i^c: $F \setminus f_i$

II-C. Relevancia, redundancia y complementariedad

Dentro del contexto de selección de características existen algunos conceptos que explican la relación existente entre una característica candidata a seleccionar (f_i), el conjunto de características ya seleccionadas (S) y la clase (C); estos conceptos corresponden a los de relevancia, redundancia y complementariedad; los cuales exploraremos con mayor detalle a continuación.

II-C1. Relevancia: La relevancia se define como $I(f_i; C)$ y se interpreta como la información que comparten la característica candidata f_i con la clase C. Se pueden clasificar las características en distintos niveles de relevancia como:

- Fuertemente relevantes: $I(f_i; C|f_i^c) > 0$, es decir, la característica f_i contiene información única de la clase, que no contienen el resto de las características.
- Débilmente relevantes: $\exists S \subset f_i^c \quad I(f_i; C|f_i^c) = 0 \wedge I(f_i; C|S) > 0$, es decir, se tiene que la información que f_i contiene de la clase puede ser extraída de otras características de F, sin embargo se tiene que para un subconjunto S de las características f_i si añade información de la clase C.
- Irrelevantes: $\forall S \subseteq f_i^c \quad I(f_i; C|S) = 0$, es decir, f_i no contiene información de la clase.

II-C2. Redundancia: Se define la redundancia como $I(f_i; S)$, es decir, la información común existente entre la característica candidata f_i y el subconjunto de características ya elegidas S.

II-C3. Complementariedad: La complementariedad, también llamada sinergia, se define como $I(f_i; S|C)$, es decir, mide el grado de interacción entre f_i y S en el contexto de la clase C. Se tiene que si la complementariedad es mayor entonces la característica f_i interactúa con S de modo de añadir información extra para determinar la clase.

Para entender de mejor manera los conceptos de relevancia, redundancia y complementariedad se descompone la multiinformación de las siguientes maneras:

$$I(f_i; S; C) = \begin{cases} I(f_i; S|C) - I(f_i; S) \\ I(f_i; C|S) - I(f_i; C) \\ I(S; C|f_i) - I(S; C) \end{cases} \tag{11}$$

De la primera expresión de la ecuación 11 se extrae que la multiinformación se puede expresar como la diferencia entre la complementariedad y la redundancia, y un valor positivo de la multiinformación significa predominancia de la complementariedad sobre la redundancia. Se la segunda línea se extrae que la multiinformación alcanza un valor positivo cuando la información que comparten f_i y C es mayor al interactuar con S que cuando no se interactúa con S, es decir, f_i y S se complementan positivamente. La tercera línea es otra manera de verlo, se extrae que se alcanza un valor positivo cuando la información que comparten S y C es mayor cuando se interactúa con la característica f_i.

III. DEFINICIÓN DEL PROBLEMA Y MÉTODOS CLÁSICOS PARA ABORDARLOS

El problema de la selección de características es que es un problema de complejidad computacional combinatorial, ya que para encontrar el subconjunto óptimo de características mediante fuerza bruta es necesario probar todas las combinaciones de características, lo que no es factible cuando la cantidad de características es alta; es por ello que los métodos se deben orientar a encontrar soluciones subóptimas en un número de iteraciones razonable. A continuación se plantea formalmente el problema y se definen las principales formas de abordarlo, mencionando los métodos clásicos utilizados para ello.

III-A. Definición del problema

El problema de selección de características se puede formalizar mediante el siguiente problema de optimización, propuesto por Brown [3]:

$$\begin{aligned} \min_{S \subseteq F} \quad & |S| \\ \text{s.a} \quad & \min_{S \subseteq F} I(S^c; C|S) \end{aligned} \tag{12}$$

De la ecuación 12 se extrae que el problema consiste en encontrar el subconjunto de características $S \subseteq F$ que minimice la información de la clase C que contengan las características no seleccionadas, dado que ya seleccionamos las características S. En las siguientes secciones se abordará este problema más en detalle.

III-B. Estrategias de búsqueda

Debido a que el problema de encontrar el subconjunto óptimo de características es un problema de complejidad computacional combinatorial, se establecen las siguientes estrategias de búsqueda que simplifican el problema, las cuales buscan un subconjunto S de k características a partir del conjunto F de n características.

- Búsqueda secuencial hacia adelante: corresponde a partir con un subconjunto S vacío e ingresar de a una características a este subconjunto hasta ingresar todas las características deseadas. En general el método es más rápido.
- Eliminación secuencial hacia atrás: se parte con $S = F$ y se eliminan características de a una en el conjunto S hasta quedarnos solo con las características deseadas. En general el método es más lento, sobretodo cuando k es pequeño en comparación a n; pero se obtienen mejor desempeño en términos de error de clasificación, debido a que con este método se consideran las interacciones que existen entre variables, en cambio, con el método

de búsqueda hacia adelante no se toman en cuenta la relevancia de variables en el contexto de características aún no escogidas.

También existen estrategias que combinan estos dos métodos de búsqueda, o que agregan o eliminan varias características a la vez. Para este trabajo nos enfocaremos en el método de búsqueda secuencial hacia adelante.

Además estos algoritmos se caracterizan por comparar características de a pares o tríos. A continuación, en el algoritmo 1 se plantea de forma general un algoritmo para elegir un subconjunto S de k características a partir de un conjunto F de n características mediante búsqueda secuencial hacia adelante.

Algorithm 1 Algoritmo de búsqueda secuencial hacia adelante

1: Elegir $k < n$, con k el tamaño del subconjunto buscado y n la cantidad de características.
2: Inicializar $F \leftarrow$ conjunto de características iniciales, de tamaño n, $S \leftarrow$ conjunto vacío.
3: Calcular la información mutua con la variable objetivo: para cada $f_i \in S$, calcular $I(C; f_i)$
4: Selección de la primera característica: seleccionar f_i que maximiza $I(C; f_i)$; hacer $F \leftarrow F \setminus f_i$; $S \leftarrow f_i$
5: **while** (**do**) Greedy selection: Repetir mientras $|S| < k$
6: Seleccionar la siguiente característica: Escoger $f_i \in F$ que maximice la métrica de desempeño $G(f_i)$ (que depende de cada método)
7: Hacer $F \leftarrow F \setminus f_i$; $S \leftarrow f_i$
8: Retornar el conjunto S que contiene las características seleccionadas.

La idea del algoritmo planteado es elegir un subconjunto de características, tal que en cada paso de la selección se escojan las características que contengan más información que aún no se conoce, respecto a la variable objetivo.

III-C. Métodos clásicos

En el paso 7 del algoritmo se plantea la selección de una característica de acuerdo a un criterio a elegir. En la literatura se plantean distintas métricas de desempeño para esta selección, las cuales se enuncian a continuación.

Mutual Information Maximization (MIM): Planteada por Lewis [4], la métrica de desempeño viene dada por:

$$G(f_i) = I(C; f_i) \tag{13}$$

Como se puede observar, este término corresponde a la relevancia de la característica f_i respecto a la clase C.

Mutual Information Feature Selection (MIFS):

Planteada por Battiti [5], en este caso la métrica de desempeño $G(f_i)$ es:

$$G(f_i) = I(C; f_i) - \beta \sum_{f_s \in S} I(f_s; f_i) \tag{14}$$

La métrica R se divide en dos términos, el primero está relacionado a la relevancia de la característica f_i respecto a la variable objetivo C. El segundo término está relacionado a la redundancia que tiene la característica f_s respecto a las características seleccionadas anteriormente. La variable $\beta > 0$ es un parámetro definido por el usuario que da un balance entre la importancia que se le da a la redundancia y la importancia que se le da a la relevancia al momento de seleccionar una característica.

MIFS-U

Propuesto por Kwak y Choi [6], se plantea en este caso que la métrica de desempeño $G(f_i)$ es:

$$G(f_i) = I(C; f_i) - \beta \sum_{f_s \in S} \frac{I(C; f_s)}{H(f_s)} I(f_s; f_i) \tag{15}$$

Mínima redundancia máxima relevancia (mRMR)

Propuesto por Peng [7], se plantea en este caso que la métrica de desempeño $G(f_i)$ está dada por:

$$G(f_i) = I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} I(f_s; f_i) \tag{16}$$

Este caso es un caso especial de **MIFS**, en el sentido de que $\beta = \frac{1}{|S|}$. La principal ventaja de elegir β de esta manera es que para el caso de MIFS y MIFS-U; los términos de relevancia y redundancia no son comparables, debido a que el segundo término corresponde a una suma acumulativa de informaciones mutuas, es por eso que a medida hemos elegido más características el valor de redundancia aumentara respecto al valor de relevancia y solo estaremos eligiendo las características que sean menos redundantes.

Joint Mutual Information (JMI)

Se presenta en [2], la métrica está dada por:

$$G(f_i) = I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} I(f_s; f_i) + \frac{1}{|S|} \sum_{f_s \in S} I(f_s; f_i | C) \tag{17}$$

El primer término corresponde a la relevancia, mientras que el segundo término a la redundancia, ponderado por un factor normalizador. Por último se incluye un término asociado a la complementariedad, también con su respectivo factor normalizador asociado.

Conditional Infomax Feature Extraction (CIFE)

Presentado en [8], la métrica viene dada por:

$$G(f_i) = I(C; f_i) - \sum_{f_s \in S} I(f_s; f_i) + \sum_{f_s \in S} I(f_s; f_i | C) \tag{18}$$

Está métrica es casi equivalente a JMI, con la diferencia de que el factor normalizador es igual a 1.

Normalized mutual information feature selection (NMIFS)

Propuesto por Estévez [9], recordemos de la definición de información mutua:

$$\text{III-C} I(f_s; f_i) = H(f_i) - H(f_i|f_s) = H(f_s) - H(f_s|f_i) \tag{19}$$

Se puede deducir de la ecuación que:

$$0 \leq I(f_s, f_i) \leq min\{H(f_i), H(f_s)\} \tag{20}$$

Esto quiere decir que la información mutua entre dos variables está acotada superiormente por el mínimo de las entropías de ambas variables. A raíz de que este mínimo puede variar mucho entre dos variables, es necesario normalizar el término $I(f_s, f_i)$ antes de aplicarlo a un conjunto de variables, se define entonces la información mutua normalizada entre dos variables como:

$$NI(f_i, f_s) = \frac{I(f_i, f_s)}{min\{H(f_i), H(f_s)\}} \tag{21}$$

A partir de esto, se plantea la métrica de desempeño $G(f_i)$ de la siguiente manera:

$$G(f_i) = I(C; f_i) - \frac{1}{|S|} \sum_{f_s \in S} NI(f_s; f_i) \tag{22}$$

IV. FORMALIZACIÓN MATEMÁTICA

Debido a que los métodos presentados hasta el momento se basan en su mayoría en heurísticas es necesario presentar un marco que unifique estos métodos y les de formalidad matemática, para ello utilizamos los procedimientos descritos en [3] y [8].

Como fue mencionado en la sección anterior, el problema de selección de características puede ser abordado como el problema de optimización de la ecuación 12.

Si usamos la regla de la cadena sobre $I(F; C)$ se tiene que:

$$I(F; C) = I(S; C) + I(S^c; C|S) \tag{23}$$

Debido a que $I(F; C)$ es un término constante, entonces el problema de minimización de $I(S^c; C|S)$ es equivalente al problema de maximización de $I(S; C)$. El problema de minimizar $I(S^c; C|S)$ es conocido como el criterio de minimizar la información mutua condicional, mientras que maximizar $I(S; C)$ es conocido como el criterio de máxima dependencia.

Para describir el proceso de selección de características mediante búsqueda secuencial hacia adelante introducimos primero la notación a utilizar:

- S_t: subconjunto de las variables seleccionadas en el instante t.
- f_i: característica candidata a añadir al subconjunto S_t.
- S_{t+1}: subconjunto de características en el instante $t+1$. Se tiene $S_{t+1} \leftarrow S_t, f_i$.
- S^c_{t+1}: Complemento de S_{t+1}, es decir, $F \setminus S_{t+1}$. Se puede escribir también como $S^c_t \setminus f_i$

Desarrollemos a continuación la expresión de mínima información mutua condicional:

$$\min_{f_i \subseteq S^c_t} I(S^c_{t+1}; C|S_{t+1}) \tag{24}$$

$$= \min_{f_i \subseteq S^c_t} I(S^c_t \setminus f_i; C|\{S_t, f_i\}) \tag{25}$$

$$= \min_{f_i \subseteq S^c_t} I(S^c_t; C|S_t) + \min_{f_i \subseteq S^c_t} - I(f_i; C|S_t) \tag{26}$$

Debido a que el lado izquierdo es constante para todo valor de f_i, nos quedamos solo con el lado derecho.

$$= \max_{f_i \subseteq S^c_t} I(f_i; C|S_t) \tag{27}$$

Por otro lado, desarrollando el criterio de máxima dependencia, se tiene que:

$$\max_{f_i \subseteq S^c_t} I(S_{t+1}; C) \tag{28}$$

$$= \max_{f_i \subseteq S^c_t} I(\{S_t, f_i\}; C) \tag{29}$$

$$= \max_{f_i \subseteq S^c_t} I(S_t; C) + \max_{f_i \subseteq S^c_t} I(f_i; C|S_t) \tag{30}$$

Debido a que el lado derecho de 30 es independiente de f_i, solo nos queda el lado izquierdo:

$$= \max_{f_i \subseteq S^c_t} I(f_i; C|S_t) \tag{31}$$

Como se extrae de las ecuaciones 27 y 31 el algoritmo de búsqueda secuencial hacia adelante tiene por objetivo seleccionar la característica que maximice $I(f_i; C|S_t)$ en cada iteración. A partir de las distintas expansiones de la multi-información, presentes en la ecuación 11, la expresión de $I(f_i; C|S_t)$ se puede expresar de la siguiente manera:

$$I(f_i; C|S_t) = I(f_i; C) - I(f_i; S_t) + I(f_i; S_t|C) \tag{32}$$

De la ecuación 32 notamos que el primer término del lado derecho corresponde a la relevancia de f_i respecto a C, el segundo término corresponde a la redundancia de f_i respecto a las características S ya seleccionadas, mientras que el tercer término corresponde a la complementariedad entre f_i y S en el contexto de C. Sin embargo esta ecuación presenta la dificultad de que tanto para la redundancia como para la complementariedad se deben calcular información mutua en espacios multidimensionales. Para simplificar este problema se puede realizar la siguiente expansión de la redundancia, con $|S_t| = p$:

$$\begin{aligned}
I(f_i; S_t) = &\quad I(f_i; s_1) + &I(f_i; s^c_1|s_1) \\
I(f_i; S_t) = &\quad I(f_i; s_2) + &I(f_i; s^c_2|s_2) \\
&... \\
I(f_i; S_t) = &\quad I(f_i; s_p) + &I(f_i; s^c_p|s_p) \\
\hline
I(f_i; S_t) = &\frac{1}{|S_t|} \sum_{j=1}^{p} I(f_i; s_j) + \frac{1}{|S_t|} \sum_{j=1}^{p} I(f_i; s^c_j|s_j)
\end{aligned} \tag{33}$$

Y de forma equivalente para la complementariedad se puede llegar a la ecuación:

$$I(f_i; S_t|C) = \frac{1}{|S_t|} \sum_{j=1}^{p} I(f_i; s_j|C) + \frac{1}{|S_t|} \sum_{j=1}^{p} I(f_i; s^c_j|\{C, s_j\}) \tag{34}$$

Reemplazando las ecuaciones 33 y 34 en la ecuación 32 se llega a:

$$I(f_i;C|S_t) = I(f_i;C)$$
$$-(\tfrac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j) + \tfrac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j^c|s_j))$$
$$+(\tfrac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j|C) + \tfrac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j^c|\{C,s_j\})) \quad (35)$$

Por último, esta ecuación se aproxima considerando solo dependencias de orden bajo entre características, es decir, se considera que las características actúan de a pares. Formalmente, asumiendo independencia estadística

$$p(f_i|S_t) = \prod_{s_j \in S_t} p(f_i|s_j)$$
$$p(f_i|\{S_t,C\}) = \prod_{s_j \in S_t} p(f_i|\{s_j,C\}) \quad (36)$$

se obtiene la siguiente aproximación:

$$I(f_i;C|S_t) \approx I(f_i;C) - \frac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j) + \frac{1}{|S_t|}\sum_{j=1}^{p} I(f_i;s_j|C) \quad (37)$$

De la ecuación 37 se pueden derivar algunos de los criterios heurísticos más utilizados. Por ejemplo, si se utiliza solo el primer término se obtiene el criterio "MIM". Por otro lado, si se utilizan solo los dos primeros términos de la ecuación 37 se obtiene el criterio de "mínima redundancia máxima relevancia" (mRMR). Por otro lado, considerando los dos primeros términos pero cambiando la constante de normalización del segundo por un parámetro β se obtiene el criterio de "mutual information feature selection (MIFS)"; mientras que para los criterios MIFS-U y NMIFS se utilizan constantes de normalización dependientes del término.

Si consideramos los tres términos de la ecuación 37 obtenemos el criterio de "JMI"; mientras que si consideramos también los tres términos, pero reemplazando las constantes de normalización $1/|S_t|$ por 1, obtenemos el criterio "CIFE".

V. RESULTADOS EXPERIMENTALES Y DISCUSIÓN

A continuación, para comparar los distintos métodos de selección de características, en cuando a desempeño de selección, como tiempo de ejecución, se implementó en python todo un sistema de selección de características, el cual trabaja con variables discretas. Para esto se implementa la discretización de variables continuas en bines; el cálculo de entropías, entropías conjuntas y entropías condicionales de variables discretas; el cálculo de informaciones mutuas e informaciones mutuas condicionales de variables discretas, y además la metodología de selección de características mediante búsqueda secuencial hacia adelante, incluyendo los criterios "mRMR", "MIFS", "MIFSU", "NMIFS", "MIM", "CIFE" y "JMI".

V-A. Wine Dataset

Para comenzar, se plantea un experimento en un dataset pequeño; este corresponde al Wine dataset, los datos corresponden a análisis químicos de vinos cultivados en la misma región de Italia pero de tres cultivos diferentes. El análisis químico determinó la cantidad de 13 componentes encontrados en cada uno de los tres tipos de vino. El problema corresponde entonces a un problema de clasificación con 3 clases, en que se tienen 13 inputs. Además se tienen 178 muestras. Entre las entradas se encuentran: alcohol, ácido málico, magnesio, fenoles totales, intensidad del color, entre otros. Como parte del preprocesamiento, se discretizan cada una de las variables de entrada en 10 bines.

El objetivo de este primer experimento es comparar el desempeño de clasificación (accuracy) de un clasificador ante distinto número de variables de entrada, seleccionando estas variables de entrada mediante los métodos planteados de selección de características. El clasificador utilizado para esto corresponde a Random Forest, debido a que no es tan dependiente de sus hiperparámetros como otros clasificadores. Como métrica de desempeño se utilizó accuracy promedio de validación cruzada, con 5-fold. Por otro lado, cabe recordar el supuesto de los métodos de filtro, en que se asume que el conjunto se inputs a seleccionar con el algoritmo es independiente de un clasificador y las métricas desprendibles de este; en consecuencia, tampoco es necesario un clasificador optimizado a modo de comparar los sets de características obtenidos. Se presentan en la figura 2 y 3 las métricas de desempeño al seleccionar las características con los distintos métodos.

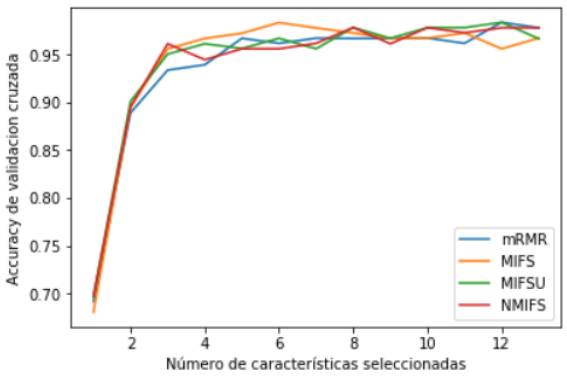


Figura 2. Accuracy promedio obtenido mediante métodos mRMR, MIFS, MIFS-U y NMIFS

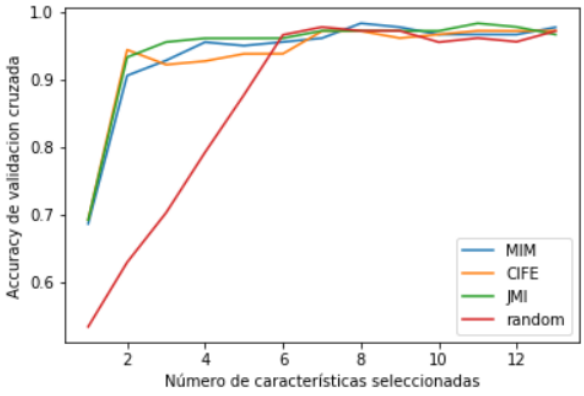


Figura 3. Accuracy promedio obtenido mediante métodos MIM, CIFE, JMI y aleatorio

A partir de estos resultados se pueden extraer varias discusiones. La primera es que en general los métodos de selección planteados se comportan similar en cuanto a performance de clasificación, a excepción del método ÇIFE"(figura 3). Un punto muy importante a recalcar es que como se puede observar en las figuras 2 y 3, muchas veces no es necesario considerar todos los datos de entrada para alcanzar buen desempeño de clasificación, en el caso de este dataset, con 6 características basta para alcanzar el mismo desempeño que con las 13 características del dataset completo. También se debe recalcar que los métodos realmente mejoran el desempeño de elegir características respecto a elegir características por ejemplo de modo aleatorio, como se ve en la figura 3.

Otro punto a recalcar es la ventaja de utilizar métodos de selección de características para obtener un mejor entendimiento de los datos, por ejemplo, de los métodos planteados se extrae que dentro de las características más importantes a la hora de afrontar este problema de clasificación son las características 6, 12 y 9; esto se extrae de que en la mayoría de los métodos estas características eran las primeras en ser seleccionadas. Estas características corresponden a flavonoides (pigmentos naturales presentes en los vegetales), prolina (un aminoácido que constituye la mayoría de las proteínas) y la intensidad de color, respectivamente.

Respecto al tiempo de selección, se presenta en la figura V-B el tiempo que se demoran los distintos métodos en la selección de características, dependiendo del número de características a seleccionar. Se extrae de la figura que el método MIM es el que tarda menos, puesto que solo considera la relevancia al momento de la selección de características; por otro lado los metodos MIFS, mRMR, MIFS-U y NMIFS están en un punto intermedio, debido a que consideran tanto relevancia como redundancia; finalmente los métodos CIFE y JMI son los que más tardan en la selección, puesto que consideran relevancia, redundancia y complementariedad, cálculos que deben ser realizados en forma separada pues capturan distintas interacciones entre las características analizadas.

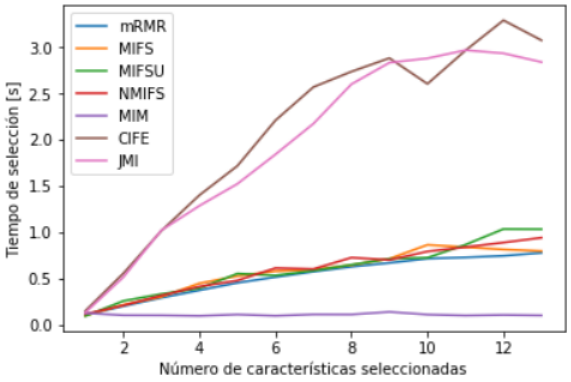


Figura 4. Tiempo en la selección para los distintos métodos

V-B. Breast cancer dataset

Esta base de datos tiene por objetivo el diagnóstico de cáncer de mama. Las características fueron extraídas a partir de una imagen digitalizada de una aspiración con aguja fina de una masa mamaria. Estas describen características de los núcleos celulares presentes en la imagen, tales como radio, textura, perímetro, área. El dataset se conforma por 30 atributos y 569 muestras. Además se tienen dos clases, correspondiente a un diagnóstico positivo y negativo de cáncer de mama.

Para términos de análisis, se realizó selección de características y se comparan los métodos "MIM", "MIFS" y "JMI" en términos de accuracy de validación cruzada, se presentan estos resultados en la figura V-B.

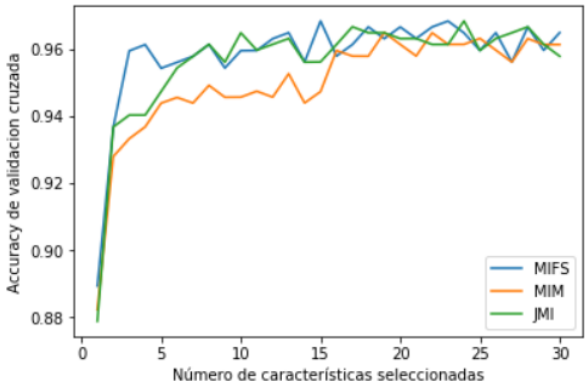


Figura 5. Accuracy de validación cruzada, métodos MIM, MIFS y JMI

De la figura lo primero que se debe destacar es que con el método MIFS y 4 características seleccionadas se alcanza un desempeño similar al obtenido con todas las características, es decir, con 4 características estamos capturando la información de la clase que tienen las 30 características. Otro punto a destacar es que el criterio "MIM" tiene un crecimiento lento en el accuracy a medida se seleccionan más características; esto se debe a que este criterio solo considera las relevancias de las características, pero no considera la redundancia existente con las características ya seleccionadas hasta el momento, lo cual hace que se capture información extra de la clase de manera más lenta. Se destaca también que al menos para este dataset la complementariedad no juega un papel relevante en la selección.

En términos del tiempo de selección, se muestra en la figura el tiempo que demoran los distintos métodos implementados en seleccionar dado número de características. Se aprecia en esta figura que se satisface el mismo comportamiento observado en el gráfico de la figura , estableciendo consistencia entre el comportamiento de las curvas y el método empleado en la selección.

Por otro lado, el tiempo de ejecución puede ser estudiado como dependiente de dos parámetros: el número de features deseados en el conjunto resultante k (o equivalentemente $|S|$) y el tamaño N del conjunto inicial de características $|F|$. El análisis de tiempo de ejecución realizado anteriormente mediante las figuras y considera el primer caso, en que se tiene un conjunto inicial, con N características fijas y se varía el tamaño k del subconjunto de características seleccionadas.

Para estudiar el comportamiento del tiempo de selección de features como función de la variable descrita en el segundo caso se muestra en la figura . Dado un valor de k características seleccionadas fijo, se observa que el tiempo que demoran los distintos métodos en seleccionar dicho número de características crece en forma lineal con $|F|$, no así los

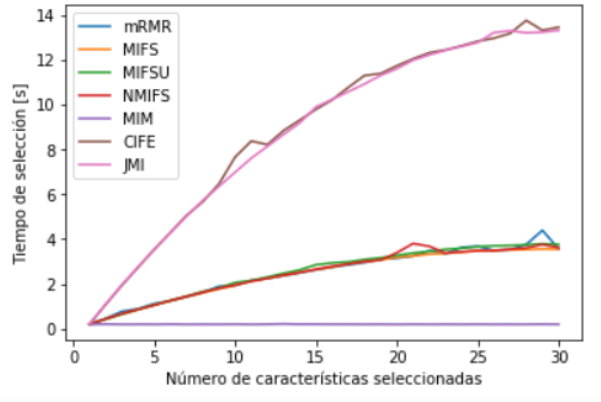
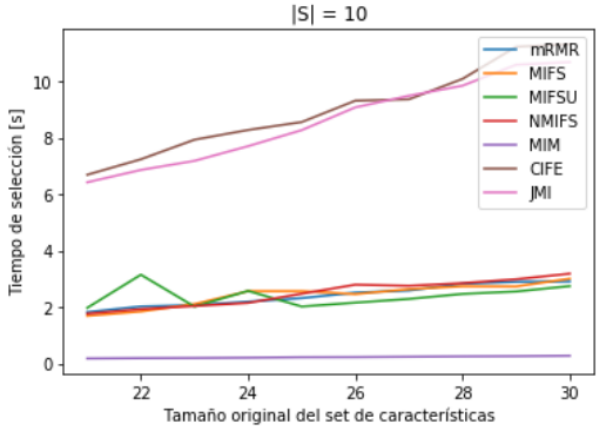
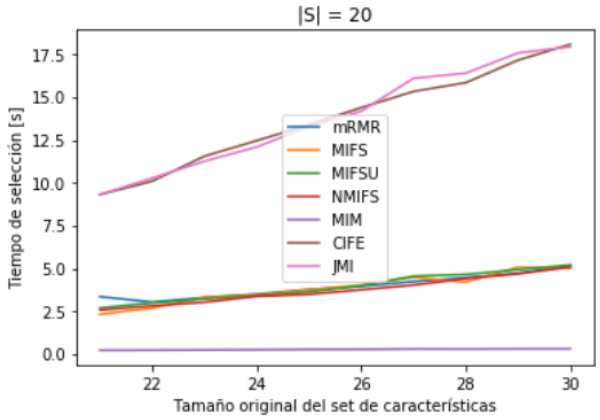


Figura 6. Tiempo de selección para los distintos métodos.

casos estudiados anteriormente (con dependencia en k) de comportamiento asintótico. Además, se aprecia que a medida que se incrementa el número deseado de features, también crece la pendiente de las curvas descritas.



(a) $|S| = 10$



(b) $|S| = 20$

Figura 7. Tiempo en la selección de características para distintos conjuntos iniciales, dado $|S|$ deseado.

Ambas dependencias pueden ser formalizadas matemáticamente y unidas mediante una única ecuación que capture todos los comportamientos establecidos para el tiempo de ejecución. Mediante un análisis del orden del algoritmo se llega a que el orden del mismo depende de k y de N mediante la ecuación $O(k \cdot N - \frac{k^2}{2})$, siendo N el tamaño original del conjunto. Además, cabe destacar que esta ecuación sustenta uno de los principales problemas de este tipo de algoritmos, considerando que en la mayoría de los casos con alta complejidad dimensional es preferible y razonable que el set escogido mantenga cierta complejidad, no reduciéndolo demasiado. En tal escenario tanto k como N son de orden considerable, y su producto domina el orden del algoritmo.

VI. CONCLUSIÓN

A lo largo de esta experiencia se revisaron distintos métodos de selección de características basados en información mutua presentes en la literatura. Se planteó un marco teórico en el cual se revisaron conceptos básicos de teoría de la información, además de los conceptos de relevancia, redundancia y complementariedad. Se trabajó en particular con métodos de selección de características de búsqueda secuencial hacia adelante; planteando el algoritmo general de resolución de este tipo de búsqueda, y revisando los métodos clásicos y las distintas métricas planteadas en la literatura para abordar este tipo de problemas, entre las que se pueden mencionar Mutual information maximization (MIM), Mutual Information Feature Selection (MIFS), entre otras; cabe destacar que todas estas métricas nacen de heurísticas. A continuación se realiza una formalización matemática del problema, planteando el problema de selección como un problema de optimización, y llegando a que la característica óptima a seleccionar en cada paso de la selección es aquella que maximice relevancia menos redundancia más complementariedad, y que mediante una aproximación que asume dependencias de orden menor entre variables se llega a una expresión de la cual derivan la mayoría de heurísticas.

Respecto a los resultados experimentales, se llegó a que los distintos métodos propuestos pueden realizar una selección efectiva de características, observándose que en muchos de los casos basta con seleccionar un subconjunto de las características para capturar la información de la clase que contienen todas las características. Se comparó también con una selección de características al azar, lo que vendría siendo el peor caso de selección, y se comprobó que el desempeño del método es muy superior a esta selección al azar.

Respecto a la comparación entre distintas métricas, se obtuvo que la métrica "MIM" es la de peor desempeño, esto debido a que al considerar solo la relevancia de una característica al momento de realizar la selección, se tiende a seleccionar características redundantes entre sí, por lo que la información capturada sobre la clase tiende a aumentar más lentamente respecto a la cantidad de características seleccionadas. Por otro lado, los métodos basados en relevancia y redundancia ("MIFS", "MIFS-U", "NMIFS" y "mRMR") presentaron en general desempeños similares entre si, siendo estos métodos los que en general presentaron mejores resultados considerando un trade-off entre el accuracy de clasificación alcanzado por las características seleccionadas y el tiempo de ejecución del algoritmo. Por otro lado los métodos que consideraban relevancia, redundancia y complementariedad fueron "JMI" y "CIFE"; el primero alcanzó un mejor desempeño que el segundo, debido a que se consideran las constantes de normalización de los términos de relevancia y redundancia; sin embargo los resultados de "JMI" fueron similares a los de los métodos "mRMR" y "MIFS", esto se explica debido

a que la complementariedad tiene una mayor importancia al seleccionar varias características a la vez, y no solo una a la vez como en el método de selección hacia adelante, por lo cual la complementariedad no resulta ser un gran aporte en la selección de características.

Al considerar los tiempos de ejecución de los algoritmos se concluye que los tiempos no son considerablemente grandes al tener datasets con pocas características, sin embargo, al tener datasets con muchas características (que en general esto es cuando el problema de selección de características se vuelve relevante) los tiempos aumentan bastante; por lo que se deben buscar formas de optimizar el algoritmo para disminuir sus tiempos de cómputo.

Como conclusiones generales se menciona que el trabajo realizado es un aporte relevante a las técnicas de aprendizaje de máquinas y aprendizaje estadístico, debido a que presenta un modo de extraer conocimiento útil acerca de los datos, pudiendo determinar de manera sencilla en que características fijarse al momento de realizar una clasificación o que características son más importantes a la hora de determinar una característica asociada a ellas; además de disminuir el costo computacional de entrenar modelos de aprendizaje y lidiar con la maldición de la dimensionalidad; que muchas veces resulta en que los modelos presenten un peor desempeño.

Como trabajo propuesto se plantea la comparación de estos métodos con métodos "wrapper" o "embedded"; un método de selección a priori de las características k óptimas a seleccionar y la optimización del algoritmo, para lograr un menor tiempo de cómputo.

REFERENCIAS

[1] C. Shannon, .ª Mathematical Theory of communication", The Bell System Technical Journal, Vol. 27, pp 379-423, July 1948.

[2] W. McGill, "Multivariate Information Transmission", Psychometrica, Vol. 19, pp 97-116, 1954.

[3] G. Brown, A. Pocock, MJ. Zhao, M. Luján, Çonditional likelihood maximisation: A unifying framework for information theoretic feature selection", J Mach Learn Res, Vol. 13, pp 27-66, 2012.

[4] D. Lewis, "Feature selection and feature extraction for text categorization", Proceedings of speech and natural language workshop, Morgan Kaufmann, pp 212–217, 1992.

[5] R. Battiti, Üsing mutual information for selecting features in supervised neural net learning", IEEE Transactions on Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.

[6] N. Kwak, C.H. Choi, "Input feature selection for classification problems", IEEE Transactions on Neural Networks, vol. 3, no. 1, pp. 143–159, Jan. 2002.

[7] H. Peng, F. Long, C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no 8, pp. 1226–1238, Aug. 2005.

[8] D. Ling, X. Tang, Çonditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion", In: Leonardis A., Bischof H., Pinz A. (eds) Computer Vision – ECCV 2006. ECCV 2006. Lecture Notes in Computer Science, vol 3951. Springer, Berlin, Heidelberg, 2006.

[9] P. Estévez, M. Tesmer, C. Pérez, J. Zurada, "Normalized mutual information feature selection", IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189-201, february 2009.

[10] J. Vergara, P. Estévez, .ª review of feature selection methods based on mutual information", Neural Computing and Applications, vol. 24, pp. 175-186, 2014.