Selección de atributos

Hugo Andrés Dorado B.

Muchas mediciones

- ¿Necesitamos tanta información de variables en para nuestro algoritmo de aprendizaje?

- ¿Puedo mejorar mi algoritmo de aprendizaje sí remuevo algunas variables, como escojo cuales?

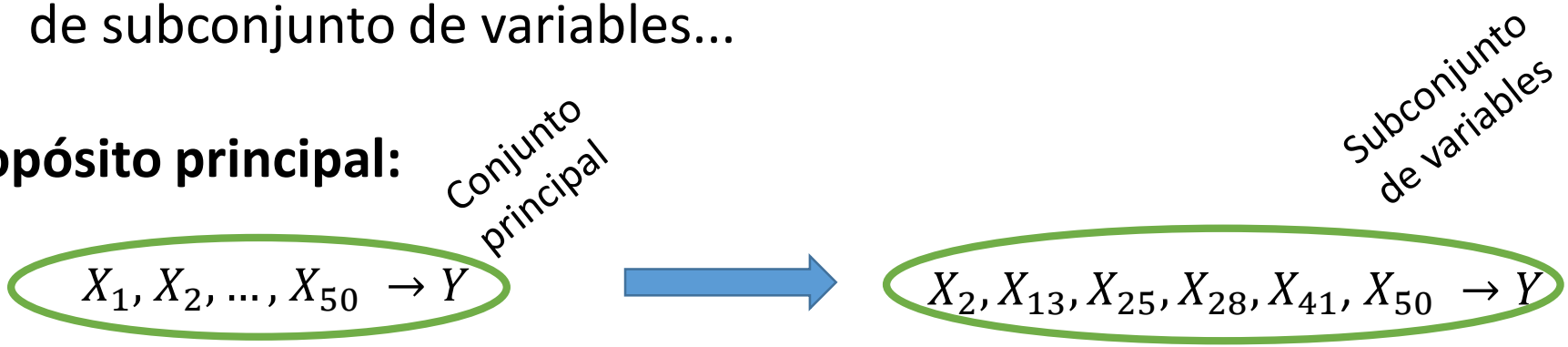
- ¿Cómo puedo reducir el tiempo de ejecución de mi algoritmo de aprendizaje?

Selección de atributos

Otros nombres:

Selección de variables, selección de característica o selección de subconjunto de variables...

Propósito principal:



En este proceso se escoge un subconjunto de variables relevantes de acuerdo a ciertos criterios para un modelo el cual se está a punto de construir.

Porque hacer selección de atributos.

- ✓ Remover y reducir efectos de variables poco importantes.
- ✓ Reducir la dimensionalidad.
- ✓ Mejorar el desempeño predictivo.
- ✓ Facilitar la interpretación de resultados.
- ✓ Construir modelos más eficientemente.

Campos de aplicación en la selección de atributos

Minería de texto

Industria

Genética

...

Algunos conceptos

- Un atributo es *irrelevante* si no afecta de ninguna forma al desempeño del modelo.
- Un atributo es *redundante* si no añade nada nuevo al modelo.
- Un atributo se considera *relevante* si no es irrelevante o redundante.

Ejemplo de selección de atributos

Supongamos que tenemos 3 atributos A, B, C y un clasificador M, entonces queremos predecir a T.

Selección de atributos	Clasificación	Desempeño	
{A,B,C}	M	98.0%	
{A,B}	M	92.1%	
{A,C}	M	98.0%	→ Mejor desempeño
{B,C}	M	56.3%	
{A}	M	97.5%	→ Simplicidad
{B}	M	90.3%	
{C}	M	91.2%	
.	M	85.1%	

Pasos en la selección de atributos

1. Generación de subconjuntos candidatos (estrategia de búsqueda o **función criterio**).

2. Evaluación de subconjuntos posibles.

3. Criterio de parada. (Número de iteraciones, número de atributos, poca mejora al añadir o quitar, alcanzar el desempeño deseado.)

4. Validación de los resultados (Atributos conocidos como relevantes, comparando el error en la clasificación o una línea base sin la selección de atributos).

Clasificación de métodos de selección de atributos

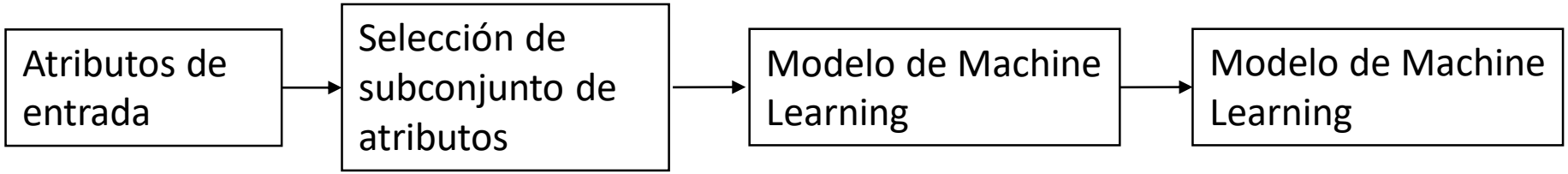
Se distinguen por la forma de evaluar atributos en tres clases:

- **Filtros:** La selección de atributos es independiente del algoritmo de aprendizaje, usando medidas de distancia, información o dependencia.

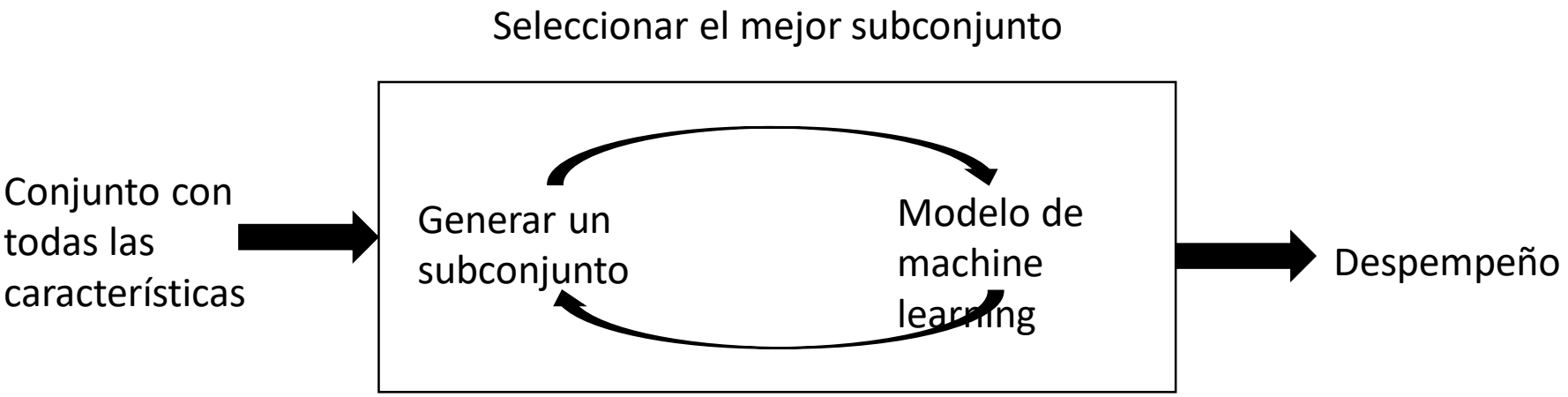
- **Wrappers:** La estrategia de búsqueda utilizada es el propio conjunto de reglas generadas por el algoritmo de aprendizaje que posteriormente se usará en el modelo.

- **Híbridos:** usan una combinación de los dos criterios de evaluación en diferentes etapas del proceso de búsqueda.

Algoritmos tipo filtros



Algoritmos tipo wrapper



Algoritmos tipo híbrido

Se realiza de la siguiente forma:

- Empieza con un subconjunto de variables (puede ser vacío).
- Adicionar de un atributo y evaluar todos los elementos con los nuevos atributo con un algoritmo de filtro.
- El mejor subconjunto se evalúa con un modelo de machine learning y conserva el mejor (de todos los subconjuntos que salen para adicionar).
- Continua con el siguiente atributo candidato.

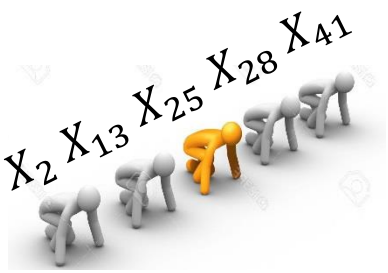
Para detenerse puede usarse como referencia la no mejora

Estrategias de búsqueda

Pretende guiar la forma en que se selecciona los subconjuntos posibles

¿Porque implementar una estrategia de búsqueda?

Para k atributos existen 2^k subcojuntos posibles, entonces a veces son demasiadas pruebas (k=30 => 2^{30} = 1.073.741.824)…



Punto de partida

Puede ser un subconjunto vacío, o el subconjunto completo, una cantidad estimada de atributos o aleatoria.

Estrategias de búsqueda clásicas

- Forward: Se parte de un subconjunto vacío de atributos, y se va adicionando atributos (Rápido pero a veces poco eficiente) .

$$\emptyset \Rightarrow X_2 \Rightarrow X_2X_5 \Rightarrow X_2X_5X_7 \Rightarrow X_2X_5X_7X_{15}$$

- Backward: Se eliminan progresivamente variables. (Contrario al anterior)

$$X_1X_2 \dots X_{50} \Rightarrow X_1X_3 \dots X_{50} \Rightarrow X_1X_3 \dots X_{49} \Rightarrow X_1X_3 \dots X_{49}$$

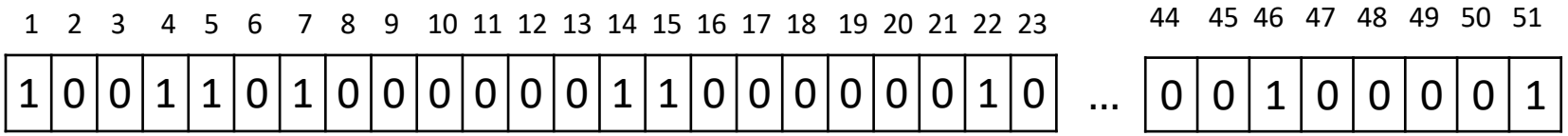
- Bi direccional: Se pueden añadir o quitar atributos a partir de un subconjunto inicial.

$$\emptyset \Rightarrow X_2 \Rightarrow X_2X_5 \Rightarrow X_2X_5X_7 \Rightarrow X_2X_5X_{15}$$

Otras estrategias de búsqueda

- Búsqueda aleatoria, donde luego se le une un algoritmo de optimización. (Ej: Ranmdon restar hill-climbing) ó (Simulated anneling).

- Se pueden usar diversas variantes de búsqueda: *local search, tabú search, ant colony optimization, algoritmos genéticos, swarm optimization, etc.*



- Existen búsquedas completas (No necesariamente exhaustivas), heurísticas o aleatorias.

Evaluación de subconjuntos

- Se puede hacer de manera independiente, como es el caso de los filtros; en los cuales se usan medidas de distancia, dependencia o consistencia. (Correlación, información mutua, variación, entropía cruzada..)

- De manera dependiente, como es el caso de los wrappers y se evalúa a partir del modelo de machine learning empleado. (luego un indicador de precisión tal como Kappa, instancias correctamente clasificadas, AIC…)

- En el caso de clustering se utilizan medidas de calidad de agrupamiento, por ejemplo un coeficiente entre la inercia inter clase e intraclase.

Ejemplo de caso

Investigación para determinar factores relevantes en la evaluación nutricional de los niños

- Se pretende aplicar selección de atributos, a una base de datos que contenía variables involucradas en el estado nutricional de niños de 6 a 11 años.
- El propósito de precisar cuál de los métodos aplicados determinaba los factores que más aportaron a la evaluación nutricional.

Resultados

Tabla 1: Atributos obtenidos por cada método de selección empleado.

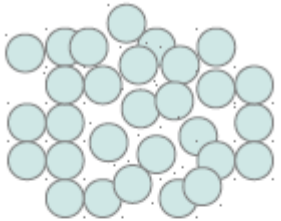
Métodos de Selección de Atributos Utilizados							
Cfs	**Consistency**	**Classifier**	**Wrapper**	**ChiSquared**	**GainRatio**	**InfoGain**	**OneR**
Edad OtrosVeg Proteínas Triglic. Colest.	Edad OtrosVeg Proteínas Triglic.	Edad Lácteos Proteínas Triglic. Colest.	Leche VegHoja OtrosVeg Proteínas Triglic. Colest. Hb.	Triglic. Colest. Edad Proteínas OtrosVeg	OtrosVeg Proteínas Triglic. Colest. Edad	Triglic. Colest. Edad Proteínas OtrosVeg	Triglic. Colest. OtrosVeg. Hb. Vísceras

Tabla 2: Porcentajes de individuos mal clasificados por J48 antes y después de la selección de los atributos

Porcentajes de individuos mal clasificados por J48								
Antes	**Después de la selección de los atributos**							
J48	***Cfs***	***Consistency***	***Classifier***	***Wrapper***	***ChiSquared***	***GainRatio***	***InfoGain***	***OneR***
14.75%	15.10%	16.18%	12.59%	13.67%	15.46%	15.46%	15.46%	16.90%

Atributos	Frecuencia
TRIGLICÉRIDOS	8
COLESTEROL	7
PROTEÍNAS	7
EDAD	6
OTROSVEG.	6
HB	2

Ejemplo en WEKA



Selección de atributos

Conclusiones

- Los métodos de selección de atributos son un paso importante previo al procesamiento de modelos en machine learning, mejoran la eficiencia, precisión y pueden ahorrar costos de medición.

- Existen distintas maneras de realizar la selección de atributos, entre ellas los filtros, wrappers e híbridos; los primeros pueden ser más rápidos pero pueden dejar características irrelevantes, los segundos más lentos pero consideran efectos entre conjuntos de variables; los últimos una combinación de los 2 anteriores.

- La selección de atributos, tiene aplicación en muchos campos tales como la genética, minería de texto, industria y marketing.

Bibliografía

- *Eduardo Morales*, Jesús González. Selección de atributos. INAOE. Mayo, 2010.
- libGen—Descarga de libros
- S. B. Kotsiantis, “Feature selection for machine learning classification problems : a recent overview,” 2011.
- W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, “Feature Selection, Encyclopedia of Systems Biology,” no. 1, 2013, pp. 1889–2113.
- Ramos, R. M., Palmero, M. R. M. R., Ávalos, R. G., & Lorenzo, M. M. G. (2007). Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños. *Gaceta Médica Espirituana*, *9*(1), 1.