Reducción de la dimensionalidad: Selección de Características

Minería de Datos

José T. Palma

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

DIIC, UMU, 2015

Contenidos de la presentación

1. Introducción
2. Selección de Características
   - Generación de subconjuntos
3. Métodos basados en filtros
   - Medidas de relevancia
   - Clasificadores de características (Rankers)
4. Métodos basados en envoltura (wrappers)

Introducción

- Como ya es conocido en la actualidad, las datos disponibles para análisis se van acumulando a una velocidad sin precedentes.
- Además, la minería de datos no tiene que ser el único objetivo de la recogida de datos.
- Por lo tanto, el preprocesamiento, y en especial la reducción de la dimensionalidad es un aspecto crucial para una minería de datos eficiente.
- Concretamente, la selección de características es un aspecto crucial para reducir el tamaño de los datos.

Introducción: ¿Por qué reducir la dimensionalidad? I

- La complejidad de la mayoría de algoritmos de aprendizaje depende de la dimensión del conjunto de entrada (número de variables) y del número de instancias.
  - **La maldición de la dimensionalidad:** A medida que aumentamos el número de variables se reduce la densidad de los datos, con lo que la determinación de los hipersuperficies de clasificación se vuelve más difícil.
  - La precisión de las consultas y su eficiencia se degrada rápidamente a medida que aumenta la dimensionalidad.
  - Modelos simples son más robustos en conjuntos pequeños.

Introducción: ¿Por qué reducir la dimensionalidad? II

- La dimensión intrínseca del problema suele ser pequeña.
  - Sólo un subconjunto de las variables de entrada suele tener relevancia para el proceso de aprendizaje.
- Cuando tenemos menos variables para explicar los datos:
  - nos podemos hacer una mejor idea de los procesos que generan dichos datos, y
  - esto facilita la extracción de conocimiento.
- Al representar los datos en unas pocas dimensiones facilitamos su representación gráfica y su análisis visual.

Métodos I

- Existen dos tipos de métodos para la reducción de la dimensionalidad: *Extracción de Características* y *Selección de Características.*
- **Extracción de Características**.
  - **Objetivo:** Encontrar k dimensiones que sean combinación de las d $(d > k)$ dimensiones originales.
  - Básicamente consiste en encontrar una transformación desde un espacio de dimensión d a un espacio con menos dimensiones, k.
  - Hacen uso de todas las dimensiones originales.
  - Supervisados: maximizar la discriminación entre clases.
  - No supervisados: minimizar la pérdida de información.

Métodos II

- Técnicas:
  - Análisis lineal discriminante (LDA)
  - Análisis discriminante generalizado (GDA)
  - Análsis de componentes principales (PCA) y su versión con funciones Kernel.
  - Modelos de variables latentes basados en procesos gaussianos constreñidos
  - t-distributed stochastic neighbor embedding (t-SNE)
  - Uniform Manifold Approximation and Projection (UMAP)

Métodos III

- **Selección de Características**.
  - **Objetivo:** Encontrar que k dimensiones, de las d dimensiones originales, aportan la mayor cantidad de información, descartando el resto $(d - k)$.
  - Básicamente, se trata de encontrar el subconjunto óptimo de dimensiones de acuerdo con alguna determinada función objetivo.
  - Sólo se selecciona un subconjunto de las dimensiones originales.
  - Se trabaja directamente con las dimensiones originales sin transformar.

Selección de Características

- La **Selección de Características** trata de encontrar el subconjunto de características optimo de acuerdo con un determinado criterio de selección.
- En este caso no se realiza ningún proceso de transformación de las características.
- Dicho subconjunto óptimo de características debe garantizar que el proceso al que se aplique la selección se pueda desarrollar con totales garantías, minimizando el número de características.

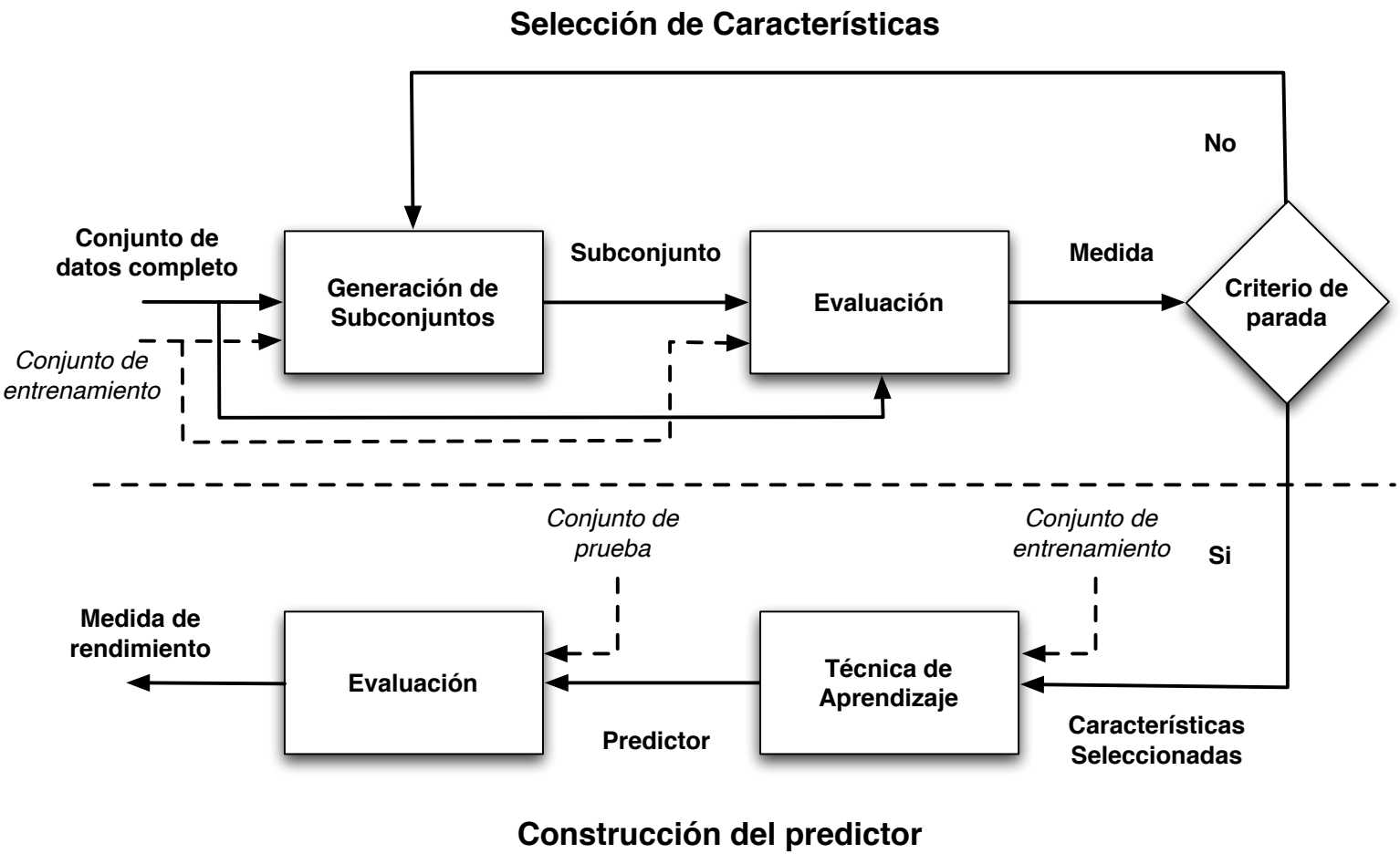
Subconjunto óptimo

- El **subconjunto óptimo** de características es el subconjunto mínimo que permite construir una hipótesis consistente con los datos de entrenamiento.
  - Podemos utilizar todos los datos o sólo los de entrenamientos.
  - El subconjunto se puede sobreajustar a los datos de entrenamiento.
- Sea F el conjunto de todas las caracerísticas.
- El subconjunto óptimo es conjunto mínimo de características G tal que $P(C|G)$ es igual o lo más próxima posible a $P(C|F)$.
  - Esta definición se apoya en todo el conjunto de datos,
  - o bien, sólo cuando los datos de entrenamiento están disponibles.

Selección de características: Métodos I

- El proceso de selección de características se desarrolla en dos fases:
  1. Generación de subconjuntos de características. Se van generando los distintos subconjunto de características candidatos a ser el subconjunto óptimo.
  2. Se evalúan cada uno de los subconjuntos candidatos hasta que se cumpla algún criterio de parada.
  3. Finalmente, se procede a generar el predictor teniendo en cuenta sólo las características seleccionadas.

Selección de características: Métodos II



Selección de características: Métodos III

- Dependiendo de cómo se realice la evaluación de cada uno de los subconjuntos tenemos:
  - **Métodos basados en filtros**, que evalúan la relevancia de las características teniendo en cuenta sólo las propiedades intrínseca de los datos.
  - **Métodos basados en envoltura (wrappers)**, en los que cada subconjunto candidato es evaluado a través de la eficacia de un predictor.
  - **Métodos empotrados (embedded)**, en los que el proceso de selección de variables está integrado en la técnica de construcción del predictor.

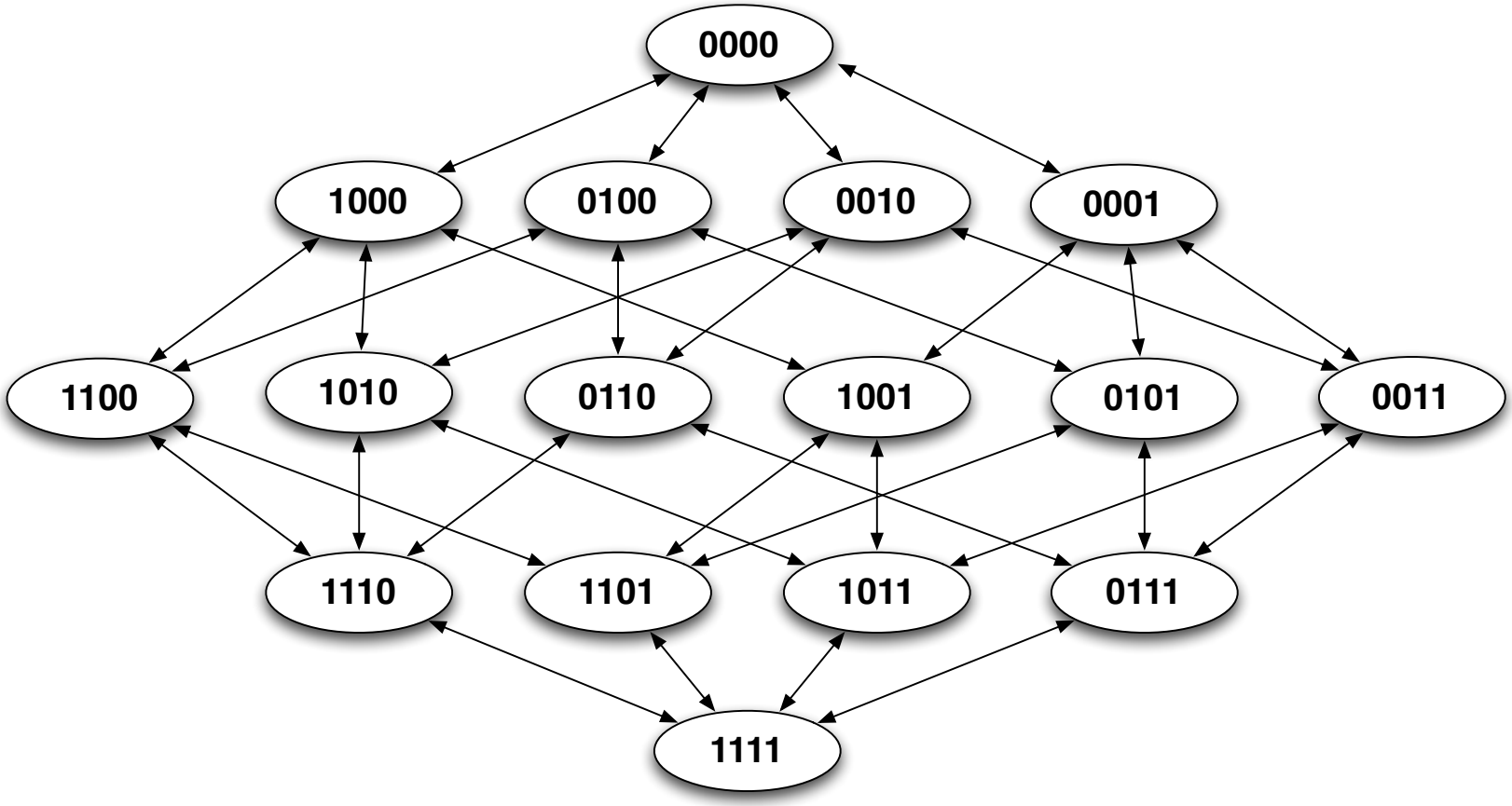
Generación de subconjuntos

- El proceso de selección de características se puede plantear como un proceso de búsqueda en el espacio de posibles subconjuntos de características.
- Una búsqueda exhaustiva, en la mayoría de los casos, no es computacionalmente posible.
  - con $n = 12$ características tendríamos 4096 subconjuntos y para $n = 100$ características tendríamos 10^{100}.
  - Incluso en el caso de que estuviéramos buscando exactamente un subconjunto de tamaño m el número de posibles subconjuntos sería:

$$\binom{n}{m} = \frac{n!}{(n-m)!m!}$$

que puede ser mucho menor que 2^n, pero desde el punto de vista computacional muy grande.

Generación de subconjuntos



Ejemplo de espacio de estado para 4 características.

Generación de subconjuntos: Estrategias de búsqueda I

- Se pueden utilizar diferente estrategias de búsqueda en el espacio de posibles subconjuntos:
  - **Exhaustiva**. Se barre todo el espacio de posibles subconjutos:
    - Se puede recorrer el espacio en profundidad o en anchura.
    - Sólo es posible para pocas características.
    - Es la única forma de garantizar encontrar el subconjunto óptimo.
  - **Heurística**. Disponen de alguna información sobre qué subconjunto es el más prometedor.
    - No garantizan que el subconjunto encontrado sea el óptimo.
    - Normalmente encuentran una buena solución en un tiempo razonable.

Generación de subconjuntos: Estrategias de búsqueda II

- **Aleatoria**. Se parte de una configuración inicial formada por un conjunto finito de posibles subconjuntos.
  - Mediante una pequeña transformación, se va modificando la configuración inicial para dirigir la búsqueda hasta la solución final.
  - No garantiza la solución óptima.

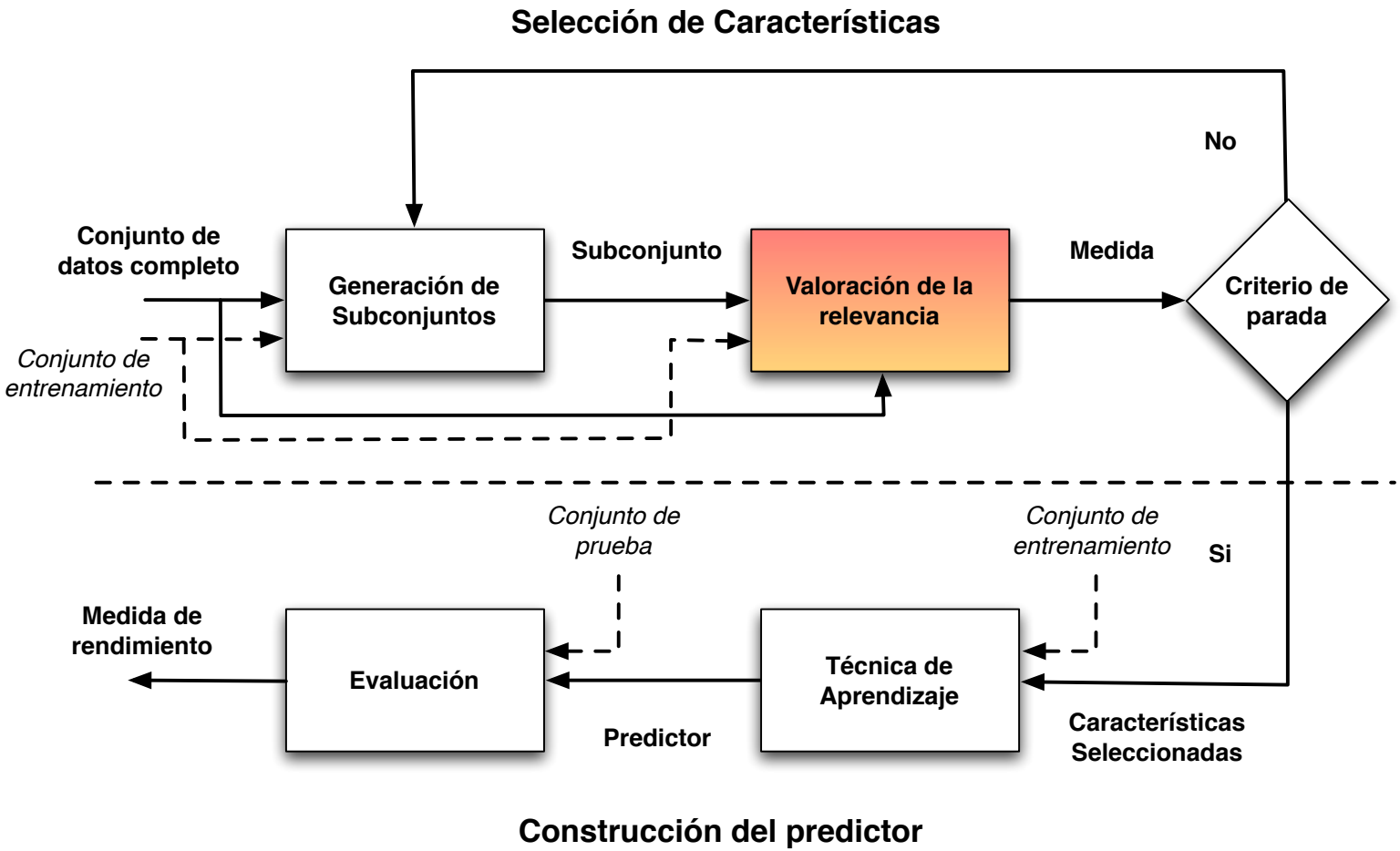
Generación de subconjuntos: Dirección de Búsqueda

- Para definir una estrategía de búsqueda primero debemos definir **la dirección de búsqueda**:
  - **Hacia delante (Forward)**. Se empieza con el conjunto vacío y se van añadiendo una característica no seleccionada cada vez.
  - **Hacía atrás (Backward)**. Se empieza con todas las características y se va eliminando una característica cada vez.
  - **Bidireccional.** Se comienza por los dos extremos del espacio de búsqueda y se realiza de forma paralela una búsqueda hacia delante y otra hacia detrás.

Selección de características: Filtros

- Los métodos basados en filtros evalúan la relevancia de las características teniendo en cuenta sólo las propiedades intrínsecas de las mismas.
- Son independientes de la técnica de clasificación que se va a utilizar.
  - La búsqueda en el espacio de subconjunto de características está desacoplado del de las hipótesis.
- Suelen ser rápidos, bastante escalables y no se ven sesgados por la influencia de una determinada técnica de clasificación.

Selección de características: Filtros



Esquema general de un método de selección basado en filtros

Selección de características: Medidas de relevancia

- La valoración de la idoneidad de un subconjunto de caracteristicas depende de la medida utilizada:
  - Medidas basadas en la cantidad de información.
  - Medidas basadas en distancias.
  - Medidas basadas en consistencia.

Medidas de relevancia: Cantidad de información I

- Esta medida está basada en la ganancia de información proporcionada por las distintas características.
- La ganancia de información se puede interpretar como la reducción de la incertidumbre al clasificar, dado un conjunto de elementos definidos a través de un conjunto de características.
- Supongamos:
  - Un conjunto de datos definido a través de n características $X = \{x_1, x_2, \ldots, x_n\}$.
  - Cada elemento de X está etiquetado como perteneciente a una clase del conjunto $C = \{c_1, c_2, \ldots, c_l\}$.

Medidas de relevancia: Cantidad de información II

- La incertidumbre inicial teniendo en cuenta sólo la información procedente de la distribución de las clases se puede medir mediante la entropia.

$$E(C) = -\sum_{i=1}^{I} P(c_i) \log_2 P(c_i)$$

- Mediante la entropía condicional podemos medir la incertidumbre asociadas a las clases teniendo en cuenta la información proporcionada por el conjunto datos X.

$$E(C|X) = -\sum_{j=1}^{n} P(x_j) \left(\sum_{i=1}^{I} P(c_i|x) \log_2 P(c_i|x) \right)$$

Medidas de relevancia: Cantidad de información III

- La ganancia de información teniendo en cuenta la información proporcionada por X es:

$$IG(C|X) = E(C) - E(C|X)$$

- $IG(C|X)$ nos da una medida de la capacidad del conjunto de características del conjunto X a la hora de predecir las clases.
- El conjunto de características de X es totalmente irrelevante si la ganancia de información es igual a 0.
- Es poco práctico en problemas de alta dimensionalidad y pocos datos, por la dificultad de estimar las probabilidades condicionales.

Medidas de relevancia: distancias

- Al utilizar una medida basada en distancia lo que se intenta es medir la separación entre clases.
- Elementos de distintas clases deben estar separados en el espacio.
- Por lo tanto, un conjunto de características es mejor que otro si los datos definidos a partir de ellas tienen una distancia entre clases mayor.
- En el capítulo sobre agrupamiento analizaremos distintas medidas de este tipo.

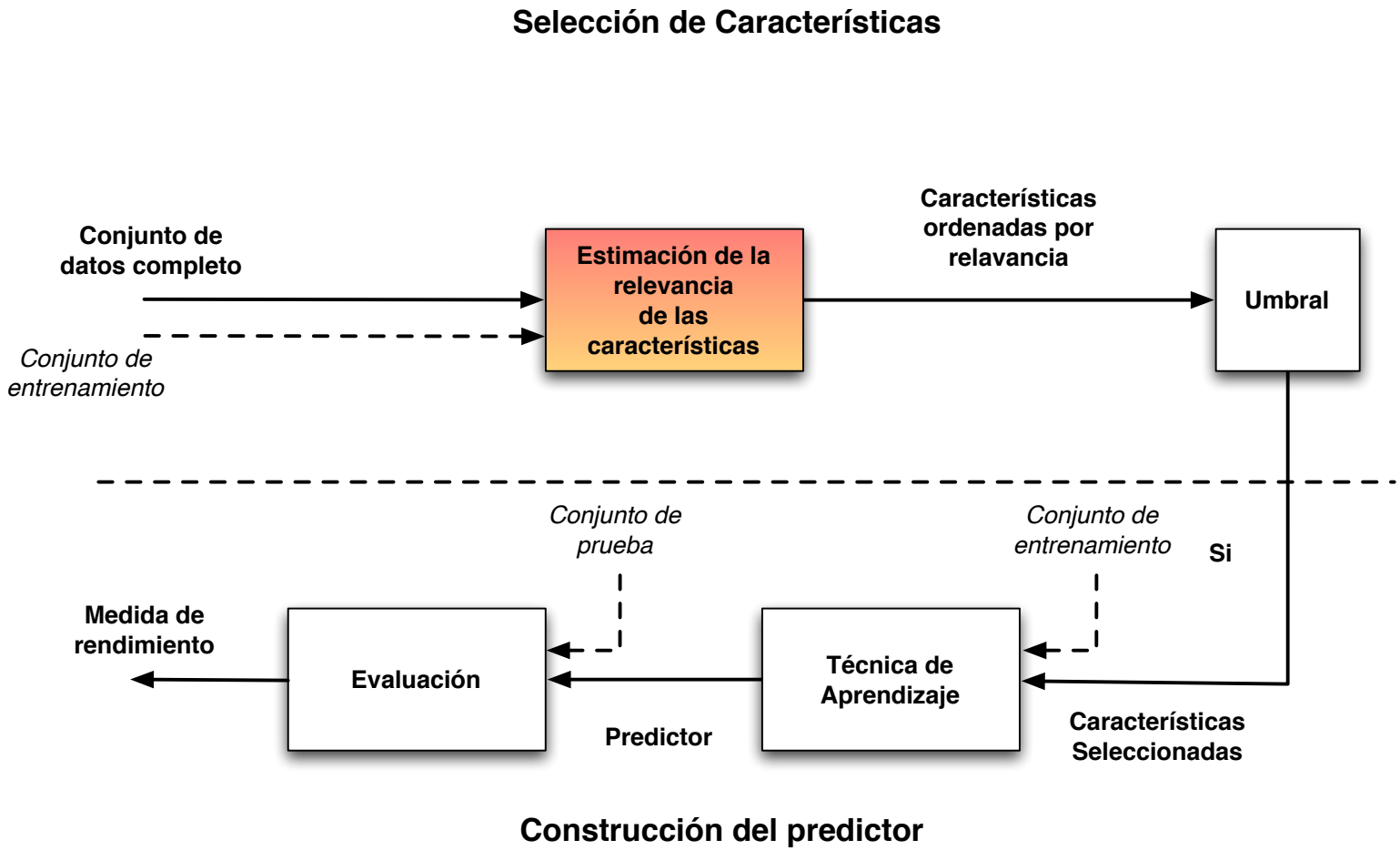
Medidas de relevancia: consistencia

- El objetivo es encontrar el conjunto de mínimo de características que separan las clases de la forma más consistente que el conjunto de datos permita.
- Una inconsistencia aparece cuando tenemos dos instancias iguales, de acuerdo con el subconjunto de características evaluado, que pertenecen a clases distintas.
- Un conjunto de características será mejor que otro si el ratio de inconsistencias que produce es menor.

Selección de características: Clasificadores de características (Rankers)

- Existe un caso especial de métodos basados en filtros que eliminan la búsqueda de subconjuntos.
- Al eliminar la búsqueda lo único que se hace es aplicar una medida que indique la relevancia de cada característica por separado.
- Una vez obtenida la lista, se seleccionan aquellas características que están mejor clasificadas.
- Este tipo de técnicas son muy rápidas y fácilmente escalables.
- Tiene el inconveniente de que en algunos casos es difícil establecer el umbral de corte.

Selección de características: Rankers



Esquema general de un método de un filtro de tipo ranker

Selección de características: Rankers

- En estos casos, las medidas de relevancia se suelen basar en test estadísticos.
  - **Paramétricos**: t-test, ANOVA, Test de Welch, ...
  - **No paramétricos**: Test de rangos con signos de Wilcoxon, test de suma de rangos con signos de Wilcoxon, test de Kruskal-Wallis.
- El problema de los rankers es que sólo permiten una análisis univariable:
  - Cada característica es considerada por separada y no se tienen en cuenta las dependencias entre ellas.
- Para abordar análisis multivariable debemos acudir a los filtros con búsqueda en el espacio de subconjuntos de características.

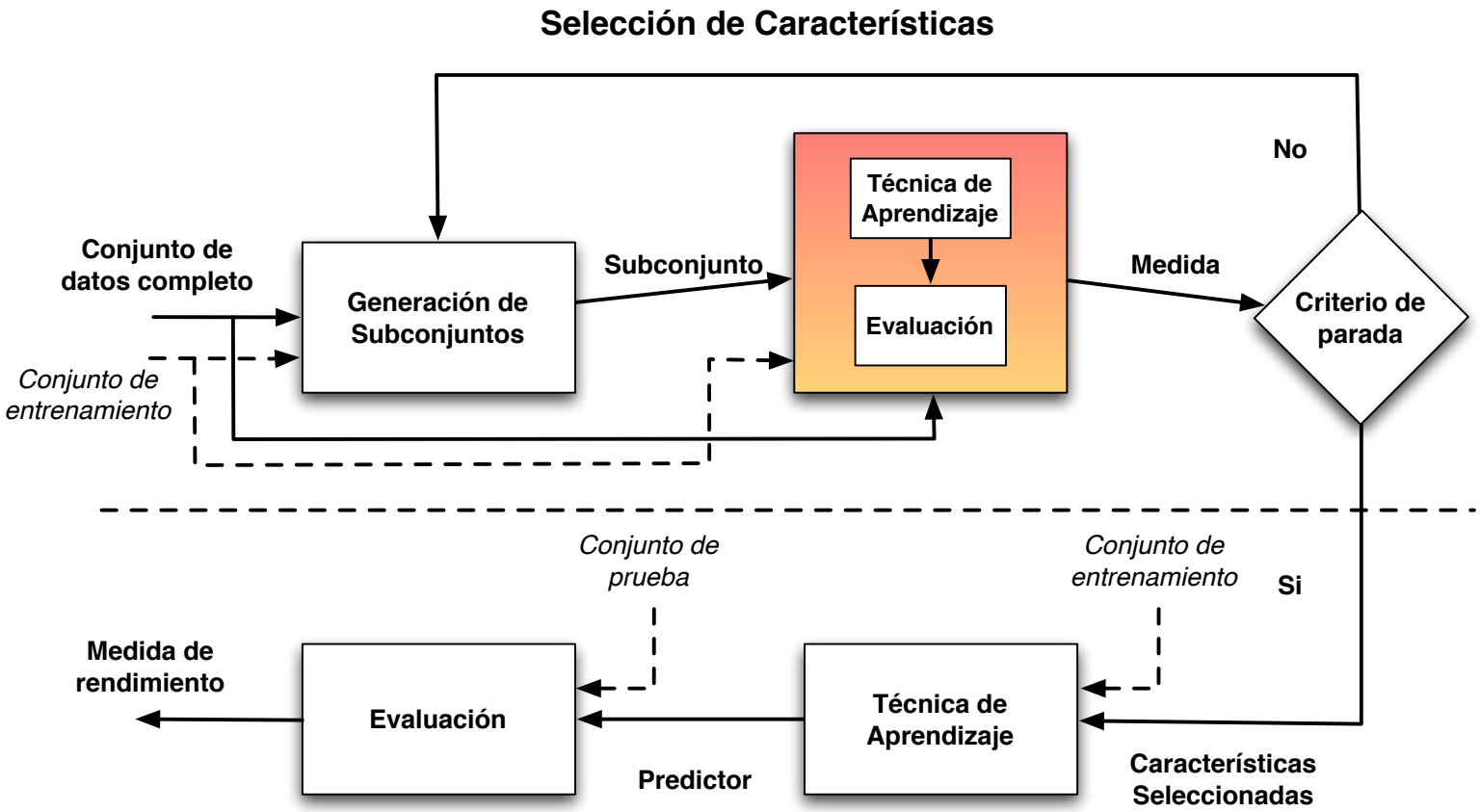
Filtros: Algoritmos representativos

- **Rankers**: Relief [Kira and Rendell, 1992] calcula la relevancia de las características apoyándose en medidas de distancia.
- **Filtros**:
  - Focus [Almuallim and Dietterich, 1994]: búsqueda exhaustiva con medida de consistencia.
  - CFS (Correlation-Based Feature Selection) [Hall, 1999]: búsqueda heurística (ramificación y poda) con una medida basada en la correlación.

Selección de características: Wrappers

- Este tipo de métodos, la evaluación de cada uno de los subconjuntos candidatos se realiza mediante la construcción de un clasificador (o regresor).
  - Por lo tanto, como medida de evaluación utilizan la capacidad predictiva del clasificador.
- De esta forma, se selecciona el subconjunto de características que produce el mejor clasificador.
- Ventajas:
  - Producen mejores resultados al estar orientados al problema de clasificación.
- Desventajas:
  - Tienen un mayor riesgo de sobreajuste que los filtros.
  - Son más costosos computacionalmente.

Selección de características: wrappers



Esquema general de un método de un wrapper

Wrappers: esquema con eliminación recursiva

Algoritmo Eliminación recursiva de características

1: Construir un modelo con todas las características
2: Evaluar el modelo
3: Calcular la relevancia de las características
4: Crear una lista con las características ordenadas de mayor a menor relevancia.
5: **para** $size = n$ to 1 **hacer**
6: Crear un conjunto S_{size} con las $size$ características más relevantes
7: Construir un modelo utilizando las características S_{size}
8: Evaluar el modelo
9: [Opcional] Recalcular la relevancia de las características
10: **fin para**
11: Crear una lista con todos los S_i y el resultado de la evaluación
12: Determinar el subconjunto óptimo S_{opt}

Wrappers: esquema con eliminación recursiva

- El mismo esquema se puede adaptar para una búsqueda hacía delante de subconjuntos.
  - Se empieza por un conjunto de tamaño 1.
  - Se van agregando las características más relevantes.
- La principal desventaja de este tipo de métodos, a parte de su coste computacional, es el sobreajuste.
  - Para evitarlo, hay esquemas que incluyen un bucle externo para llevar a cabo un remuestreo.

Wrappers: eliminación recursiva con remuestreo

Algoritmo Eliminación recursiva de características con remuestreo

para *Cada iteración de remuestreo* **hacer**
    Crear los conjunto de entrenamiento E y prueba T
    Construir un modelo sobre E con todas las características
    Evaluar el modelo en T
    Calcular la relevancia de las características
    Crear una lista con las características ordenadas de mayor a menor relevancia.
    para $size = n$ to 1 **hacer**
        Crear un conjunto S_{size} con las *size* características más relevantes
        Construir un modelo utilizando las características S_{size}
        Evaluar el modelo
        [Opcional] Recalcular la relevancia de las características
    fin para
fin para
Crear una lista con todos los S_i y el resultado de la evaluación
Determinar el subconjunto óptimo S_{opt}
Crear un modelo con las variables S_{opt} y con el conjunto de entrenamiento original.

Wrappers: Algoritmos representativos

- OBLIVION [Langley and Sage, 1994]: búsqueda voraz y árboles de decisión.
- RFE+SVM [Guyon *et al.*, 2002]: eliminación recursiva de características y SVM.
- FFE + NNets [Goutte, 1997]: búsqueda hacia delante y redes neuronales.
- GA + C4.5 [Abbasimehr and Alizadeh, 2013]: búsqueda aleatoria y C4.5.
  - También se pueden encontrar muchos ejemplos basados en colonias de hormigas y enfriamiento simulado.
- También es posible implementar modelos de tipo ranker con predictores que calculen la relevancia de las características.

Selección de características: conclusiones I

- En la actualidad, es crucial la aplicación de técnicas de reducción de la dimensionalidad antes de abordar la creación de modelos predictivos.
- Estas técnicas se pueden agrupar en dos grandes grupos: técnicas de **extracción de características** y **selección de características**.
- En este capítulo nos hemos centrado en las técnicas de selección características, que se centran en encontrar un conjunto óptimo de características de acuerdo con algún criterio.

Selección de características: conclusiones II

- Dependiendo del criterio elegido las técnicas puedes ser de tipo filtro o de tipo envoltura (wrapper):
  - Las técnicas de tipo filtro sólo se apoyan en las propiedades intrínsecas de las características.
  - Las técnicas de tipo wrappers se basan en la construcción de clasificadores (o regresores) para evaluar la idoneidad de los conjuntos de características.
- Un aspecto importante en ambos tipos de técnicas es la técnica de búsqueda utilizada para recorrer el espacio de búsqueda de posibles subconjuntos de características.
  - Búsquedas exhaustivas, heurísicas o aleatorias.

Referencias I


H. Abbasimehr and S. Alizadeh.
A novel genetic algorithm based method for building accurate and comprehensible churn prediction models.
International Journal of Research in Industrial Engineering,, 2(4):1–14, 2013.


Hussein Almuallim and Thomas G Dietterich.
Learning boolean concepts in the presence of many irrelevant features.
Artificial Intelligence, 69(1):279–305, 1994.


Ethem Alpaydin.
Introduction to machine learning.
MIT press, 2014.


Krzysztof J Cios, Witold Pedrycz, and Roman W Swiniarski.
Data mining methods for knowledge discovery, volume 458.
Springer Science & Business Media, 2012.


Richard O Duda, Peter E Hart, and David G Stork.
Pattern classification.
John Wiley & Sons, 2012.

Referencias II


C. Goutte.
Extracting the relevant decays in time series modelling.
In *Proceedings of the VII IEEE Workshop, Neural Networks for Signal Processing*, 1997.


Isabelle Guyon and André Elisseeff.
An introduction to feature extraction.
In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and LotfiA. Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 1–25. Springer Berlin Heidelberg, 2006.


Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik.
Gene selection for cancer classification using support vector machines.
Machine Learning, 46(1):389–422, 2002.


Mark A Hall.
Correlation-based feature selection for machine learning.
PhD thesis, The University of Waikato, 1999.

Referencias III


Kenji Kira and Larry A Rendell.
A practical approach to feature selection.
In *Proceedings of the ninth international workshop on Machine learning*, pages 249–256, 1992.


Vipin Kumar and Sonajharia Minz.
Feature selection: A literature review.
Smart Computing Review, 4(3):211–229, June 2014.


Langley and Stephanie Sage.
Oblivious decision trees and abstract cases.
In *Proc. AAAI-94 Workshop on case-based reasoning*, pages 113–117, 1994.


Yvan Saeys, Iñaki Inza, and Pedro Larrañaga.
A review of feature selection techniques in bioinformatics.
bioinformatics, 23(19):2507–2517, 2007.