LIDIA
Laboratorio de I+D en Inteligencia Artificial

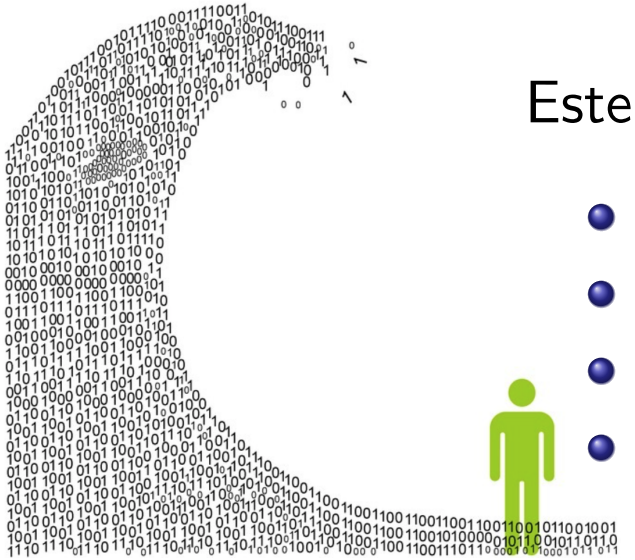
UNIVERSIDADE DA CORUÑA

CITIC

Método de selección de características distribuido basado en medidas de complejidad de datos

L. Morán-Fernández, V. Bolón-Canedo & A. Alonso-Betanzos

En la era del **Big Data** muchos conjuntos de datos presentan una peculiaridad común: el elevado número de **características**.



Este problema puede afectar a...

- Complejidad temporal
- Requisitos de almacenamiento
- Falta de interpretación
- Generalización del error (ruido y sobreajuste)

SATURACIÓN OXÍGENO

VELOCIDAD

ECG

CO_2

PULSO

POSTURA

EMG

CAÍDA

RESPIRACIÓN

TEMPERATURA DE LA PIEL

ACELERACIÓN

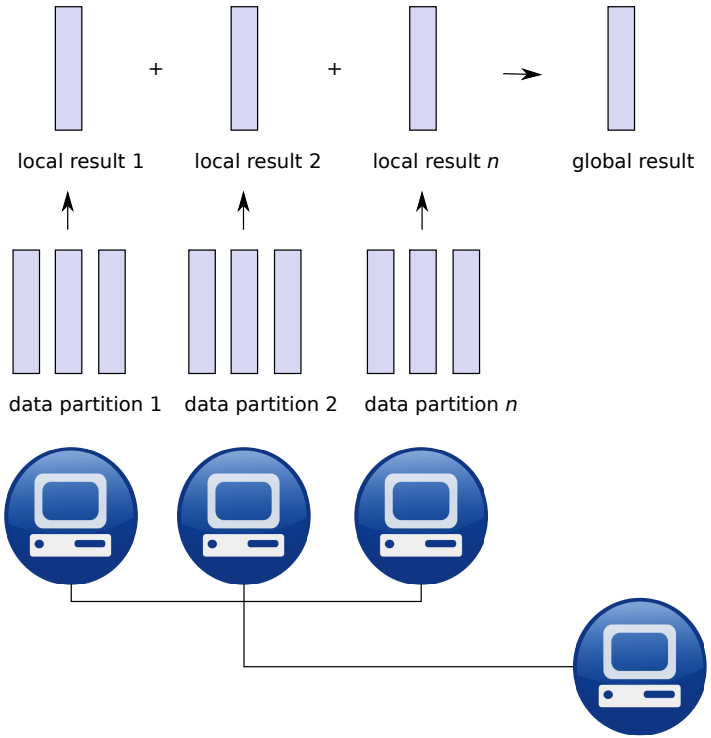
La **Selección de características** es el proceso de detectar las características relevantes e ignorar las irrelevantes y redundantes.

- Ventajas:
  - Rendimiento de los algoritmos de Aprendizaje automático
  - Compresión, visualización y conocimiento de los datos
  - Reducción datos, limitación de los requisitos de almacenamiento y reducción de costes

- Los datos pueden estar distribuidos en múltiples localizaciones
  - No es económico e incluso legal mantenerlos en la misma localización
- La mayoría de los métodos de selección existentes **no escalan bien** y por ello, una posible solución es **distribuir** los datos

El objetivo del proceso selección de características distruido es **reducir el tiempo computacional** mientras se **mantiene el rendimiento de clasificación**.

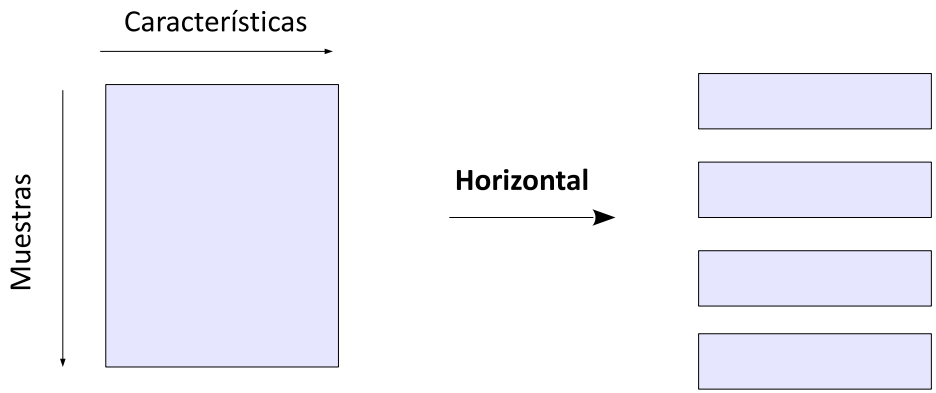
Metodología



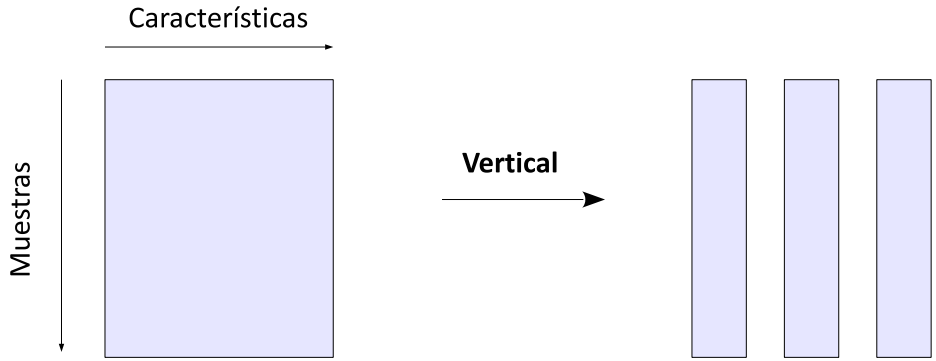
1. Partición del conjunto de entrenamiento
2. Aplicación del algoritmo de selección de características a los subconjuntos
3. Combinación de los resultados en un único subconjunto de características

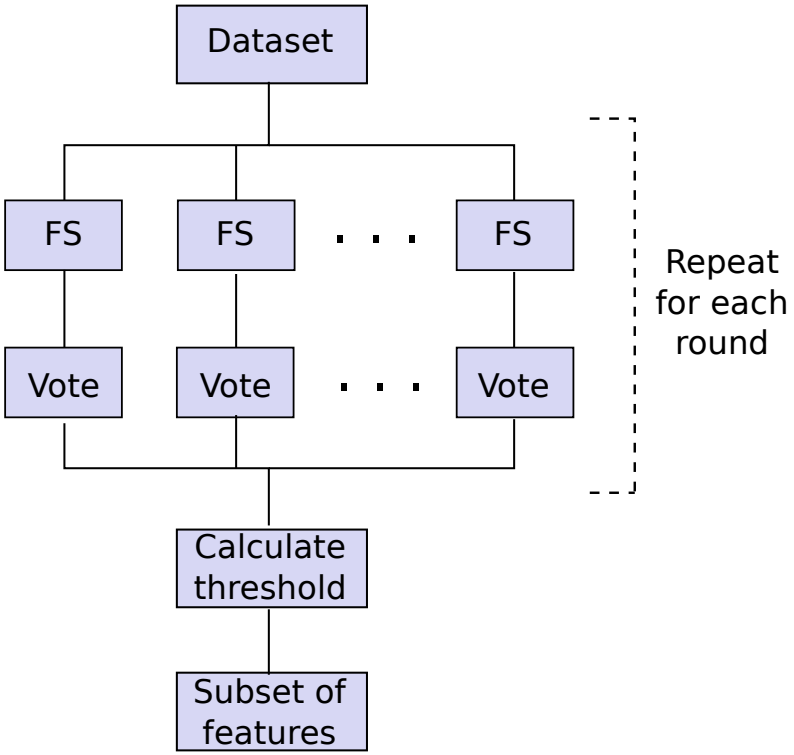
Tipos de partición

- Por **muestras**



- Por **características**





- Cada vez que una característica es seleccionada para ser eliminada, recibe un voto
- Umbral: mejor valor para el número de votos dependiendo de su efecto en el conjunto de entrenamiento
- Para calcular el umbral:
  $e[v] \leftarrow \alpha \times \mathit{error} + (1 - \alpha) \times \mathit{\%nCaract}$

Problemas

- Metodología dependiente del clasificador elegido
- Sobrecarga en el tiempo de ejecución

Nuestra propuesta

- Combinar las salidas parciales utilizando medidas de complejidad de datos

Problemas

- Metodología dependiente del clasificador elegido
- Sobrecarga en el tiempo de ejecución

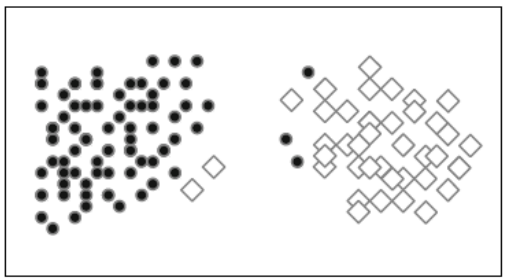
Nuestra propuesta

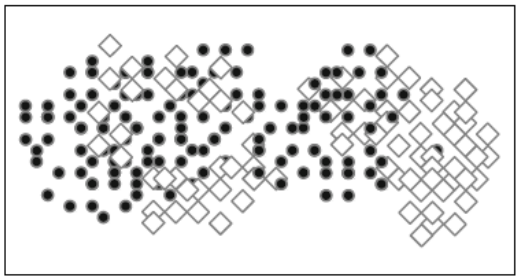
- Combinar las salidas parciales utilizando medidas de complejidad de datos

Medidas de complejidad de datos

- Medidas de solapamiento en características de clases diferentes
  - Razón discriminante de Fisher
  - Longitud de la región de solapamiento

- Medidas de separabilidad de clases
  - Media de la distancias de vecinos más cercanos intra/inter clases





$$e[v] \leftarrow \alpha \times \textit{medidaComplejidad} + (1 - \alpha) \times \%\textit{nCaract}$$

Justificación

Conservar características que contribuyen a reducir la complejidad del conjunto de datos y descartar aquellas que la incrementen.

Conjuntos de datos:

	#Características	#Muestras		#Clases
		Entrenamiento	Test	
Connect4	42	45038	22519	3
Isolet	617	6238	1236	26
Madelon	500	1600	800	2
Ozone	72	1691	845	2
Spambase	57	3067	1534	2
Mnist	717	40000	20000	2
Breast	24481	78	19	2
Gli85	22283	56	29	2
CLL-SUB-111	11340	74	37	3
Lung cancer	12600	136	68	5
11-Tumors	12534	114	58	11

Métodos de selección de características

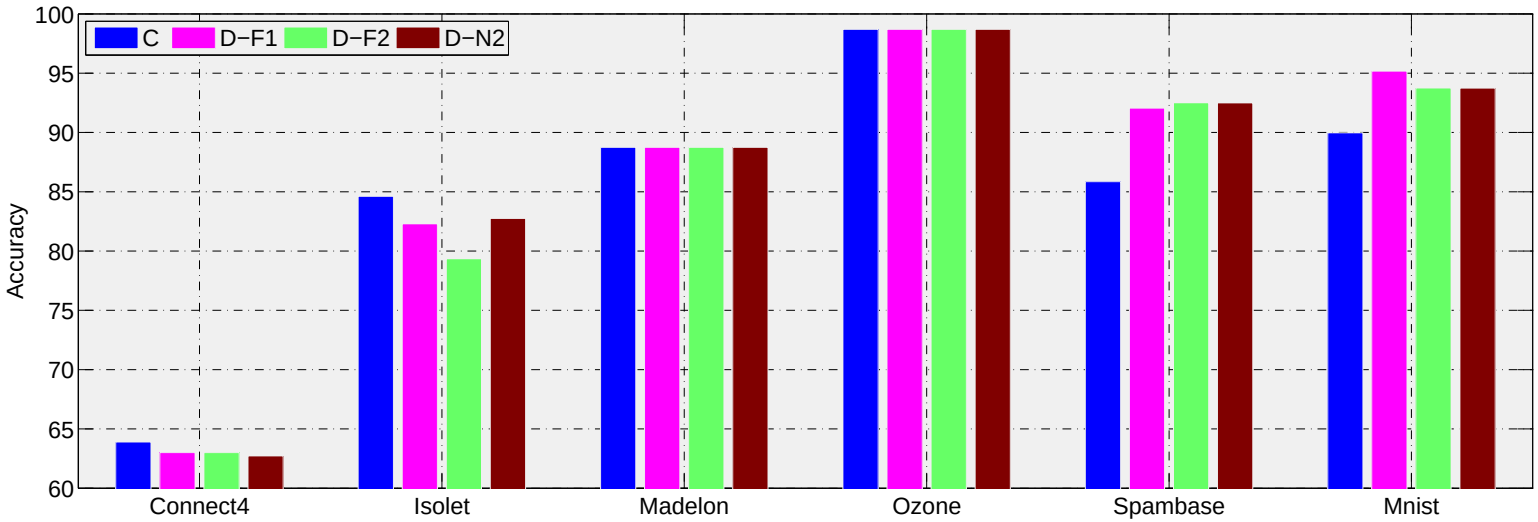
- Correlation-based Feature Selection
- Consistency-based filter
- INTERACT
- Information Gain
- ReliefF

Clasificadores

- C4.5
- Naive Bayes
- 1-NN
- SVM

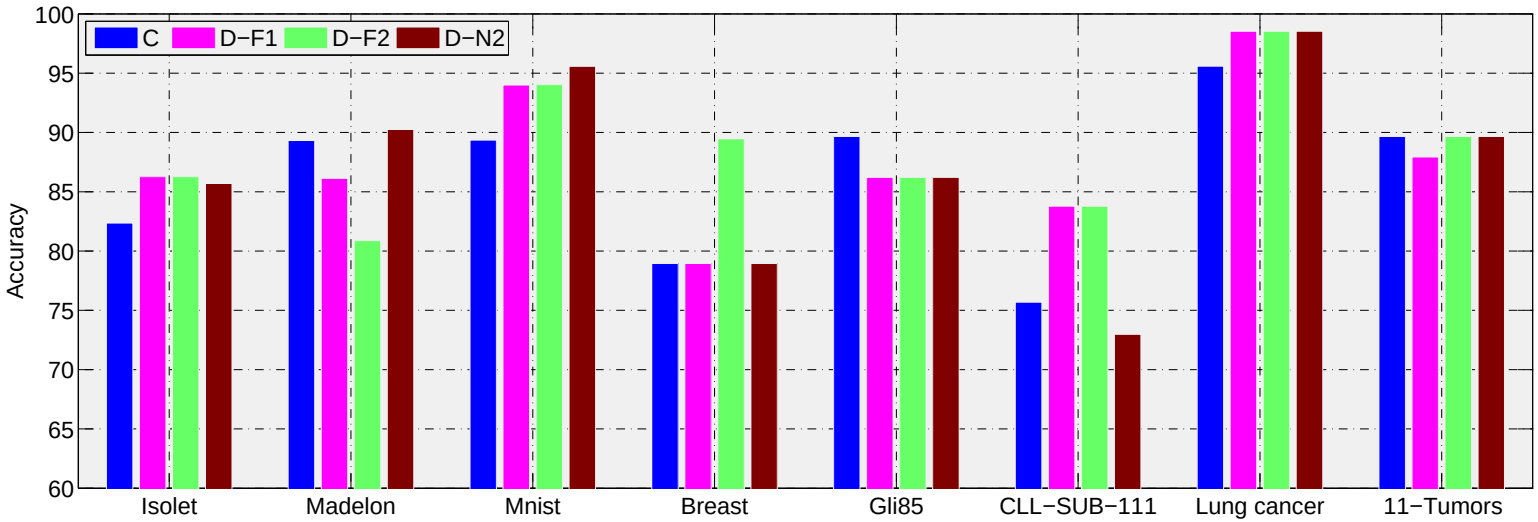
Distribución horizontal

- Centralizado (C) vs. Distribuido (D)



Distribución vertical

- Centralizado (C) vs. Distribuido (D)



Resumen

	Precisión		Tiempo de ejecución		
Método	C	D	C	D	Speed up
CFS	75.80	76.25	3322.18	34.42	96.52
INT	74.60	76.68	370.36	30.09	12.31
Cons	70.16	76.18	618.31	29.20	21.17
IG	75.59	77.24	162.57	31.29	5.19
ReliefF	76.09	77.35	2978.31	330.88	9.01
Media	74.45	**76.74**	1490.35	**91.18**	16.35

Nuestro enfoque distribuido

- Independencia del clasificador ✓
- Reducción del tiempo de ejecución ✓
- Mantener, e incluso superar, la precisión de clasificación ✓

Referencias

- **Centralized vs. distributed feature selection methods based on data complexity measures.** Knowledge-Based Systems, 2017.

- **A time efficient approach for distributed feature selection partitioning by features**. Conferencia de la Asociación Española para la Inteligencia Artificial, 2015.

¡Muchas gracias!

WGML 2017, Santiago de Compostela