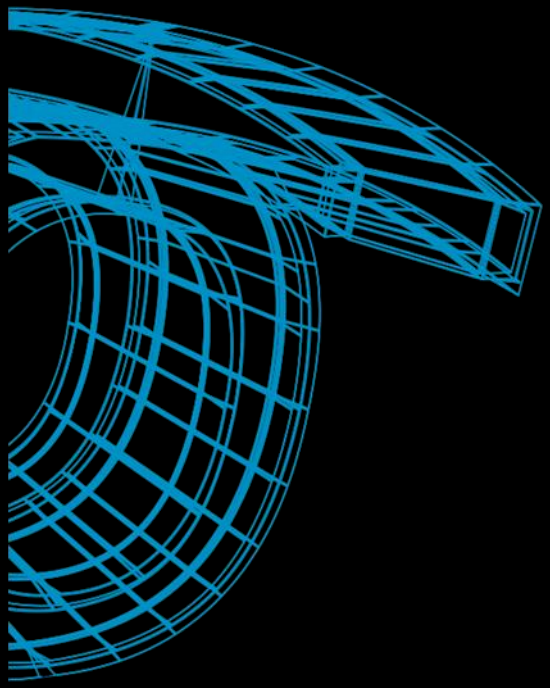
vicomtech
ik4 research alliance



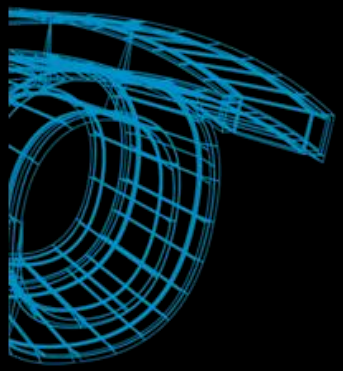
The Random Subspace Method for Constructing Decision Forests

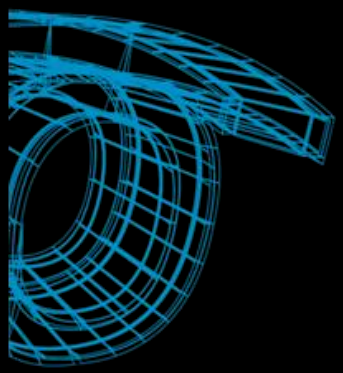
(IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 20, NO. 8, AUGUST 1998)

Iñigo Barandiaran

vicomtech
ik4 research alliance

Índice



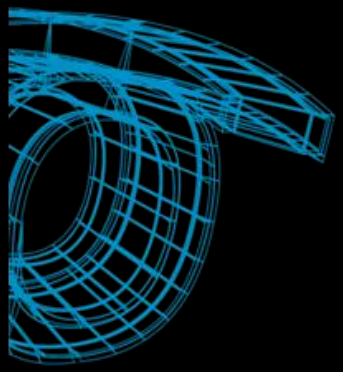


Introducción

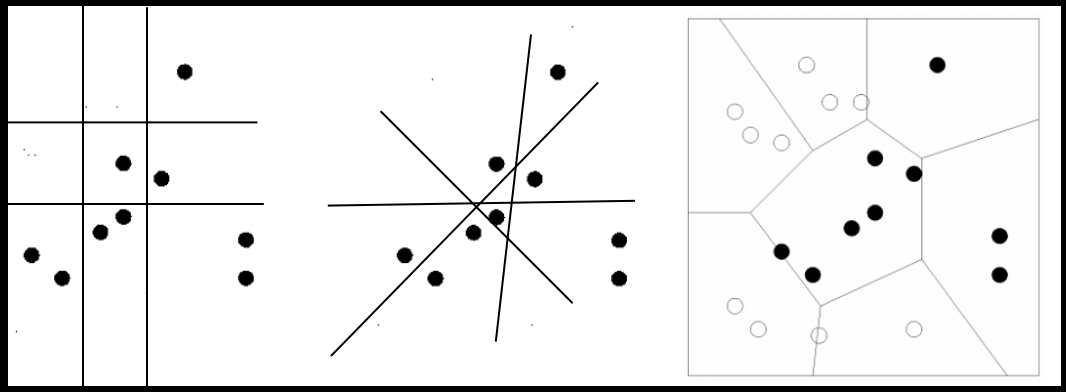
- El autor tiene una de las primeras publicaciones relacionadas con Random subspaces en "*Recognition of Harwritten Digits by Learning Vector Quantization*"
- Leo Breiman basó su artículo "Random Forest" en los desarrollos de éste autor.

vicomtech
ik4 research alliance

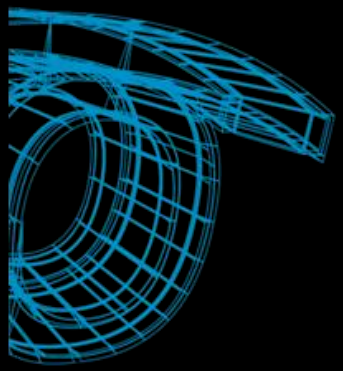
Introducción



- Métodos de construcción de árboles
  - Mecanismos de *split*
    - *Axis-parallel linear splits*
    - *Oblique linear splits*
    - *Piecewise Linear splits (voronoi tessellation)*



  - Criterios de Parada
    - Máximo número de niveles (profundidad)
    - Clase única
    - Mínimo número de instancias.

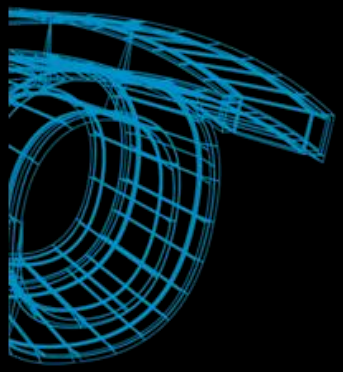


Objetivo

- Proponer un nuevo mecanismo para la construcción de árboles, dentro de un contexto de combinación de clasificadores (bosques).
  - Tratando de obtener 100% de precisión en el Training y una función de generalización monota decreciente a medida que crece el bosque
  - Que proporcione árboles diversos.
  - Evaluar como afectan diferentes factores del clasificador a los valores de precisión o generalización.

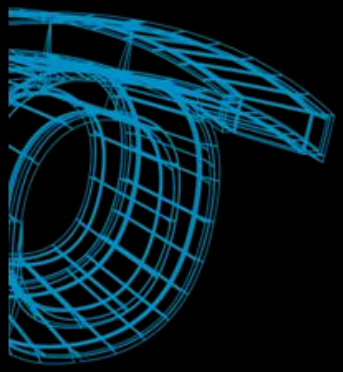
vicomtech
ik4 research alliance

Implementación



- Seleccionar un subconjunto de características en cada *split* durante el crecimiento del árbol.
- El resto de *features* no seleccionadas se asignan a un valor constante (0).
- Las instancias se proyectan sobre dicho espacio de características.
- Los árboles se entrenan sobre dicho sub-espacio.

"*Hence, while most other classification methods suffer from the curse of dimensionality , this method can take advantage of high dimensionality. And contrary to the Occam's Razor, our classifier improves on generalization accuracy as it grows in complexity*"



Implementación

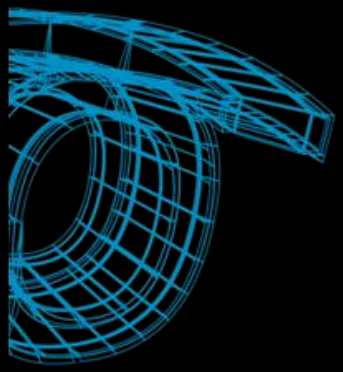
- En cada hoja se almacena la probabilidad condicionada de pertenencia a cada clase.

$$P\left(c \middle| v_j(x)\right) = \frac{P\left(c \middle| v_j(x)\right)}{\sum_{k=1}^{n_c} P\left(c_k, v_j(x)\right)}$$

- La combinación de clasificadores(función de decisión) se realiza mediante mayoría.

$$g_c(x) = \frac{1}{n_t} \sum_{j=1}^{n_t} \hat{P}\left(c \middle| v_j(x)\right)$$

- No comenta nada sobre *pruning* o no.



Implementación

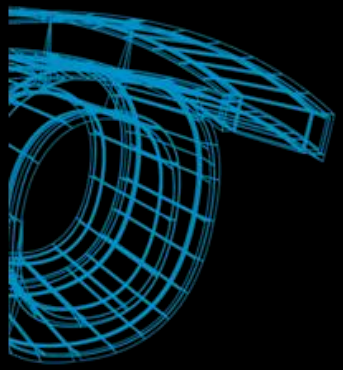
- La similaridad entre miembros la calcula, evaluando sobre un conjunto fijo de *n* ejemplos de test como:

$$\hat{s}_{i,j} = \frac{1}{n}\sum_{k=1}^{n} f(x_k),$$

$$f(x_k) = \begin{cases} 1 & \text{if } c_i(x_k) = c_j(x_k) \\ 0 & \text{otherwise} \end{cases}.$$

vicomtech
ik4 research alliance

Evaluación



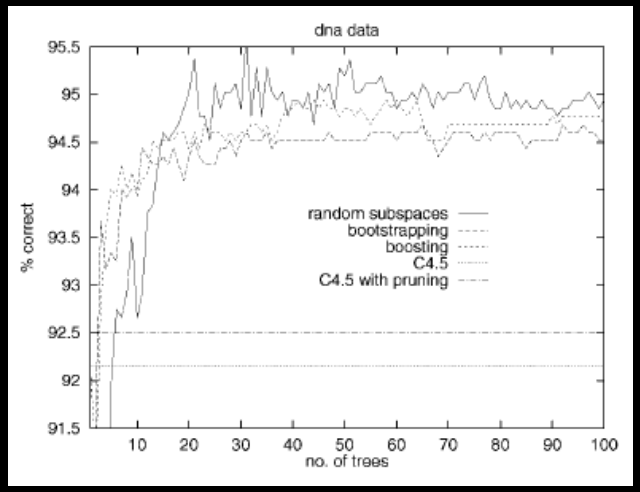
- 4 Data Sets, con diferentes nº de clases, instancias y dimensionalidad.

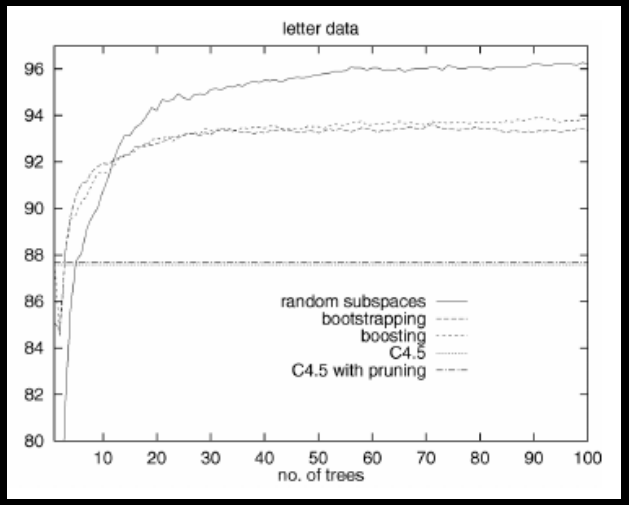
name of data set	no. of classes	no. of feature dimen-sions	no. of training samples	no. of testing samples
dna	3	180	2,000	1,186
letter	26	16	15,000	5,000
satimage	6	36	4,435	2000
shuttle	7	9	43,500	14,500

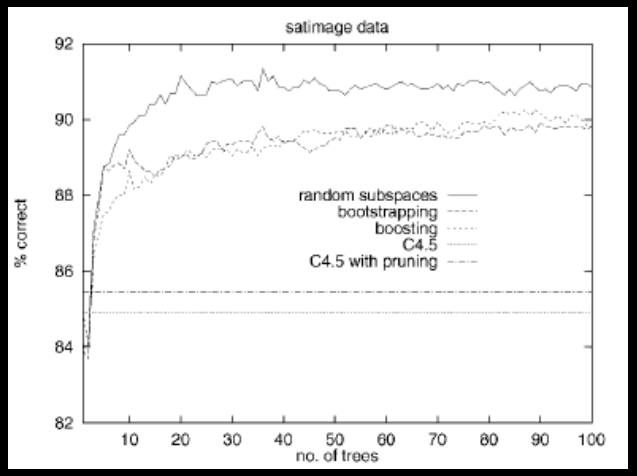
- Comparación con respecto a un solo árbol
  - Comparar precisión (*accuracy*) con respecto a un solo árbol con C4.5.
  - Comparación con respecto a *boostraping* y *Boosting* (*AdaBoost*)
  - Evaluación en relación a uso de diferentes mecanismos de *Split*.
  - Evaluación en relación al número de características aleatorias (*subespacios*).

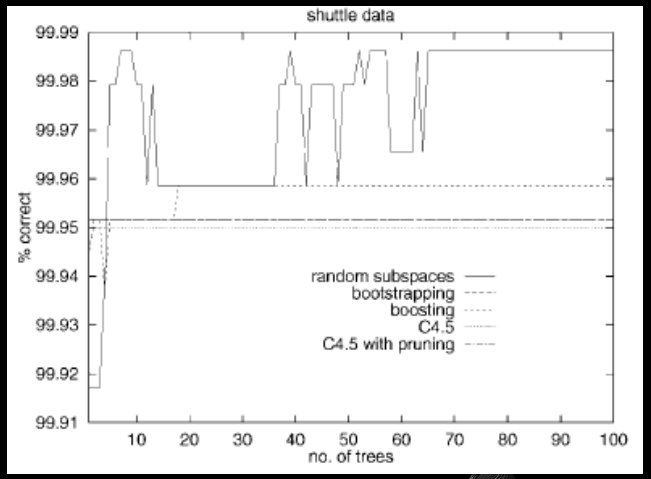
vicomtech
ik4 research alliance

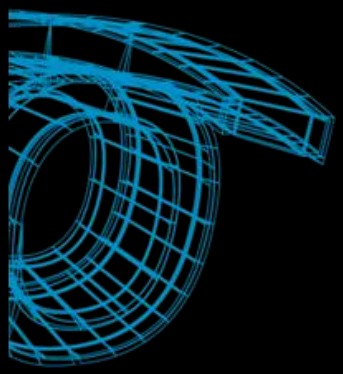
Evaluación (*single Tree, Boostrap*)









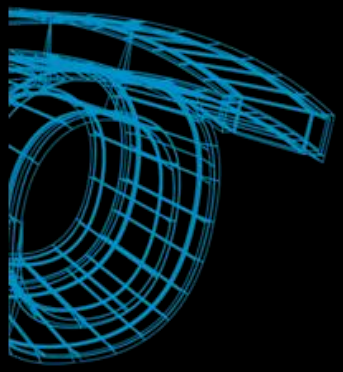


Evaluación (*single Tree, Boostrap*)

- *Random Subspaces* ofrece mejores resultados con respecto a un solo C4.5, pero cuando la dimensionalidad es reducida, la ventaja no es significativa.
- El autor sugiere expandir los vectores de características, generando características adicionales (diferencia, suma, producto entre características), en casos de baja dimensionalidad. No valida esta idea.
- Para nº bajo de árboles, C4.5 ofrece mejores resultados. A medida que incrementa el número de árboles, la aproximación de *Random Subspaces* mejora.

vicomtech
ik4 research alliance

Evaluación (*single Tree, Boostrap* – Tree Agreement)

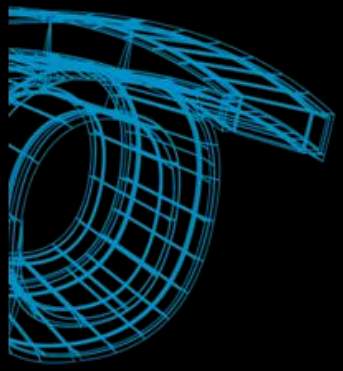


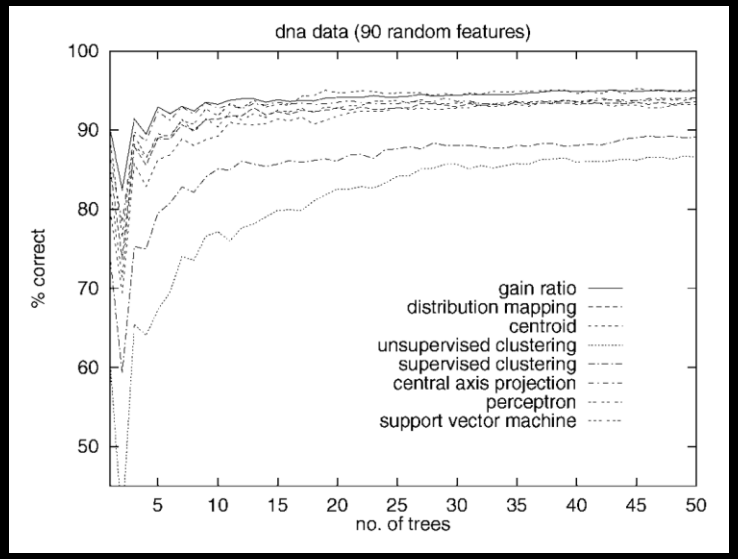
construction method	data set			
	dna	letter	satimage	shuttle
random subspaces	0.7540	0.6595	0.8228	0.9928
bootstrapping	0.9081	0.8197	0.8294	0.9998
boosting	0.8969	0.7985	0.8205	0.9996

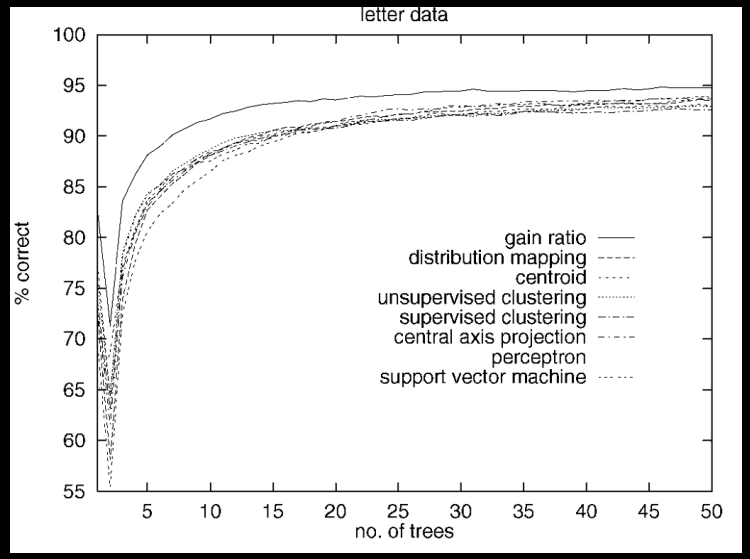
- *Random Subspaces* presenta la mayor diversidad media del conjunto, salvo en data sets de baja dimensionalidad.

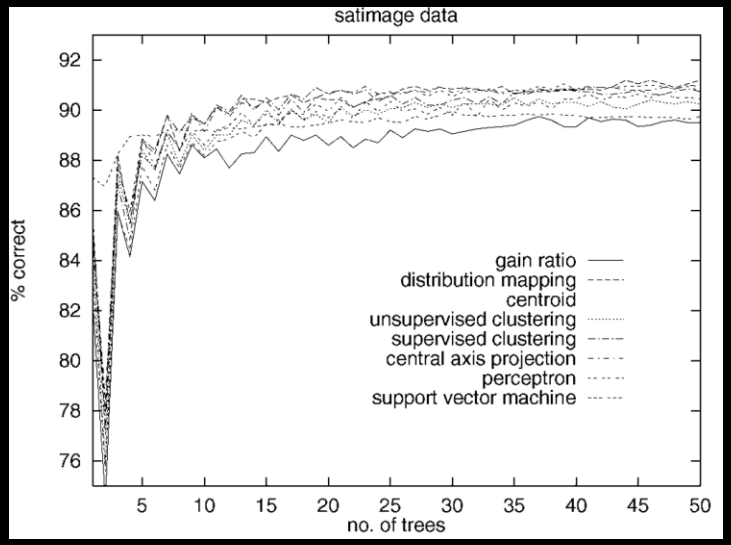
vicomtech
ik4 research alliance

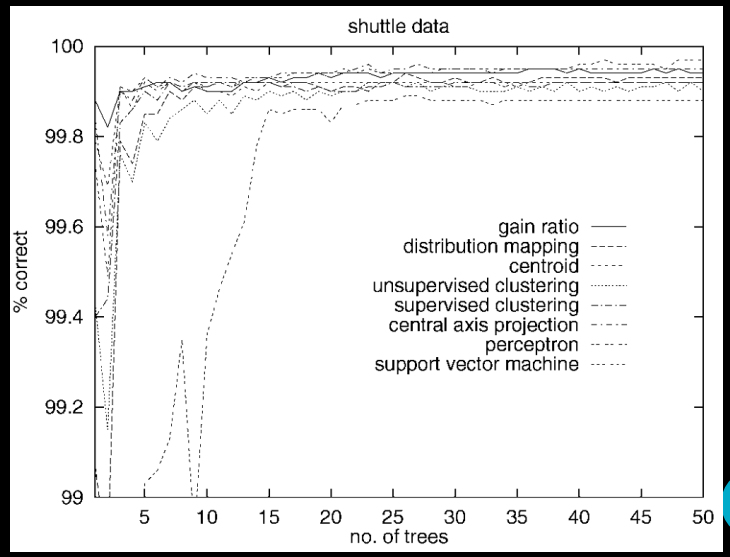
Evaluación (Split)

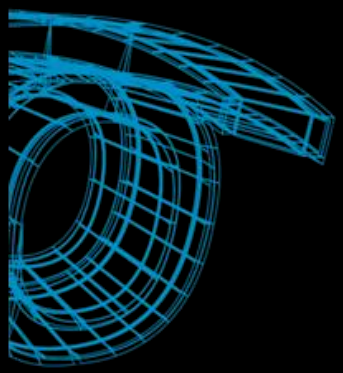








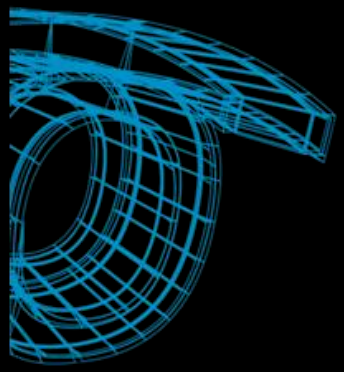




Evaluación (Split)

- Los efectos en el tipo de *split* no son significativos en ninguna de sus diferentes versiones. El autor sostiene que las diferencias vienen marcadas por los datos (no existe un mecanismo de *split* mejor que otro).

Evaluación (Split, Tree Agreement)



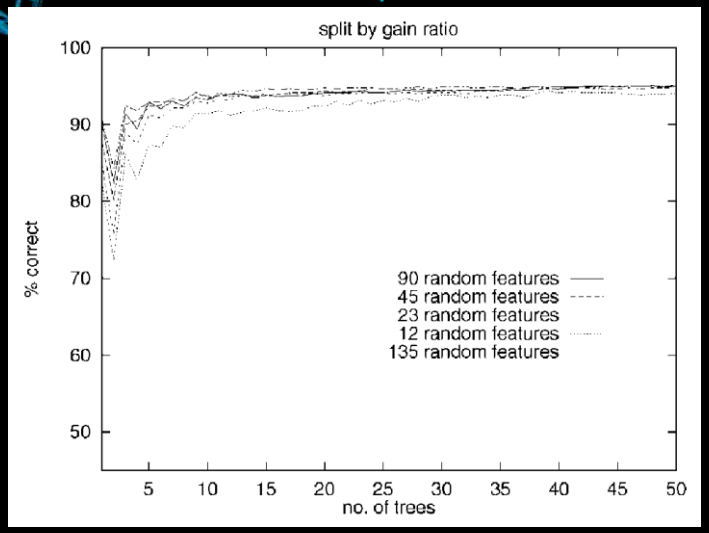
splitting function	data set			
	dna	letter	satimage	shuttle
gain ratio	0.8804	0.7378	0.7980	0.9975
dist. mapping	0.7728	0.5861	0.8188	0.9864
centroid	0.7143	0.6593	0.8310	0.9975
unsup. clustering	0.5337	0.5701	0.8110	0.9885
sup. clustering	0.6137	0.6320	0.8378	0.9812
c. axis projection	0.8368	0.6481	0.8440	0.9911
perceptron	0.7913	0.5758	0.8200	0.9950
support vectors	0.7907	0.4827	0.9205	0.8720

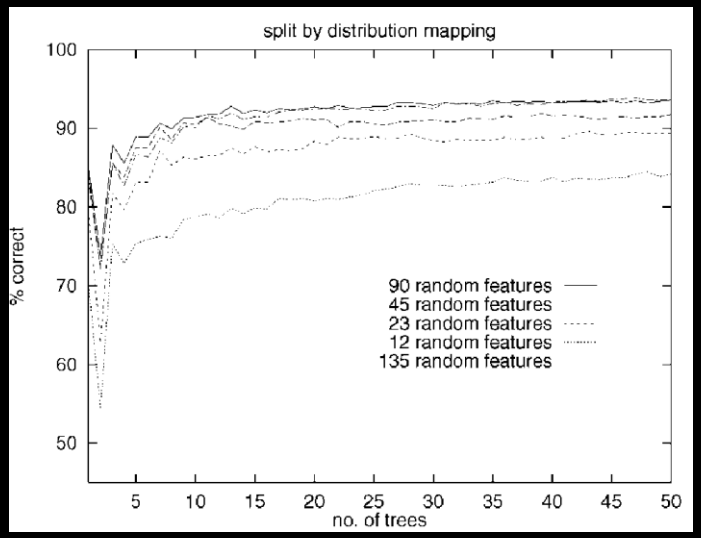
- *Gain Ratio* muestra, en general, los peores resultas de similaridad, mientras que *Unsupervised Clustering* los mejores.
- Aun así, al igual que las tablas de *accuracy*, no existe un mecanismo de *split* que proporcione los mejores resultados en todos los casos (dependencia de los datos)

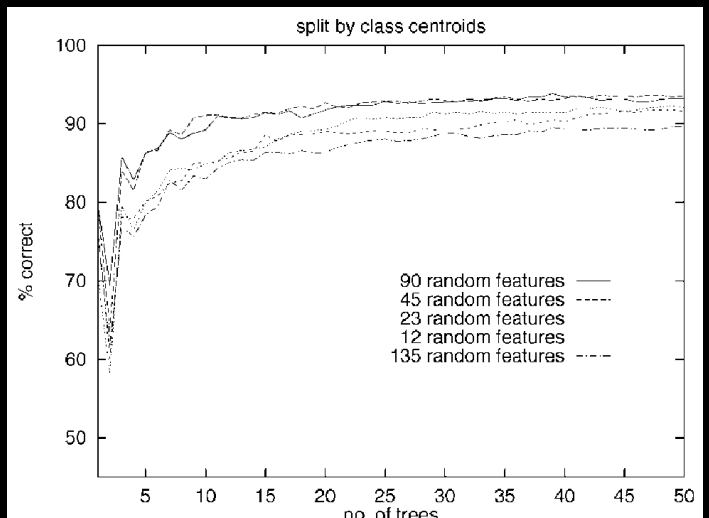
vicomtech
ik4 research alliance

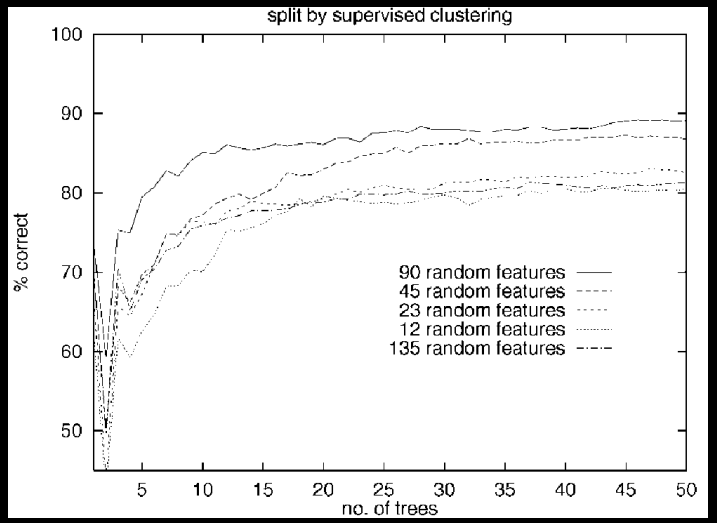
Evaluación (Nº de features)

- Data set de mayor dimensionalidad "dna"



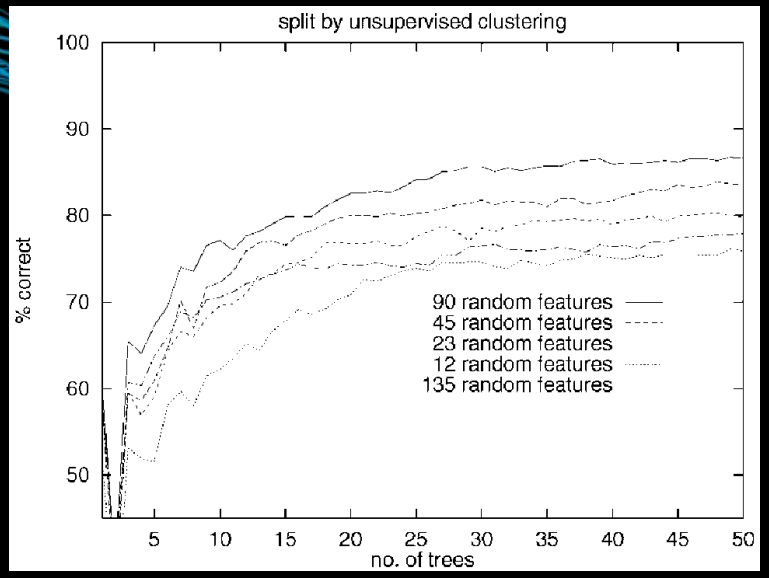


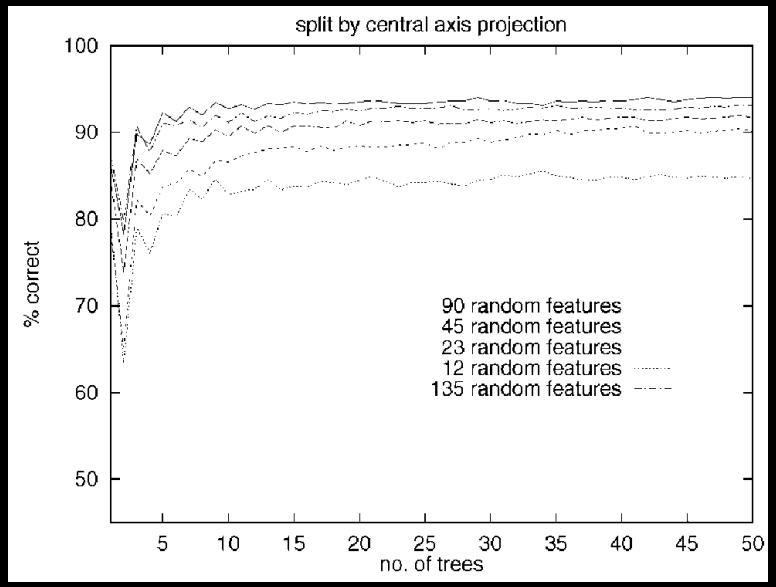


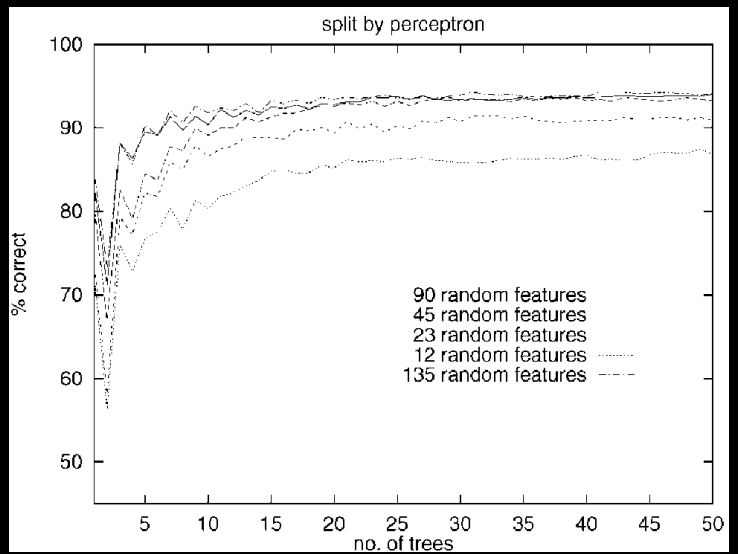


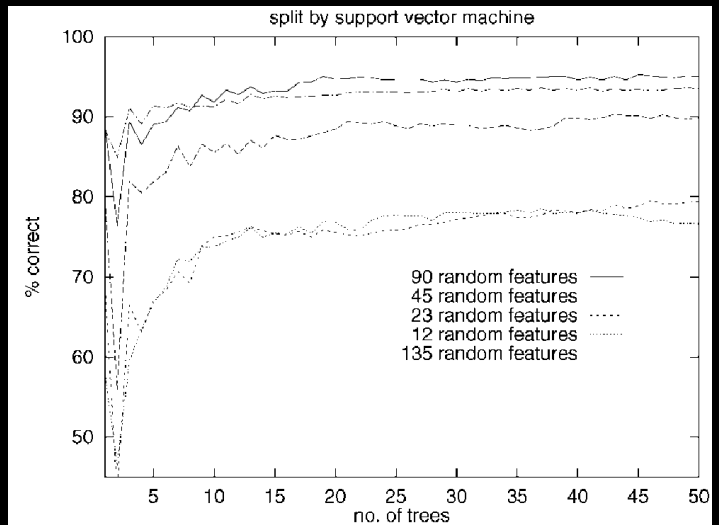
vicomtech
ik4 research alliance

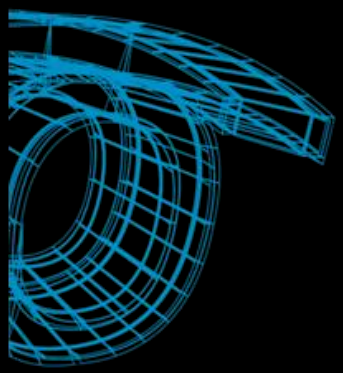
Evaluación (Nº de features)







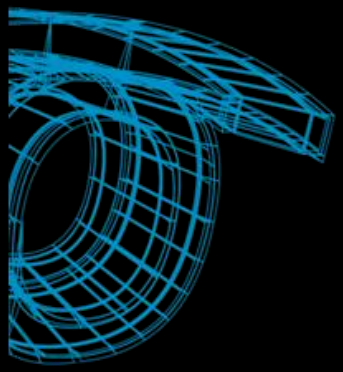




Evaluación (Nº de Features)

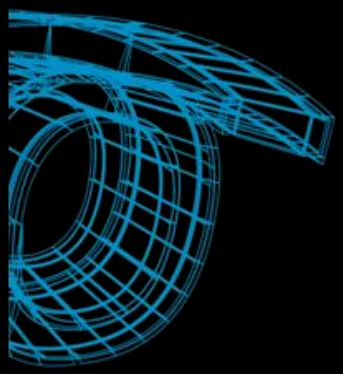
- *Gain Ratio* parece ser el mecanismos de Split más insensible al número de características.
- Un número de la mitad de la dimensionalidad (180) ofrece los mejores resultados.

Evaluación (Nº de Features, Tree Agreement)



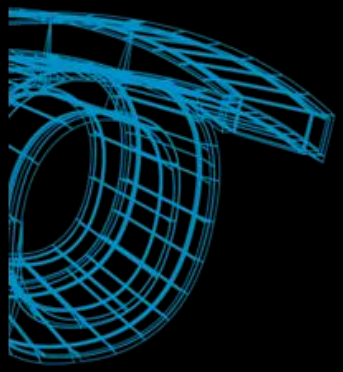
splitting function	number of random features				
	12	23	45	90	135
gain ratio	0.7070	0.7773	0.8498	0.8804	0.8849
dist. mapping	0.5855	0.6555	0.7216	0.7728	0.7530
centroid	0.6168	0.6385	0.6717	0.7143	0.6567
unsup. clustering	0.4618	0.4723	0.4862	0.5337	0.5451
sup. clustering	0.5089	0.5253	0.5507	0.6137	0.5571
c. axis projection	0.6412	0.7082	0.7714	0.8368	0.8224
Perceptron	0.6174	0.6811	0.7313	0.7913	0.7968
support vectors	0.4985	0.5198	0.6281	0.7907	0.8144

- La similaridad aumenta con el número de características.
- *Gain Ratio* genera los árboles más similares en general, mientras que *Unsupervised Clustering* los más diversos.
- La precisión del bosque viene marcada por la precisión de cada individuo y por la similaridad entre ellos.
- Optimizar uno solo de ellos no tiene por que repercutir positivamente en la precisión del bosque.



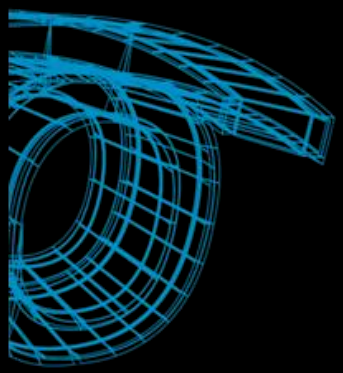
Evaluación (Nº de Features, Tree Agreement)

- Los métodos *boostraping o boosting* generan mejores individuos, pero el bosque es peor que aquellos entrenados con *Random Subspaces*.
- El autor propone usar la medida de similaridad para ordenar los árboles. Puede ser útil para aplicaciones que requieren compromiso entre precisión y tiempo.



Conclusiones

- Método de contrucción de árboles basado en generación de subespacio de características con selección aleatoria.
- Combinación de árboles mediante
- Evaluación del clasificador en varios DataSets con respecto a algoritmos como *C4.5, boostraping, Boosting (Full feature set)*
- Mejoras significativas de precisión con respecto a los algoritmos citados.
- La complejidad del bosque (nº de árboles) no disminuye la precisión de generalización (test set), mientras que en el training set se mantiene estable.
- La construcción de los árboles se puede realizar con diferentes mecanismos de split, siendo éstos más dependientes de los datos que en la construcción de los árboles (la precisión sigue la misma tendencia a medida que sube el nº de árboles en todos los mecanismos de *split*).



Conclusiones

- Los mejores resultados se obtienen empleando aproximadamente un nº de características equivalente a la mitad del total de dimensiones.
- El método propuesto parace comportarse mejor cuando el dataset contiene características de muchas dimensiones y con muchas instancias por clase.