Jesús García Herrero

CLASIFICADORES KNN –II: APRENDIZAJE DE PROTOTIPOS

En esta clase se completa el tema anterior de aprendizaje basados en instancias, enfatizando el aprendizaje de ejemplares. En particular, se destaca la equivalencia de las distancias a vecinos más próximos con la construcción de fronteras de decisión basadas en ejemplares, y se profundiza en la búsqueda de los prototipos útiles para determinar estas fronteras de decisión a través del algoritmo IBK, eliminando prototipos ruidosos y utilizando para clasificar o predecir los que tienen más calidad.

Aprendizaje basado en ejemplares

Jesús García Herrero
Universidad Carlos III de Madrid

Universidad
Carlos III de Madrid

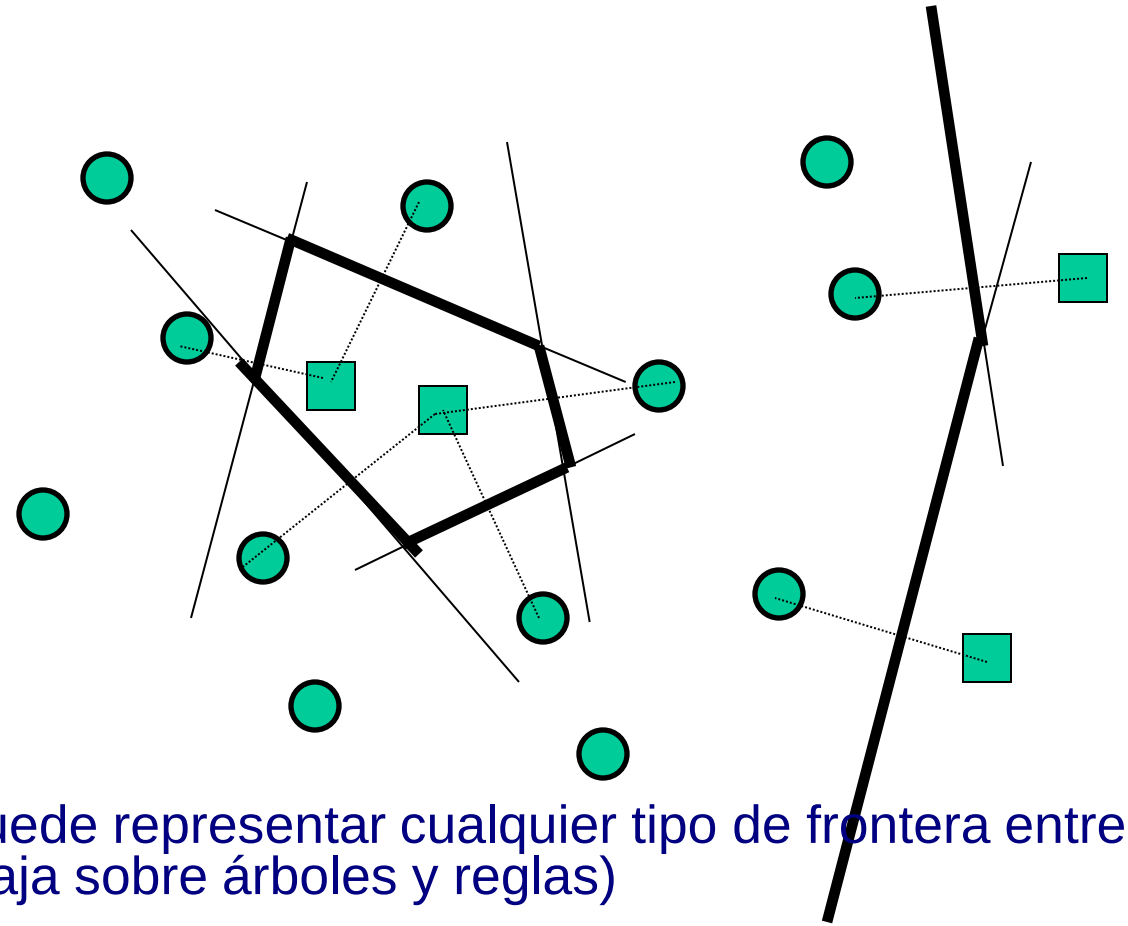
Aprendizaje Basado en Ejemplares

- Modelización de algunos procesos de aprendizaje humano (memorización y similitud)
- Espacio de hipótesis: conjunto de ejemplares
- Características de los ejemplares:
  - Los conceptos se representan por conjuntos de ejemplares sin información sobre las condiciones necesarias y/o suficientes
  - La representación es explícitamente disyuntiva
  - Las propiedades de un concepto son función de las propiedades de los ejemplares
- Similitud con el Razonamiento Basado en Casos (CBR)

Aprendizaje Vago

- Se almacena todo el conjunto de ejemplos de entrenamiento. Clasificar es buscar el ejemplo más “parecido” al que estamos analizando.
  - Coste de clasificación alto (análisis diferido)
  - Utilizan todos los atributos y ejemplos
  - No generalizan. No se describen explícitamente patrones
  - El mayor problema estriba en como analizar el “parecido”
    - Mediante función distancia: vecino más próximo (nearest neighbour)
    - Cómo considerar distancias con varios atributos (pesos)
    - Cómo considerar atributos nominales

Regiones de clasificación



- Se puede representar cualquier tipo de frontera entre clases (ventaja sobre árboles y reglas)

Método kNN

- Aprendizaje: guarda todas las instancias
- Clasificación: elegir aquella clase más común entre los k vecinos más cercanos, utilizando una distancia. Suaviza el ruido (sobreadecuamiento)
- Distancia Euclidea:

$$d(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_{m=1}^{F} (x_{im} - x_{jm})^2}$$

  - Si la clase es continua: elige la media de las clases de los k vecinos más cercanos
  - Si los atributos tienen diferentes rangos:
    - normalizar:

$$a_i = \frac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}$$

    - distancia estadística (Mahalanobis)

$$d(\vec{X}_i, \vec{X}_j) = \sqrt{\sum_{l=1}^{F} \frac{(x_{il} - x_{jl})^2}{\sigma_l^2}}, \quad d(X_i, X_j) = \sqrt{(\vec{X}_i - \vec{X}_j)^t S^{-1} (\vec{X}_i - \vec{X}_j)}$$

Aprendizaje IBL

Método kNN

- Atributos nominales

$$d(x_{il}, x_{jl}) = \begin{cases} 1, & \text{si } x_{il} \neq x_{jl} \\ 0, & \text{si } x_{il} = x_{jl} \end{cases}$$

- Se puede dar mayor preferencia a los vecinos más cercanos dentro de los k : se multiplica el voto de cada vecino por:

$$1/d^2(\vec{x}_i, \vec{x}_j)$$

- Para las clases continuas, se multiplica cada voto por esa cantidad y se divide por la suma de esas distancias a los k vecinos más cercanos

Ponderación de atributos

- La normalización de atributos no es adecuada cuando varía la importancia de los atributos
  - Se incorporan pesos con la importancia en cada dimensión

$$d(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_{m=1}^{F} w_m^2 (x_{im} - x_{jm})^2}$$

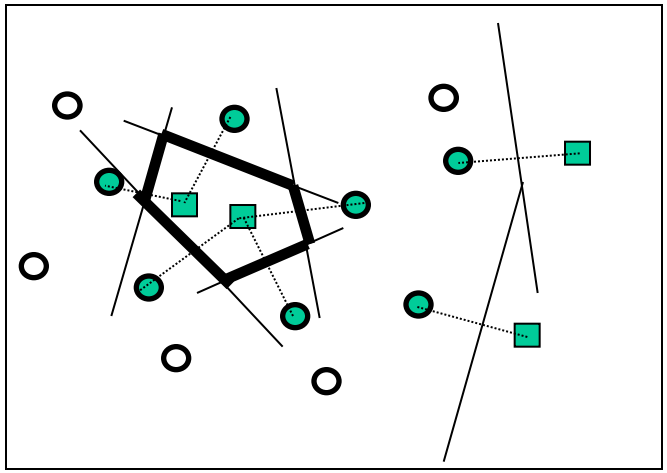
  - Actualización de pesos con ejemplo x: según el comportamiento de la clasificación (con el ejemplar más próximo, y):
    - Para cada atributo i, la diferencia $|x_i - y_i|$ mide el peso del atributo:
    - Si es pequeño, el atributo contribuye mucho
      - Se incrementa el peso si se acierta y decrementa si falla
      - El incremento es inversamente proporcional a $|x_i - y_i|$
- La ponderación proporciona protección contra atributos ruidosos/irrelevantes

Método kNN

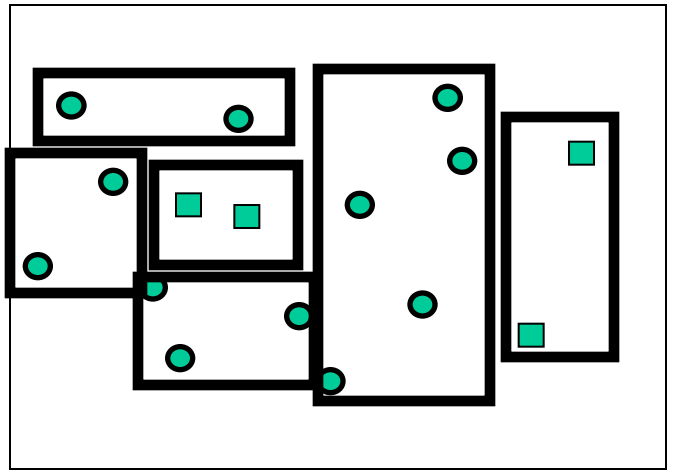
- Inconvenientes:
  - Se ralentiza al aumentar el número de ejemplos de entrenamiento
  - k?? Si k es bajo, es sensible al ruido, si k aumenta, pierde detalles
  - Mal comportamiento si los atributos tienen importancia distinta
    - Muy sensible a atributos irrelevantes ruidosos
  - No se realiza generalización implícita

- Mejoras posibles (Aha 92)
  - Reducción del número de ejemplos. Selección de prototipos/ejemplares
  - Ponderación de atributos en el cálculo de la distancia
  - Generalización de los ejemplares en reglas

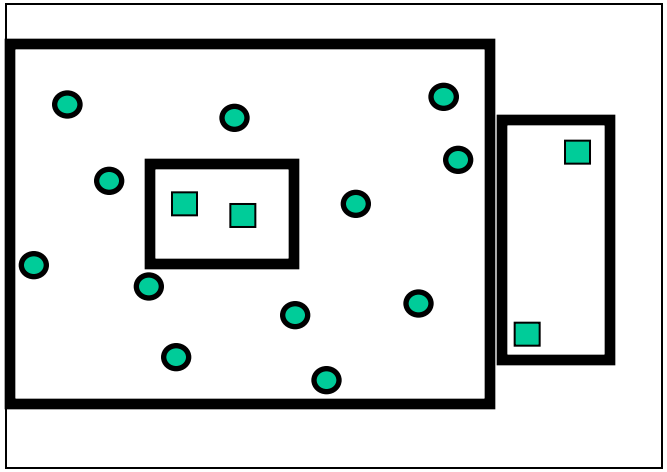
Posibilidades de generalizar



Selección de prototipos

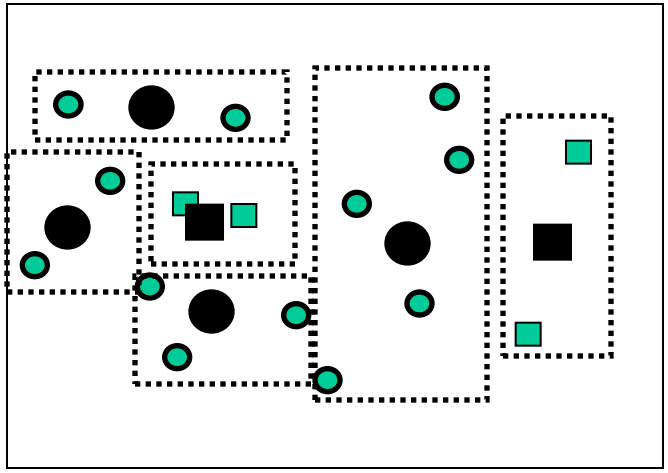


Ajuste de regiones: reglas



Regiones anidadas (excepciones)

Aprendizaje IBL



Centroides

Otros Modelos

- Modelo de Mejores-Ejemplos (Smith y Medin, 81)
  - Asume que existe un prototipo para cada clase
  - Un prototipo es un conjunto de ejemplares de la clase
  - Los prototipos son los ejemplares que contienen más valores de los atributos en común que el resto
  - Clasifica una instancia en la clase con el prototipo más similar

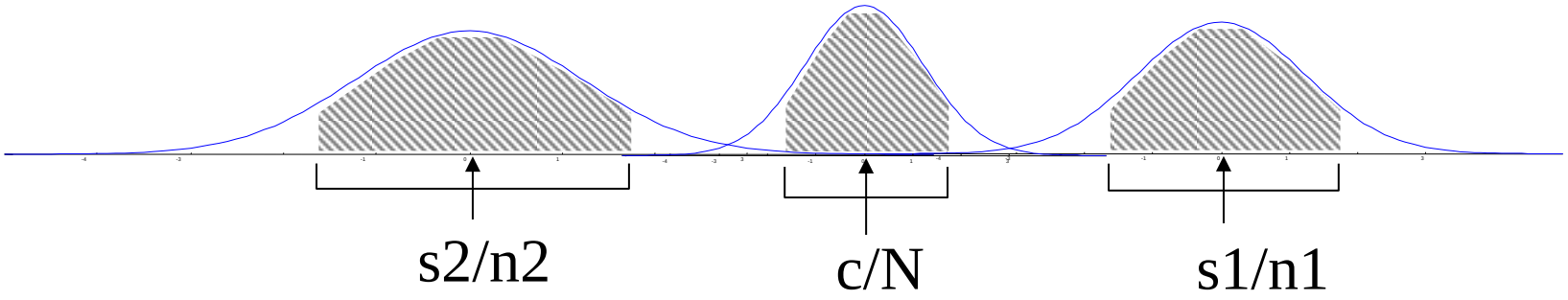
  - Reducción del número de ejemplares
    - Puede clasificarse cada ejemplo de entrenamiento con respecto a los vistos, y descartar los clasificados correctamente
      - Idealmente sólo habría un ejemplar por cada clase
      - Problema del ruido: los ejemplos ruidosos tienden a acumularse y degradar las prestaciones
      - Problema de eliminación de ejemplares importantes después

Tipos de Modelos

- Modelo de Selección-Ejemplos (Kibler y Aha, 87)
  - Salvan sólo una parte de las instancias
  - No asumen que existen los prototipos de cada clase
    - Método de crecimiento: almacena sólo las instancias que no se clasifican correctamente
    - Método de decrecimiento: almacena todas las instancias inicialmente y borra, por turno, aquellas que se clasifiquen correctamente
  - El orden de presentación de las instancias es importante
  - Ponderación de atributos
  - Ventajas:
    - Menor espacio
    - Mejores clasificadores, al rechazar las instancias atípicas (posible ruido)
  - Inconveniente: Mayor complejidad computacional

Otros Modelos

- Poda de ejemplares ruidosos
  - Se analiza el comportamiento de los ejemplares de cada clase para eliminar los que no clasifican correctamente
  - Método IB3:
    - Se mide la tasa de acierto de un ejemplar (s/n), y la probabilidad "por defecto" de la clase (c/N), definiéndose umbrales mediante intervalos de confianza



    - Se mantienen dos umbrales: selección para predicción (5%) y eliminación (12%)
    - Los ejemplos entre ambos umbrales se actualizan cada vez que son los más próximos al ejemplo a clasificar