LAZY LEARNING

Si cerca di salvare i dati di training così come sono e usarli direttamente in fase di classificazione.
Questi metodi spostano la complessità in fase di predizione.

K-NEAREST NEIGHBOR

Questo metodo corrisponde nel considerare ogni tupla del dataset come un punto nello spazio. Utilizza poi questi punti per cercare i K vicini più vicini alla tupla da classificare.
Per poter definire un vicino usa una metrica di distanza (come la distanza euclidea).
Una volta definiti i K vicini sceglie la classe di maggioranza fra quei vicini.

Come calcolare la distanza fra valori non numerici?

Semplicemente ponendo a zero la distanza fra valori uguali e 1 fra valori diversi.

Se qualche attributo è vuoto o nullo!

Quando il valore manca a una delle due tupla viene considerata la differenza massima. Che è 1 per i valori categorici ma anche nei valori numerici se mancano entrambi i valori.

Determinare K

A tentativi, e scegliere l'accuratezza migliore

Complessità:

Se nel nostro DB abbiamo n tupla e f attributi la complessità va calcolata in due fasi:

1) Complessità della ricerca di tutte le distanze: $n \cdot f$
2) Complessità dell'ordinamento della distanze: $n \cdot \log n$

TOTALE: $n \log n$.

La complessità dell'ordinamento può essere migliorato utilizzando strutture dati come un albero di ricerca. Oppure utilizzando algoritmi in parallelo.
Altri modi per migliorare la complessità sono i metodi di Editing

EDITING METHODS

L'obiettivo è ridurre il rumore fra classi.

WILSON EDITING

Usa K-NN per pulire la sovrapposizione fra classi.

1) Trova i K-NN di x_i in D (escluso x_i)
2) Etichettare x_i con la classe di maggioranza in K-NN
3) Eliminare tutti i punti che al punto 2 sono stati etichettati con una classe ma in realtà nel dataset ne hanno un'altra.

MULTI-EDIT

Viene applicato Wilson più volte in vari sotto-insiemi di D (n insiemi con $n \geq 3$).
Poi ricomporre D' con l'unione degli n sottoinsiemi.

CITATION EDITING

Per ogni elemento x_i del dataset D:

1) Trovare i K-NN di x_i
2) Trovare i c CITERS di x_i, ovvero i c oggetti che hanno nel proprio K-NN x_i.
3) Etichettare x_i con la classe di maggioranza degli oggetti presenti nel K-NN e C-NN.
4) Non considerare gli oggetti misclassificati

SUPERVISED CLUSTERING

Rimpiazzare il dataset D con il risultato di un algoritmo di clustering supervisionato.
Nei cluster trovati eliminare i mismatch.

VALUTAZIONE DI UN METODO DI EDITING

Si usa una metrica detta TRAINING SET COMPRESSION RATE (TSCR)

$$TSCR = \left(1 - \frac{\varepsilon}{n}\right) \cdot 100$$

ε: numero di esempi nel set editato
n: numero di esempi nel set originale

Indica di quanto D è stato ridotto.