Data Mining e Scoperta di Conoscenza

Progetto 13

Utilizzo della tecnica K-Nearest Neighbor per la stima di valori mancanti

1. Si estenda Weka con la nuova classe **`weka.filters.unsupervised.IBkFilter.java`** che utilizzi l'algoritmo **k-Nearest Neaighbor** per la stima dei valori mancanti. Nel dettaglio, l'algoritmo deve implementare:
    i. Diverse funzioni di distanza da affiancare alla distanza Euclidea per la determinazione dei neighbor. Le funzioni di distanza richieste sono:
        1. Coseno
        2. Jaccard
        3. Dice
        4. Mahalanovis
        5. Minkowski con parametro p variabile.
2. Si usi l'algoritmo implementato sulla repository di dati **UCI** fornita con il progetto e si analizzi come varia l'accuratezza predittiva se i dati sono processati con il filtro descritto. In particolare:
    a. Valutare come varia l'accuratezza di diversi tipi di classificatori (alberi di decisione, classificatori bayesiani, reti neurali) qualora:
        i. Non trattassimo il problema dei valori mancanti;
        ii. Sostituissimo i valori mancanti di un attributo con il valore più frequente calcolato sull'intero data set di training;
        iii. Utilizzassimo i dati con il filtro **IBk**.
3. Ripetere gli esperimenti del punto 2 utilizzando i dati genetici contenuti nella repository **microarray_project_data.zip** forniti con il progetto. L'accuratezza va valutata confrontando il dataset con una sua versione perturbata (ottenuta cioè eliminando casualmente valori). Come variano le statistiche all'interno del dataset?

NOTE PER L'ESECUZIONE DEL PROGETTO

1. Scrivi un rapporto di circa 10 pagine in cui
    a. Descrivi analiticamente l'algoritmo che hai implementato.
    b. Commenti le parti essenziali del codice Java che hai scritto, e metti in un'appendice l'intero codice
    c. commenti e illustri graficamente e quantitativamente gli esperimenti effettuati.
2. Prepara delle slides Powerpoint (non più di 10 slides) in cui riassumi gli esiti del progetto