Data Mining e Scoperta di Conoscenza

Progetto 11

Miglioramento della tecniche Nearest Neighbor utilizzando l'indicizzatore MTree

1. Si scriva un programma Java che implementi l'indice **MTree** i cui dettagli sono riportati nell'articolo "*M-tree: An Efficient Access Method for Similarity Search in Metric Spaces*" fornito con il progetto. L'indice, che permette la ricerca efficiente di tuple all'interno di uno spazio metrico, deve essere utilizzato per velocizzare la ricerca dei neighbors di un'istanza in un approccio di classificazione di tipo Instance-Based. Nel dettaglio, l'implementazione dell'indice MTree deve essere resa indipendente dalla metrica scelta per valutare la similarità fra istanze. In particolare, viene richiesto di implementare le seguenti definizioni di distanza :
    a. Jaccard
    b. Mahalanobis
    c. Minkowski
2. Usando l'algoritmo implementato al punto 1, si costruisca un **classificatore IBk**, analizzandone le performance sui dati contenuti nella repository UCI fornita con il progetto. Nel dettaglio si valutino:
    a. le prestazioni (in termini di velocità) del classificatore su datasets di dimensioni diverse;
    b. le performance (in termini di accuratezza) utilizzando le diverse nozioni di metriche implementate al punto 1;
    c. le performance (in termini di accuratezza) rispetto ad altri classificatori forniti in Weka (Naive Bayes, IBk con distanza Euclidea,....)

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