Data Mining e Scoperta di Conoscenza

Progetto 1

Confronto di Algoritmi di Decision-Tree Learning

1. Si estenda Weka con due nuove classi
    a. **`weka.classifiers.trees.CART.java`**, che implementa l'algoritmo CART. Nel dettaglio, l'algoritmo deve implementare:
        i. L'indice di Gini come criterio di split per lo sviluppo dell'albero
        ii. Il pruning con la tecnica dell'albero pruned minimale che si discosta di al più ε dall'errore minimo per la prevenzione dell'overfitting
        iii. L'imputazione con surrogati per la sostituzione dei valori mancanti
        iv. (OPZIONALE) il trattamento dei valori numerici
    b. **`weka.classifiers.trees.CHAID.java`** che implementa l'algoritmo CHAID. Nel dettaglio, l'algoritmo deve implementare:
        i. Il test del Chi quadro come criterio di split per lo sviluppo dell'albero
        ii. L'early stopping per la prevenzione dell'overfitting
        iii. (OPZIONALE) il trattamento dei valori numerici tramite l'F-test.
2. Si usino gli algoritmi implementati sulla repository di dati UCI fornita con il progetto e si analizzi l'accuratezza predittiva dei due metodi.
    a. Come influenza il preprocessing l'errore medio di classificazione?
    b. Come variano gli algoritmi al variare dei parametri?
    c. Come varia la velocità di apprendimento in funzione del numero di tuple usate come training?
3. Si confrontino le performances degli algoritmi implementati con una tecnica tradizionale in Weka (ad esempio l'algoritmo J48 per il decision-tree learning).

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