Data Mining e Scoperta di Conoscenza

Progetto 7

Realizzazione di uno Spam Filter

Si realizzi un classificatore che sia capace di separare spam da non-spam emails. Il classificatore analizzerà ogni email in arrivo classificandola adeguatamente e facendo in modo di fornire solo le emails ritenute non-spam.
In base allo scenario descritto, dopo aver costruito il classificatore sulla base delle 8.000 emails fornite nel dataset *data_dmc2003_train.txt*, si chiede di:

1. minimizzare il numero di spam emails (fra le 11.177 emails fornite per la classificazione nel dataset *data_dmc2003_class.txt*) che superano il filtro. Il vincolo posto è il seguente: tra le emails filtrate, è consentito l' 1% massimo di non-spam emails.
2. studiare la performance della classificazione con particolare riferimento alle curve di ROC:
   - Come variano le performance del classificatore al variare dell'insieme di addestramento?
   - Come varia l'accuratezza.
   - Si studi l'effetto dell'analisi delle componenti principali sull'errore
   - Si determinino delle trasformazioni dei dati che permettano un miglioramento generale delle performances del classificatore.

I datasets completi, oltre ad un file contenente una spiegazione dettagliata degli attributi ivi contenuti, vengono forniti contestualmente al progetto.

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. Prepare delle slides Powerpoint (non più di 10 slides) in cui riassumi gli esiti del progetto