Data Mining e Scoperta di Conoscenza

Progetto 6

Realizzazione di un task per l'analisi del "Comportamento di clienti in acquisti effettuati via mail"

Si realizzi un modello di classificazione il cui scopo è, a partire da un set di 20.146 clienti che acquistano a distanza usando la posta, predire la frequenza con la quale i clienti rimandano indietro gli oggetti acquisiti, classificando i clienti stessi come "high rate customer" oppure "low rate customer". Più in dettaglio, sono definiti "high rate customer" clienti in cui la percentuale di oggetti restiutiti sugli oggetti consegnati è >=40%. Sono definiti, invece "low rate customer" clienti in cui la percentuale di oggetti restituiti sugli oggetti consegnati è <=18%. Gli altri consumatori non appartengono a nessuna di queste due categorie.
In osservanza allo scenario descritto, si chiede di:

1. generare una etichetta di membership per codificare l'appertenenza delle istanze contenute nell'insieme di training *dmc2004_train.txt* ad una delle due classi specificate.
2. generare un modello di mining da applicare ai 20.146 clienti da classificare (contenuti in *dmc2004_class.txt*) in modo di assegnarli ad una delle due classi specificate.
3. Analizzare il modello e confrontarlo con altri modelli utilizzando un'analisi ROC e tramite Lift-charts.

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