Database e Data Mining

Esercitazione N. 3

L'esercitazione ha come obiettivo l'analisi di tre dataset mediante le tre tecniche classiche di data mining:

1. regole di associazione
2. clustering / text mining
3. classificazione

1. Estrazione delle regole di associazione

Dataset: *Market Basket Data*

Attività da svolgere

1. Analisi preliminare del dataset:
    a. Importare il dataset
    b. Valutare eventuale pre-processing dei dati
    c. Analizzare la distribuzione dei dati. Eventualmente plottare un grafico sugli item singoli piu' frequenti
2. Estrazione itemset:
    a. Estrarre gli itemset con diverse soglie di supporto
    b. Individuare una soglia ottimale, valutando il anche il numero di itemset estratti
3. Estrazione delle regole di associazione:
    a. Estrarre le regole di associazione a partire dagli itemset
    b. Individuare una soglia di confidenza ottimale
4. Analisi:
    a. Analizzare le regole estratte sia mediante metriche quantitative che valutazioni di dominio
    b. Ipotizzare un possibile utilizzo della conoscenza estratta

2. Clustering / Text mining

Dataset

- *feed rss tecnologia: http://www.repubblica.it/rss/tecnologia/rss2.0.xml*
- *feed rss sport: http://www.repubblica.it/rss/sport/rss2.0.xml*
- *fedd rss politica: http://www.repubblica.it/rss/politica/rss2.0.xml*

Attività da svolgere

1. Generazione del dataset:
    a. Utilizzare l'appropriato operatore per scaricare i 3 feed rss
    b. Selezionare l'attributo titolo da ciascun feed
    c. Aggiungere l'attributo categoria e assegnare come valore la categoria del relativo feed
    d. Valutare eventuale pre-processing dei dati per la manipolazione dei documenti (titoli dei feed)
2. Clustering
    a. Applicare uno o più algoritmi di clustering in modo da clusterizzare i titoli dei feed nelle 3 categorie
3. Analisi
    a. Confrontare la categoria originale con il risultato del clustering
    b. Valutare le performance ottenute

3. Classificazione

Dataset: *Wine Data (traning e test)*

Attività da svolgere

1. Analisi preliminare del dataset:
    a. Importare il dataset
    b. Valutare eventuale pre-processing dei dati in funzione dell'algoritmo di classificazione utilizzato
2. Costruzione del modello:
    a. Costruire un modello di classificazione, con diversi algoritmi, al variare dei parametri disponibili
    b. Validare i modelli creati
    c. Individuare il modello migliore: algoritmo e configurazione dei parametri
3. Classificazione
    a. Applicare il modello al dataset di test
    b. Valutare le prestazioni ottenute