DATA MINING

PERCHÉ IL DATA MINING?

Nella nostra epoca dobbiamo interfacciarci con un'enorme quantità di informazioni. Se prima il problema principale dell'informazione era l'acquisizione di dati, ora è la comprensione di quali di questi siano importanti per l'ottenimento dell'informazione
Il data mining è un processo tramite il quale, partendo da dati già acquisiti e manipolandoli, si estrae la conoscenza richiesta
Il data mining si divide nelle seguenti fasi:

- Data source, dai quali otteniamo i dati;
- Data preprocessing, dove si ha una prima manipolazione dei dati, per integrarli, ridurli, ad esempio;
- Data exploration, nel quale si eseguono le vere analisi sui dati;
- Data mining, nella quale si risale all'informazione;
- Data presentation, nella quale tali informazioni vengono visualizzate e presentate agli utenti;
- Decision making, nel quale si prendono decisioni sulla base dei dati ottenuti

Nel data mining, dopo un processo di classificazione o generazione di classi (clustering), si eseguiranno analisi di correlazione ed associazione.
Si presta anche particolare attenzione agli outliers, istanze molto discordanti dal generico andamento.

DATA SETS, DATA OBSECTS E ATTRIBUTI

I data sets o insieni di dati sono raggruppabili in quattro categorie:

- numerici, come matrici di dati, matrici numeriche, vettori di frequenza, dati di transizione;
- grafici e siti internet;
- ordinati, in cui è importante l'ordine conferito ai dati, come video, sequenze di dati, come temporali e genetici.

master
COPY
COPISTERIA
050/8312126 388/9837745

- spaziali, multimediali ed immagini

Caratteristiche generiche di questi data set sono:
- la dimensione: più dati ci sono nel set, più difficile è il confronto tra due (curse of dimensionality);
- sparsità: anche solo la presenza di data set è importante nell'analisi;
- risoluzione, dipende dalla scala di visualizzazione;
- distribuzione, come la loro centralità o dispersione

Un data object costituisce un'istanza di data set. Esso è caratterizzato da attributi, ossia campi di dati contenenti caratteristiche del data object

Essi sono di tre tipi:
- nominali, caratterizzati da nomi di cose; assumono un numero limitato di valori, i quali non hanno ordine;
- binari, che assumono valori 0 e 1; tali valori possono essere simmetrici, ossia ~~or~~ avere la stessa importanza (genere), o asimmetrici (risultato dell'analisi medica di un paziente), in tal caso 1 è più importante di 0;
- ordinali, con un preciso ordine di significato, tra i quali non risulta tuttavia nota l'intensità della differenza tra due valori
- numerici, valori interi o reali, suddivisi tra intervalli e rapporti; entrambi sono misurabili con unità di misura, tuttavia nei primi non esiste lo zero assoluto della misura, mentre nei secondi sì; in questi ultimi è quindi possibile dividere per lo 0 della misura e quantificare con precisione la distanza tra due misure

Gli attributi si definiscono discreti o continui, se il loro dominio ha la cardinalità dei numeri interi o reali. Gli attributi binari sono un particolare caso di variabile discreta.

DESCRIZIONE STATISTICA DEI DATI: ANDAMENTO MEDIO

Una descrizione statistica dei dati ottenuti si

legati alla moda dalla relazione

media − moda = 3(media − mediana)

Se i tre valori sono uguali, la distribuzione di dati è detta simmetrica. Se moda < mediana < media, è positivamente obliqua, mentre se media < mediana < moda, allora è negativamente obliqua

MISURA DELLA DISPERSIONE DEI DATI: IL GRAFICO A SCATOLA

Per descrizione statistica della dispersione di dati, si usano i K-percentuali. Essi sono i valori per cui il K% dei dati è ad essi inferiore. Rilevanti sono il 25-esimo e 75-esimo percentuale, Q_1 e Q_3, detti quartili. IQR è l'intervallo interquartile $Q_3 - Q_1$.

Tramite essi si costruisce il grafico a scatola. I dati sono rappresentati da una scatola, i cui estremi sono Q_1 e Q_3, ed ~~il valore medio~~ la mediana è posto marcato con una linea all'interno della scatola. Vi sono poi due baffi (whiskers) associati al massimo e minimo dei valori. Gli outliers sono posti oltre una specifica soglia, ad esempio $1,5 \times IQR$ dalla mediana.

Con questo grafico è immediato effettuare confronti tra le dispersioni di diversi set di dati.

MISURA DELLA DISPERSIONE DEI DATI

Altri indici statistici della dispersione dei dati sono:

- varianza, definita come

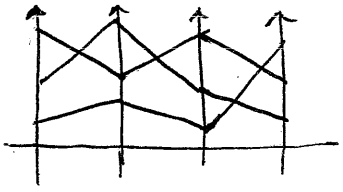
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

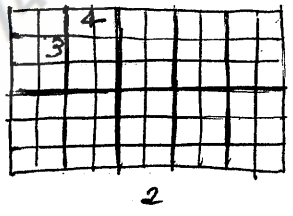
- deviazione standard, radice quadrata di s^2 o μ^2

Tra $\mu \pm \sigma$ è contenuto il 68% dei valori assunti, tra $\mu \pm 2\sigma$ il 95% e tra $\mu \pm 3\sigma$ il 99,7%

Un altro tipo di tecnica di visualizzazione è a proiezioni geometriche
Una di queste sono gli scatter plot, visualizzati a matrici per mappare tutte le possibili relazioni
Un altro tipo sono le coordinate parallele. Se più assi paralleli rappresentano le variabili vengono collegati i valori della stessa istanza



Per set di dati a grandi dimensioni viene usata una tecnica basata sull'inseguimento di proiezioni.
Tecniche basate su icone visualizzano i dati tramite forme di vario aspetto, colore e grandezza. Alcune di queste sono le facce di Chernoff e le Stick Figures.
Le tecniche di visualizzazione gerarchiche usano una partizione gerarchica di sottospazi
Nell'impilamento dimensionale, uno spazio bidimensionale è suddiviso in partizioni e ciascuna riga o colonna, anche inglobante, corrisponde ad un attributo.



Questa rappresentazione è buona per set di dati a ordinali dimensione minore di 9
e pone grandi problemi di dimensionamento della tabella
Nel grafo ad albero si realizza partizionando lo schermo in regioni cui sono associati i valori degli attributi; ogni regione ha diverse dimensioni a seconda dei valori.
Gli infocubi sono cubi partizionati in altri cubi che occupano più spazio, all'aumentare del valore assunto dall'istanza
Un'ultima tecnica per visualizzare dati non numerici è quella dei tag cloud, insiemi di varie etichette, in cui è rilevante il font e la dimensione delle parole usate.