Data minig

Limiti di OLAP

Data l'elevata quantità e la complesstà delle relazioni, le informazioni non sono completamente identificabili.
Si predispone una situazione in cui i dati rimangono inutilizzati o sotto-utilizzati (molti dati ma poche informazioni).
Gli strumenti OLAP non sono più sufficienti perchè operano per supportare processi decisionali, quindi sviluppano percorsi di analisi da ipotesi che è l'utente a formulare.

Fasi del processo di mining

Il data mining è spesso definito KDD (Knowledge Discovery in Databases) e viene diviso nelle seguenti fasi elementari:

- **Pulizia dei dati:** vengono eliminte le incorretteze.
- **Integrazione dei dati:** uniformare i dati.
- **Selezione dei dati**.
- **Trasformazione dei dati:** riorganizzazioen dei dati.
- **Data mining:** il vero e proprio processo di analisi.
- **Valutazione dei pattern:** l'insieme delle condizioni viene ridotto a quelle interessanti.
- **Presentazione delle conoscenze**.

Le prime fasi coincidono con il popolamento del DWH, può essere considerato un'evoluzione delle indagini OLAP.

Da OLAP a OLAM

OLAM (On Line Anaklytical Mining)
Partendo dai DWH abbiamo dati ben strutturati, puliti e completi. Ciononostante il processo di mining non può essere interamente automatico, infatti i pattern rilevati potrebbero essere troppi e non interessanti, il processo di datat mining deve quindi essere interattivo con gli utenti che specifico la direzione in cui indagare.
Un processo interattivo permette di affinare le ricerche.

Architettura dei sistemi di Data Mining

L'architettura dei sistemi di data mining si appoggia ai seguenti componenti:

- **Data warehouse**
- **Base di conoscenza (knowledge base):** l'insieme di regole e conoscenze note, verranno utilizzate per guidare le ricerche.
- **Motore di data mining (data mining engine):** l'insieme delle funzioni di analisi dei dati.

- **Valutazione delle condizioni (pattern evaluation):** i moduli che fanno focalizzare la ricerca sulle condizioni interessanti.
- **Sistema di presentazione:** l'interfaccia con la quale l'uetnte fa le ricerche.

I 4 principi di analisi

Staticstiche elementari e analisi relative

Generalizzazione

Deve fornire una visione ad lato livello tramite l'accorpazione di concetti e riassumendo caratteristiche di base.
Il principio di base è che gli elementi che un utente può analizzare devono essere un numero limitato.
Un diffuso tipo di generalizzazione è l'aggregazione dei sistemi OLAP, i sistemi di data mining apmplificano il potenziale mettendo a disposizione anche delle metodologie di induzione.

1. Caratterizzazione

   Serve a comprendere le caratteristiche di una classe, che siano queste (caratteristiche) di tendenza o di dispersione.
   Viene spesso rappresentata con tabelle, grafici e boxplot.

2. Discriminazione

   Con questa modalità invece, le caratteristiche di una classe vengono messe a confronto con quelle di un'altra classe ad essa paragonabile.
   Viene quindi eseguito un confronto diretto sulle tabelle o sui grafici.

Analisi associative

Meccanismi che permettono di identificare stuazione che ne implicano altre con un'elevata frequenza.
Devono essere individuati pattern che rappresentano implicazioni logiche come $A \rightarrow B$.
La significatività de un'associazione viene definita con due parametri:

1. **Confienza:** misura la certezza di un pattern, è definita come P(A|B).
2. **Supporto:** la frequenza con cui il pattern è stato verificatoi nel DB, è definito come la percentuiale degli elemrni che verifica la regola.

Classificazione e predizione

Prevediamo degli evnti futuri oppure facciamo delle inferenze su dei valori mancanti.

- **Classificazione:** definiamo i criteri che permettono di assegnare un soggetto ad una classe.
- **Predizione:** calcolo di funzionui di tendenza interpolando dati noti.

Classificazione

Specifica quali sono le classi obbiettivo della classificazione, quali sono i dati su cui costruire il modello e a quale classe appartengono.
Le tecniche per costruire i classificatori sono diverse:

- Funzioni matematiche
- Regole associative
- Alberi decisionali
- Reti della verità bayesiane
- Reti neurali

1. Alberi decisionali

   Un albero di decisioe è una struttura semi alberi in cui:

   - I nodi interni sono attributi del soggetto.
   - Gli archi in un uscita da un nodo sono i valori che l'attributo può avere.
   - Le foglie sono le classi.

   L'albero è di fatto una struttura *if-then-else* che va letta dalla radice alle foglie.

2. Predizione

   La predizione permette di prevedere (ma pensa un po') valori non ancora noti di un dominio continuo.
   La costruzione delle funzioni di tendenza avviene tramite interpolazione dei putni noti (regressione), esistono diversi modelli di regressione:

- **Lineare semplice:** $Y = q + mX$
- **Multilineare:** $Y = q + m_1X_1 + m_2X_2 + m_3X_3$
- **Non lineare:** $Y = q + m_1X + m_2X^2 + m_3X^3$

Meccanismi di clustering

I meccanismi di clustering, come i classificatori, ripartiscono i dati in classi differenti senza però conoscere le classi, solo sulle affinità che gli elementi hanno. Esistono diverse tecniche:

- **Partizionamento:** l'utente indica quante classi esistono.
- **Classificazione gerarchica:**
  - Aggregativa quando iterativamente aggrego gli elemnti in base alle similitudini.
  - Divisiva quando spezzo gli insiemi in sottoinsiemi di elementi caratterizzati.
- **Valutazione della densità:** gli elementi vengono divisi in base alla loro posizione nell'iperpiano.

Ricerca degli outlier

I metodi di clustering hanno come effetto secondario l'identificazione degli outlier, ovvero degli elementi che si discostano dai raggruppamenti.
L'analisi degli outlier è simile alla procedura di clustering ma si concentra sull'identificazione degli elementi he si discostano maggiormente. La ricerca si avvale di:

- **Metodi statistici:** utilizzabili solo quando si conosce la distribuzione (statistica) dei dati.
- **Metodi basati sulla distanza:** ricerca degli elementi con distanza maggiore.
- **Metodi basati sulla deviazione:** ricerca di elementi che diavo dal gruppo.