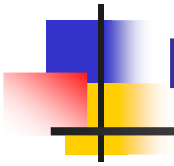
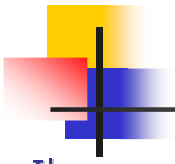


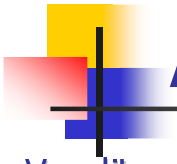
Data mining



Data Mining

Il processo di Data Mining

- consente di estrarre automaticamente informazione da un insieme di dati
- l'informazione è nascosta a causa di
  - la quantità di dati: ad. es. transazioni delle carte di credito, delle compagnie telefoniche, ...
  - la loro complessità: ad es. occorre integrare sorgenti di informazioni diverse fra loro, non ci sono noti i fattori che influenzano quello che si cerca, ...
  - la velocità a cui arrivano: ad es. per le carte di credito possono essere decine di transazioni al secondo,...
- E' l'ultimo stadio del processo di analisi (si usa a valle degli OLAP)
- Può fornire un importante ritorno economico



Applicazioni

Vendita al dettaglio e marketing

- Scoperta delle abitudini dei clienti
- Scoperta delle associazioni fra le caratteristiche demografiche dei clienti
- Predizione della risposta alle campagne pubblicitarie
- Analisi delle associazioni fra i prodotti acquistati (market basket)

Banche

- Uso fraudolento delle carte di credito
- Individuare i clienti che stanno per cambiare carta di credito, i clienti fedeli,...
- Determinare la quantità d'uso della carte di credito per gruppi di clienti

Applicazioni II

Assicurazioni

- Analisi delle richieste di risarcimento
- Predirre quali clienti possono essere interessati a nuove tipologie di polize
- Predirre il rischio associato ad una polizza con nuovo cliente

Medicina

- Predirre il rischio di una malattia associato ad ogni paziente
- Predirre la migliore cura per un determinato paziente

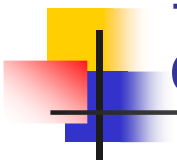
Applicazioni III

Bioinformatica

- Predirre la cancerogenità di una molecola
- Predirre l'efficacia di una molecola nella cura di una certa malattia
- Scoprire gruppi di molecole simili per le quali ci si aspetta propretà simili

Applicazioni web

- In un servizio dedicato al cinema (libri, giochi, ..) , suggerire agli utenti nuovi film da vedere (libri da acquistare, giochi da provare,...)
- Individuare nel web le comunità che sono interessate allo stesso argomento
- In un forum di discussione individuare gli eventi, cioè i momenti in cui cambia drasticamente l'argomento di cui si discute



Il processo di knowledge discovery e quello di data mining

Il processo di knowledge discovery è suddiviso nelle seguenti fasi

- Selezione dei dati
  Si scelgono i dati da analizzare. Essi possono provenire da un OLTP o da un OLAP
- Ripulitura dei dati e trasformazione
  Occorre ripulire i dati e prepararli per le operazioni successive. Spesso le tabelle sono denormalizzate e combinate in un'unica tabella
- Data mining
  Si applicano tecniche di apprendimento automatico, clustering, ...
- Valutazione e interpretazione
  Nella maggior parte dei casi i risultati prodotti dal data mining non sono abbastanza affidabili da essere usati direttamente. Essi devono essere valutati e interpretati.

Tecnologie per il data mining

- Si usano tecniche provenienti dall'intelligenza artificiale
  - tali tecniche sono adattate per migliorarne le prestazioni su grandi quantità di dati
- Esistono numerosi tool per il data mining, ma ....
  - ogni applicazione ha una soluzione differente
  - per trovare una buona soluzione occorrono degli "artigiani" che selezionino la strada giusta fra un ampio insieme di tecnologie
- Le tecnologie per il data mining
  - permettono di scoprire informazione che in altri modi non è accessibile: sapere qualcosa che nessuno sa può essere un vantaggio enorme
  - sono molto costose da implementare

Tipologie di applicazioni

Analisi delle associazioni

- individuare le regole nascoste del tipo: l'evento A implica l'evento B
  - ad es. chi compra una stampante di solito compra anche il toner

Problemi di classificazione o regressione

- a partire da un insieme di esempi si apprende a classificare un oggetto
  - ad es. si vuol classificare un nuovo utente di un'assicurazione come utente ad alto rischio o meno: addestra un modello con gli esempi dei vecchi clienti

Problemi di clustering

- Si cerca di organizzare automaticamente gli eventi/oggetti di un database
  - ad es. si vuol identificare le molecole con un proprietà farmacologiche simili

Scoperta degli eventi che deviano dal comportamento normale

- Si cerca di individuare gli eventi, gli oggetti i comportamenti anomali
  - ad es. si vuol individuare le frodi su una carta di credito

Analisi delle associazioni: il problema del carrello

Il problema del carrello del supermercato

- Data la registrazione delle "transazioni" di un supermercato:
  - una transazione è un insieme di oggetti acquistati contemporaneamente da un utente
- trovare gli oggetti che più di frequente sono stati acquistati insieme
  - ad es. farina e lievito oppure farina, lievito, latte

TID	CID	Data	Prod.	Q.tà
111	201	5/1/05	farina	2
111	201	5/1/05	lievito	1
111	201	5/1/05	latte	3
111	201	5/1/05	carne	6
112	105	7/1/05	farina	1
112	105	7/1/05	lievito	1
112	105	7/1/05	latte	2
113	106	7/1/05	farina	2
113	106	7/1/05	latte	1
114	201	8/1/05	farina	3
114	201	8/1/05	lievito	2
114	201	8/1/05	carne	6
114	201	8/1/05	vino	6

Regole di associazione

Ricerca delle regole di associazione

- Consiste nell'identificare le regole di implicazione fra gli eventi $H \Rightarrow T$
  - Ad es., $\{farina\} \Rightarrow \{lievito\}$
- Per ogni regola $H \Rightarrow T$ si definiscono
  - $supporto(H \Rightarrow T) = supporto(H \cup T)$
    Ad. es. $supporto(\{farina\} \Rightarrow \{lievito\}) = 0.75$
  - $confidenza(H \Rightarrow T) = supporto(H \Rightarrow T) / supporto(H)$
    Ad. es. $confidenza(\{farina\} \Rightarrow \{lievito\}) = 0.75$

TID	CID	Data	Prod.	Q.tà
111	201	5/1/05	farina	2
111	201	5/1/05	lievito	1
111	201	5/1/05	latte	3
111	201	5/1/05	carne	6
112	105	7/1/05	farina	1
112	105	7/1/05	lievito	1
112	105	7/1/05	latte	2
113	106	7/1/05	farina	2
113	106	7/1/05	latte	1
114	201	8/1/05	farina	3
114	201	8/1/05	lievito	2
114	201	8/1/05	carne	6
114	201	8/1/05	vino	6

Classificazione (regressione)

In cosa consiste

- consiste nell'inferire una proprietà di un oggetto sulla base di alcune sue caratteristiche
  - ad es. si vuol inferire il rischio di un utente di una polizza
- la proprietà da inferire può essere un valore numerico qualsiasi (regressione) o apparttenere ad un insieme finito (classificazione)

Nel nostro caso

- Spesso si crea una tabella che contiene tutte le proprietà necessarie all'inferenza
- La proprità da inferire è un attributo della tabella

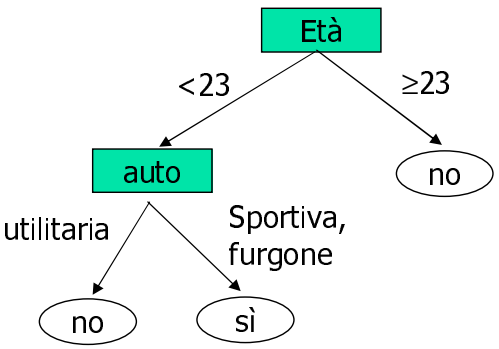
POLIZZE(**id, nome,età,auto_o_furgone,cavalli,attività**,……,**altoRischio**)

Caratteristiche

Proprietà da predirre

Alberi di decisione

- Rappresentano un insieme di regole che permettono di fare la predizione automaticamente
  - Ogni nodo interno rappresenta un test e i suoi rami indicano le risposte
  - Ogni foglia rappresenta una decisione
- Sono costruiti automaticamente usando i dati disponibili
  - ad le caratteristiche di rischio dei vecchi clienti dell'assicurazione



Clustering

In cosa consiste

- mira a suddividere un insieme di oggetti in modo che
  - oggetti nello stesso gruppo siano simile
  - oggetti in gruppi diversi siano dissimili
- Il raggruppamento viene attuato con tecniche di apprendimento non supervisionato

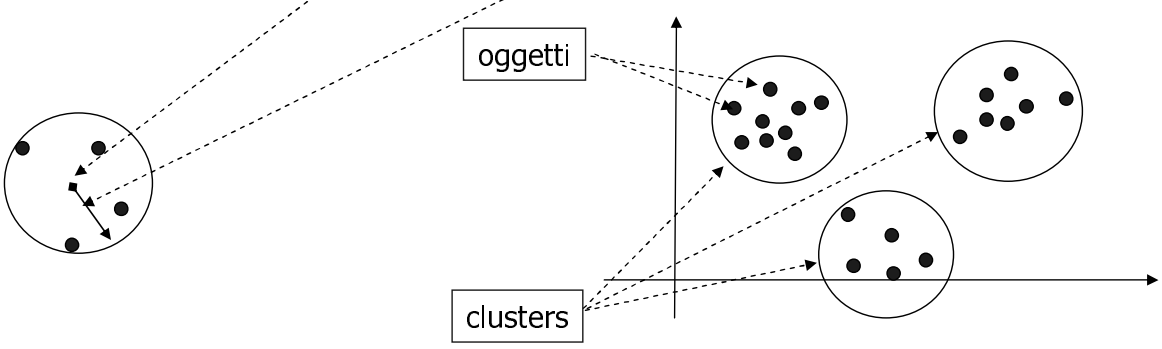
Applicazioni

- Individuazione di molecole con proprietà curative simili
- Raggruppamento di utenti in base al loro comportamento su un sito
- Raggruppamento di utenti in base alle loro caratteristiche sociali ed economiche

Clustering II

Gli algoritmi tipici di clustering

- gli oggetti da organizzare sono punti in uno spazio n-dimensionale
- esiste una misura che definisce la distanza fra gli oggetti
- l'algoritmo deve individuare delle sfere che racchiudano gli oggetti
- ogni cluster ha un centro e un raggio



Strumenti per il data mining

Strumenti costruiti appositamente

- alcuni produttori costruiscono strumenti ad hoc per il data mining, capaci di prendere dati da sorgenti diverse
  - ad. es. SAS Enterprise Miner, SPSS Clementine, CART (Salfort Systems), Megaputer PolyAnalyst, ANGOSS KnowledgeStudio

Strumenti associati ai DBMS

- i maggiori produttori di DBMS offrono anche strumenti per il data mining
  - IBM Intelligent Miner
    Supporta numerosi algoritmi per la ricerca di regole di associazione, la classificazione, la regressione e il clustering
  - Microsoft Analysis Service
    Supporta gli alberi di decisione, il clustering ….