



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 proprietà 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

9

Analisi delle associazioni: il problema del carrello II

Si definisce una misura

- Il supporto di un insieme di oggetti S: la percentuale delle transazioni in cui S è presente
 - ad es.
supporto({farina, lievito})=75%,
supporto({farina, lievito,latte})=50%

Usare l'SQL può non essere efficiente

```
select T1.prod, T2.prod, count(*) as N
from Transazioni as T1, Transazioni as T2
where T1.TID=T2.TID and not (T1.prod=T2.prod)
GROUP BY T1.prod, T2.prod
HAVING N >= 3
```

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

Trova gli insiemi di due oggetti con supporto maggiore del 75%

Franco Scarselli

Sistemi per basi di dati 2005-2006

10

Un algoritmo efficiente

- si trovano tutti gli insiemi più frequenti con n elementi utilizzando quelli con n-1 elementi
- Occorre scandire la tabella delle transizioni n volte: molto più efficiente che usare un join fra n tabelle con un raggruppamento!!

```
Leggi le transizioni e calcola supporto( $\{o\}$ ) per ogni oggetto o
for each oggetto o begin
    if supporto( $\{o\}$ ) >= 75% then inserisci  $\{o\}$  in  $I_0$ ;
end

k=1;
repeat
    Costruisci  $G_k$  inserendovi tutti gli insiemi  $A \cup B$ ,  $A \in I_0$ ,  $B \in I_{k-1}$ 
    Leggi le transizioni e calcola supporto(C) per ogni  $C \in I_0$ 
    for each  $C \in I_0$  begin
        if supporto(C) >= 75% then inserisci C in  $I_k$ ;
    end
end
```

Franco Scarselli

Sistemi per basi di dati 2005-2006

11

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. $supporto(\{farina\} \Rightarrow \{lievito\}) = 75\%$
 - $confidenza(E_1 \Rightarrow E_2) = supporto(H \Rightarrow T) / supporto(H)$
Ad. es. $confidenza(\{farina\} \Rightarrow \{lievito\}) = 0.75$

TID	CID	Data	Prod.	Q.t à
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

Franco Scarselli

Sistemi per basi di dati 2005-2006

12

Regole di associazione II

Trovare le regole di associazione

- Si tratta di trovare tutte le regole R per le quali $supporto(R) > min_sup$ e $confidenza(R) > min_con$
- Si risolve basandosi sull'algoritmo per la ricerca degli insiemi frequenti

```
Trova tutti gli insiemi frequenti  $S$  per i quali  $supporto(S) > min\_sup$   
for each  $S$  begin  
    Dividi  $S$  in tutte possibili coppie di insiemi  $A, B, S = A \cup B$   
    for each  $A, B$  begin  
        if  $confidenza(A \Rightarrow B) \geq min\_con$   
            then  $A \Rightarrow B$  è una delle regole trovate  
        end  
    end  
end
```

Franco Scarselli

Sistemi per basi di dati 2005-2006

13

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 appartenere ad un insieme finito (classificazione)

Nel nostro caso

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

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

Caratteristiche

Franco Scarselli

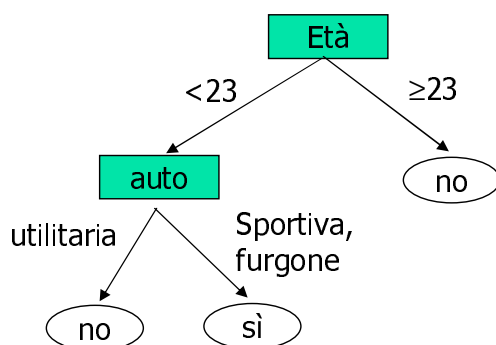
Sistemi per basi di dati 2005-2006

Proprietà da predire

14

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



Alberi di decisione

Esistono algoritmi efficienti adatti a lavorare con grandi insiemi di dati

La realizzazione dell'albero si basa su due fasi

- Costruzione
- Raffinamento

Costruzione

- Si cerca un buon criterio C per dividere il dataset in due sottoinsiemi D1, D2
- Si costruisce un nodo che usa il criterio C e si applica ricorsivamente l'algoritmo a D1 e D2

Raffinamento

- L'albero costruito viene semplificato eliminando i rami meno importanti



Reti neurali

Cosa sono

- sono modelli parametrici in grado di implementare una funzione $f_w(x_1, x_2, \dots, x_n)$
 - w sono i parametri da apprendere
 - x_1, x_2, \dots, x_n le caratteristiche note (ad es. età, auto posseduta, ... di un cliente)
 - $f_w(x_1, x_2, \dots, x_n)$ calcola la proprietà da predire
- i parametri sono appresi da esempi

Osservazione

- Poichè l'apprendimento può richiedere molto tempo con le reti neurali, non si usano tutti i dati disponibili, ma sono un sottoinsieme selezionato un modo casuale



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**

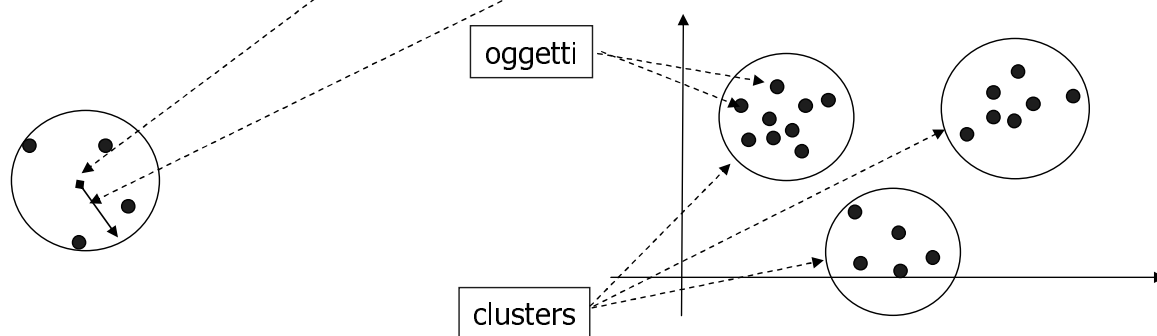
Applicazioni

- 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**



Franco Scarselli

Sistemi per basi di dati 2005-2006

19

L'algoritmo BIRCH

- Occorrono **algoritmi** di clustering specifici per trattare **grandi moli di dati**
- L'algoritmo BIRCH legge una sola volta tutti gli oggetti e produce k clusters $(c_1, r_1), (c_2, r_2), \dots, (c_k, r_k)$
 - c_1, c_2, \dots, c_k sono i centri dei clusters, r_1, r_2, \dots, r_k i raggi
 - k è scelto tale che la definizione dei clusters possa essere tenuta in memoria
 - esiste un parametro ϵ che definisce la massima dimensione di un cluster

repeat

leggi il record corrente A e trova il cluster i più vicino ad A
prova ad inserirvi A e calcola il nuovo centro nc_i e la distanza di A a nc_i
if $d < \epsilon$ inserisci A nell' i -esimo cluster
else crea un nuovo cluster con centro A
end

if si è raggiunto il massimo numero di clusters
incrementa ϵ ed, eventualmente, fonda i clusters;
vai prossimo record

until si sono letti tutti i record

Franco Scarselli

Sistemi per basi di dati 2005-2006

20



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 (Salford 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 Server Intelligent Miner
Supporta gli alberi di decisione e il clustering



..... the end

- seconda prova intermedia mercoledì 7 dicembre ore 10-12
- consegna progetto, recupero prove intermedie, verbalizzazioni in qualsiasi giorno di ricevimento
- settori di ricerca in cui potrebbero essere disponibili tesi:
data mining e apprendimento automatico