



# 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

3



# Applicazioni II

## Assicurazioni

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

## Medicina

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

Franco Scarselli

Sistemi per basi di dati 2005-2006

4



# Applicazioni III

## Bioinformatica

- Predire la cancerogenità di una molecola
- Predire 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

5



# 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.

Franco Scarselli

Sistemi per basi di dati 2005-2006

6



# 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

7



## 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

8

# 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.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

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



# Analisi delle associazioni: il problema del carrello III

Soluzioni inefficienti per il problema del carrello (con molti dati)

- Si implementa e si ottimizza un'interrogazione SQL come una qualsiasi interrogazione
  - Inefficiente perché richiede l'implementazione di join e raggruppamento: due join per gruppi di due oggetti, tre join per gruppi di tre oggetti .....
- Si scorrono le transazioni in maniera sequenziale, tenendo dei contatori che contano tutti i gruppi di oggetti incontrati
  - inefficiente, perché i possibili gruppi possono essere numerosissimi: la struttura dati per memorizzare i contatori potrebbe essere più grande della memoria principale

Franco Scarselli

Sistemi per basi di dati 2005-2006

11



## Un algoritmo efficiente: Apriori

- Si trovano tutti gli insiemi più frequenti con n elementi utilizzando quelli con n-1 elementi
  1. Al primo passo si scandisce la tabella delle transizioni e si contano i singoli oggetti più comprati: si sceglie l'insieme  $I_1$  di oggetti che contengono quelli il cui supporto è maggiore di una certa soglia T
  2. Si scandisce la tabella della transizioni e contano le occorrenze delle coppie che contengono oggetti di  $I_1$ ; si sceglie l'insieme  $I_2$  di coppie che contengono quelle il cui supporto è maggiore di T
  3. Per trovare gli insieme di n oggetti si ripete 2 (si ottiene  $I_n$  combinando gli oggetti di  $I_{n-1}$ )
- Al passo 2 si considera solo un sottoinsieme delle coppie (gruppi): se la soglia è alta, la struttura dati sta sicuramente in memoria
- Nota che ... non possono esistere se un gruppo di oggetti ha supporto maggiore di T anche tutti i sottogruppi hanno supporto maggiore di T
- Per gruppi di n oggetti, occorre scandire la tabella delle transizioni n volte!

# Un algoritmo efficiente: Apriori

Ricerca di oggetti con supporto maggiore del 75%

```

Leggi le transizioni e calcola  $supporto(\{o\})$  per ogni oggetto o
for each oggetto o begin
    if  $supporto(\{o\}) \geq 75\%$  then inserisci  $\{o\}$  in  $I_1$ ;
end

k=2;
repeat
    Costruisci  $G_k$  inserendovi tutti gli insiemi  $A \cup B, A \in I_{k-1}, B \in I_{k-1}$ 
    Leggi le transizioni e calcola  $supporto(C)$  per ogni  $C \in I_k$ 
    for each  $C \in I_k$  begin
        if  $supporto(C) \geq 75\%$  then inserisci C in  $I_k$ ;
    end
end

```

Franco Scarselli

Sistemi per basi di dati 2005-2006

13

## 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

14

# 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  $confidenza(A \Rightarrow B) > min\_con$   
      then  $A \Rightarrow B$  è una delle regole trovate  
    end  
  end
```

Franco Scarselli

Sistemi per basi di dati 2005-2006

15

## 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

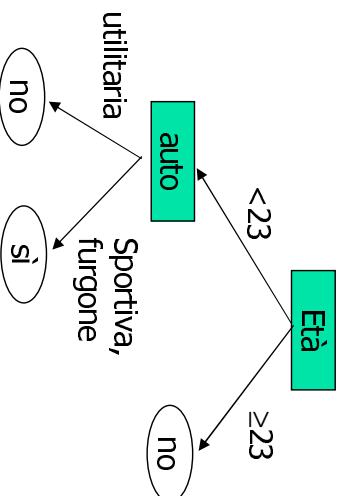
Proprietà da predire

16



# 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



Franco Scarselli

Sistemi per basi di dati 2005-2006

17

## Alberi di decisione II

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
  - Esistono criteri diversi per la divisione in sottoinsiemi: ad esempio si può scegliere l'attributo che massimizza l'information gain
  - un buon criterio dovrebbe minimizzare la profondità dell'albero
- 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

18



## Alberi di decisione III

```
buildTree(Nodo n, Partizione P, criterio di scelta S )
1.  Applica S a P per trovare il criterio di divisione
2.  If (esiste un buon criterio di divisione){
3.      crea due figli n1, n2
4.      dividi P in P1, P2
5.      buildTree(n1,P1,S)
6.      buildTree(n2,P2,S)
}
```



## Scelta del criterio

- Cosa succede se la tabella è troppo grande e non entra in memoria principale ?
- Si costruisce l'**AVC set** (Attribute Value Class label): contiene le occorrenze di ogni coppia attributo-classe
- L'AVC set è tipicamente molto più piccolo della tabella
- L'AVC set viene usato per calcolare il criterio
  - Si osservi che questo è possibile perché il criterio usa un solo attributo e un solo valore

# AVC set: un esempio

## POLIZZE

età	auto	alto_rischio
20	furgone	sì
30	sportiva	sì
20	utilitaria	no
20	utilitaria	sì
40	furgone	sì
20	sportiva	sì

```
SELECT età, alto_rischio , count(*)  
FROM polizze  
GROUP by età, alto_rischio  
  
SELECT auto, alto_rischio , count(*)  
FROM polizze  
GROUP by auto, alto_rischio
```

## AVC set

età	alto_rischio	occorrenze
20	sì	3
20	no	1
30	sì	1
40	sì	1

auto	alto_rischio	occorrenze
furgone	sì	2
utilitaria	sì	1
utilitaria	no	1
sportiva	sì	2

Franco Scarselli

Sistemi per basi di dati 2005-2006

21

# Reti neurali

## Cosà 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

Franco Scarselli

Sistemi per basi di dati 2005-2006

22

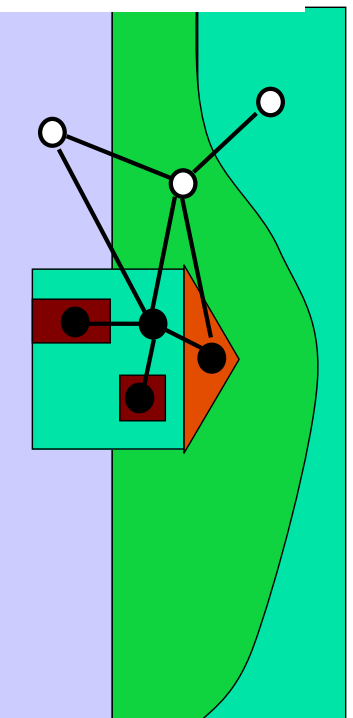
# Reti neurali per l'elaborazione di grafi: Graph Neural Networks

- Sono un'evoluzione delle reti neurali che possono essere applicate su insieme di grafi
  - Il dataset rappresenta un insieme di oggetti (nodi) correlati da relazioni (archi)
  - Ogni nodo ha delle etichette che rappresentano caratteristiche dell'oggetto
  - si tratta di predire una proprietà degli oggetti usando sia le caratteristiche degli oggetti che le loro relazioni

## Esempio

Localizzare una casa, distinguendo le regioni (nodi neri) che la compongono dalle altre regioni (nodi bianchi)

L'apprendimento usa immagini di case e non



23

## Relational learning

- Il relational learning consiste nel prevedere un attributo di una relazione utilizzando gli altri attributi e i collegamenti con le altre relazioni
- Il relational learning può essere affrontato con tecniche di programmazione logica e con le graph neural networks

**Esempio: prevedere il livello di rischio di una polizza**

Assicurati			Polizze			Rimborsi	
nome	età		tipo	oggetto assicurato	livello rischio	motivo	q.tà
paperino	35		RCA	fiat punto	?	incidente	100
pipipo	40		vita	Pippo	?	atti vand.	300
			RCA	cinquecento	?		

# 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

Franco Scarselli

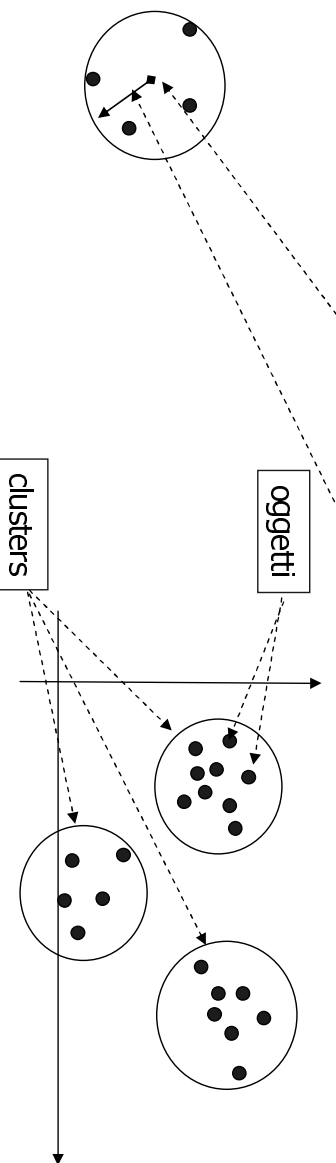
Sistemi per basi di dati 2005-2006

25

## 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

26

# Un algoritmo efficiente: BIRCH

- Occorrono algoritmi di clustering specifici per trattare grandi moli di dati
- Gli algoritmi classici (ad esempio, k-means) non sono adatti perché prevedono un alto numero di epoche di apprendimento: ogni epoca implica una lettura del training set
- 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  $\epsilon$  che definisce la massima dimensione di un cluster

Franco Scarselli

Sistemi per basi di dati 2005-2006

27

# Un algoritmo efficiente: BIRCH

```
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  $\epsilon$  ed, eventualmente, fondi i clusters;
    vai prossimo record
until si sono letti tutti i record
```

Franco Scarselli

Sistemi per basi di dati 2005-2006

28



# 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