

Pattern Frequenti e Regole di Associazione Confidenti

Caso di studio di Metodi Avanzati di
Programmazione

AA 2016-2017

Data Mining

Lo scopo del **data mining** è l'*estrazione* (semi) automatica di *conoscenza* nascosta in voluminose basi di dati al fine di renderla disponibile e direttamente utilizzabile



Aree di Applicazione

1. previsione

utilizzo di valori noti per la previsione di quantità non note (es. stima del fatturato di un punto vendita sulla base delle sue caratteristiche)

2. classificazione

individuazione delle caratteristiche che indicano a quale gruppo un certo caso appartiene (es. discriminazione tra comportamenti ordinari e fraudolenti)

3. segmentazione

individuazione di gruppi con elementi omogenei all'interno del gruppo e diversi da gruppo a gruppo (es. individuazione di gruppi di consumatori con comportamenti simili)

4. associazione

individuazione di elementi che compaiono spesso assieme in un determinato evento (es. prodotti che frequentemente entrano nello stesso carrello della spesa)

5. sequenze

individuazione di una cronologia di associazioni (es. percorsi di visita di un sito web)

...

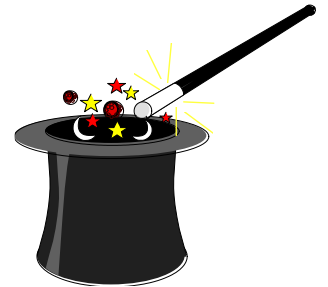
Pattern frequenti

Dati:

- una collezione D di transazioni
- dove, ogni transazione è un vettore di coppie attributo-valore (item)

Lo scopo è:

- Identificare gli insiemi di item (itemset o pattern) che occorrono con una frequenza minima in D



Pattern frequenti

- Sono inizialmente definiti nel **market basket analysis** (http://it.wikipedia.org/wiki/Market_basket_analysis).
- **Motivazione**: scoprire regolarità in un data base di transazioni di un cliente
- *Quali prodotti SPESSO compaiono sullo stesso scontrino emesso da un supermercato (sono comprati insieme)?*



Esempio

- Nel 2% degli scontrini di un supermercato sono registrati pannolini, omogeneizzati e birra

PANNOLINI,

OMOGENEIZZATI,

BIRRA



Regole di associazione

- Più interessante,

Il 98% degli scontrini in cui sono registrati pannolini e omogeneizzati registrano anche l'acquisto di birra

PANNOLINI, OMOGENEIZZATI → BIRRA

Regola di associazione che correla la presenza di PANNOLINI e OMOGENEIZZATI (*antecedente*) alla presenza di BIRRA (*conseguente*).

Regole di associazione

In generale, una regola di associazione è nella forma

$$X \rightarrow Y (s\%, c\%)$$

- X è l'*antecedente* della regola
- Y è il *conseguente* della regola
- X e Y sono insiemi di item tali che $X \cap Y = \emptyset$
- La percentuale $s\%$ denota il *supporto* della regola. Esso stima $p(X \cup Y)$
- La percentuale $c\%$ denota la *confidenza* della regola. Essa stima $p(Y|X)$

Regole di associazione: *Confidenza*

- Una regola deve avere una minima confidenza specificata dall'utente
- **PANNOLINI , OMOGENEIZZATI → BIRRA**
ha una confidenza del 90% se il 90% degli scontrini che includono pannolini e omogeneizzati, includono anche la birra.
- In generale,

$$c(X \rightarrow Y) = \frac{p(X \cup Y)}{p(X)} = \frac{\text{numero di transazioni in cui si osserva X e Y}}{\text{numero di transazioni in cui si osserva X}}$$

Regole di associazione:

Supporto

- Una regola deve avere un minimo *supporto* specificato dall'utente
- **PANNOLINI , OMOGENEIZZATI & BIRRA**
devono comparire insieme in un minimo numero di scontrini per avere una qualche valenza statistica.
- In generale,

$$s(X \rightarrow Y) = \frac{p(X \cup Y)}{D} = \frac{\text{numero di transazioni in cui si osserva X e Y}}{\text{numero totale di transazioni}}$$

Scoperta di regole di associazione:

Definizione del problema

- Dati:
 - un database di transazioni;
 - un valore di minimo supporto ($0 < \text{minS} \leq 1$);
 - un valore di minima confidenza ($0 < \text{minC} \leq 1$).
- Trovare tutte le *regole di associazione* in D che siano **frequenti** (supporto maggiore o uguale di minS) e **confidenti** (confidenza maggiore uguale di minC)

Scoperta di regole di associazione:

Esempio

Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

Min. supporto 0.5
Min. confidenza 0.5

Frequent pattern (minS=0.5)	support
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5
Rain, Normal	0.5
Rain, Cool	0.5
Cool, Normal	0.5
Hot, High	0.5
Rain, Cool, Normal	0.5

Per la regola di associazione

Rain → Normal

$\text{supporto}(\text{Rain} \rightarrow \text{Normal}) = 3/6 = 0.5$

$\text{confidenza}(\text{Rain} \rightarrow \text{Normal}) = 3/4 = 0.75$

Scoperta di regole di associazione:

Decomposizione del problema

1. Trovare i pattern frequenti
2. Usare i pattern frequenti per generare le regole di associazione confidenti

Scoperta di regole di associazione:

1. individuare i pattern frequenti

Nota: Ciascun sottoinsieme di un pattern frequente
DEVE essere frequente

- Se “PANNOLINO,OMOGENEIZZATO,BIRRA” è frequente allora anche “OMOGENEIZZATO, BIRRA” deve essere frequente
- Ogni scontrino che contiene
{ pannolino,omogeneizzato,birra } contiene anche
{ omogeneizzato,birra }

Conseguenza: *Monotonia del supporto*

Se un pattern è infrequente allora raffinamenti dello stesso pattern sono anche infrequenti

Scoperta di regole di associazione:

1. individuare i pattern frequenti

Metodo:

1. Scoprire pattern di lunghezza k a partire da pattern FREQUENTI di lunghezza $k-1$
2. Testare i pattern candidati in D

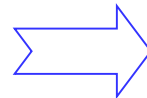
Scoperta di regole di associazione:

1. individuare i pattern frequenti

Min. supporto 0.5

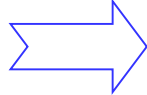
Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

K=3



Frequent pattern (minS=0.5)	s
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5
Rain, Cool	0.5
Rain, Normal	0.5
Hot, High	0.5
Cool, Rain	0.5
Cool, Normal	0.5
High, Hot	0.5
Normal, Rain	0.5
Normal, Cool	0.5
Rain, Cool, Normal	0.5
Rain, Normal, Cool	0.5
Cool, Rain, Normal	0.5
Cool, Normal, Rain	0.5
Normal, Rain, Cool	0.5
Normal, Cool, Rain	0.5

Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

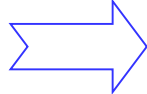
K=1


Frequent pattern (minS=0.5)	s
Sunny	0.33
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5

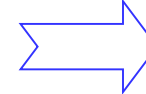
Dr. A. Appice

Min. supporto 0.5

Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

K=1


Frequent pattern (minS=0.5)	s
Sunny	0.33
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5

K=2


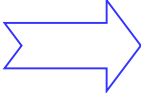
Dr. A. Appice

Frequent pattern (minS=0.5)	s
Rain,Hot	0.166
Rain,Cool	0.5
Rain,High	0.166
Rain, Normal	0.5
Hot, Rain	0.166
Hot, High	0.5
Hot, Normal	0
Cool, Rain	0.5
Cool, High	0
Cool, Normal	0.5
High, Rain	0.166
High, Hot	0.5
High, Cool	0
Normal, Rain	0.5
Normal, Hot	0
Normal, Cool	0.5

Min. supporto 0.5

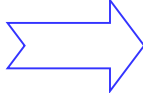
Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

K=1



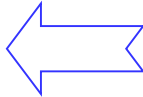
Frequent pattern (minS=0.5)	s
Sunny	0.33
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5

K=2



Frequent pattern (minS=0.5)	s
Rain,Hot	0.166
Rain,Cool	0.5
Rain,High	0.166
Rain, Normal	0.5
Hot, Rain	0.166
Hot, High	0.5
Hot, Normal	0
Cool, Rain	0.5
Cool, High	0
Cool, Normal	0.5
High, Rain	0.166
High, Hot	0.5
High, Cool	0
Normal, Rain	0.5
Normal, Hot	0
Normal, Cool	0.5

K=3



Frequent pattern (minS=0.5)	s
Rain,Cool, High	0
Rain, Cool, Normal	0.5
Rain, Normal, Hot	0
Rain, Normal, Cool	0.5
Hot, High ,Sunny	0.33
Hot, High, Rain	0.166
Cool, Rain, High	0
Cool, Rain, Normal	0.5
Cool, Normal, Sunny	0
Cool, Normal, Rain	0.5
High, Hot, Sunny	0.33
High, Hot, Rain	0.166
Normal, Rain, Hot	0
Normal, Rain, Cool	0.5
Normal, Cool, Sunny	0
Normal, Cool, Rain	0.5

Min. supporto 0.5

Dr. A. Appice

Scoperta di regole di associazione:

2. Derivare regole confidenti

- **Nota:** pattern frequenti \neq regole di associazione
- Un ulteriore passo è richiesto per scoprire le regole di associazione
- Per ciascun pattern frequente ***P***,

Per ciascun sottoinsieme non vuoto ***X*** di ***P***,

- Sia ***Y*** = ***P*** - ***X***
- $X \Rightarrow Y$ è una regola di associazione se e solo se
confidenza ($A \Rightarrow B$) \geq minC,

dove:

- $\text{supporto}(A \Rightarrow B) = \text{supporto}(AB)$ e
- $\text{confidenza}(A \Rightarrow B) = \text{supporto}(AB) / \text{supporto}(A)$

Day	Outlook	Temperature	Humidity
D1	Sunny	Hot	High
D2	Sunny	Hot	High
D3	Rain	Hot	High
D4	Rain	Cool	Normal
D5	Rain	Cool	Normal
D6	Rain	Cool	Normal

Frequent pattern (minS=0.5)	s
Rain	0.66
Hot	0.5
Cool	0.5
High	0.5
Normal	0.5
Rain, Cool	0.5
Rain, Normal	0.5
Hot, High	0.5
Cool, Rain	0.5
Cool, Normal	0.5
High, Hot	0.5
Normal, Rain	0.5
Normal, Cool	0.5
Rain,Cool, Normal	0.5
Rain, Normal, Cool	0.5
Cool, Rain, Normal	0.5
Cool, Normal, Rain	0.5
Normal, Rain, Cool	0.5
Normal, Cool, Rain	0.5

Frequent pattern (minS=0.5, minC=1)	c
Rain → Cool	0.75
Rain → Normal	0.75
Hot → High	1
Cool, → Rain	1
Cool → Normal	1
High → Hot	1
Normal → Rain	1
Normal → Cool	1
Rain → Cool, Normal	0.75
Rain, Cool → Normal	1
Rain → Normal, Cool	0.75
Rain, Normal → Cool	1
Hot → High, Sunny	0.66
Hot, High → Sunny	0.66
Cool, → Rain, Normal	1
Cool, Rain → Normal	1
Cool → Normal, Rain	1
Cool, Normal → Rain	1
Normal → Rain, Cool	1
Normal, Rain → Cool	1
Normal → Cool, Rain	1
Normal, Cool → Rain	1

Apriori: *pseudo-code*

R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, 1994.

(http://en.wikipedia.org/wiki/Apriori_algorithm)



<http://rakesh.agrawal-family.com/>



<http://www.rsrikant.com/>

Apriori: *pseudo-codice*

```
frequentPatternDiscovery(D, minS) → P
begin
  P =  $\emptyset$ 
  L1 = {1-item che compaiono in minS × |D| transazioni di D}
  K = 2
  while LK-1 ≠  $\emptyset$  do
    begin
      CK = candidati generati da Lk-1 aggiungendo un nuovo item
      LK =  $\emptyset$ 
      for each (p ∈ Ck) do
        if (supporto(p, D) ≥ minS) then
          LK = LK ∪ p
      P = P ∪ Lk
      K = K + 1
    end
  return P
end
```

Apriori: *pseudo-codice*

ConfidentAssociationRuleDiscovery(*D*, *P*, *minC*) \rightarrow *AR*

begin

AR = \emptyset

for each (*p* \in *P*) **do**

for each (*j*=1 to *j*<*p*.LENGTH) **do**

// *p*=*p*[1],*p*[2],...,*p*[*p*.LENGTH]

begin

ar = *p*[1],...,*p*[*j*] \rightarrow *p*[*j*+1],..., *p*[*p*.LENGTH]

if (confidenza(*ar*,*D*) \geq *minC*) **then**

AR = *AR* \cup *ar*

end

return *AR*

end

Problema :

Gestione di attributi numerici

- Attributi discreti (genere, stato civile)
- Attributi numerici (età, reddito)
 - Discretizzare l'attributo numerico

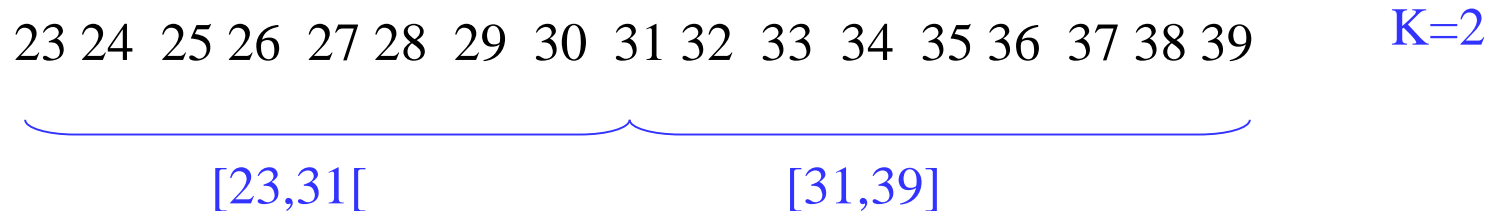
Id	Età	Stato civile	Numero di Auto di proprietà
1	23	celibe	1
2	25	coniugato	1
3	29	celibe	0
4	34	coniugato	2
5	39	coniugato	2

$$\text{minS} = 0.3 \quad \text{minC} = 0.5$$

Età=[23...31] → Stato Civile =celibe (s=0.33, c=0.66)

Discretizzazione in uguale ampiezza

- Sia dato il numero di intervalli k
- Si ricavano k intervalli di uguale ampiezza



Caso di studio

Progettare e realizzare un sistema **client-server** denominato “APRIORI”.

Il server include funzionalità di **data mining** per la scoperta di pattern frequenti e regole di associazione confidenti.

Il client consente di usufruire del servizio di scoperta remoto e visualizza la conoscenza (pattern e regole) scoperta.

Istruzioni

1. Il progetto dello A.A. 2016/17 denominato RULE, è valido solo per coloro che superano la prova scritta o prove in itinere entro il corrente A.A.
2. Ogni progetto può essere svolto da gruppi di **al più TRE** (3) studenti.
3. Coloro i quali superano la prova scritta devono consegnare il progetto **ENTRO** la data prevista per la corrispondente prova orale.
4. Il voto massimo assegnabile al progetto è 33.
5. Il voto finale sarà stabilito sulla base della media aritmetica del voto attribuito allo scritto e del voto attribuito al progetto. Un voto superiore a 30 equivale a 30 e lode.



Istruzioni

Non si riterrà sufficiente, e come tale non sarà corretto, un progetto non sviluppato in tutte le su parti (client-server, parte grafica, accesso al db, serializzazione,...)