Metodi di Esplorazione e Sommazione

Metodi descrittivi d´apprendimento

Regolarità locale

Regole d´Associazione

Mining di Sottogruppi

Modelli globali

Clustering

Regole d´Associazione

Obiettivo generale
Compito della Ricerca
Attributi Binari
Apriori algoritmo
Insiemi Frequenti
Problemi
Supplementi

Obiettivi di Regole d´Associazione

Obiettivo generale: Sopportare esplorazioni

Trovare associazioni fra valori di variabili
dipendenze fra variabili: associazioni, correlazioni

Analizzare transazioni: merce comprate in stessa transazione parole in un testo, errori in un rete, ... "items" ricorrendo insieme

Se si comprano pizza e birra, è probabile che si compra anche chips.

Ordinare merce in mercato. Offrire prodotti ai clienti...

Introdotto da Agrawal, Mannila 1996 e poi soggetto predominante di data mining, ma non dominante in pratica

Compito: Regole d´Associazione

Sia I un insieme d´oggetti (items)
e T un insieme di transazioni $t \in T: \; t \subset I$
s_{min} minimale frequenza (supporto) e c_{min} minimale confidenza
$0 < s_{min}, c_{min} \leq 1$ (scelti dall´utente)

Il compito è: trova tutte le regole $r := \; X \rightarrow Y$

$X \subset I, \; Y \subset I \quad X \cap Y = \varnothing$

$s(r) := |\{t \in T \mid X \cup Y \subset t\}| \; / \; |T| \geq s_{min}$

$c(r) := |\{t \in T \mid X \cup Y \subset t\}| \; / \; |\{t \in T \mid X \subset t\}| \geq c_{min}$

valori tipici: $s_{min} = 0.01$ $c_{min} = 0.5$ {pizza, birra} $\rightarrow$ {chips}

almeno il 1 percento dei clienti hanno comprato pizza, birra, chips
almeno il 50 percento dei clienti che hanno compato pizza, birra hanno anche comprato chips

Regole d´Associazione: Attributi binari

Sia $R=\{A_1,...A_p\}$ uno schema d´attributi binari
e r una relazione su R

$X \subset R$, $Y \subset R$, $X \cap Y = \varnothing$

s_{min} minimale frequenza (supporto) e c_{min} minimale confidenza
$0 < s_{min}, c_{min} \leq 1$ (scelti dal uttente)

Il compito è: trova tutte le regole $r := X \rightarrow Y$

$X=\{A_{i1},...,A_{ik}\}$ $Y=\{B\}$ $r := A_{i1}=1 \wedge ... \wedge A_{ik}=1 \rightarrow B=1$

attributi categoriali: selettori A=a
mà: esplosione di *frequent set*

Algoritmo *Apriori*

APRIORI(I, T, s_{min}, c_{min})

trova L:= insiemi-frequenti(I, T, s_{min}) frequent sets

trova R:= regole (L, c_{min})

Un insieme $X \subset I$ è frequente:= $s(X) := |\{t \in T \mid X \subset t\}| \ / \ |T| \geq s_{min}$

Il spazio degli insiemi $X \subset I$ è parzialmente ordinato

$X < Y ::= X \subset Y$ $X \cup \{a\}$ successore (diretto) di X

1) X non è frequente, Y successore di X (X < Y) => Y non è frequente

2) $c(X \rightarrow Y) = s(X \cup Y) / s(X)$

3) $c(X \rightarrow Y \cup Z) \leq c(X \rightarrow Y)$

Insiemi Frequenti

insiemi-frequenti(*I*, *T*, s_{min})

k:=1, $C_k := \cup \{i\}$ $i \in I$, L_k:=pota(C_k,T)
while $L_k \neq \emptyset$
 C_{k+1}:= genera-candidati(L_k)
 L_{k+1}:=pota(C_{k+1},T)
 k :=k+1
return $\cup L_j$ j=1,...,k

pota (C_k,T) verifica per ogni insieme $X \in C_k$, se X é frequente

genera-candidati(L_k) include gli insiemi X con k+1 oggetti, così che ogni sottoinsieme Y di X con k oggetti è elemento di L_k

Problemi

Efficienza:

una transazione contiene pochi oggetti --->
numero d´insiemi frequenti non è esponenziale e cala con k

mà: nessun frequent set con più di ca. 15 oggetti:
se esiste un frequent set con K oggetti, poi ci sono almeno 2^K set

Abbondanza delle regole

migliaia di regole

Ridondanza delle regole

p.e. una regola più speciale ha una confidenza più bassa

Valutazione d´una regola è troppo semplice

confidenza: c_{min} indipendente dalla frequenza della conclusione

Supplementi

Efficienza
minimizzare gli scan (passi) per la banca dati
organizzazione d´insiemi frequenti con alberi di hash (nodi: items)
campionamento, integrazione nella banca dati

Abbondanza delle regole
tassonomie d´oggetti

Ridondanza delle regole
sistema minimale di regole

Valutazione d´una regola è troppo semplice
valutazioni statistiche

Regole di sequenze
insieme ordinato d´oggetti (tempo): web pagine visitate, errori in rete

Clustering

Sono dati *N* vettori di *d* dimensioni (*d* attributi e *N* oggetti)
Trova spartizione *ragionevole* dei *N* esempi in *c* sottoinsiemi

Gli oggetti d´un sottoinsieme (cluster) sono i più simili possibili
e oggetti di differenti cluster sono i più dissimili possibili

Il metodo deve scoprire le classi, raggruppando oggetti simili
(=con valori simili degli attributi) nella stessa classe/cluster

c dato o scoperto dall´algoritmo

Non esistono attributi target

Usato come
- descrizione sommaria
- fondamento per modelli predittivi più accurati

Clustering: Esempio

Ci vuole pochi cluster omogenei e non un gran numero d´oggetti

oggetti: i clienti d´un´impresa
cluster: gruppi omogenei di clienti
azioni di marketing specifici per i gruppi

I gruppi (cluster) non solo vengono definiti di liste d´oggetti appartenenti mà idealmente sono descritti da comprensibili e compatte caratteristiche

gente con figli abitando in campagna
gente anziana con reddito alto
laureati lavorando in servizio pubblico

Clustering: Metodi

- Metodi Numerici

variabili numeriche sono dominanti
distanza, centroido, media, correlazioni ...

- spartizioni
- gerarchie
- componenti di densità (modelli probabilistici)

- Metodi Concettuali

variabili categoriali sono dominanti
descrizioni caratteristiche vengono derivate

Spartizioni

Sia X un insieme d´oggetti e $S \subseteq X$ un insieme d´esempi
dist: $X \times X \rightarrow \Re^+$ una funziona di distanza

q: $2^{2^X} \rightarrow \Re^+$ una funziona di qualità (per un insieme di sottoinsiemi)

Il compito di analisi di clustering:

Trova $C=\{C_1,...,C_k\}$ $C_i \subseteq S$ così che q(C) sia massimo

tipicamente una spartizione: $C_i \cap C_j = \emptyset$ $\cup C_i = S$

q viene definito in base a *dist*

p.e. $q := -\sum d(C_i)$ $d(C_i) := \sum dist(x,y)/ m_i$ $x,y \in C_i$ $m_i := |C_i|(|C_i|-1)/2$
o la distanza media dal centro (centroido)

Spartizioni: Problemi

Come definire la funziona di distanza?

Come pesare gli attributi?

Il numero delle spartizioni è super-esponenziale,
solo una ricerca semplice (greedy search)

Spartizioni: k-Means

k viene dato dall´utente

supposizione: si può definire il centro z d´un insieme C d´oggetti

col minima somma di distanze a tutti gli oggetti di C, non necessariamente $z \in C$

p.e. $z(C) := \sum c / |C| \quad c \in C$ c vettore numerico

k-Means(S,k) (tipicamente con convergenza rapida)

$Z = \{z_1, \ldots z_k\}$ *k* esempi selezzionati casualmente

while qualità diventa migliorata

$C_i := \{s \in S | i = \text{argmin}_{i=1,\ldots,k} \text{ dist}(s, z_i)\} \quad i=1,\ldots,k$

$z_i := z(C_i) \quad i=1,\ldots,k$

end

return $\{C_1, \ldots, C_k\}$

Qualità = (negativa) somma delle medie di distanze ai centri

Problema: convergenza in un minimo locale; diverse iniziazioni

outlier formano cluster singolari

Spartizioni: Componenti algoritmici

Funziona di qualità: come è buona la spartizione?

omogeneità dei cluster / similarità degli oggetti

p.e. errore quadrato (distanza dal centro)

o medio errore

Algoritmo di ricerco: *k* centri dei cluster con ottima qualità

p.e. iterativo:

aggrega oggetti al prossimo centro, calcola nuovo centro

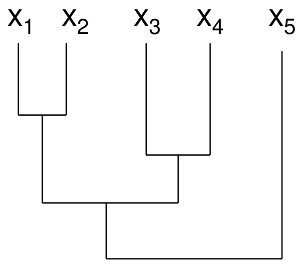
Rappresentazione del cluster: centro del cluster

Gerarchie

Matrice di distanza: NXN paia d´oggetti

Seria di clustering: bottom up, da cluster d´un solo oggetto a un solo cluster

Dendogrammo:



Agglomerazione

I cluster più simili vengono agglomerati ad un nuovo cluster

Distanza di due cluster con un solo oggetto:= distanza dei due oggetti

$\text{dist}(C_1 \cup C_2, C_3) = f(\text{dist}(C_1, C_3), \text{dist}(C_2, C_3))$

f= min: single linkage clustering, distanza basata su prossimi oggetti: catene
f= max: complete linkage clustering cluster compatti
f= average: average linkage clustering

Problema: $O(N^2)$ Dati non voluminosi, p.e. $N < 10^4$

Metodi di spartizioni: quasi $O(N)$

Scalabilità: campionamento

Clustering di Densità

Supposizione
k cluster di oggetti
Distribuzione di S è una combinazione di k componenti distribuzioni (somma lineara, pesata)
p.e distribuzioni multinomiali (normali) di variabili categoriali (continue)
I parametri delle distribuzioni sono sconosciuti (p.e. media, varianza)

Funziona di qualità:
Likelihood: probabilità dei dati osservati dipendente di parametri
Trova parametri, così che la probabilità sia massima
Parametri: parametri delle distribuzioni, pesi, k
Problema d´ottimizzazione nonlineara difficile: EM algoritmo expectation max.

Probabilità per un oggetto di appartenere ad un cluster,
nessun assegnazione ad un solo cluster

Autoclass (Cheeseman)

Densità: Pros e Cons

+ Non richiede una misura di distanza

– Supposizioni su modello probabilistico sono ragionevoli?
    A priori conoscenze sull´applicazione sono necessarie per specificare la forma funzionale delle distribuzioni. Può essere difficile, ma obbliga utente di specificare esplicitamente le supposizioni.

+ Funziona di qualità è una misura naturale
    probabilità dei dati osservati dipendente dei parametri

– Spazio dei parametri è voluminoso

Clustering Concettuale

Aggiungere ad un metodo numerico di clustering un metodo d´apprendimento di concetti, derivando per ogni cluster una descrizione che generalizza i suoi oggetti.
Ma esiste una buona descrizione per ogni cluster numerico?

Metodi concettuali di clustering ricercano cluster in uno spazio di descrizioni
Automatica interpretazione d´un clustering

Michalski, Stepp: Star, Cluster/2

Origine:
apprendimento supervisionato di concetti disgiuntivi da esempi solo positivi

Componenti: rappresentazione, qualità, ricerca

Valutare un Clustering predictability

Categorizzare un oggetto nuovo:
determinare il cluster a cui appartiene l´oggetto
clustering concettuale: verificare la descrizione
(clustering numerico: distanza ai centri del clustering)

Predire i valori degli attributi per un membro di un cluster
p.e. attributi sconosciuti d´un oggetto classificato (valori mancanti)
gerarchie offrono migliori predizioni

trade-off per una valutazzione di un attributo, dato un clustering:
predictability: dato il cluster a cui appartiene oggetto, predire valore mancante per un attributo
predictiveness: determinare cluster con il valore d´un attributo

Valutare clustering: valutare predictability per tutti gli attributi

Clustering Concettuale: Componenti

Rappresentazione

cluster=sottogruppo (paia d´attributo-valore)
cluster probabilistico (Autoclass, Cobweb): distribuzioni probabilist. (valori attr.)

spartizioni (Cluster/2; Michalski, Stepp) / gerarchie/ intersezioni fra cluster overlap

Qualità

predictability vs. predictiveness (Autoclass, Cobweb)
funzioni lexicografiche di valutazzione (LEF lexicographic evaluation functions):
criteri ordinati, granularità (Michalski)

Ricerca

incrementiva e partizionante:
associare un cluster per un altro oggetto, partizionare cluster se necessario
dipende dal collocamento d´oggetti: raffinamento iterativo necessario
STAR: Operatori di specializzazione (p.e. aggiungi selettore) / generalizzazione