Mining di Sottogruppi

Linguaggi per definire sottogruppi

Ricerca in spazi di sottogruppi

Compiti analitici

Funzioni di qualità

Visualizzazione

Controllare ridondanze in set di sottogruppi

Scalabilità

Sottogruppo

Sottogruppo = sottoinsieme di una popolazione
definito con un linguaggio d´interrogazione

Selettore: clausola condizionale d´un attributo A: $A \in V$ $V \subset$ Dominio(A)
forma congiuntiva di selettori
Sesso=maschio & età=18-25

Estensioni: interni disgiuntivi: Nazionalità = austriaca | svizzera
tassonomie: Nazionalità = europea

Interrogazioni multirelazionali:
persona. età=1-6, vicino(persona,impianto-nucleare),
impianto. età =vecchia

Pattern di Sottogruppo

Pattern di sottogruppo
descrive un sottogruppo con una statistica **significativa**

Statistica significativa:
Tasso (percentuale) d´un valore d´una variabile discreta di target
media d´una variabile di target

Tasso di disoccupazione più alto per giovani maschi con bassa educazione

Giovani povere donne sono infettate con AIDS più dei corrispondenti maschi

La mortalità di cancro di polmone è aumentata per le donne durante gli ultimi 10 anni

Obiettivi dell´approccio

Esplorazione e induzione descrittiva
dipendenze fra target e molte variabili esplicative
identificare sottogruppi adatti per azioni

Classificazione
p.e. regole di classificazione

Ottimizzazione
sottogruppi con qualità ottimale

Esplorazione e Induzione Descrittiva

Ricerca trova molti sottogruppi di sovrapposizione

Intersezioni fra sottogruppi descritti da variabili correlate

Cluster di sottogruppi simili
- Profilo con utilità descrittiva
- Descrizioni dalle viste differenti (includendo altri variabili correlati)

Identificare i sottogruppi più rilevanti
- adatti per azioni
- causali (manipolazione del valore d´una variabile influenza target)
- o dependenza fra target e sottogruppo esiste soltanto a causa di multicorrelazioni con altri sottogruppi nascosti?

Linguaggi per Sottogruppi

Selettori

Linguaggi multirelazionali

Grafo di Relazioni

Aggregazioni di relazioni 1-n

Materializzazione adeguata

Linguaggi per Sottogruppi: Granularità

Granularità di selettori A=v (selettività adatta)

dipende dall´applicazione, ma generalmente è meglio evitare:
sottogruppi troppo speciali, ricerca troppo *greedy*

Attributi A con molti valori
- A categoriale: usa tassonomia su dom(A)
- A continuo (o ordinato): discretizza valori, intervalli A=[v,w)
  - globali: derivati in preprocessing
    - intervalli semplici: uguale lunghezza / frequenza
    - criteri d´ottimizzazione per intervalli, p.e. minima varianza o somma di quadrati
    - intervalli definiti con misure di dispersione (media, varianza)
  - locali: derivati durante ricerca raffinando un sottogruppo parentale
    - qualità non è molto meglio, complessità è più alta (intervalli simili, ma differenti)
  - supervisionati: considerando variabile di target

Linguaggi per Sottogruppi: multirelazionali

Banca dati contiene parechie tabelle, p.e.
persone, regioni, ospedali, diagnosi, terapie, impianti industriali

Una tabella primaria viene selezzionata, p.e. persone

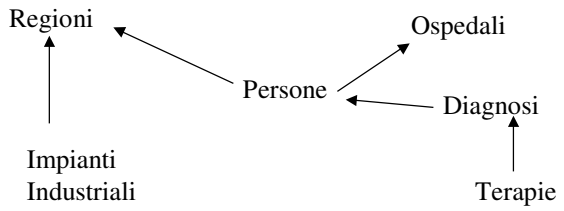
Sottogruppi d´oggetti della tabella primaria vengono definiti con attributi di tutte le tabelle

Attributi delle tabelle secondarie vengono associati agli oggetti della tabella primaria collegando (join) le tabelle e usando funzioni d´aggregazione

persone anziane abitando in regioni industriali con una diagnosi di cancro

join: persone → regioni, persone ← diagnosi, regioni ← impianti industriale
attributi: età(persone), tipo(diagnosi), id(impianti)
aggregazioni: esistenza (count>0)

Grafo di Relazioni



nodo: tabella
arco: relazione fra due classi d´oggetti (tabelle) (sottoinsieme del prodotto)
foreign key (p.e. Terapie→Diagnosi), predicati, predicati spaziali

cammino dall´oggetto primario attraverso oggetti secondari:
persone.età= anziano
persone.età= anziano, diagnosi.tipo=cancro
persone.età= anziano, diagnosi.tipo=cancro, regioni.tipo=industriale
persone.età= anziano, diagnosi.tipo=cancro, regioni.tipo=industriale, terapie.sum(costo)=alto

Aggregazioni

Aggregazioni per relazioni 1 - n

1 persona - n diagnosi

esistenza: la persona ha (almeno) una diagnosi (Count>0)

persone con diagnosi (sottogruppo di persone)
persone anziane con diagnosi di cancro (una delle diagnosi è cancro)

$O=O_1x...O_k$ p.e. O_1= persone O_2= diagnosi O_3= terapie

S_1= sottoinsieme di O definito da selettori, p.e. età=anziano, tipo=cancro, costo=alto

S_2= sottoinsieme di S_1 definito da prodotto delle relazione (2 foreign keys D->P,T->D)

proiezione di S_2 su O_1

Funzioni d´aggregazione: count, sum, max, min, avg
sum(Terapie.costo)>100.000 min(Diagnosi.tempo)>2 tutti conosciuti da 3+ anni

Variabili distinguendo oggetti: persone con una diagnosi di cancro e una diagnosi di cardio e la diagnosi di cancro conosciuta prima (ILP)

Sottogruppi Multirelazionali e preprocessing

Joining parecchie tabelle costa tempo

Preprocessing, poi usare il join materializzato in algoritmi di data mining

Grafo di relazioni con le condizioni di aggregazione come background knowledge per restringere le possibilità illimitate di sottogruppi multirelazionali

Amministrazione automatica adatta del preprocessing, usando il grafo di relazioni

Specializzazione: relazioni spaziali (dati spaziali)

Metodi di Ricerca

Spazi di sottogruppi

Euristica strategia di ricerca

Ricerca completa

Vantaggi di beam search

Ricerca in Spazi di Sottogruppi

Sottogruppi sono parzialmente ordinati (generalità delle descrizioni)
sesso=maschio più generale di *sesso=maschio & età=giovane*

Livello k dello spazio: sottogruppi con k selettori

Aggiungere un selettore congiuntivo genera sottogruppo più speciale
tassonomie, interni disgiunti: altri operatori di raffinamento/espansione
monotonia d´operatori, problema: funzioni d´aggregazione (multirelazionali)

Greedyness d´un operatore: supporto di sottogruppi cala troppo presto
qualità si aggrava, poi: tassonomie, non-greedy funzioni di qualità

Strategie di Ricerca

Strategia Iterativa, Generale a Speciale, di Generare e Verificare

Iterazione k: un sottoinsieme d´ipotesi su livello k-1 viene raffinato
operatori di raffinamento generano ipotesi più speciali
alternative di selezionare il sottoinsieme da raffinare
i successori generati vengono valutati

Gli spazi sono molto voluminosi (molti attributi, tabelle)

Strategie euristiche, p.e. beam esteso:

includi w_1 migliori sottogruppi di livello k in risultati
cluster w_2 migliori sottogruppi di livello k
seleziona sottogruppo migliore di ogni cluster e ordina questi
raffina w_3 migliori sottogruppi

Funzioni di qualità per ordinare i sottogruppi

Vincoli (constraints): Minimo supporto, qualità, ...

Strategie di Ricerca Completa

Best-n trova i n migliori sottogruppi

Voluminosi spazi di ricerca:

- Vincoli (constraints): minimo supporto
- feature selection (selezionare attributi importanti in preprocessing)

Stimati ottimistici potare sotto-spazi con qualità inferiore all´attuale ennesima

Soddisfazione di vincolo tutti i sottogruppi osservando i vincoli

Spartizioni sottogruppi disgiunti

Ricerca in spazi di parechi dimensioni

- sottogruppi di target (p.e. gerachie di prodotti, malattie)
- sottogruppi di segmenti (p.e. gerarchia di regioni)
- sottogruppi di variabili descrittive (indipendenti, di spiegazione)

Vantaggi di beam search

Forte riduzione del spazio di ricerca

recall alto: non molti sottogruppi di alta qualità mancano

- p.e. due selettori: manca solo se tutti i due non sono buoni

Clustering evita ricerca troppo stretta

Scalabilità alta

Tipi di Compiti d´Analisi

Compito analitico

Singola Popolazione

Confrontare parechie Popolazioni

Compiti analitici

Dimensioni per un raggruppamento di compiti analitici

Numero (n) di popolazioni da confrontare

tipicalmente n=1: tutti gli oggetti appartengono ad una popolazione

n=2, p.e. sani e ammalati: le differenze fra le popolazioni vengono studiate

n punti di tempo: cambiamenti fra 2 o tendenze in una sequenza delle popolazioni

Variabile di target selezionata?

si: tipo della variabile di target binario, categorico, ordinale, continuo

no: p.e. insiemi frequenti

Una singola Popolazione

Tipo della variabile di target è:

binario: tipico pattern di regola

categorico:

binarizzare implicitamente selezionando un valore di target e confrontandolo contro gli altri sottogruppi con un´alto tasso di oggetti target = sottogruppi dipendenti del gruppo target

sottogruppi con una distribuzione differente della variabile di target

ordinale / continuo: una misura sommaria, p. e. media è studiata

sottogruppi con un´alta media della variabile di target

nessuna variabile di target selezionata:

derivazione di tutti i frequenti insiemi

Parecchie Popolazioni

Capire le differenze fra parecchie popolazioni di contrapposizione

non molto studiato in data mining

non coinvolga una variabile di target (mà: popolazioni → targets)

Mining contrast sets (Bay, Pazzani)

sottogruppi che differiscono significativamente nelle loro distribuzioni attraverso le popolazioni

per almeno 2 delle n popolazioni le differenze del supporto del sottogruppo sono significative

Emerging patterns (Li, Dong, Wong)

sottogruppi con i supporti che aumentano da una popolazione ad un altro (2 populazioni)

emerging patterns con supporto zero in una popolazione (Bailey)

emerging trends (n punti di tempo)

applicato ai dati genetici

con una variabile di target

sottogruppi per cui la distribuzione della variabile di target differisce significativamente attraverso le popolazioni

Qualità di Ipotesi

Funzioni di qualità

Esempio: qualità di una regola

Criteri fondamentali

Fasci (famiglie) di Funzioni

Iso-Linea

Assiomi di monotonia

Symmetrie

Equivalenze

Proprietà di Potamento e Campionamento

Funzioni di Qualità

Misurare come un ipotesi (p.e. un sottogruppo) è interessante

specialmente la **significatività statistica**

misure statistiche dipendano dal compito analitico
test statistico, ipotesi nulla: sottogruppo non è interessante proprietà d´una distribuzione
i dati per il sottogruppo rifiutano ipotesi nulla

p.e. tasso di target è significativamente differente in un sottogruppo S

p := tasso di target in S

p_o:= tasso di target in Popolazione P

n := |S| N := |P|

$$\frac{p - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}\sqrt{\frac{N}{N-n}}$$

ipotesi nulla: tasso di target in S = tasso di target in P\S

Esempio di una Qualità: Q_1

$$Q_1 := \frac{p - p_0}{\sqrt{p_0(1-p_0)}} \sqrt{n} \sqrt{\frac{N}{N-n}}$$

$p-p_0$

p_0: $1/\sqrt{(p_0(1-p_0))}$

n: $\sqrt{n}$

n/N: $\sqrt{(N/(N-n))}$

Q_1 unisce 4 criteri

Simmetrica (S -> P \ S)

uguale con - chi-2 test di dipendenza fra S e T(gruppo di target)
- coefficiento di correlazione variabili binari S, T

in caso di una continua variabile di target: media m, m_o e varianza (p, p_o, $p_o(1-p_o)$)

Criteri fondamentali / Fasci di Funzioni

Criteri fondamentali

p.e. $p-p_0$, n, p_0, n/N per una regola S -> T

trade-offs fra criteri: non possibile ottimizzare tutti, p.e. fra *n* e $p-p_0$

possibilità varie di combinare criteri fondamentali

p.e. Q_2=TP/(FP+g) TP=veri positivi, FP=falsi positivi

Q_3=(n/N) ($p-p_0$)

per confrontare varie funzioni: trasformare e usare gli stessi criteri:

Q_2=p/(1-p+g/n)

Fasci di Funzioni tengono un parametro, come g in Q_2

Il trade-off (fra generalità n/N e precisione p) può essere aggiustato con questo parametro

Iso-Linea

Iso-linea in spazio di due dimensioni p, n/N
linea dei punti (p, n/N) con la stessa qualità

comparando le iso-linee di Q_1 con quelli di Q_2, Q_3 -->
Q_2 favorisce grande p
Q_3 favorisce grande n

ROC spazio di due dimensioni TP-rate, FP-rate
area sotto l´involucro convesso
ROC per validare un classificatore: area è adeguata
ROC per validare un insieme di sottogruppi

Assiomi di Monotonia

p.e. Q = Qualità di regole

Q sale con p, se n e p_0 sono fissi
Q sale con n, se p e p_0 sono fissi
Q scende con n, se pn e p_0 sono fissi
Q scende con p_0, se p e n sono fissi

Q_1, Q_2, Q_3 osservano questi assiomi

Symmetrie

p.e. Q = Qualità di una regola $S \rightarrow T$

Simmetrica relativamente a

S e P\S (sottogruppo complementare)

$Q(S \rightarrow T) = Q(P\backslash S \rightarrow T)$

$Q(S \rightarrow T) = - Q(P\backslash S \rightarrow T)$ antisimmetrica

S e T : $Q(S \rightarrow T) = Q(T \rightarrow S)$

Q_1 è antisimmetrica relativamente a S e P\S

Q_1 è simmetrica relativamente a S e T

Q_2 e Q_3 non sono simmetriche

dipende dall´applicazione se simmetrie sono importanti

Equivalenze

Approcci statistici differenti possono fornire la stessa funzione come per regole $S \rightarrow T$ e le variabili binari associate a S e T

Funzioni possono anche essere equivalenti

p.e. equivalente relativamente all´ordinamento

Q_1 eq-o Q_2: $Q_1(S_1 \rightarrow T) > Q_1(S_2 \rightarrow T) \Leftrightarrow$
$Q_2(S_1 \rightarrow T) > Q_2(S_2 \rightarrow T)$

Q_1 e Q_2 generano lo stesso ordinamento

importante per beam e best-n ricerca

Proprietà di Potamento e Campionamento

Potare sotto-spazi di ipotesi che non possono contenere una soluzione è importante per l´efficienza della ricerca

p.e. stimati ottimistici della qualità per best-n ricerca
limiti superiori per funzioni di qualità in sotto-spazi
sotto-spazio di tutti i successori di un nodo

Usare un campionamento d´una populazione molto grande può accelerare un algoritmo
Quanti oggetti deve contenere un campionamento, affinchè una qualità calcolata col campionamento sia quasi così buona come questa calcolata con tutta la popolazione?

Tavole di Contingenza

Tab 1. Contingency table for target *ill=yes* and subgroup *unemployed=yes*.

Parent subgroup, e.g. gender=male and age=[40,55)		**Target**		
		ill=yes	¬ ill=yes	
Subgroup refinement	unemployed=yes	$\mathbf{n_{11}}$	n_{12}	$\mathbf{n_{1+}}$
	¬unemployed=yes	n_{21}	n_{22}	n_{2+}
		$\mathbf{n_{+1}}$	n_{+2}	**N**

Per calcolare la qualità di un sottogruppo si usa una tavola di contingenza.
n_{11} , n_{1+} vengono (efficientemente) calcolati per ogni raffinamento d´un sottogruppo parentale (aggiungendo tutti i selettori possibili)
Target continuo: media del target nelle celle
Parechie popolazione: terza dimensione con n popolazioni

Visualizzazione

Varie giunte fenestre per i sottogruppi trovati

Tabella di sottogruppi (definizione, qualità, criteri elementari)

Grafica (plot) di sottogruppi (p, n/N spazio; ROC spazio; iso-linee)

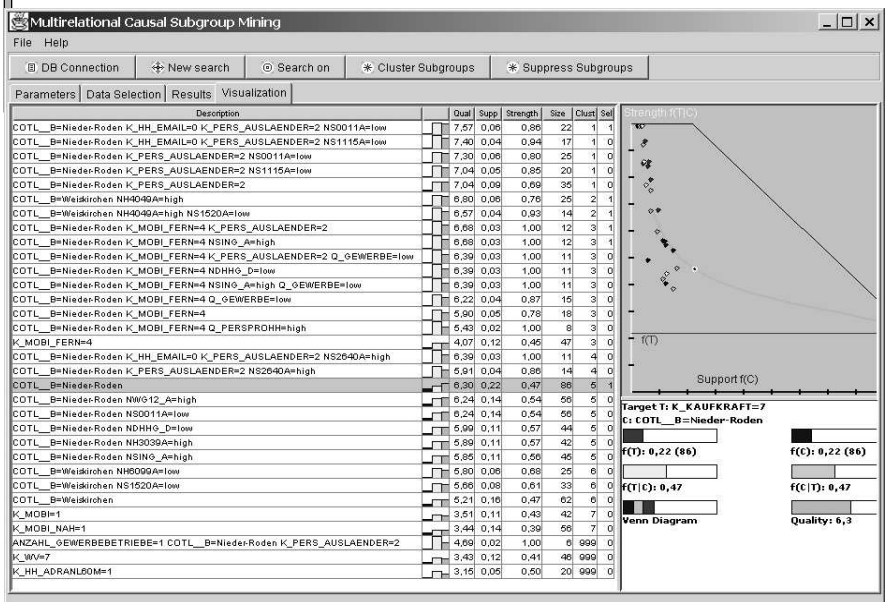
Visualizzazione di un singolo (selezionato) sottogruppo

Struttura di clustering dei sottogruppi

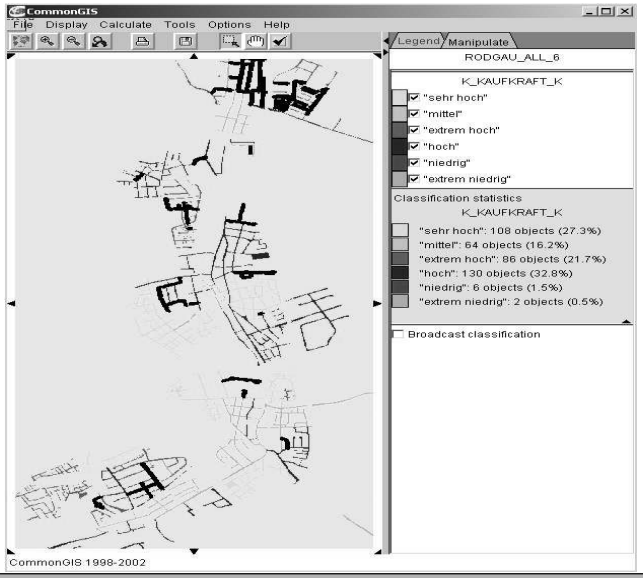
Struttura causale dei sottogruppi

Dati spaziali: mappa geografica

Visualizzazione: Esempio



Visualizzazione: GIS mappa



Controllare Ridondanze in Set di Ipotesi

Multicorrelazioni fra variabili indipendenti

Raggruppare ipotesi verificate

Clustering di ipotesi verificate

Clustering vs. Selezionare

Azioni su ipotesi verificate

Metodi causali

PC algoritmo

Obiettivi di analisi causali

Tre obiettivi più o meno fastidiosi

Multicorrelazioni fra Sottogruppi

Tipicamente 20-200 sottogruppi trovati in ricerca

Intersezioni fra questi sottogruppi

Variabili descrittive sono correlate

Ridurre la complessità dei risultati
cluster di sottogruppi simili, ordinare, sopprimere ridondanti, selezionare causali sottogruppi

Multicorrelazioni, var. nascoste (variabili rilevanti mancano), proxy var.
Tipici problemi per analisi di dipendenze

Raggruppare Risultati

Ordinare
qualità e misure diverse, livello di ricerca, selettori

Clustering

similarità(S_1,S_2):= $|S_1 \cap S_2| \; / \; |S_1 \cup S_2|$

complete linkage: $sim(C_1,C_2)$= min similarità(S_1,S_2) $S_i \in C_i$

Clustering di Sottogruppi: Specialità

Oggetti da cluster sono sottoinsiemi (sottogruppi)

Cluster è un insieme di sottoinsiemi

Relazioni speciali fra cluster (C1 < C2 "subset", disgiunti(C1,C2))

Descrizione di un cluster (derivata dalle descrizioni dei sottogruppi)

Omogeneità di un cluster ($|\cap S_i| \ / \ |\cup S_i|$) , eterogeneità di due cluster

Qualità di un cluster (estensione, kernel, sottogruppo rappresentativa)

Clustering vs. Selezionare

Clustering offre profili:
- sottogruppi simili / correlati in un cluster
- utente vide correlazioni fra selettori e può selezionare fra alternative descrizioni

Automaticamente selezionare i sottogruppi rilevanti
- trovare un sistema di sottogruppi che sono
  - i più disgiunti possibili
  - con una estensione combinata la più ampia possibile
  - con una qualità combinata la più alta possibile
    - p.e. copertura degli oggetti target
- trade-offs fra questi obiettivi

Selezione euristica (boosting-like): iterazione dei passi
- seleziona ottimo sottogruppo (con un specifico criterio di qualità)
- manipola pesi degli oggetti gia coperti

Capacità per Azioni

Sottogruppi che possono essere usati per azioni

Una dipendenza esiste fra un sottogruppo e un gruppo di target

Per azioni utili è spesso necessario che il sottogruppo è causale

azione: mandare un´offerta

dipendenza fra sottogruppo (che contiene molti oggetti di target) e il gruppo di target (clienti) è sufficiente per raggiungere molti potenziali clienti

azione: cambiare una proprietà di un oggetto, p.e. bere moderatamente vino

causalita fra sottogruppo (bere vino) e target (nessun infarto) è necessario

una terapia (variabile, sottogruppo) è causale per sopravivere (variabile di target), se la probabilità di sopravivere per un paziente diventa maggiore con questa terapia (esperimento ideale)

Metodi causali dalla Statistica: dati esperimentali controllati secondo piano

Metodi possono anche essere usati per dati d´osservazioni?

Metodi Causali

- Metodi Bayesiani (p.e. Heckermann, Pearl)
- Metodi di vincoli
  - da certi sistemi di dipendenze e indipendenze delle variabili
  - si può restringere possibili relazioni causali
  - usati in data mining (p.e. Glymour, Cooper, Silverstein)

Testing dependence and (conditional) independence of subgroups

Independent(S1,S2)	::=	Chi-2(S1,S2) < c1 (e.g. c1=1)
Dependent(S1,S2)	::=	Chi-2(S1,S2) > c2 (e.g. c2=9)
CondInd(S1,S2|S3)	::=	Chi-2(S1,S2|S3=0) + Chi-2(S1,S2|S3=1) < 2c1

PC Algoritmo

- Spirtes, Glymour
- inferenza di una rete causale da sistemi di (in)dipendenze
- sottogruppo visto da variabile binaria
- passo centrale:

Let:
S1, S2, S3 be groups (subgroups or target group) with: Dep(S1,S2), Dep(S2,S3), Indep(S1,S3).

Then:
CondInd(S1,S3|S2) <=> S1 →S2 →S3 or S3 →S2 →S1 or S1 ← S2 →S3

Not CondInd(S1,S3|S2) <=> S1 →S2 ← S3

Obiettivi di Analisi Causali

- Opprimere qualche sottogruppo non-causale
- Identificare qualche sottogruppo causale
- Derivare una reta dei sottogruppi causali per il gruppo di target

I sottogruppi vengono trovati con un algoritmo di data minig
- L´analisi causale è aggiunta (postprocessing)
- L´analisi causale è integrata nella ricerca di sottogruppi

Opprimere qualche sottogruppo non-causale

Opprime (S_1,S_2) := CondInd$(S_2,T|S_1)$ S_1,S_2 sottogruppi, T target gruppo

indipendenza condizionale è una condizione sufficiente per opprimere questa condizione è ragionevole:

tasso del valore di target (gruppo di target T) in S_2 =: $t(S_2)$ $(> t(P))$
P popolazione

$t(S_2)$ può essere spiegata dal tasso alto in $S_2 \cap S_1$:
$t(S_2 \cap S_1) = t(S_1) > t(P)$ $t(S_2 \backslash S_1) = t(P \backslash S_1) < t(P)$

"=" significa: non signifivativamente differente (test statistico)

Identificare qualche sottogruppo causale

PC algoritmo (Spirtes et al.) ---->
CCC regola (Cooper) [anche per variabili nascoste]
tutti le tre variabili sono correlate
CCU regola (Silverstein) [nessun variabile nascosta]
un paio delle tre variabili non è correlato

Input:
target group T and set of subgroups **S**={S_i} which are dependent of T
subset **W** of subgroups $S_0 \in$ **S** that are known not to be caused by any other subgroup or by T
Output: a list L of possible causal relationships between S_i and T.

For all $S_0 \in$ **W**:
For all $S_i \in$ **S** ($S_i \neq S_0$):
If Dependent(S_0,S_i) and Independent(S_0,T|S_i) then include S_i into L

Identificare una Rete Causale

1) ricerca di sottogruppi
2) rete probabilistica per i sottogruppi

- usa solo i sottogruppi gia identificati come sottogruppi dipendenti
- ma usa solo i variabili binari, però con congiunti
- non usa i variabili categoriali (come in reti probabilistiche puri)

p.e. per classificazioni: usa solo i parenti diretti di T

problema: algoritmo non è robusto
attendibilità dei test: un test falso implica allineamenti falsi che si propagano
test falsi: indipendenze false (chi-2 < 1)

Scalabilità di Mining di Sottogruppi

Metodi Scalabili

Integrazione efficiente nella base di dati

Approcci di Campionamento

Calcolare statistiche sufficienti nella base di dati

SQL e grandi sistemi delle tabulazioni

Strategia per costruire efficienti SQL-Interrogazioni

Tempi di interrogazione

Interrogazioni multirelazionali

Quando campioni sono utili?

Campionamento sequenziale

Possibilità per la Scalabilità di un Algoritmo

Molti variabili (>100), moltissimo oggetti (> 1 milione),
varie tabelle:
che fare per ricevere risultati in tempo?

- Integrazione efficiente nella base di dati
- Campionamento degli oggetti di analisi
- Metodi paralleli
- Metodi con risoltati in qualunque momento

Integrazione efficiente nella base di dati

Moltissimi ipotesi da validare

Ogni ipotesi ha bisogno di aggregazioni di parecchie variabili per una selezione dei tuples (frequenze, medie, varianze)

Pochi passi (scans) della base di dati devono essere organizzate efficientemente fornendo le informazioni sufficienti per valutare molte (tutte le) ipotesi (gli aggregazioni per tutte le ipotesi)

Integrare un algoritmo di ricerca nella base di dati:

- algoritmo di ricerca completo corre nella base di dati
- algoritmo interroga iterativamente la base di dati (p.e. usando SQL) per fornire le statistiche sufficienti per le ipotesi del seguente set di ipotesi da validare

Approcci di Campionamento

Usano solo una parte casuale della base di dati

p.e. trovare le n migliori ipotesi
approssimando le qualità delle ipotesi sufficientamente esatte
e iterativamente usando una sequenza di parti casuali

le qualità sono aggiornate in ogni iterazione con i dati supplementari dalla parte casuale seguente

necessario soltanto per il numero in diminuzione di ipotesi per cui le approssimazioni precedenti non sono ancora abbastanza esatte per includerle o escluderele fra le n migliori

Approcci di campionamento possono essere combinati con metodi di integrazione

Integrazione efficiente nella base di dati: statistiche sufficienti

Architettura di due strati
calcolare i sistemi di tavole di contigenze in base di dati (SQL)
algoritmo di ricerca in cliente (JAVA)

una singola interrogazione fornisce gli aggregazioni sufficienti
per valutare tutti i sottogruppi di successione (raffinando una ipotesi)

Un passo per una interrogazione
Algoritmo di ricerca riceve solo dati aggregati
SQL: portabilità, dati distribuiti, approcci paralleli

SQL e grandi sistemi delle tabulazioni con medie, frequenze

Compito di calcolo per il caso di una singola relazione

$\{A_i$; $i=1, ..., k\}$ set di attributi

T target attributo (categoriale / continuo)

$\{P_j$; $j=1, ..., w\}$ set di sottogruppi parentali (su ogni livello)

Compito di calcolo :

$\{A_i$ X T | P_j ; $i=1,...,k; j= 1,...,w\}$ sistema di tabulazioni (frequenze o medie per continuo T)

Opzioni di SQL per sistemi delle tabulazioni

Group-by Option (i,j)

...union select count(*) from table where P_j group by A_i ,T
union select count(*) from table where P_j group by A_{i+1} ,T...

resp. (continuous T):
...union select sum(T) from table where P_j group by A_i ...

Case Option (i,j,l)

...,sum(case when $A_i=a_{il}$ then 1 else 0), sum(case when $A_i=a_{il}$ and T=t then 1 else 0), sum(case when $A_i=a_{il+1}$ then 1 else 0),sum(case when $A_i=a_{il+1}$ and T=t then 1 else 0), ...
from table where P_j

resp. ...,sum(case when $A_i=a_{il}$ then 1 else 0), sum(case when $A_i=a_{il}$ then T else 0) ...

Strategia per costruire SQL-Interrogazioni

Group-by Opzione è preferibile:

- in core databases (piccolo, formato medio): nessuna pena per passi multipli
- attributi con molti valori (p.e. > 10): struttura del sistema
- nessun limitazioni per la lunghezza di interrogazioni

Case Opzione è preferibile :

- grandi basi di dati: soltanto un passo
- Livelli di ricerca > 0: stessa where condizione non ripetuto
- manipolazioni dei valori di attributi possibili: intervalli (discretizzazione), sottostringhe (gerarchie)

Esempi di Tempi di Interrogazione

Oracle su PC normale

beam 2 livelli, ampiezza=10

100.000 esempi

100 attributi binari

group-by: 2 minuti case: 1 minuto

20 attributi categoriali 20 valori per attributo

group-by: 3 secondi case: 3 minuti

2.000.000 esempi

100 attributi binari

group-by: 4 ore case: 20 minuti

Interrogazioni Multirelazionali

Selezionare sottogruppo parentale e generare tutti i successori

sottogruppo parentale

singola relazione: semplice congiunzione di selettori (SQL: where)

parechie relazioni: join di relazioni con aggregazioni

```
select ED.T,ROAD.A  from
        (select distinct ED.KEY,ED.T,ROAD.A   from  ED,ROAD
         where  MDSYS.SDO_FILTER(ED.GEOM,ROAD.GEOM,´mask=anyinteract´...))
group by ED.T, ROAD.A

1000 EDs, 4000 A_ROADs :    1 sec
1000 EDs, 4000 A_ROADs, 5000 B_ROADs:    14 secs  (joining 3 tables)
```

Espressione involuta di select

proiettare: attributo di chiave della tabella primaria, attr. di target, attr. di espansione (aggregare) e tabulazione incrociata

Fattori per Tempi di Ricerca: Quando Campioni sono Utili?

Fattori primari:

N (esempi)

numero di attributi e loro valori discretizzati: S selettori

Grandezza dello spazio di ipotesi: S, strategia di ricerca

beam: lineare in S

per un fisso S: tempo lineare in N fino a $N < N_{tecnologia}$

p.e. PC, Oracle di standard, S<300: N< 100.000

tempo di ricerca: pochi minuti

applicazioni interattive possibili

N> 100.000 approcci supplementari necessari:

Campionamento: $N_{campione}$<100.000

Strategie di *best-n* e Sequenze di Campioni

Compito: Trovare i *n* migliori ipotesi (p.e. sottogruppi)
anche per ricerca di beam: i *w* migliori su ogni livello

usa un primo campione di 10.000 o 50.000 oggetti
Solo per pochi ipotesi si non può ancora decidere se queste entrano nelle *n* migliori (o non entrano)

Per le ipotesi indecise usa un secondo campione

Tabulalazioni incrociati possono essere calcolati cumulativi (incrementalmente)
Solo poche iterazioni sono necessarie

Campionamento Sequenziale

Campioni danno garanzie stocatiche, p.e.

che le *n* migliori ipotesi vengono identificate
con una confidenza di (p.e.) 95%: non per tutte le ricerche
confidenza più alta: un più grande campione necessario

Approccio generale: Campionamento Sequenziale Wald, Domingo
precisione 1-ε, ε= errore di qualità paragonando q_{vera} e $q_{campione}$
confidenza δ:
iterativamente determinare la grandezza del campione cumulato (iterazioni) per l´approssimazione (con un errore < ε e una confidenza > δ)

applicato anche per mining di sottogruppi (Scheffer e Wrobel)

Approssimativo best-n mining di sottogruppi

Utente specifica ε e δ (p.e. ε=0.05, δ=0.95)

trova *n* ipotesi (sottogruppi) così che con confidenza δ non esiste un´ ipotesi altra con qualità più alta di $q_{min} + \varepsilon$

q_{min} è la qualità più bassa delle *n* ipotesi

Sia q una funzione di qualità, δ la confidenza e un campione con *m* oggetti

Una funzione $E(m,\delta)$ che limita $|q-q_m|$ stocasticamente viene derivato per q: $P(|q-q_m| < E(m,\delta)) < \delta$

E tipicamente contiene $\sqrt{m}$ nel denominatore, così $E<\varepsilon$ se m abbastanza grande

esperimenti mostrano: cumulati campionamenti di 10.000 oggetti sono spesso sufficienti : m-cumulato viene determinato dai dati (iterazioni necessarie)

(m molto più basso di m da campioni tradizionali non sequenziali)

Conclusioni

Componenti per mining di sottogruppi possono essere adattati per altri approcci

Applicazioni multirelazionali hanno bisogno di una mistura di materializzazione di preprocessing e join dinamici

Condizione di esistenza può essere supplementata da funzioni di aggregazione (altra di count>0)

Strategie euristiche sono necessarie per voluminosi spazi di ipotesi (beam search con clustering integrato)

Compiti analitici conosciuti in analisi di dati, p.e. contrast analysis, sono trattabili da compiti di mining di sottogruppi

Metodologia delle funzioni di qualità per selezionare funzioni adeguate

Visualizzazione con fenestre giunte. Grafiche interattive con ROC curve e GIS mappe

Clustering e analisi causali generano risultati per progettare azioni

Combinazione di integrazione nella base di dati e campionamento sequenziale facilita scalabilità