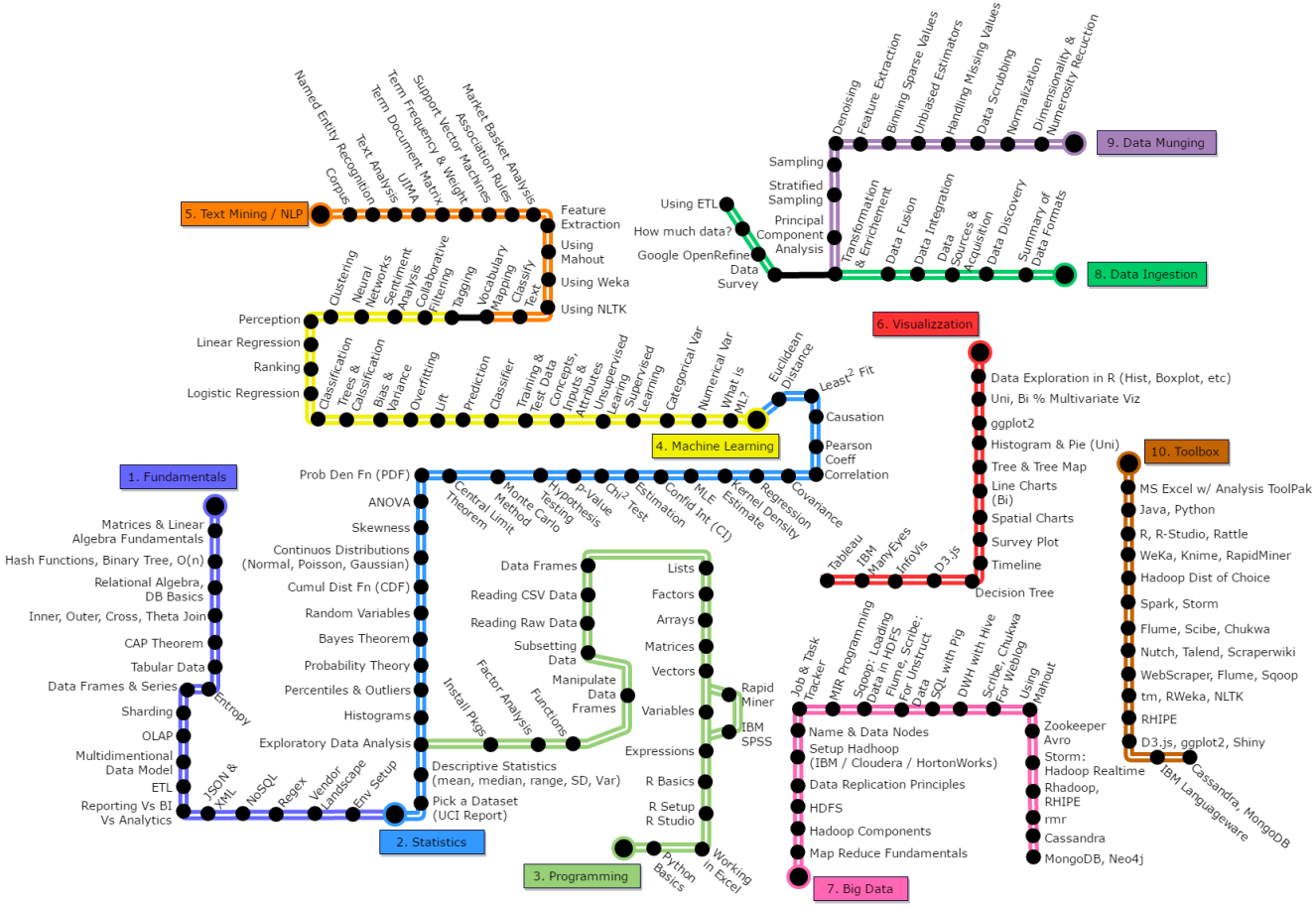


Fondamenti di Data Science e Machine Learning

Dipendenze Funzionali Rilassate

Prof. Giuseppe Polese, aa 2024-25

Outline

- Dipendenze Funzionali
- Dipendenze Funzionali Rilassate
- Criteri di rilassamento
- Definizione generale
- Proposte in letteratura
  - Dipendenze funzionali che rilassano sull'extent
  - Dipendenze funzionali che rilassano sul confronto
  - Dipendenze funzionali ibride
- Open Challenges

Dipendenze Funzionali

- Una dipendenza funzionale (FD – Functional Dependency) tra due insiemi di attributi $X,Y \subseteq R$ specifica un *vincolo* sulle tuple che possono formare uno stato di relazione r di R
  - indicata con $X \rightarrow Y$
- Il vincolo è che per ogni coppia di tuple t_1 e t_2 in r per le quali vale che

$$t_1[X] = t_2[X]$$

allora deve valere anche

$$t_1[Y] = t_2[Y]$$

Dipendenze Funzionali

- Ciò significa che i valori della componente *Y* di una tupla in *r* dipendono da, o sono *determinati da*, i valori della componente *X*
- In alternativa, che i valori della componente *X* di una tupla *determinano* univocamente (o *funzionalmente*) i valori della componente *Y*
- L'insieme di attributi di *X* è detto *parte sinistra (LHS – Left Hand Side)* della dipendenza funzionale (FD), e *Y* è detto *parte destra (RHS – Right Hand Side)*

Dipendenze Funzionali: Esempio

- Esempio: consideriamo la seguente istanza *r* di dello schema R(CAP, Città, Indirizzo)

CAP	Città	Indirizzo
84123	Salerno	Piazza Mazzini
84123	Salerno	Piazza Mazzini
84123	Salerno	via G.Vicinanza
84126	Salerno	Piazza Vicinanza

- su questa istanza di DB vale la seguente dipendenza funzionale

$$\{\text{Città}, \text{Indirizzo}\} \rightarrow \text{CAP}$$

Dipendenze Funzionali

- Recentemente, c'è stato un rinnovato interesse verso le dipendenze funzionali, dovuto alla possibilità di poterle utilizzare in diverse operazioni avanzate sui database
  - Data cleaning
  - Query relaxation
  - Record matching
- Tuttavia, c'è stata la necessità di rilassare alcuni vincoli della definizione di dipendenza funzionale canonica

Dipendenze Funzionali: Esempio

- Esempio: Riguardo lo schema visto nel precedente esempio
  - sia la città che l'indirizzo potrebbero essere memorizzati utilizzando diverse abbreviazioni

CAP	Città	Indirizzo
84123	Salerno	Piazza G. Mazzini
84123	SA	P. Mazzini
84123	Salerno	via G. Vicinanza
84126	SA	Piazza Vicinanza

Dipendenze Funzionali: Esempio

- Esempio: Riguardo lo schema visto nel precedente esempio
  - sia la città che l'indirizzo potrebbero essere memorizzati utilizzando diverse abbreviazioni

CAP	Città	Indirizzo
84123	Salerno	Piazza G. Mazzini
84123	SA	P. Mazzini
84123	Salerno	via G.Vicinanza
84126	SA	Piazza Vicinanza

$$\{\text{Città}_{\approx}, \text{Indirizzo}_{\approx}\} \rightarrow \text{CAP}_{EQ}$$

Dipendenze Funzionali Rilassate

- In letteratura sono state introdotte nuove definizioni di dipendenza funzionale: dipendenze funzionali rilassate (RFD)
  - sebbene si voglia mantenere il concetto di vincolo tra i dati, è importante ammettere eccezioni o condizioni che restringono l'insieme delle tuple per cui deve valere una dipendenza, o catturare vincoli esistenti sfruttando la similitudine e non l'esatta uguaglianza tra i valori degli attributi
- Inoltre, è importante estendere il concetto di dipendenza al fine di mettere in corrispondenza non più soltanto attributi di una stessa relazione, ma anche attributi di schemi differenti

Criteri di rilassamento

- Prima di andare a fornire la definizione formale di dipendenza funzionale rilassata è importante capire i criteri di rilassamento sui quali si basano le nuove definizioni di FD
- Esistono due principali criteri di rilassamento
  - Confronto tra i valori degli attributi (o attribute comparison)
  - Grado di soddisfacibilità (o extent)

Attribute comparison

- Si vuole ammettere la possibilità che una RFD possa essere soddisfatta considerando anche coppie di tuple con valori simili (il grado di similitudine da rispettare può essere definito anche sui singoli attributi)
  - Rilassamento sull'attribute comparison
- In genere, si applica questo tipo di rilassamento in una delle due seguenti modalità
  - Usando un confronto approssimato (approximate match)
  - Usando un criterio di ordinamento
- Esse possono essere generalizzate utilizzando il concetto di vincolo

Vincolo

- Un vincolo è definito come una restrizione su alcuni valori
- Nel contesto delle RFD un vincolo
  - esprime la "vicinanza" di due valori in uno specifico dominio
  - è rappresentato da una funzione ϕ
- Dati due attributi A e B su un dominio D

$$\phi(A,B)$$

- valuta la similarità/distanza di A e B a seguito di una possibile modifica dei loro valori

Vincolo

- ϕ può essere definita in termini di una metrica di similarità $\approx$
  - Edit distance
  - Jaro distance, ecc.
- in modo tale che

$$a \approx b$$

- è vero se a e b sono abbastanza "vicini" rispetto ad una soglia pre-definita
- oppure, ϕ può essere definita in termini di operatore di matching $\rightleftharpoons$
  - Il quale, prima modifica i valori a e b e poi confronta i valori derivati attraverso l'operatore di uguaglianza

Extent

- Si vuole ammettere la possibilità che una RFD possa essere soddisfatta anche per un sottoinsieme di tuple di un'istanza di database (non necessariamente per tutte)
  - Rilassamento sull'extent
- In genere, si applica questo tipo di rilassamento in una delle due seguenti modalità
  - Utilizzando una misura di copertura (coverage measure)
  - Utilizzando una condizione

Coverage Measure

- Una misura di copertura (coverage measure) Ψ definita su una RFD φ quantifica il grado di soddisfacibilità di φ su un'istanza di relazione r
  - Può essere definito come una funzione Ψ
- Dati due insiemi di attributi X e Y che rappresentano il LHS e il RHS di una una RFD φ, allora

$$\Psi : \mathrm{dom}(X) \times \mathrm{dom}(Y) \rightarrow \Re$$

- misura l'ammontare di tuple in r che soddisfano o violano φ
  - g3 error
  - Confidenza, ecc.

Confidenza

- La misura di confidenza valuta il massimo numero di tuple

$$r_1 \subseteq r$$

- tale che è φ valida per r_1
- Esempio: $\{\text{Città}, \text{Indirizzo}\} \rightarrow \text{CAP}$

CAP	Città	Indirizzo
84123	Salerno	Piazza G. Mazzini
84123	SA	P. Mazzini
84123	Salerno	via G.Vicinanza
84126	SA	Piazza Vicinanza

- A questa dipendenza funzionale può essere associata una confidenza pari a 3 tuple su 4 tuple totali

Condizione

- Un'insieme di condizioni c definito su una RFD φ permette di determinare il sottoinsieme di tuple per cui la dipendenza è valida
  - Il dominio di applicabilità può essere così definito

$$D_c = \left\{ t \in D \mid \bigwedge_{i=1}^{k} c_i(t[A_i]) \right\}$$

  - dove ogni c_i rappresenta un predicato su $D(A_i)$ che filtra le tuple su cui si applica φ

Definizioni Generale di RFD

- Formalmente, dati:
  - uno schema di database R definito su un insieme fissato di attributi
  - due schemi di relazione di R
    - $R_1 = \{A_1, \ldots, A_k\}$, e
    - $R_2 = \{B_1, \ldots, B_m\}$

- una RFD φ definita su R è denotata con

$$\mathbf{D}_{c_1} \times \mathbf{D}_{c_2} : (\mathbf{X}_1, \mathbf{X}_2)_{\Phi_1} \xrightarrow{\Psi \geq \epsilon} (\mathbf{Y}_1, \mathbf{Y}_2)_{\Phi_2}$$

Definizioni Generale di RFD

$$\mathbf{D_{c_1} \times D_{c_2} : (X_1, X_2)_{\Phi_1} \xrightarrow{\Psi \geq \epsilon} (Y_1, Y_2)_{\Phi_2}}$$

- dove:
  - $D_{c_1} \times D_{c_2} = \left\{(t_1, t_2) \in D(R_1) \times D(R_2) \mid \left(\bigwedge_{i=1}^{k} c_{1_i}(t_1[A_i])\right) \wedge \left(\bigwedge_{j=1}^{m} c_{2_j}(t_2[B_j])\right)\right\}$ definisce l'insieme di tuple su cui si applica φ
  - ϕ_1 e ϕ_2 sono insiemi di vincoli su X e Y
  - Ψ è la coverage measure definita su D_{c_1} x D_{c_2}

RFD: Semantica

- Date due istanze di relazioni $r_1 \subseteq D_{c_1}$ e $r_2 \subseteq D_{c_2}$ su (R_1, R_2)
- La coppia (r_1, r_2) soddisfa la RFD φ se e soltanto se
- per ogni $(t_1, t_2) \in (r_1, r_2)$

se

$\phi[X_1, X_2]$ indica true per ogni vincolo $\Phi_1 \in \phi$

allora quasi sempre

$\phi[Y_1, Y_2]$ indica true per ogni vincolo $\Phi_2 \in \phi$

- quasi sempre significa che il grado di soddisfacibilità misurato da Ψ risulta essere maggiore o uguale ad ε

Descrizione di FD

- Utilizzando la definizione generale di RFD è possibile descrivere anche le dipendenze funzionali canoniche

$$\mathbf{D_{TRUE}: X_{EQ} \xrightarrow{\Psi_{err(0)}} Y_{EQ}}$$

  - $X_1 = X_2 = X$ e $Y_1 = Y_2 = Y$
    - Si applica su una singola istanza di database
  - $D_{c_1} \times D_{c_2} = D_{TRUE}$ indica una tautologia
    - Non si restringe il dominio di applicabilità
  - $\Psi_{err(0)}$ indica che la dipendenza deve valere per tutte le tuple
  - EQ rappresenta il vincolo di uguaglianza

Proposte in letteratura

- Analizziamo le più importanti proposte in letteratura sulla base dei criteri di rilassamento presentati
- Ogni nuova definizione di dipendenza funzionale può essere descritta utilizzando la definizione generale e rientra in una delle seguenti classi
  - RFD che rilassano sull'extent
  - RFD che rilassano sul confronto
  - RFD ibride

RFD che rilassano sull'extent

Rilassare sull'extent significa che una dipendenza funzionale deve essere valida per "quasi" tutte le tuple o per un sottoinsieme di esse

RFD che rilassano sull'extent

- Una dipendenza funzionale potrebbe non valere per tutte le tuple a causa di
  - Errori
  - Valori mancanti
  - Utilizzo di differenti formati di dati
  - Specificità di dominio che ammettono la presenza di outlier
- Le RFD che appartengono a questa categoria differiscono nel metodo attraverso il quale permettono di definire il sottoinsieme di tuple

RFD che rilassano sull'extent

- Le RFD che appartengono a questa categoria differiscono nel metodo usato nel definire il sottoinsieme di tuple
  - Attraverso la coverage measure

$$D_{TRUE} : X_{EQ} \xrightarrow{\Psi(X,Y)} Y_{EQ}$$

    - dove Ψ è definita in termini di
      - Probabilità
      - Cardinalità di dominio
      - Impurità
      - g3 error e/o confidenza

RFD che rilassano sull'extent

- Le RFD che appartengono a questa categoria differiscono nel metodo usato nel definire il sottoinsieme di tuple
  - Attraverso una condizione

$$D_c : X_{EQ} \xrightarrow{\Psi_{err(0)}} Y_{EQ}$$

    - dove D_c è un sottoinsieme identificato da una sequenza di predicati in c che possono essere definiti in termini di
      - Vincoli
      - Pattern Tableau

AFD

- AFD (Aproximate Functional Dependency)
- Una RFD che rappresenta dipendenze funzionali che valgono per "quasi" ogni tupla
- Per quantizzare il "quasi" si possono utilizzare diverse misure tra cui il g3 error
  - Il minimo numero di tuple che devono essere rimosse da un'istanza di relazione r per far si che una dipendenza funzionale $\phi: X \rightarrow Y$ valga

$$\Psi(X,Y) = \frac{\min\{|r_1| \text{ t.c. } r_1 \subseteq r \text{ e } X \rightarrow Y \text{ vale in } r \setminus r_1\}}{|r|}$$

AFD

- AFD (Approximate Functional Dependency)
- Formalmente, data una soglia di errore ε
  - Con $0 \leq \varepsilon \leq 1$
- Una AFD può essere definita come

$$D_{TRUE}: X_{EQ} \xrightarrow{\Psi(X,Y) \leq \varepsilon} Y_{EQ}$$

AFD: Esempio

- In un database clinico
  - È improbabile avere due pazienti ricoverati nell'ospedale che hanno lo stesso nome
  - Bisogna comunque considerare quei pochi casi di omonimia che potrebbero verificarsi
- In questo dominio vale la seguente AFD

$$D_{TRUE}: Nome_{EQ} \xrightarrow{\Psi(X,Y)\leq 0,02} GruppoSanguigno_{EQ}$$

AFD: Esempio

SIN	Nome	Data di Nascita	Sesso	Gruppo Sanguigno
087-34-7789	Andrea White	1935-03-14	F	0+
087-11-3455	Mary Brown	1930-08-31	F	A+
089-65-3325	Bill Mc Gregor	1970-12-21	M	0+
091-87-9437	John Smith	2005-11-01	M	AB+
092-12-1439	Eric Ford	1929-10-19	M	A-
…				
097-19-7367	Mary Brown	1990-03-20	F	AB+

$$D_{TRUE}: Nome_{EQ} \xrightarrow{\Psi(X,Y)\leq 0,02} GruppoSanguigno_{EQ}$$

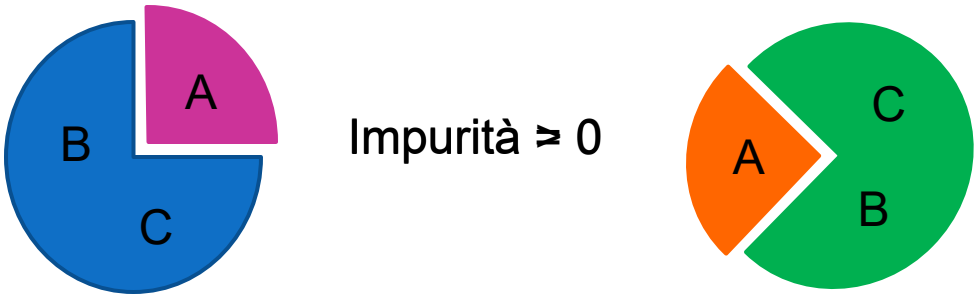
AFD

- AFD (Aproximate Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	g3 error Confidenza τ association Information dependency
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Gestione degli errori/outlier Query answering Query rewriting Decomposizione orizzontale
Problemi analizzati	Discovery (Inferenza) dai dati

PuD

- PuD (Purity Dependency)
- Una RFD che generalizza il concetto di dipendenza funzionale mediante l'utilizzo della misura di impurità
  - Date due partizioni Π_s' e Π_s'' di un insieme S la misura di impurità avrà valore 0 se e soltanto se ogni blocco di Π_s' è incluso in uno ed un solo blocco di Π_s''
  - La misura di impurità crescerà nel valore quando i blocchi di Π_s' si intersecano con più blocchi di Π_s''



PuD

- PuD (Purity Dependency)
- Formalmente, siano
  - r un'istanza di database
  - X e Y due insiemi di attributi
  - π_X e π_Y due partizioni dell'insieme delle tuple di r indotte dai valori di X e Y in r
- Una PuD può essere definita come

$$\mathbf{D_{TRUE}: X_{EQ} \xrightarrow{\theta(\pi_X, \pi_Y) \leq \varepsilon} Y_{EQ}}$$

  - dove θ è una funzione concava e sub-additiva che calcola la più grande misura di impurità dei blocchi di π_X rispetto a quelli di π_Y

PuD: Esempio

- Nella relazione che contiene i dati relativi alle medicine di un database clinico

Produttore	Nome	Categoria	…	Ingrediente Attivo	Prescrizione
Angelini	Aulin	NSAID		Nimesulide	Si
Dompè	Oki	NSAID		Ketoprofen	Si
Lisapharma	Arfen	NSAID		Ibuprofen	Si
…					
Zambon It	Spidifen	NSAID		Ibuprofen	No

- vale la seguente PuD

$$D_{TRUE}: IngredienteAttivo_{EQ} \xrightarrow{\theta(\pi_{IngredienteAttivo}, \pi_{Prescrizione}) \leq 0.09} Prescrizione_{EQ}$$

PuD

- PuD (Purity Dependency)
  - Alcune sue caratteristiche sono

Extent	Impurità
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Classificazione approssimata
Problemi analizzati	-

NuD

- NuD (Numerical Dependency)
- Una RFD che rilassa il concetto di dipendenza funzionale mediante l'utilizzo di un vincolo sulla cardinalità di dominio
  - data uno schema di relazione R, un'istanza r di R e $X,Y \subseteq attr(R)$
  - una NuD specifica che ogni proiezione di tupla $t[X]$ è associata al più a k differenti tuple su Y, per qualche costante k
- Formalmente, una NuD può essere definita come
- dove

$$D_{TRUE}: X_{EQ} \xrightarrow{card(X,Y) \leq k} Y_{EQ}$$

$$card(X,Y) = \left| \pi_Y \left(\sigma_{(X=t[X])}(r) \right) \right|$$

NuD: Esempio

- In un database clinico
  - Se ad ogni reparto dell'ospedale possono essere assegnate al più 10 stanze
- vale la seguente NuD

$$\mathbf{D_{TRUE}: StanzaCheckIn_{EQ} \xrightarrow{card(StanzaCheckIn,\#Stanza)\leq 10} \#Stanza_{EQ}}$$

NuD

- NuD (Numerical Dependency)
  - Alcune sue caratteristiche sono

Extent	Cardinalità di dominio
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Progettazione di database
Problemi analizzati	Problema di Implicazione (regole di inferenza)

RFD basate su probabilità

- Esistono varie RFD basate sulla probabilità
- PD (Partial Determination)
- soft FD (soft Functional Dependency)
- pFD (probabilistic Functional Dependency)
  - Sfruttano il concetto di probabilità per effettuare l'approssimazione sull'extent
- Le loro definizioni sembrano essere equivalenti, tuttavia
  - sono state introdotte in periodi differenti
  - sono state introdotte per obiettivi differenti
  - utilizzano algoritmi differenti di validazione e di inferenza

RFD basate su probabilità

- Formalmente, una RFD basata sulla probabilità può essere definita come

$$D_{TRUE}: X_{EQ} \xrightarrow{\rho(X,Y) \geq 1-\varepsilon} Y_{EQ}$$

  - dove $\rho(X,Y)$ rappresenta la probabilità che prese due tuple scelte casualmente queste abbiano lo stesso valore di Y, purché abbiano gli stessi valori di X

RFD basate su probabilità: Esempio

- Nella relazione che contiene i dati relativi alle medicine di un database clinico
  - La presenza di possibili errori nei dati sull'attributo Nome di medicine che hanno nomi simili come Daflon vs. Deflan o Lanoxin vs. Laroxyl fa si che la FD canonica Nome→Produttore non valga
- Tuttavia, vale la seguente RFD basata sulla probabilità

$$\mathbf{D_{TRUE}: Nome_{EQ} \xrightarrow{\rho(Nome,Produttore)\geq 0.97} Produttore_{EQ}}$$

RFD basate su probabilità

- PD (Partial Determination)
  - Alcune sue caratteristiche sono

Extent	Probabilità Compressione
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Compressione dei database
Problemi analizzati	-

RFD basate su probabilità

- soft FD (soft Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	Probabilità
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Query Optimization Compressione delle tabelle
Problemi analizzati	Discovery (Inferenza) dai dati

RFD basate su probabilità

- pFD (probabilistic Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	Probabilità
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Data Integration
Problemi analizzati	Validazione delle dipendenze Discovery (Inferenza) dai dati

CD

- CD (Constrained functional Dependency)
- Una RFD che permette di specificare il sottoinsieme di tuple per il quale una dipendenza funzionale è valida attraverso un vincolo
- Formalmente, data una classe di vincoli $\mathcal{L}$
- Una CD può essere definita come

$$\mathbf{D_c}: \mathbf{X_{EQ}} \xrightarrow{\Psi_{\mathbf{err(0)}}} \mathbf{Y_{EQ}}$$

  - dove $D_c \subseteq dom(R)$ rappresenta le tuple che soddisfano il vincolo $c \in \mathcal{L}$

CD: Esempio

- Nella relazione che contiene i dati relativi ai medici di un database clinico
  - Supponiamo che solo i medici specializzati in “Pediatria” vale che lo stipendio dipende dall’anzianità di lavoro
- In questo dominio vale la seguente CD

$$\mathbf{D_c: Esperienza_{EQ} \xrightarrow{\Psi_{err(0)}} Stipendio_{EQ}}$$

  - dove

$$D_c = \{t \in dom(Dottore) \mid t[Specializzazione] = \text{‘Pediatria’}\}$$

CD: Esempio

ID	Nome	Specializzazione	…	Esperienza	Stipendio
1	George Johnson	Neurologia		1 anno	$118,000
2	Joe House	Cardiologia		10 anni	$314,000
3	Derek Williams	Pediatria		2 anni	$156,000
4	Henry Jones	Neurologia		1 anno	$158,000
5	Robert White	Pediatria		2 anni	$156,000
…					
30	Victor Sanchez	Radiologia		5 anni	$225,000

$$\mathbf{D_c: Esperienza_{EQ} \xrightarrow{\Psi_{err(0)}} Stipendio_{EQ}}$$

- dove

$D_c = \{t \in dom(Dottore) \mid t[Specializzazione]='Pediatria'\}$

CD

- CD (Constrained functional Dependency)
  - Alcune sue caratteristiche sono

Extent	Vincolo
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Query optimization
Problemi analizzati	Problema di Implicazione (regole di inferenza)

CFD

- CFD (Conditional Functional Dependency)
- Una RFD che permette di specificare il sottoinsieme di tuple per il quale una dipendenza funzionale è valida attraverso una condizione
  - Simile alle CD
- Tuttavia, le condizioni risultano essere meno generali di quelle viste per le CD dato che si basano solo sull'operatore di uguaglianza
- Inoltre, il dominio di applicabilità viene definito mediante il concetto di pattern tableau

Pattern Tableau

- Una pattern tableau T_r rappresenta una tabella che contiene gli attributi di una data relazione R, i cui valori per ogni attributo $A \in attr(R)$ possono essere
  - una costante 'a' in $dom(A)$, oppure
  - una variabile senza nome '-' che prende valori da $dom(A)$
- Esso estrae un sottoinsieme di tuple da dom(R) vincolando dati semanticamente correlati secondo specifiche in esso definite

Pattern Tableau: Esempio

- Il seguente pattern tableau T_r

Specializzazione	Esperienza	Stipendio
Pediatria	-	-

  - estrae tutte le tuple che hanno come valore di Specializzazione = 'Pediatria' e un qualsiasi valore in dom(Esperienza) e dom(Stipendio)

CFD

- CFD (Conditional Functional Dependency)
- Formalmente, data una CFD può essere definita come

$$D_{T_r} : X_{EQ} \xrightarrow{\Psi_{err(0)}} Y_{EQ}$$

- Nota: l'esempio precedente può essere definito anche in termini di CFD

CFD: Esempio

ID	Nome	Specializzazione	…	Esperienza	Stipendio
1	George Johnson	Neurologia		1 anno	$118,000
2	Joe House	Cardiologia		10 anni	$314,000
3	Derek Williams	Pediatria		2 anni	$156,000
4	Henry Jones	Neurologia		1 anno	$158,000
5	Robert White	Pediatria		2 anni	$156,000
…					
30	Victor Sanchez	Radiologia		5 anni	$225,000

$$D_{T_r} : \text{Esperienza}_{EQ} \xrightarrow{\Psi_{err(0)}} \text{Stipendio}_{EQ}$$

- dove $T_r =$

Specializzazione	Esperienza	Stipendio
Pediatria	-	-

CFD

- CFD (Conditional Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	Pattern Tableau
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Data Cleaning Data Quality Error detection Conflict resolution
Problemi analizzati	Problema di Implicazione (regole di inferenza) Controllo della consistenza Scoperta di violazioni Validazione delle dipendenze Discovery (Inferenza) dai dati

RFD basate su CFD

- Negli anni sono state prodotte diverse estensioni di CFD
- eCFD (extended Conditional Functional Dependency)
  - Estendono le condizioni con la disgiunzione e la disuguaglianza
- CFD^p (CFD with built-in predicates)
  - Permettono la specifica di pattern di valori con predicati $\in \{<, \leq, >, \geq, \neq\}$

RFD basate su CFD

- eCFD (extended Conditional Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	Pattern Tableau
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Data Cleaning Data Quality
Problemi analizzati	Problema di Implicazione (regole di inferenza) Controllo della consistenza Scoperta di violazioni

RFD basate su CFD

- CFD^p (CFD with built-in predicates)
  - Alcune sue caratteristiche sono

Extent	Pattern Tableau
Tipo di confronto	Match esatto
Modelli supportati	Relazionale
Domini applicativi	Data Quality
Problemi analizzati	Problema di Implicazione (regole di inferenza) Controllo della consistenza Scoperta di violazioni

RFD che rilassano sul confronto

Rilassare sul confronto significa che possono essere utilizzati paradigmi di matching approssimato per comparare i valori degli attributi sia sull'LHS che sull'RHS

RFD che rilassano sul confronto

- L'obiettivo è quello di catturare relazioni semantiche tra gruppi di valori che possono essere considerati "simili" piuttosto che identici
- La struttura generale di questo tipo di RFD è

$$\mathbf{D_{TRUE}}: \mathbf{X}_{\Phi_1} \xrightarrow{\Psi_{\mathbf{err}(0)}} \mathbf{Y}_{\Phi_2}$$

  - dove Φ_1 e $\Phi 2$ sono insiemi di vincoli che esprimono la "vicinanza" di due valori in uno specifico dominio

RFD che rilassano sul confronto

- Le RFD che appartengono a questa categoria differiscono nel metodo usato per definire gli insiemi di vincoli
  - Attraverso funzioni di matching approssimato
    - Closeness function
    - Differential function
    - Matching operator
    - Metric distance
    - Similarity function
    - Tolerance relation
  - Attraverso criteri di ordinamento
    - Order relation
    - Temporal constraint

MFD

- MFD (Metric Functional Dependency)
- Una RFD che generalizza le dipendenze funzionali ammettendo un confronto tra i valori delle tuple dell'RHS che tollera piccole differenze controllate da una metrica
  - La metrica che vale sul dominio di un attributo Y è così definita

$$\phi: \mathbf{dom}(Y) \times \mathbf{dom}(Y) \rightarrow \mathcal{R}$$

  - Attraverso essa è possibile calcolare la distanza massima tra ogni coppia di valori nell'insieme di tutti i possibili valori V, come

$$\Delta_{\phi}(V) = \max_{a,b \in V} \phi(a,b)$$

MFD

- MFD (Metric Functional Dependency)
- Formalmente, siano
  - r un'istanza di database
  - X e Y due insiemi di attributi
  - $\Delta_\phi(V)$ la funzione che calcola la distanza massima tra una coppia di valori
  - ε una soglia di tolleranza
- Una MFD può essere definita come

$$D_{TRUE}: X_{EQ} \xrightarrow{\Psi_{err(0)}} Y_{\max_{s \in \pi_X} \Delta_\phi(s[Y]) \leq \varepsilon}$$

MFD: Esempio

- Nella relazione che contiene i dati relativi ai medici di un database clinico
  - Supponiamo che la differenza nello stipendio tra i dottori con uguale specializzazione ed esperienza sia minore di 5,000$
- In questo dominio vale la seguente MFD

$$\mathbf{D_{TRUE}: (Specializzazione, Esperienza)_{EQ} \xrightarrow{\Psi_{err(0)}} Stipendio_{\max_{s \in \pi_X} \Delta_\phi(s[Y]) \leq 5{,}000}}$$

MFD: Esempio

ID	Nome	Specializzazione	…	Esperienza	Stipendio
1	George Johnson	Neurologia		1 anno	$218,000
2	Joe House	Cardiologia		10 anni	$314,000
3	Derek Williams	Pediatria		2 anni	$156,000
4	Henry Jones	Neurologia		1 anno	$222,000
5	Robert White	Pediatria		2 anni	$156,000
…					
30	Victor Sanchez	Radiologia		5 anni	$225,000

$$\mathbf{D_{TRUE}: (Specializzazione, Esperienza)_{EQ}} \xrightarrow{\Psi_{err(0)}} \mathbf{Stipendio}_{\max_{s \in \pi_X} \Delta_\phi(s[Y]) \leq 5{,}000}$$

MFD

- MFD (Metric Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Metric distance
Modelli supportati	Relazionale
Domini applicativi	Merging di sorgenti Data Quality
Problemi analizzati	Validazione delle dipendenze

ND

- ND (Neighborhood Dependency)
- Una RFD che è stata introdotta per esprimere le regolarità presenti all'interno dei dati
- La sua definizione utilizza il concetto di closeness function (funzione di vicinanza) che viene associata ad ogni attributo A
  - Restituisce un numero compreso tra 0 e 1 per esprimere quanto due valori di un attributo possono essere considerati simili

ND

- ND (Neighborhood Dependency)
- Essa sfrutta anche il concetto di predicato di vicinanza
  - Esso associa ogni attributo A ad una soglia α che permette di valutare una coppia di valori sulla base della funzione di vicinanza associata ad A
    - Denotato con A^{α}
- Formalmente, una ND può essere definita come

$$\mathbf{D_{TRUE}}: \mathbf{X}_{\left(\theta_{A_1} \geq \alpha_1 \wedge \ldots \wedge \theta_{A_n} \geq \alpha_n\right)} \xrightarrow{\Psi_{err(0)}} \mathbf{Y}_{\left(\theta_{B_1} \geq \beta_1 \wedge \ldots \wedge \theta_{B_m} \geq \beta_m\right)}$$

ND: Esempio

- Nella relazione che contiene i dati relativi alle cartelle cliniche
  - Si può ipotizzare che i pazienti che hanno un'età simile e diagnosi simili vengano ricoverati in reparti simili
- In questo dominio vale la seguente ND

$$\mathbf{D_{TRUE}}: \ (\mathbf{Diagnosi}, \mathbf{Anni})_{(\theta_{\mathbf{Diagnosi}} \geq 0.85 \ \wedge \ \theta_{\mathbf{Anni}} \geq 0.8)} \xrightarrow{\Psi_{\mathbf{err}(0)}} \mathbf{RepartoCheckIn}_{(\theta_{\mathbf{RepartoCheckIn}} \geq 0.9)}$$

ND

- ND (Neighborhood Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Closeness Function
Modelli supportati	Relazionale
Domini applicativi	Data Mining
Problemi analizzati	Discovery (Inferenza) dai dati

SFD

- SFD (Similarity Functional Dependency)
- Una RFD che sfrutta il concetto di relazione di tolleranza associata ad ogni attributo A
  - Soddisfa le proprietà di simmetria, riflessività e transitività
- Una relazione di tolleranza Θ permette di ricavare blocchi di tolleranza, che rappresentano i sottoinsiemi in cui vale Θ tra ogni coppia di valori
  - Per ogni attributo A, una relazione di tolleranza Θ_A è definita come

$$t_i \, \theta_A t_j \Leftrightarrow \left| t_i(A) - t_j(A) \right| \leq \varepsilon$$

  - e deve valere per ogni coppia di tuple $(t_i, t_j) \in r$

SFD

- SFD (Similarity Functional Dependency)
- Analogamente
  - Per ogni insieme di attributi X, una relazione di tolleranza Θ_X è definita come

$$t_i \, \theta_X t_j \Leftrightarrow t_i \, \theta_A t_j$$

  - per ogni $A \in X$

- Formalmente, una SFD può essere definita come

$$D_{TRUE} : X_{\theta_X} \xrightarrow{\Psi_{err(0)}} Y_{\theta_Y}$$

SFD: Esempio

- Nella relazione che contiene i dati relativi ai medici

ID	Nome	Specializzazione	…	Stipendio	Tasse
1	George Johnson	Neurologia		$218,000	$62,500
2	Joe House	Cardiologia		$314,000	$94,200
3	Derek Williams	Pediatria		$156,000	$39,500
4	Henry Jones	Neurologia		$222,000	$63,000
5	Robert White	Pediatria		$156,000	$39,500
…					

- vale la seguente SFD

$$D_{TRUE}: \text{Stipendio}_{\theta_{\text{Stipendio}}} \xrightarrow{\Psi_{err(0)}} \text{Tasse}_{\theta_{\text{Tasse}}}$$

SFD

- SFD (Similarity Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Relazione di Tolleranza
Modelli supportati	Relazionale
Domini applicativi	Analisi del comportamento
Problemi analizzati	-

TMFD

- TMFD (Type-M Functional Dependency)
- Una RFD definita per i dati multimediali
  - È parametrizzata attraverso funzioni di distanza, le quali vengono usate per confrontare gli attributi multimediali
- Inoltre, utilizza una funzione che permette di riassumere i risultati delle funzioni di distanza dei singoli attributi applicate su una coppia di tuple (tuple distance function)

$$\chi(t_1, t_2) = \theta(\phi_1(a_1, b_1), \ldots, \phi_n(a_n, b_n))$$

  - dove φ_i è una funzione di distanza definita su $\mathrm{dom}(A_i)$, e
  - Θ rappresenta una funzione di aggregazione Θ: $[0,1]^n \rightarrow [0,1]$

TMFD

- TMFD (Type-M Functional Dependency)
- Siano
  - t_1,t_2 due tuple di un'istanza di relazione r che sono simili entro una soglia ε' per l'insieme di attributi X e una soglia ε'' per l'insieme di attributi Y
    - in generale, denotato con $t_1 \cong_{(\chi,\varepsilon)} t_2$ se e soltanto se $\chi(t_1, t_2) \leq \varepsilon$
- Una TMFD può essere definita come

$$\mathbf{D_{TRUE}}: \mathbf{X}_{\chi_1 \leq \varepsilon'} \xrightarrow{\Psi_{\mathbf{err}(0)}} \mathbf{Y}_{\chi_2 \leq \varepsilon''}$$

TMFD: Esempio

- Nella relazione che contiene i dati relativi alle cartelle cliniche di un database clinico
  - È possibile definire una dipendenza per correlare gli elettrocardiogrammi (ECG) con i battiti cardiaci (Battiti)
  - Tuttavia è necessario utilizzare diverse funzioni di similarità sui singoli attributi: FRATTALE e BC (Battiti cardiaci), rispettivamente
- In questo dominio vale la seguente TMFD

$$\mathbf{D_{TRUE}}: \mathbf{ECG}_{(\mathbf{FRATTALE},\varepsilon')} \xrightarrow{\Psi_{\mathbf{err}(0)}} \mathbf{Battiti}_{(\mathbf{BC},\varepsilon'')}$$

TMFD

- MFD (Metric Functional Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Similarity Function
Modelli supportati	Relazionale
Domini applicativi	Normalizzazione di DB multimediali
Problemi analizzati	-

MD

- MD (Matching Dependency)
- Una RFD che mira a risolvere il problema di identificare quali record di istanze di database differenti rappresentano la stessa entità nel mondo reale
  - Permettendo di tollerare la diversità dei valori degli attributi dovuta agli errori di inserimento e all'utilizzo di diversi formati di rappresentazione
- Questa RFD mette a confronto coppie di relazioni (R_1, R_2)

MD

- Formalmente, data una coppia di relazioni (R_1, R_2)
- una MD può essere definita come

$$\mathbf{D_1 \times D_2 : (X_1, X_2)_{\approx}} \xrightarrow{\Psi_{\mathbf{err(0)}}} \mathbf{(Y_1, Y_2)_{\rightleftharpoons}}$$

- dove
  - $X_1: A_1, \ldots A_n$ e $X_2: B_1, \ldots B_n$ rappresentano l'insieme di attributi X su $D_1 = dom(R_1)$ e $D_2 = dom(R_2)$ rispettivamente
  - $Y_1: E_1, \ldots E_m$ e $Y_2: F_1, \ldots F_m$ rappresentano l'insieme di attributi Y su $D_1 = dom(R_1)$ e $D_2 = dom(R_2)$ rispettivamente
  - $\approx_j$ rappresenta un predicato di similarità definito sui domini $R_1[A_j]$ e $R_2[B_j]$ rispettivamente
  - $\rightleftharpoons$ rappresenta un matching operator, il quale indica che i valori di $R_1[E_i]$ e $R_2[F_i]$ sono identici a seguito di modifiche

MD: Esempio

- Consideriamo la relazione paziente di un DB medico

SIN	Nome	DataDiNascita	Sesso	…	Indirizzo
087-34-7789	Andrea White	1935-03-14	F		987 Jefferson, NV
087-11-3455	Mary Brown	1930-08-31	F		55 Fifth AV, NV
089-65-3325	Bill Mc Gregor	1970-12-21	M		100 Canal ST, NJ
…					

- e la relazione cliente di un DB di assicurazioni mediche

#Polizza	Cliente	DataNascita	…	Domicilio	Premio
35677651	Andreaw White	03-14-1935		987 Jefferson AV, NV	$2,400
35677712	M. Brown	08-31-1930		55 Fifth AV, NV	$2,900
35677754	B. Gregor	12-21-1970		100 Canal Street, NJ	$1,000
…					

MD

- In questo dominio vale la seguente MD

$$\mathbf{D_{Paziente} \times D_{Cliente}}:$$
$$(\{\mathbf{Nome}, \mathbf{DataDiNascita}\}, \{\mathbf{Cliente}, \mathbf{DataNascita}\})_{\approx}$$
$$\xrightarrow{\Psi_{\mathbf{err}(0)}} (\mathbf{Indirizzo}, \mathbf{Domicilio})_{\rightleftharpoons}$$

  - la quale indica che quando il nome di un paziente è simile al cliente di una polizza e anche la data di nascita è simile
  - allora il loro indirizzo è lo stesso, a meno di dover effettuare delle piccole modifiche per renderlo identico

MD: Esempio

- Consideriamo la relazione paziente di un DB medico

SIN	Nome	DataDiNascita	Sesso	…	Indirizzo
087-34-7789	Andrea White	1935-03-14	F		987 Jefferson, NV
087-11-3455	Mary Brown	1930-08-31	F		55 Fifth AV, NV
089-65-3325	Bill Mc Gregor	1970-12-21	M		100 Canal ST, NJ
…					

- e la relazione cliente di un DB di assicurazioni mediche

#Polizza	Cliente	DataNascita	…	Domicilio	Premio
35677651	Andreaw White	03-14-1935		987 Jefferson AV, NV	$2,400
35677712	M. Brown	08-31-1930		55 Fifth AV, NV	$2,900
35677754	B. Gregor	12-21-1970		100 Canal Street, NJ	$1,000
…					

MD

- MD (Matching Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Similarity Function Matching Operator
Modelli supportati	Relazionale
Domini applicativi	Record Matching Data Cleaning Data Quality Entity Resolution
Problemi analizzati	Problema di Implicazione (regole di inferenza) Validazione delle dipendenze Discovery (Inferenza) dai dati

CoD

- CoD (Comparable Dependency)
- Una RFD che generalizza il concetto di dipendenza funzionale nel contesto dei dataspace eterogenei
  - Permette di gestire il confronto tra attributi con nomi differenti attraverso un operatore di confronto degli attributi
  - Copre la semantica di un'ampia classe di RFD, quali FD, MFD, e MD
- Questa RFD mette a confronto coppie di attributi in un dataspase S

CoD

- CoD (Comparable Dependency)
- L'operatore di confronto, denotato con $\leftrightarrow_{ij}$ compara due attributi A_i, A_j in un dataspace S secondo la seguente semantica
  - Equality operator: $A_i = A_j$
  - Metric operator: $A_i \approx_\varepsilon A_j$ (se la distanza tra due valori calcolata mediante una metrica è $\leq \varepsilon$)
  - Matching operator: $A_i \leftrightharpoons A_j$

CoD

- CoD (Comparable Dependency)
- In generale, una funzione di confronto è definita come

$$\Theta(A_i, A_j): [A_i \leftrightarrow_{ii} A_i, A_i \leftrightarrow_{ij} A_j, A_j \leftrightarrow_{jj} A_j]$$

- e specifica un vincolo sui valori confrontabili degli attributi A_i e A_j
- Per verificare che una funzione di confronto sia valida è necessario che uno dei seguenti confronti sia vero
  - $t_1[A_i] \leftrightarrow_{ii} t_2[A_i]$
  - $t_1[A_i] \leftrightarrow_{ij} t_2[A_j]$
  - $t_2[A_i] \leftrightarrow_{ij} t_1[A_j]$
  - $t_1[A_j] \leftrightarrow_{jj} t_2[A_j]$

CoD

- Formalmente, dato un dataspace S
- una CoD può essere definita come

$$\mathbf{D_{TRUE}}: (\mathbf{X_1}, \mathbf{X_2})_{\theta(\mathbf{X_1},\mathbf{X_2})} \xrightarrow{\Psi_{\mathbf{err(0)}}} (\mathbf{Y_1}, \mathbf{Y_2})_{\theta(\mathbf{Y_1},\mathbf{Y_2})}$$

- dove
  - $\theta(X_1, X_2) = \wedge\theta(A_i, A_j)$ e $\theta(Y_1, Y_2) = \wedge\theta(B_k, B_l)$ con
    - $A_i \in X_1$
    - $A_j \in X_2$
    - $B_k \in Y_1$
    - $B_l \in Y_2$
  - $\theta(A_i, A_j)$ e $\theta(B_k, B_l)$ sono funzioni di confronto nel dataspace S

CoD: Esempio

- In un database clinico che contiene le cartelle cliniche di diversi ospedali vale la seguente CoD

$$\mathbf{D_{TRUE}}: (\#\mathbf{Stanza}, \#\mathbf{Letto})_{\theta(\#\mathbf{Stanza},\#\mathbf{Letto})} \xrightarrow{\Psi_{\mathbf{err}(0)}} (\mathbf{Sesso}, \mathbf{Genere})_{\theta(\mathbf{Sesso},\mathbf{Genere})}$$

  - Dato che ogni stanza di un ospedale identifica univocamente il sesso del paziente che la occupa

CoD

- CoD (Comparable Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Similarity Function Matching Operator Metric Function
Modelli supportati	Dataspace
Domini applicativi	Query answering Object identification
Problemi analizzati	Scoperta di violazioni Validazione delle dipendenze

DD

- DD (Differential Dependency)
- Una RFD che permette di esprimere vincoli di differenza sui valori degli attributi attraverso una funzione di distanza
- Una funzione di distanza $\phi[B]$ su un attributo B specifica un vincolo sulla differenza che due tuple t_1 e t_2 devono avere su B
  - Un tale vincolo quando è soddisfatto è denotato con $(t_1, t_2) \asymp \phi[B]$
  - I vincoli che possono essere specificati dall'operatore rientrano nell'insieme $\{=, <, \leq, >, \geq\}$ e sono associati ad una soglia ε

DD

- Formalmente, data una relazione R
- una DD può essere definita come

$$D_{TRUE}: X_{\phi_L} \xrightarrow{\Psi_{err(0)}} Y_{\phi_R}$$

- dove
  - ϕ_L e ϕ_R rappresentano funzioni di differenza sugli insiemi di attributi X e Y rispettivamente
- Indica che per ogni coppia di tuple per cui la differenza sugli attributi in X soddisfa il vincolo specificato da $\phi_L[X]$, allora la loro differenza sugli attributi in Y deve soddisfare il vincolo $\phi_R[Y]$

DD: Esempio

- Nel contesto di un database clinico per scoprire le anomalie nell'esecuzione dei controlli medici rispetto alla DataDiPrescrizione si può utilizzare la seguente DD

$$\mathbf{D_{TRUE}}: \mathbf{DataDiPrescrizione}_{\phi_L} \xrightarrow{\Psi_{err(0)}} \mathbf{DataDiEsecuzione}_{\phi_R}$$

- dove
  - ϕ_L[DataDiPrescrizione]=[DataDiPrescrizione(=0)]
  - ϕ_R[DataDiEsecuzione]=[DataDiEsecuzione ($\leq$5)]
- Indica che i controlli medici prescritti nello stesso giorno non devono essere eseguiti ad una distanza maggiore di cinque giorni

DD: Esempio

- Consideriamo la relazione controlli clinici di un DB medico

ID	Nome	…	DataDiPrescrizione	DataDiEsecuzione
1	Colesterolo		2014-03-03	2014-03-06
2	HIV		2014-03-07	2014-03-14
3	HCV		2014-03-03	2014-03-05
4	AIDS		2014-03-05	2014-03-14
5	Mammografia		2014-03-06	2014-03-08
6	Ecocardiogramma		2014-03-21	2014-03-25
…				

DD

- DD (Differential Dependency)
  - Alcune sue caratteristiche sono

Extent	-
Tipo di confronto	Differential Function
Modelli supportati	Relazionale
Domini applicativi	Query Optimization Data Partition Record Linkage
Problemi analizzati	Problema di Implicazione (regole di inferenza) Controllo della consistenza Scoperta di violazioni Validazione delle dipendenze Discovery (Inferenza) dai dati