Data Mining e Bioimmagini

Regole Associative

Prof. Pierangelo Veltri – 06/11/2023- Autori: Panarello, Saeedzadeh - Revisionatori: Panarello Saeedzadeh

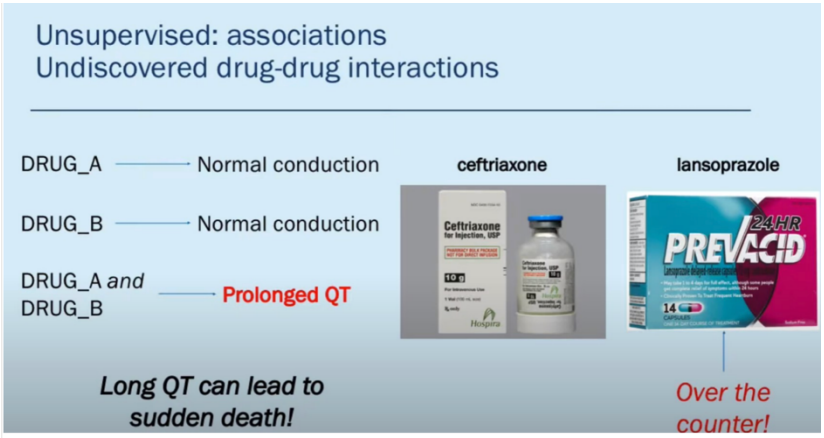
Le date degli appelli sono 02/02/2024 e 23/02/2024. La modalità di esame non è ancora stata decisa, potrebbe esserci un esercizio scritto ed un orale con un progetto (non lo sa manco lui).

REGOLE ASSOCIATIVE

Il data mining è il processo di estrazione di conoscenza e di informazioni utili da grandi quantità di dati memorizzati nei database. Una tecnica di data mining prevede l'uso delle **regole associative**, ossia delle condizioni che descrivono relazioni di associazione rilevanti tra gli attributi del dataset. Si parte da un dataset e si cerca di identificare degli items (un evento che accade) e le loro caratteristiche.

Esempio 1

Si ha a disposizione un dataset medico-clinico su un farmaco A ed uno B e la combinazione dei due farmaci. In particolare, si vuole sapere se ci sono delle associazioni tra gli effetti collaterali dei due farmaci soprattutto quando sono combinati. Se, per esempio, abbiamo 100 analisi (in termine tecnico **transazione**, ossia combinazioni di eventi) e si vede che il numero di pazienti che hanno preso la combinazione dei due farmaci ed hanno presentato effetti collaterali raggiunge una certa soglia, si può stabilire la regola associativa secondo cui la transazione farmaco A e B porta ai seguenti effetti collaterali.



Lo scopo di tale tecnica è quella di estrarre nuove informazioni da quelle contenute nel dataset tramite la formulazione di regole associative. Inoltre, se le associazioni tra gli eventi sono numerose (cioè gli eventi in questione si presentano insieme con una certa frequenza), diventa rilevante studiarne la correlazione.

Esempio 2

Si consideri il caso in cui si voglia studiare gli eventi correlati ad una combinazione di acquisti (i dati possono quindi provenire da dataset diversi, come diversi scontrini); dall'analisi degli scontrini si osserva una correlazione tra l'acquisto di pannolini e birre. Questa informazione può essere usata sia a scopo predittivo che di marketing cosicché i pannolini e le birre saranno posizionati in scaffali non troppo distanti e ad altezza di adulto perché è quello il target della vendita (in questo caso si fa customizzazione).

Mining di regole associative

- Dato un insieme di transazioni, trovare le regole che segnalano la presenza di un elemento sulla base della presenza di altri elementi nella transazione

Transazioni del carrello della spesa

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Alcuni esempi di regole:

{ Diaper } → { Beer },
{ Milk, Bread } → { Eggs, Coke },
{ Beer, Bread } → { Milk },

L'implicazione indica co-occorrenza e non causalità!

ATTENZIONE: gli item sono modellati da variabili binarie asimmetriche. Un item è presente oppure è assente nella transazione; la sua presenza è considerata un evento più importante della sua assenza

Formulare regole associative significa estrarre delle informazioni a partire da eventi.
Ad esempio, tramite un'analisi di mercato si può estrarre la regola pane → burro oppure [cipolle, patate] → pomodoro, intendendo che chi compra il pane generalmente acquista il burro e che chi compra cipolle e patate tende a comprare anche pomodori.

Quindi si prendono i dati, si estraggono delle regole associative e le si usano per fare previsione, marketing o customizzazione. In ambito medico una possibile applicazione è predire il numero di tecnici di cui si necessita in radiologia sulla base delle correlazioni tra tipi di malattie e numero di pazienti in pronto soccorso.

SUPPORTO E CONFIDENZA
Quando si analizza un dataset alla ricerca di regole associative, si è soliti calcolare due valori di metrica:

- **Supporto:** probabilità di transazioni con X sul totale
- **Confidenza:** probabilità delle transazioni con X che hanno anche Y

Esempio 1
Si consideri in un super market di avere un totale di 100 transazioni (scontrini) di cui 20 presentano l'acquisto del pane e in 9 di queste 20 l'acquisto anche del burro. Possiamo calcolare il supporto come rapporto tra le transazioni con pane e le transazioni totali (20/100→20%) e la confidenza come rapporto 9/20 → 45%. La regola associativa che si estrae è la seguente:
Pane → burro [20%,45%]

Esempio 2
Delle regole associative si possono ricercare anche in un contesto medico-clinico in cui si vuole osservare la correlazione tra sintomatologie. Prendiamo, ad esempio, il seguente insieme di sintomi:

$$\left\{\begin{array}{c} febbre, mal\ di\ gola, mal\ di\ testa, \\ dolori\ articolari, sonnolenza, dolori\ addominali \\ nausea, disapnea, brividi, \\ prurito, perdita\ di\ capelli \end{array}\right\}$$

TID	Items
1	Febbre, difficoltà respiratorie
2	Febbre, mal di gola
3	Mal di testa, mal di gola, brividi
4	…
5	…

Si può costruire un dataset in cui si inseriscono diverse transazioni per ricercare una correlazione tra gli items.

Se si vuole estrarre delle regole associative, bisogna innanzitutto capire quante volte l'insieme di eventi di cui si vuole trovare la correlazione si presenta. In questo caso, i possibili scenari sono 2:

1. Non si hanno alcun tipo di informazioni e si vuole fare previsione; in questo caso si attua un approccio di forza bruta (descritto dopo)
2. Si hanno dei dati su cui basare l'estrazione delle regole associative

È fondamentale stabilire una soglia di affidabilità che renda la regola valida; tale soglia può essere stabilita a partire dal supporto e dalla confidenza.

In ambito biologico e medico-clinico un'altra applicazione è l'identificazione di correlazioni tra eventi che sono rari ma ricorrenti e quindi clinicamente rilevanti.

	Febbre	Dolori articolari	Mal di testa
Febbre			x
Dolore articolare	x		
Mal di testa		x	

Raro : supporto sottosoglia

ITEMSET

L'itemset è l'insieme degli item del dataset e quindi una collezione di uno o più elementi. Si parla di k-itemset per indicare un itemset che contiene k-elementi.

Da notare che il support count rimane costante nel dataset in quanto il numero di volte che si presenta un dato itemset è sempre uguale. Nell'esempio, la transazione latte, pane, pannolini compare 2 volte quindi il supporto è uguale a 2/5.

Itemset Frequenti

- **Itemset**
  - ✓ Una collezione di uno o più elementi
    - ◆ Esempio: {Milk, Bread, Diaper}
  - ✓ k-itemset
    - Un itemset che contiene k-elementi
- **Support count ( )**
  - ✓ Numero di istanze dell'itemset nell'insieme di transazioni
  - – Esempio; ({Milk, Bread,Diaper}) = 2
- **Supporto**
  - ✓ Frazione delle transazioni che contiene l'itemset :
  - – Esempio: s({Milk, Bread, Diaper})= 2/5
- **Frequent Itemset**
  - ✓ Un itemset il cui supporto è maggiore o uguale a una soglia minsup

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

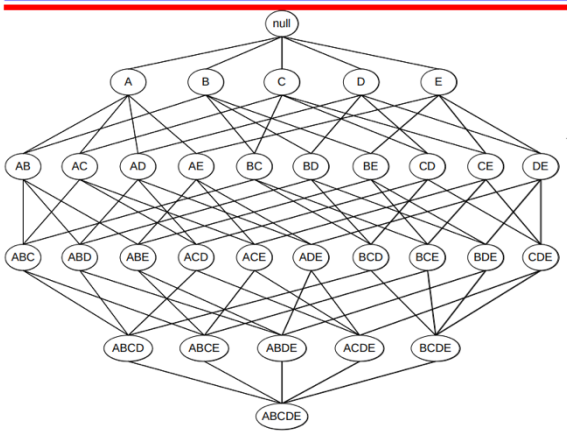
Data Mining 13

APPROCCIO BRUTO E RICERCA DEGLI ITEMSET FREQUENTI

Tale metodologia prende i valori e prova a generare delle combinazioni tra tutti gli elementi. Il numero di casi possibili è quindi esponenziale (2^d). Tuttavia, tale operazione è dal punto di vista computazionale complessa ed onerosa e perciò si procede alla riduzione dei candidati tramite due approcci:

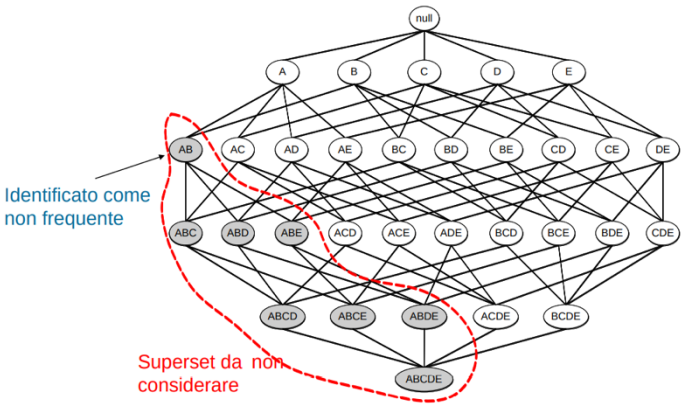
- Generare degli itemset frequenti: le combinazioni che hanno supporto al di sotto di una certa soglia vengono escluse insieme ai loro discendenti.
- Generazioni delle regole: si basa su una soglia di confidenza

Generazione di Itemset Frequenti



Dati d elmenti ci sono $\sum_{k=0}^{d}\binom{d}{k} = 2^d$ possibili item-set candidati

Riduzione del numero di candidati



Esempio

Supponiamo di definire una soglia di frequenza, in cui ci interessano solo gli elementi che compaiono almeno 3 volte; ignoriamo il set di riferimento e consideriamo i nostri elementi come eventi o valori singoli. Contiamo esattamente quanti di questi superano la soglia, in questo caso quelli che compaiono più di 3 volte. Nel processo di generazione, applichiamo un meccanismo di selezione, includendo solo gli elementi che soddisfano questa condizione. Questi elementi verranno utilizzati per creare regole associative. Escludiamo gli elementi con frequenza inferiore a 3, come quelli con frequenza 1 o 2. Continuando, esaminiamo le combinazioni di elementi, concentrandoci soprattutto sulla colonna di destra nella nostra analisi.

Applicazione del Principio Apriori

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

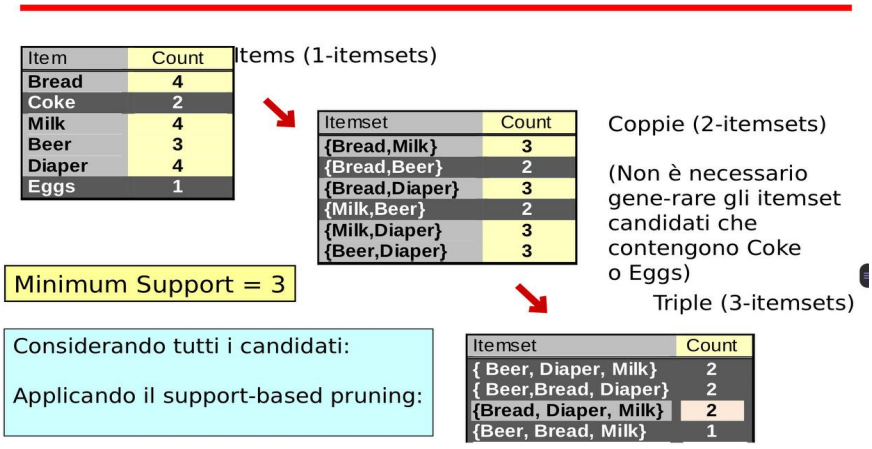
Minimum Support = 3

Considerando tutti i candidati:

Applicando il pruning support-based:

Data Mining 26

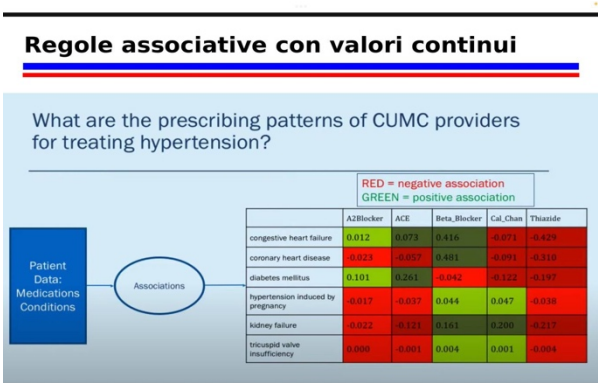
Ciò consente di registrare gli insiemi di item ottenuti dalla combinazione di due elementi, a condizione che superino la soglia di frequenza prestabilita. Usando lo stesso criterio, considerando le coppie di item presenti nel mio data set con frequenza di almeno 3, si ottiene una riduzione ulteriore del data set.



Il data set in questione è quello utilizzato per generare le combinazioni di eventi possibili. Una volta che si sono ottenute queste combinazioni con un supporto di almeno 3, si verifica quante volte queste combinazioni sono presenti nel data set di partenza. Ad esempio, se consideriamo l'insieme "birra e pane," questo appare in 4 transazioni e 2 transazioni rispettivamente. Quindi, il numero di volte in cui "birra" è presente in sequenza con "pane" è pari a 2.
Altre condizioni sono soddisfatte solo in tre casi, tranne per "birra e latte" che appare solo in due transazioni, specificamente nella terza e nella quarta.

Questa volta, se prendo gli insiemi di item composti da due elementi e li combino nuovamente con elementi che compaiono solo una volta, escludendo "bread," "beer" e "milk" con una frequenza di 2, allora, se genero una combinazione considerando solo gli elementi con frequenza 3 (come "bread and milk" e "milk and diaper"), otterrò un solo item set con support count pari a 3 nella tabella iniziale.
Ciò significa che se riduco la soglia di frequenza da 3 a 2, otterrò più combinazioni di elementi. In sostanza, questo algoritmo mi consente di estrarre un insieme di item con una certa frequenza. Una volta che ho estratto un insieme di item con la frequenza desiderata, ad esempio "beer, diaper, milk," posso derivare regole associative da esso.
Fare in questo modo significa che, partendo dagli 11 (come visto nell'esempio precedente) o 10 item di interesse nel caso delle patologie, considero le possibili combinazioni di eventi. Posso farlo basandomi su una soglia predefinita o su esperienze passate. È importante avere un set di item di riferimento.

Quindi, prendo il mio dataset di riferimento, genero una serie di combinazioni e applico il "**pruning**," cioè la selezione dei casi che soddisfano una certa soglia di conteggio. Scelgo un valore di soglia, ossia un limite, considerando solo gli item con un conteggio superiore a questo valore per generare le mie regole associative. L'estrazione di regole associative è importante non solo per dati discreti ma anche per dati continui. Ad esempio, quando hai dati numerici come le misurazioni di condizioni mediche, puoi comunque cercare relazioni tra di esse e identificare associazioni. In questo caso, puoi definire soglie per i valori numerici che ritieni rilevanti e significativi.
Per esempio, se hai dati numerici come la pressione sanguigna e stai cercando relazioni tra l'ipertensione indotta dalla gravidanza e l'uso di beta-bloccanti, puoi definire una soglia numerica per la pressione sanguigna al di sopra della quale consideri che ci sia un'associazione positiva tra le due condizioni. Utilizzando queste soglie e le misurazioni numeriche, puoi quindi eseguire un'analisi per generare regole associative.



Questo tipo di analisi è molto importante nel contesto medico e clinico, poiché può aiutare a identificare correlazioni tra diverse condizioni mediche o fattori, il che può essere cruciale per la diagnosi e il trattamento delle malattie.
Spesso, in ambito medico, si lavora con tabelle e dati numerici per cercare relazioni e associazioni significative.
La stessa cosa si può utilizzare nell'ambito, ad esempio, della classificazione. Si prendono i propri dati e si estraggono regole associative, collegandole alle condizioni di classificazione, là dove l'elemento di classificazione è legato alla fase di training.