Dispensa di Machine Learning

Federico Luzzi
Christian Uccheddu

Indice

1 Association Analysis

1.1 Introduction (*)

L'analisi di associazione ci riporta al concetto di causalità di variabili: dati certi valori di attributi cosa posso dire del valore di un altro attributo?

Le regole associative ci permettono di prendere delle scelte molto operative in diversi ambiti, in particolare in quello che viene chiamato *market busket analysis*. Il problema del carrello tratta il posizionamento di prodotti in un negozio. Si sa che all'acquisto di un certo prodotto si tende a acquistare altri prodotti connessi, quindi si cercano queste associazioni basandosi sul carrello della spesa dei clienti (appunto market basket) per poi capire come impostare la disposizione sugli scaffali.

Obiettivo: identificare quali siano gli **item associati** per poter prendere delle decisioni. In sostanza si generano **Regole associative** formate da coppie di insiemi *antecedente* e *conseguente*.

es. {Beer} → {swiss cheese}

{antecedente} → {conseguente}

NB: non è una causalità ma un'associazione

L'analisi si basa sullo studio di due diversi dataset:

- Product set: contiene informazioni legate ai prodotti come il nome e il prezzo
- Transaction set: contiene informazioni legate agli acquisti dei clienti, ogni record corrisponde ai prodotti presenti nel carrello del cliente

Si organizza il dataset delle transazioni in formato binario, ovvero ogni colonna indica se in una data transazione un certo prodotto sia o meno presente.

Transaction ID	swiss cheese	cherry coke	bio coke	Peppers	scrambled egg	Pomegranate	strawberries
0	1	0	0	0	1	0	0
1	0	1	0	0	0	0	1
2	0	0	0	1	1	0	0
3	0	0	0	0	0	1	0
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

Figura 1: esempio: transaction set binarizzato

Consideriamo:
$I = \{i_1, i_2, ..., i_d\}$ il set di tutti gli item nel market basket
$T = \{t_1, t_2, ..., t_N\}$ l'insieme di tutte le transazioni

Definizione 1.1. Una collezione di zero o più item è chiamata **Itemset**.

Definizione 1.2. Se un itemset contiene k item è detto **K-Itemset**.

Definizione 1.3. L'itemset che non contiene alcun elemento è detto **empty set**.

Definizione 1.4. La **transaction width** è il numero di item presenti in una transazione

Una transazione t_j contiene l'itemset X, se X è un sottoinsieme di t_j.

Definizione 1.5. Il **Support count** è il numero di transazioni che contentono uno specifico itemset.

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}$$

Definizione 1.6. Una **regola di associazione** viene rappresentata come:
$X \to Y$
dove X e Y sono itemset disgiunti ($X \cap Y = \emptyset$).

Una regola viene valutata in termini di *supporto* e di *confidenza*.

Definizione 1.7. Support determina quanto spesso una regola sia applicabile dato un data set:

$$s\{x \to Y\} = \frac{\sigma(X \cup Y)}{N}$$

Definizione 1.8. Confidence determina quanto frequente Y è presente in una transazione che contiene X (si assume che l'universo sia rappresentato partendo da X):

$$c\{x \to Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Es.
$X = \{swisscheese, cheddar\}$, $Y = \{dietcoke\}$
Assumiamo che:

- $\sigma(x) = 8$ (support count)

- $N = 20$ (numero di transazioni)
- $\sigma(x, y) = 6$

Allora il support e la confidence della regola $X \to Y$ sono:

$$s\{X \to Y\} = \frac{\sigma(X \cup Y)}{N} = \frac{6}{20} = 0.3$$

$$c\{x \to Y\} = \frac{\sigma(X \cup Y)}{\sigma(X)} = \frac{6}{8} = 0.75$$

Perchè utiliziamo il supporto:

- se troppo basso potrebbe esserci un'associazione casuale
- potrebbe non valere la pena seguire associazioni che si applicano in modo poco significativo dal punto di vista dei profitti

Supporto utilizzato per eliminare regole non desiderate e condivide interessanti proprietà che possono essere sfruttate per la ricerca di regole associative efficaci.

La confidenza è molto importante perchè misura l'affidabilità dell'inferenza e:

- un alta confidenza significa che Y sarà molto presente in transazioni con X
- si stima la probabilità condizionata di Y dato X

Definizione 1.9. Association Rule Mining Problem può essere formalmente definito come: dato un set di transazioni T, trovare tutte le regole con $support \geq minsup$ e $confidence \geq minconf$, dove $minsup$ e $minconf$ sono i threshold corrispondenti alle due misure.

Approccio a forza bruta di association rule non è molto praticabile in quanto i tempi di computazione aumentano in modo esponenziale:
$R = 3^d - 2^{d+1} + 1$.

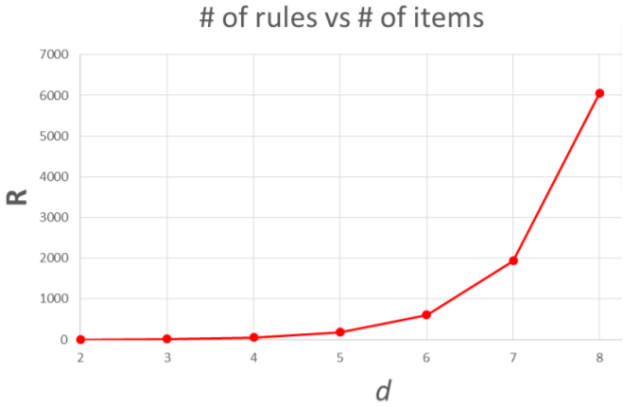


Figura 2: numero di regole calcolate in base al numero di itemset

Una strategia comunemente adottata in molti algoritmi è decomporre il problema in 2 grandi supertask:

- *Generazione dei frequenti itemset*: specifichiamo tutte e sole quelle regole per cui il supporto è maggiore del *minsup*, gli itemset generati sono chiamati **Frequent Itemset**
- *Generazione delle regole*: estriamo tutte le regole con alta confidenza (maggiore di $minconf$) dai Frequent Itemset trovati precedentemente, queste regole vengono chiamate **Strong Rules**

La complessità maggiore è richiesta dalla generazione dei Frequent Itemset.

1.2 Rule Extraction

Per comprendere l'inefficienza della generazione con l'approccio forza bruta pensiamo a questo esempio:

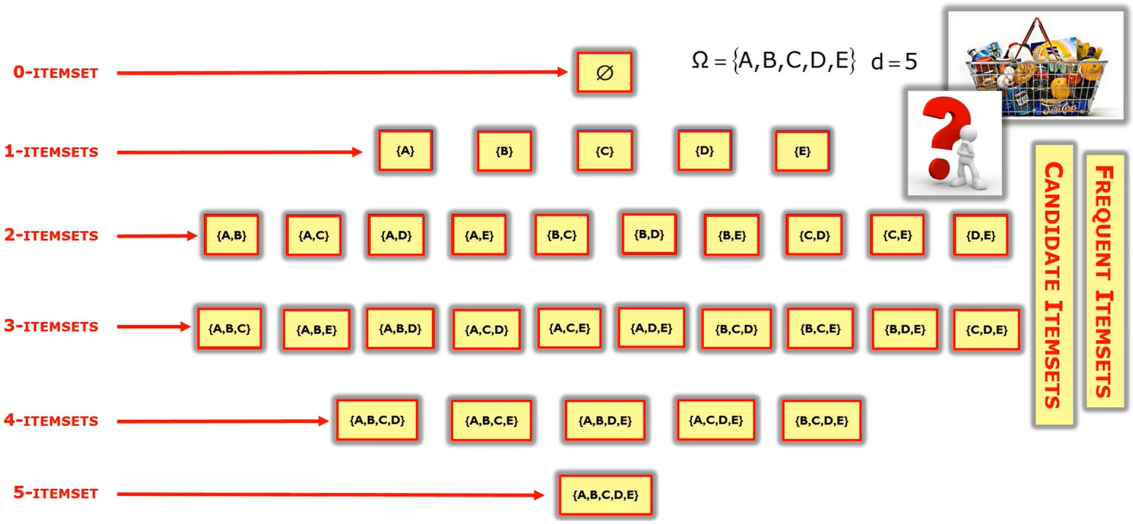


Figura 3: k-itemset brute-force

Definizione 1.10. Candidate Itemset: l'insieme di tutti gli itemset che possiamo formare. Avranno un numero di item diverso.

Nel nostro caso il numero di itemset candidato è: $M = 2^d - 1 = 2^5 - 1 = 31$

Come si può notare abbiamo un sistema a doppio cono che è tipica delle distribuzioni binomiali. Una volta che gli abbiamo considerati tutti ci interessano solo i più frequenti.

Se usassimo la forza bruta dovremmo calcolare per ogni itemset candidato il suo support count, e vedere se il suo supporto lo configura come un itemset frequente (molto dispendioso).

I confronti da effettuare sono nell'ordine di $O(NMw)$, dove:

- N = numero di transazioni
- M = numero di itemset candidato
- w = massima lunghezza delle transazioni

È decisamente troppo come numeri confronti contando che molti dei quali sono inutili o poco significativi.

Vi sono due approcci per ridurre il costo computazionale della generazione di itemset frequenti:

- ridurre il numero di candidati itemset (M). Il principio Apriori è un metodo per eliminare alcuni candidati itemset senza contare il support count.
- riduce il numero di confronti anzichè controllare tutte le possibili combinazioni, lo si fa con strutture dati avanzate

Principio Apriori se un itemset è frequente, allora tutti i suoi sottoinsiemi sono frequenti.
Quindi se una regola ha una frequenza bassa allora tutte le regole che prevedono come sottoinsieme la stessa non supereranno quella frequenza pertanto è inutile considerarle. Si procede attraverso il **pruning** dell'albero delle sequenze per queste soluzioni, viene chiamato **support-based pruning** (vedi immagine).

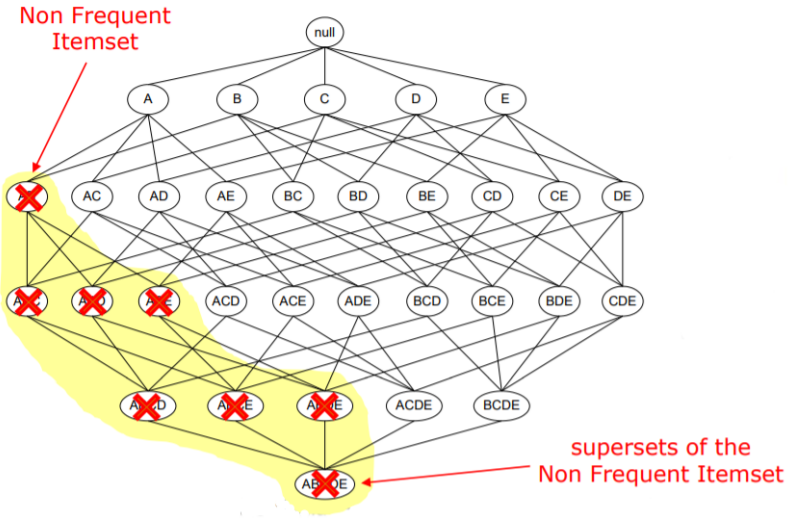


Figura 4: esempio di support-based pruning

1.2.1 Algoritmo apriori

Genera due operazioni:

- **Candidate Generation**: genera nuovi candidati k-itemset basati su (k-1)-itemset frequenti calcolati nella precedente iterazione
- **Candidate Pruning**: questa operazione elimina alcuni candidati k-itemset usando la strategia del support-based pruning

La complessità computazionale soffre di 4 limiti:

- *Support threshold*: la soglia di supporto se troppo bassa non taglio molto l'albero, però non deve essere neanche troppo elevata altrimenti non considero associazioni rilevanti.
- *Numero di item (dimensionalità)*: se il numero di item cresce, ci sarà bisogno di più spazio in memoria per registrare il support count degli

item, inoltre bisogna considerare anche il costo dell'I/O per passare i dati.

- *Numero di transazioni*: l'algoritmo scorre più volte tutta la lista di transazioni, pertanto un numero alto di transazioni inficia sui tempi.
- *Avarage transaction width*: per dataset densi la lunghezza media delle transazioni tende ad essere grande. La massima lunghezza degli itemset frequenti tende ad aumentare quindi più sequenze candidato devono essere esaminate durante la generazione e support counting. In aggiunta aumenta il numero di archi traversi nell'albero durante il support counting.

1.3 Maximal/Closed Frequent Itemsets

1.3.1 Rule Generation

Ogni k-itemset frequente, Y può generare al limite $2^k - 2$ regole di associazione.

Una regola di associazione può essere estratta partizionando l'itemset Y in 2 sottoinsiemi non vuoti $\{X\}$ e $\{Y - X\}$, tale che $X \rightarrow Y - X$ soddisfa il threshold di confidenza.

NB: Tutte le regole generate da itemset frequenti sono esse stesse frequenti.

In pratica, il numero di itemset frequenti prodotti da transazioni possono essere molto grandi. È utile identificare itemset rappresentativi e piccoli con i quali derivare gli itemset grandi. Per questo si ragiona in due rappresentazioni:

1. Maximal Frequent Itemset
2. Closed Frequent Itemset

Definizione 1.11. Maximal Frequent Itemset è definito come un Itemset Frequente per il quale nessuno dei suoi soprainsiemi immediati sono frequenti.

Maximal Frequent Itemset: per ogni nodo si verifica il vincolo della frequenza e si definisce la frontiera dove un nodo non gode più di questa proprietà, corrisponde alla massima frontiera in cui mi posso spingere.

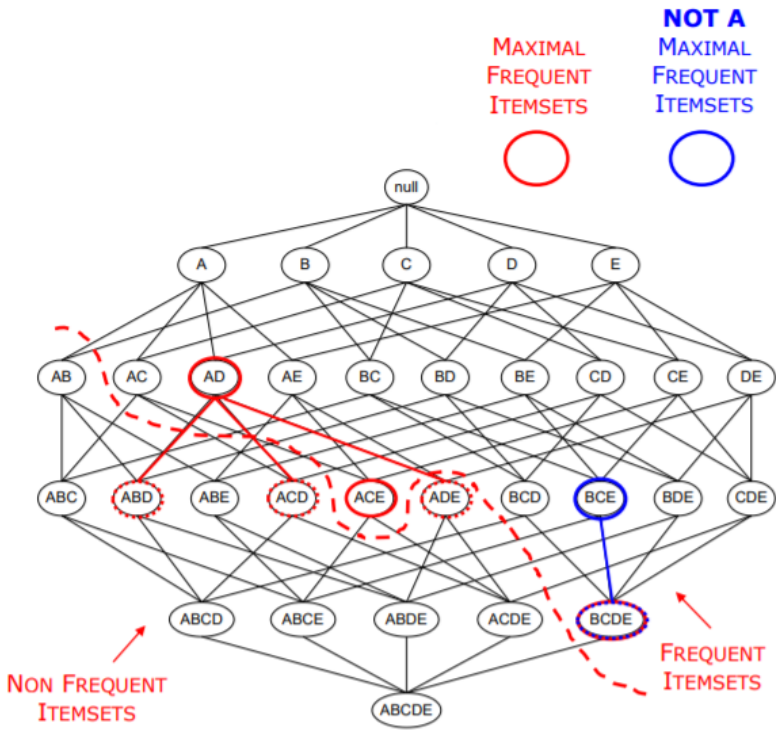


Figura 5: esempio di frontiera di maximal frequent itemset

Nella pratica: Un nodo è un Maximal Frequent Itemset se è frequente e se tutte le sue estensioni non sono frequenti.

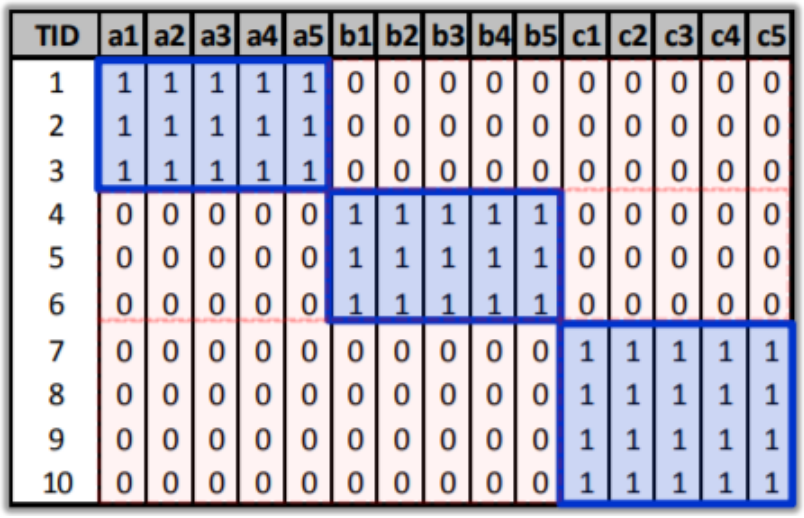
- Fornisce una rappresentazione compatta dell'insieme di itemset che cerchiamo, la più piccola espressione in cui gli itemset sono derivabili
- Calcolato dal più piccolo insieme di itemset dal quale tutti gli itemset frequenti possono essere derivati
- È praticabile solo l'algoritmo efficiente usato esplicita la ricerca dei maximal frequent itemset senza numerare tutti i suoi sottoinsiemi

Per costruzione *però* non ci dice quanto è il supporto rispetto ai suoi sottoinsiemi. In alcuni casi potrebbe servire avere una minima rappresentazione degli itemset frequenti che preservano l'informazione sul supporto.

Definizione 1.12. Un itemset X è **Closed Frequent Itemset** se nessun immediato superinsieme ha esattamente lo stesso support count di X. In ogni caso il suo spporto deve essere $\geq minsup$.

Importante qunado vi sono gruppi di prodotti venduti a blocco ignorando gli altri.

Es



TID	a1	a2	a3	a4	a5	b1	b2	b3	b4	b5	c1	c2	c3	c4	c5
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figura 6: esempio di gruppi closed frequent itemset

Come si può notare in figura ogni gruppo di variabili (Gruppo A, B e C) è perfettamente associato e non hanno bisogno di mostrare items collegati ad altri gruppi. Assumendo che il $minsup = 20\%$, il numero totale di itemset frequenti è: $3 \cdot (2^5 - 1) = 93$. In ogni caso vi sono solo 3 closed frequent itemset:

$\{a1, a2, a3, a4, a5\}$

$\{b1, b2, b3, b4, b5\}$

$\{c1, c2, c3, c4, c5\}$

Questo tipo di itemset sono utili per rimuovere **regole di associazione ridondanti**.

Definizione 1.13. Una regola di associazione $X \rightarrow Y$ è **ridondante** se esiste un'altra regola $X' \rightarrow Y'$ che rispetta certe proprietà.

- $X \subseteq X'$
- $Y \subseteq Y'$
- $s(X \rightarrow Y) = s(X' \rightarrow Y')$
- $c(X \rightarrow Y) = c(X' \rightarrow Y')$

Mostriamo ora la gerarchia dei frequent itemset:

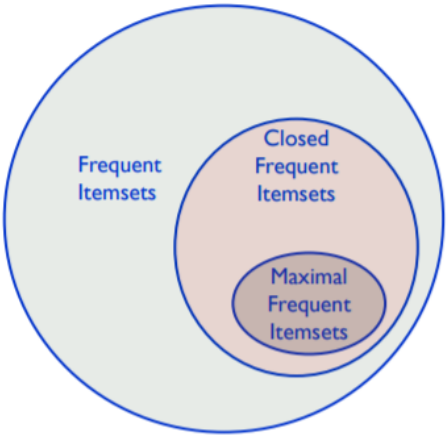


Figura 7: gerarchia dei frequent itemset

Come si può notare i Maximal Frequent Itemset sono inclusi nei Closed Frequent Itemset perchè nessun maxmal può avere lo stesso support count del suo immediato superset.

1.4 Rules Evaluation (*)

Generati gli insiemi di pattern potenzialmente utili, bisogna ordinarli in base al loro livello di attrattività per il dominio di applicazione. La valutazione della qualità viene effettuata seconodo due criteri:

- **Statistical arguments**: per patterns di items indipendenti e coperti da poche transazioni. Il problema è che *possono catturare relazioni spurie* nei dati, per evitare:
  - Objective Interestingness Measure: usare statistiche derivate dai dati per determinare quali pattern sono interessanti
  - Supporto, confidenza e correlazione
- **Subjective arguments**: un pattern è considerato non interessante a meno che riveli informazioni inaspettate riguardo ai dati e alla conoscenza: es. forte associazione tra acquisto di pannolini e birra. È difficile fare questo tipo di valutazioni, richiede una considerevole quantità di informazioni pregresse dagli esperti di domino.

Meglio cercare di applicare un approccio più *oggettivo* alla valutazione: data-driven, indipendente dal dominio in cui si richiede un minimo input dagli utenti (solo dei threshold) e calcolato basandosi sulle frequenze calcolate in una **Contingency Table**.

	B	not B	
A	f_{11}	f_{10}	f_{1+}
not A	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

Figura 8: contingency table

dove:

- f_{1+} = support count di A
- f_{+1} = support count di B
- N = nro di transazioni totali

Supponiamo che un direttore di un minimarket voglia analizzare la relazione tra le persone che bevono tè e persone che bevono caffè. La seguente contingency table è ottenuta considerando le transazioni disponibili:

	coffee	**not coffee**	
tea	*150*	*50*	**200**
not tea	*650*	*150*	**800**
	800	**200**	1,000

Figura 9: contingency table

Ora valutiamo l'associazione: tea $\rightarrow$ coffee

- $support = \frac{f_{11}}{N} = \frac{150}{1000} = 15\%$
- $confidence = \frac{f_{11}}{f_{1+}} = \frac{150}{200} = 75\%$

Ad una prima occhiata sembrerebbe che le persone che bevono tè tendono a bere anche caffè (vedi confidenza). Ma se notiamo le persone che bevono caffè, a prescindere dal tè sono l'80% ($\frac{800}{1000}$), mentre la frazione dei bevitori di tè che bevono caffè è solo il 75%.

Da questo ragionamento si può concludere che il fatto di bere tè non influisca sulle persone che bevono caffè. Infatti nonostante l'associazione abbia un alto livello di confidenza (75%) non si può ignorare il supporto dell'itemset conseguente (80%).

Pertanto, vengono definiti altri indici:

Definizione 1.14. Lift: tasso di confidenza rispetto al supporto del conseguente

$$Lift = \frac{c(A \to B)}{s(B)}$$

Definizione 1.15. Interest Factor: equivalente al *Lift* ma per attributi binari, è definito in questo modo:

$$I(A, B) = \frac{s(A, B)}{s(A)s(B)} = \frac{Nf_{11}}{f_{1+}f_{+1}}$$

I valori sono così classificati:

- $= 1$ se A e B sono indipendenti
- > 1 se A e B sono positivamente associati
- < 1 se A e B sono negativamente associati

Definizione 1.16. Analisi di correlazione (per attributi binari simmetrici): analizza la relazione tra coppie di attributi. Attributi continui possono essere analizzati con la correlazione di Pearson, la correlazione per attributi binari è misurata usando il ϕ-coefficient:

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

il valore varia da $[-1, +1]$. Se gli attributi sono statisticamente indipendenti il valore è 0.

Definizione 1.17. IS Measure (per attributi binari asimmetrici)

$$IS(A, B) = \sqrt{I(A, B)s(A, B)} = \frac{s(A, B)}{\sqrt{s(A)s(B)}}$$

Il suo valore è grande quanto l'*Interest Factor* e il supporto sono grandi. Se A e B sono indipendenti:

$$IS(A, B) = \sqrt{s(A)s(B)}$$

ha lo stesso problema della correlazione, il valore può essere grande anche per associazioni incorrelate o negativamente correlate.

Vi sono altri tipi di misure per l'analisi delle relazioni tra coppie di variabili binarie.

Definizione 1.18. Misure Simmetriche
Una misura **M** è **Simmetrica** se: $M(A \to B) = M(B \to A)$
es. Interest Factor è simmetrico

Definizione 1.19. Misure Asimmetriche
Una misura **M** è **Asimmetrica** se: $M(A \to B) \neq M(B \to A)$
es. la Confidenza è asimmetrica

Di seguito gli indici più utilizzati:

Simmetrico	Asimmetrico
Correlazione (ϕ)	Gini index
Odds ratio	Mutual information
Kappa	Certainty factor
Interest (I)	Added value
Cosine (IS)	J-measure
Jaccard	Goodman-Kruskal
Collective strength	

È importante capire che ciascuna misura è adatta per analizzare un certo tipo di associazioni come basket market analysis o document analysis. In base al caso bisogna utilizzare gli indici migliori.

NB: sono stati presentati indici relativi a valutazioni per coppie di attributi binari, ma è possibile estendere l'analisi a più di due attributi usando tabelle delle frequenti in una Contingency Table multi-dimensionale. Gli indici come il Supporto Interest Factor e IS si prestano a ciò.

1.5 Simpson's Paradox

Consideriamo la seguente situazione:

un direttore di un mini-market racconta di una curiosa scoperta, legata agli item birra e hot dogs. Una società di consulenza da lui pagata per analizzare i suoi dati di vendita ha scoperto che i clienti che comprano birra sono meno tentati di acquistare hot dogs rispetto a quelli che non acquistano la birra. Il direttore però è convinto del contrario, lo sa per esperienza professionale che chi acquista birra tende ad acquistare hot dogs.

Come si può risolvere il paradosso?
Consideriamo la seguente tabella:

	hot dogs	NOT hot dogs	Total
beer	70	98	168
NOT beer	45	55	100
	115	153	

Figura 10: es. birra - hot dog

Consideriamo le seguenti regole con le relative confidenze:

- {beer} → {hot dogs} - confidence = 42%
- {NOT beer} → {hot dogs} - confidence = 45%

Possiamo inferire che i clienti che acquistano birra sono meno inclini (42%) ad acquistare hot dog rispetto a quelli che non acquistano birra (45%).

Analizziamo ora gli stessi dati ma categorizzando se il cliente è single o no:

		hot dogs	NOT hot dogs	Total
single	beer	30	10	40
	NOT beer	5	40	45
NOT single	beer	40	88	128
	NOT beer	40	15	55

Figura 11: es. cliente single: birra - hot dog

Pertanto posso derivare queste due coppie di regole:

- Single
  - {beer} → {hot dogs} - confidence = 75%
  - {NOT beer} → {hot dogs} - confidence = 11%
- NOT Single
  - {beer} → {hot dogs} - confidence = 31%
  - {NOT beer} → {hot dogs} - confidence = 73%

Da questi dati posso affermare che: i single che acqustano birra sono più inclini (75%) ad acquistare anche hot dog rispetto a quelli che non acquistano birra (11%).

Quindi sia l'azienda di consulenza che il direttore del mini-market avevano ragione soltanto che per comprenderlo bisognava categorizzare i clienti.

Paradosso di Simpson o Yule-Simpson effect, è un paradosso della probabilità e statistica, in cui una tendenza appare in diversi gruppi di dati ma sparisce o si inverte quando questi gruppi sono combinati.

Bisogna applicare una appropriata stratificazione dei dati per evitare la generazione di associazioni spurie risultati dal paradosso di Simpson.

Es. per i dati del market-basket, la catena di supermercati dovrebbe stratificarli secondo la location del negozio, mentre i dati medici da vari pazienti dovrebbero essere stratificati secondo fattori quali l'età o il genere.