Column-aware Association Rules Mining with Dynamic Support

Problem Description

Nella ricerca delle regole di associazione vengono di solito utilizzate vincoli di soglia come il *supporto* e la *confidenza*.

Il supporto statistico misura la rilevanza statistica di una possibile regola di associazione; esso rappresenta il numero minimo di transazioni in cui sono presenti tutti i valori appartenenti ad un sottoinsieme X, denominato itemset.

T1	Milk, Ink, Juice
T2	Milk, Beer
T3	Milk, Cofee, Beer

Ad esempio si consideri l'istanza di database mostrata nella tabella x, in cui sono elencate le transazioni degli acquisti in un supermercato. Se consideriamo l'itemset composto da $X = \{Milk, Beer\}$, allora il supporto statico $supp(X)$ è pari a $2/3$; se, invece, consideriamo l'itemset composto da $X = \{Milk,\ Juice\}$, allora il suo supporto statistico $supp(X)$ è pari a $1/3$. Per cui specificando apriori il vincolo di soglia del supporto è possibile prendere in considerazione, per la definizione delle regole di associazione, soltanto quegli itemset che superano tale soglia. Ad esempio, se si vogliono ricercare regole di associazione con un supporto almeno pari al 50%, allora sei due itemset considerati in precedenza solo il primo potrà essere considerato.

La confidenza, invece, misura la significatività di una regola $X \Longrightarrow Y$, essa può essere calcolata mediante la seguente formula

$$conf(X \Longrightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Anche in questo caso, specificando apriori la soglia, è possibile valutare le sole regole ritenute significative con quel vincolo di confidenza. Nell'esempio su citato, se si impone una soglia di confidenza pari a 1 allora la regola $\{Milk\} \Longrightarrow \{Juice\}$ non potrà essere considerata valida in quanto la sua confidenza è pari a $2/3$.

In altre parole, negli algoritmi per la ricerca di regole di associazione, il vincolo di supporto viene utilizzato per individuare gli itemset, denominati frequent itemset, che hanno una certa rilevanza statistica per poter creare le regole, una volta ottenuti tali frequente itemset, le possibili regole devono superare il vincolo di significatività.

È importante notare che, nella costruzione dei frequent itemset con l'ausilio del vincolo di soglia di supporto, tutti i valori vengono considerati con la stessa probabilità di accadimento. Ad esempio un prodotto acquistato ha la stessa probabilità di accadimento di tutti altri prodotti. Purtroppo in alcuni contesti, esistono condizioni e/o casi particolari che rappresentano la presenza di un valori specifici, i quali non possono essere considerati equiprobabili.

Supponiamo di voler fare basket analysis sulle transazioni di acquisto in un supermercato dove gli elementi vengono memorizzati ad un livello di dettaglio più fine rispetto alle semplici diciture "latte, inchiostro, birra", ovvero per il latte si possono ottenere valori come "natular whole milk, whole standardised milk, whole homogenised milk, semi skimmed milk, skimmed milk, 1% fat milk", invece per inchiostro si hanno i valori "black ink, blue ink, red ink", infine per birra i valori possibili sono "light beer, dark beer" (Table y).

T1	skimmed milk, red ink, juice, light beer
T2	natural whole milk, dark beer
T3	low-fat milk, coffe, black ink
T4	skimmed milk, blue ink, light beer
T5	blue ink, light beer
T6	whole homogenised milk, dark beer

Considerare la costruzione dei frequent itemset con valori memorizzati a questo livello di dettaglio e un soglia di supporto statistico fissa potrebbe fornire informazioni grossolane che non prendono in considerazione la variabilità del supporto statistico sulla base della probabilità di accadimento. Infatti, la probabilità che venga scelta una specifica tipologia di latte, ad esempio "natural whole milk" ha una probabilità di accadimento di $P(\text{natural whole milk}) = \frac{1}{6}$, mentre la probabilità di accadimento che venga scelta una specifica tipologia di birra, ad esempio "dark beer" è di $P(dark\ beer) = \frac{1}{2}$.

Il precedente esempio fa notare come in alcuni casi specifici la misura di rilevanza statistica considerata come soglia fissa, non incorpora la possibilità di avere a che fare con valori che non siano equiprobabili. Nel caso specifico dell'esempio, è come se si considerasse la probabilità di avere "testa" nel lancio di una moneta uguale alla probabilità di ottenere "uno" nel lancio di un dado.

Quando si considerano come dati di input tabelle strutturate di database, è facile ritrovare la probabilità di accadimento di possibili valori considerando la variabilità degli stessi all'interno di una colonna. …

The Proposed Approach

Sulla base delle considerazioni fatte nella precedente sezione, si scaturisce che in contesti come quello analizzato è importante dare valore alla probabilità di accadimento di uno specifico valore rispetto alla variabilità dei possibili valori in una colonna.

Per poter effettuare ciò, non può essere considerato un vicolo di soglia fisso nel supporto statistico, piuttosto deve essere dinamico e si deve basare sulla probabilità di accadimento dei valori, che come abbiamo visto è catturabile dalla variabilità dei valori nelle colonne.

In particolare, consideriamo che la soglia di supporto statistico definita, rappresenti la soglia associabile ai valori che presentano la massima probabilità di accadimento, e che questa venga

modificata in valori specifici correnti sulla base della probabilità di accadimento del valore che si sta analizzando per poter definire la soglia di supporto statistico correlata al valore corrente. Formalmente quest'ultimo può essere definito mediante la seguente formula.

$$CurrentSupport = \frac{SpecifiedSupport * CurrentValueProbability}{MaxValueProbability}$$

Supponiamo ad esempio che si vogliano cercare regole di associazione nella tabella y con una rilevanza statistica (supporto) di almeno il 40%. Nella costruzione dei frequent itemset uno dei valori che sicuramente superano questa soglia è {dark beer} essa infatti è presente tre volte raggiungendo un supporto del 50%. È facile notare che questo, con un supporto fisso, rappresenterebbe l'unico frequent itemset preso che la supera. Andando ad analizzare le probabilità di accadimento, però, si può notare che questo frequent itemset ha una probabilità di accadimento di ½, la quale rappresenta anche la probabilità massima. Per cui è possibile considerare un supporto dinamico per quei valori che hanno una probabilità di accadimento più bassa. Consideriamo ad esempio l'itemset {natural whole milk}. Secondo la regola definita il suo supporto diventa:

$$CurrentSupport\{\text{skimmed milk}\} = \frac{2/5 * 1/6}{1/2} = \frac{2}{15} = 13\%$$

Questo implica il fatto che questo itemset può essere considerato come frequente sulla base del supporto dinamico.