Strumenti di studio del portfolio prodotti: analisi associativa

L'analisi e la gestione del portafoglio prodotti riveste un'importanza estrema, analoga a quella relativa all'analisi e gestione del portafoglio clienti. Attraverso lo studio combinato prodotti-clienti e le varie correlazioni esistenti tra la customer base e il pacchetto referenze posseduto dall'impresa questa potrà scorgere inesplorate opportunità di incentivazione commerciale o valutare la possibilità di eliminare dal portfolio assortimentale i c.d. «erosori» di margine

a cura di **Di Giovanna R. Contaldo** e **Tommaso Largo**

La possibilità di effettuare delle previsioni statisticamente rilevanti circa le dinamiche di acquisto future della propria base clienti (customer base) transattiva rappresenta un'opportunità che da sempre affascina manager e uomini di marketing. Proprio sulla stregua di questa forte esigenza, sempre più sentita dal management delle aziende industriali e di quelle distributive, che negli ultimi anni si sono sviluppate e diffuse molte tecniche di analisi ed estrazione dei dati (tecnicamente tale processo viene identificato con la locuzione anglosassone «data mining») che permettono di analizzare e «stanare» tutti i legami, più o meno visibili, esistenti in quelle che possiamo definire le dinamiche di acquisto del consumatore.

Con il presente lavoro cercheremo di approcciare in maniera elementare una delle metodologie più utilizzate per lo studio del portfolio referenze ed in particolare per l'analisi delle relazioni esistenti tra i molteplici atti di acquisto generati dalla customer base di una qualsivoglia entità economica.

Tale studio è stato identificato con l'acronimo **IAP** ovvero **Indagine Associativa di Portfolio** e, tecnicamente, rappresenta un metodo computazionale di data mining che, utilizzando la base dati delle combinazioni d'acquisto (transazioni), verificatesi in un determinato periodo di tempo ed in un certo luogo, effettuate da una clientela avente determinate caratteristiche comuni, sfrutta un algoritmo (e quindi un metodo sistematico di calcolo) formato da un'insieme di procedure ed operazioni che riducono le molteplici combinazioni d'acquisto in modo da identificare e dare risalto solo ed esclusivamente a quelle statisticamente rilevanti.

Nel nostro caso, l'Indagine Associativa di Portfolio dovrebbe evidenziare le combinazioni d'acquisto più interessanti per tutti quei soggetti che periodicamente devono prevedere le vendite di un portfolio referenze (forecasting), effettuare la pianificazione commerciale (budgeting), fissare gli obiettivi da raggiungere (targeting) ed organizzare l'attività promozionale da destinare a tali referenze.

L'Indagine Associativa di Portfolio rappresenta l'anticamera della cosiddetta **Market Basket Analysis (MBA)** che, infatti, ne rappresenta l'evoluzione in quanto individua referenze o gruppi di referenze che tendono ad essere acquistate assieme costruendo dei veri e propri modelli comportamentali associativi che lavorano sulle combinazioni d'acquisto risultanti proprio dall'indagine oggetto di studio.

L'importanza della IAP (e quindi della MBA) è evidente: conoscendo, infatti, le referenze associate è possibile riorganizzare il layout di un supermercato in maniera da facilitare la shopping experience del consumatore che, in tal guisa, troverebbe i prodotti collegati posizionati all'interno di scaffali attigui (queste analisi, infatti, sono molto utilizzate negli studi di category management con la finalità di massimizzare il margine complessivo per unità di spazio espositivo di un aggregato di referenze costituenti una categoria, p.e. primi piatti, snack salati, ecc.); inoltre, avendo una preventiva conoscenza delle combinazioni d'acquisto probabilisticamente più frequenti, è possibile incrementare l'efficacia delle attività promozionali attraverso un'opportuna strategia di bilanciamento della base promozionata che, in tal guisa, non avrà al proprio interno combinazioni di referenze che solitamente vengono acquistate assieme: se chi acquista il prodotto A, di

solito acquista anche il prodotto B, allora è inutile mettere contemporaneamente in promozione le due referenze in quanto i maggiori acquisti di B potrebbero non dipendere dalle attività promozionali destinate allo stesso B ma da quelle veicolate su A.
Chiaramente quando le due referenze oggetto dell'attività promozionale sono, ad esempio, lo «Shampoo per capelli grassi» ed il «Balsamo per capelli grassi» della stessa azienda cosmetica, il regime di complementarietà che li caratterizza rende quasi ovvia la considerazione che sarebbe alquanto inopportuno promuovere due prodotti che per reciprocità funzionale, appartenenza alla medesima linea, occasione ed intervalli di consumo saranno sicuramente acquistate assieme.
Molto meno immediato potrebbe essere, invece, dover analizzare l'esistenza di eventuali legami associativi tra prodotti molto diversi tra loro (in termini funzionali, di destinazione d'uso, di occasione ed intervallo di consumo, ecc.). Se prendessi in considerazione, ad esempio, i «crackers salati» ed il «latte» sicuramente avremmo pochi argomenti, intuitivamente logici, che ci permetterebbero di individuare una regola associativa d'acquisto tra le due referenze: eppure, in molte indagini associative di portfolio, effettuate in supermercati con superfici superiori a 1.200 mq, si è riscontrata la presenza di un legame molto forte esistente tra i due prodotti. Pare che in media, negli scontrini in cui sono presenti i crackers, per il 75/85% delle volte è anche presente il latte!
Di seguito abbiamo cercato di spiegare come, anche una Pmi (in questo caso ne abbiamo considerata una operante nel settore distributivo dei beni di largo consumo) possa, utilizzando dei rudimenti statistici, impostare, in maniera «artigianale», un'Indagine Associativa di Portfolio che, per grandi linee, potrà tornarle utile in tutta una serie di decisioni che solitamente la stessa potrebbe dover prendere, soprattutto in termini di scelte in ambito promozionale.
La decisione di studiare le transazioni di un'azienda operante nel settore distributivo sono molteplici e sono dettate fondamentalmente dalla necessità di avere sia un dataset (ovvero un insieme selezionato di dati provenienti da un database) quanto più ampio ed eterogeneo possibile sia perché risulta quasi sempre più interessante e significativo analizzare le dinamiche di acquisizione di soggetti che sono spinti all'acquisto da tutta una serie di motivazioni, più o meno tangibili, talvolta confinate nella cosiddetta «customer black box» ovvero in quell'area, pressoché sconosciuta all'azienda, in cui si formano le scelte non razionali di ogni consumatore.
È chiaro che un'azienda che si accinge ad effettuare, anche in maniera artigianale, un'analisi di questo tipo non possa essere assolutamente sprovvista di una qualche forma, ancorché embrionale, di sistema informativo. In particolare è fondamentale che la stessa sia dotata:
(1) di un sistema di identificazione e rilevamento automatizzato delle referenze (perlomeno tramite bar code);
(2) di un sistema di identificazione della customer base tramite card ad immissione manuale o magnetica (questo aspetto risulta fondamentale se, ad esempio, si volessero incrociare le informazioni derivanti dall'Indagine Associativa di Portfolio con quelle di gruppi di clienti aventi le medesime caratteristiche socio-demografiche. Al riguardo si provi ad immaginare quale ritorno si potrebbe avere in termini di cross-selling se decidessi di declinare la base promozionata in maniera differente a seconda del profilo posseduto dal gruppo di clienti-target);
(3) di un sistema POS (Point of Sales) che permetta di monitorare non solo data e valore complessivo delle singole transazioni ma soprattutto chi le compie e quali combinazioni d'acquisto perfeziona.
L'analisi e la gestione del portfolio referenze, quindi, riveste un'importanza estrema che si combina a quella del portfolio clienti per dare a quest'ultimi risposte sempre più convincenti e profilate in termini di personalizzazione dell'offerta e più in generale di marketing mix. Invero, attraverso lo studio combinato prodotti-clienti nonché delle varie correlazioni esistenti tra la customer base e il pacchetto referenze posseduto dall'impresa, questa non solo potrà scorgere nuove opportunità di business (incremento dei ricavi) ma anche valutare la possibilità di eliminare dall'assortimento qualche referenza che, indipendentemente dai ricavi che produce, tende a distruggere margini senza, ad esempio, incentivare la vendita di altri prodotti che, ad essa, dovrebbero essere associati (è il caso delle vendite sottocosto che dovrebbero essere effettuate utilizzando prodotti c.d. «civetta» al fine di incentivare la venuta del cliente nel punto vendita nonché l'acqui-

sto delle referenze statisticamente associate al prodotto «civetta» stesso e che, naturalmente, non dovrebbero essere in promozione!). Ecco, quindi, che lo studio delle relazioni prodotti-clienti acquista importanza anche ai fini della definizione ed attuazione di strategie promozionali che si basano su vendite incrociate, bandle di prodotti, ecc.

L'indagine associativa di portfolio: gli indicatori

Dal punto di vista definitorio, l'Indagine Associativa di Portfolio permette di studiare le regole associative esistenti nei processi di scelta di una particolare combinazione d'acquisto tra tutte o parte delle transazioni effettuate dalla customer base transattiva. Nella sua versione più semplice, ovvero quella binaria, essa permette di «scovare» tutte quelle relazioni con la maggiore probabilità di «accadimento combinato»: in altre parole, l'indagine permette di individuare quegli abbinamenti di referenze per le quali, dati due prodotti (A e B) ed acquistato uno dei due (ad esempio il prodotto A), l'acquisto anche dell'altro (e dunque l'acquisto congiunto di A e B) è più probabile dell'acquisto esclusivo dell'altro prodotto (nel nostro caso B).
L'analisi sopra citata si basa sull'impiego delle c.d. regole associative, uno dei più diffusi metodi computazionali di data mining, e serve proprio a misurare l'affinità associativa esistente tra due (binaria) o più referenze (multipla) sia in generale (in questo caso le informazioni non vengono incrociate con quelle relative alla customer base) sia correlando i dati relativi alle transazioni con uno o più clienti aventi un determinato profilo socio-demografico e/o psicografico.
Chiaramente prima di cominciare l'analisi è necessario predisporre la base dati (il c.d. Dataset, Tavola 1). Solitamente è necessario identificare preventivamente:
(1) un **periodo di riferimento all'interno del quale effettuare l'analisi**. Al riguardo si è soliti identificare periodi omogenei al loro interno in modo da non mescolare transazioni che solitamente vengono influenzate dalle condizioni climatiche o da eventi convenzionalmente collegati al movimento di rivoluzione terrestre (ricorrenze, ecc.). In questo caso sono state prese in considerazione le tre settimane terminanti il 28/01/2007;
(2) un **punto vendita specifico** in modo da non mescolare transazioni che potrebbero essere influenzate da eventi strettamente collegati al profilo geografico di appartenenza. Nel caso oggetto di esame è stato considerato il punto vendita identificabile con il codice P010 che riguarda una superette di 240 mq presente in provincia di Lecce;
(3) una **serie di transazioni** per le quali sarebbe opportuno conoscere: il progressivo transazione, la data di accadimento della transazione, il cliente che l'ha effettuata, le referenze che ha acquistato. In questo caso ad ogni riga corrisponderebbe una transazione e quindi una data, un cliente e una colonna per ogni referenza oggetto di analisi (nel nostro caso abbiamo preso in considerazione solo le categorie PASTA, LATTE, BISCOTTI, CRACKERS, RISO, TONNO, CAFFE e ORZO perché è su queste specifiche classi di prodotto che abbiamo intenzione di strutturare la nostre base promozionata da inserire nel prossimo volantino). Per ognuna di queste referenze, in corrispondenza di ogni transazione, dovrebbe esserci o lo «0» (referenza non acquistata) o l'«1» (referenza acquistata). Chiaramente, considerando solo il novero delle referenze scelte come oggetto dell'indagine, potrebbe accadere che, per alcune transazioni, tali categorie siano tutte in modalità «referenza non acquistata» (vedi ad esempio la transazione T003454).

TAVOLA 1 - DATASET UTILIZZATO COME BASE DATI

TRANS I	DATA	PDV	CLIENTE	PASTA	LATTE	BISCOTT	CRACKER	RISO	TONNO	CAFFE'	ORZO
T000001	08/01/2007	P010	C0001	0	1	0	1	0	1	1	0
T000002	08/01/2007	P010	C0322	1	1	0	0	0	1	0	1
T000003	08/01/2007	P010	C0311	0	1	0	0	0	1	1	1
T000004	08/01/2007	P010	C0007	0	1	1	1	0	1	0	1
T000005	08/01/2007	P010	C0010	0	0	0	0	0	0	1	1
T000006	08/01/2007	P010	C0034	1	0	1	0	1	0	0	1
T000007	08/01/2007	P010	C0001	1	0	0	1	1	0	1	0
T000008	08/01/2007	P010	C0034	1	1	1	1	1	0	0	0
T000009	08/01/2007	P010	C0311	0	1	1	0	1	1	0	0
T000010	08/01/2007	P010	C0001	0	1	0	0	0	1	0	1
T000011	08/01/2007	P010	C0007	1	0	0	0	0	1	1	1
T000012	08/01/2007	P010	C0999	0	0	1	1	0	1	1	0
T000013	08/01/2007	P010	C0002	1	0	1	1	0	1	1	0
T000014	08/01/2007	P010	C0034	0	1	1	1	0	0	0	1
T000015	09/01/2007	P010	C0001	0	0	0	1	0	0	0	1
T000016	09/01/2007	P010	C0311	1	1	0	0	1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...
T003454	27/01/2007	P010	C0939	0	0	0	0	0	0	0	0
T003455	27/01/2007	P010	C0311	0	1	1	0	1	1	0	0
T003456	27/01/2007	P010	C0845	1	0	1	1	0	1	1	0

A questo punto utilizzando tre semplici indicatori è possibile misurare la forza della regola associativa che lega le referenze oggetto di indagine e che, nell'analisi associativa binaria, verranno combinate a gruppi di due: questi indicatori prendono il nome di **INDICE DI SUPPORT**, **INDICE DI CONFIDENCE** ed **INDICE DI LIFT** (detti anche indici di «interesse statistico» di una regola, Hand-Mannila-Smyth) e, fondamentalmente, l'applicazione e l'interpretazione di tali indicatori sarà l'unico ambito di analisi che affronteremo in questa sede, lasciando ulteriori approfondimenti a lavori futuri.

Indice di support

L'**indice di support** è dato dal rapporto tra il numero di transazioni in cui si verifica congiuntamente l'acquisto di entrambi i prodotti considerati (A e B) e il numero totale di transazioni, ovvero il rapporto tra la frequenza relativa delle osservazioni che soddisfano la regola associativa ed il numero delle osservazioni totali. In formula:

$$\text{Support } \{A \rightarrow B\} = \frac{{}^{N}A \rightarrow B}{N}$$

${}^{N}A \rightarrow B \Rightarrow$ Numero di transazioni in cui è stato acquistato sia A che B;

$N \Rightarrow$ Numero di Transazioni Totali

La freccia indica il verso del legame ovvero spiega l'associazione tra A e B quando il consumatore ha già scelto di acquistare A. Chiaramente, questo primo indice è fondamentalmente simmetrico in quanto il risultato non cambierebbe se decidessimo di invertire il senso della freccia.

Dal punto di vista interpretativo esso identifica la percentuale media di volte in cui, dato un numero N di transazioni, i consumatori hanno acquistato la combinazione (A, B). Traslando prospetticamente il significato dell'indicatore, potremmo dire che lo stesso rappresenta la probabilità che si verifichi questa combinazione d'acquisto in futuro.

Solitamente l'indice di support viene utilizzato per scremare le combinazioni in modo da eliminare quelle poco significative in quanto di raro accadimento (di solito viene fissata una determinata soglia t di esclusione).

Indice di confidence

Chiaramente l'indice di support non rappresenta un indicatore sufficiente, esso solitamente viene combinato con un altro molto significativo che si è soliti applicare in seconda istanza una volta effettuata la scrematura, ovvero l'**indice di confidence**. Questo, in effetti, esprime la probabilità che, acquistato A, il medesimo cliente acquisti anche B: quanto più alto è l'Indice di Confidence tanto più forte sarà la

regola associativa che lega le due referenze. Esso è dato dal rapporto tra il numero di volte in cui si verifica l'acquisto congiunto di A e B e il numero di volte in cui si verifica, in genere, l'acquisto di A. In formula:

$$\text{Confidence } \{A \rightarrow B\} = \frac{{}^{N}A \rightarrow B}{{}^{N}A} = \frac{\text{Support } \{A \rightarrow B\}}{\text{Support } \{A\}}$$

${}^{N}A \rightarrow B \Rightarrow$ Numero di transazioni in cui è stato acquistato sia A che B;

${}^{N}A \Rightarrow$ Numero di Transazioni in cui è stato acquistato A

Chiaramente, in questo caso, l'indice non è simmetrico in quanto la probabilità di acquistare una referenza B quando si è acquistata A, rapportata all'insieme di volte in cui viene acquistata A, può essere diversa dalla probabilità di acquistare A quando ho acquistato B, rapportata all'insieme di volte in cui viene acquistato B. Tale indicatore genera un risultato compreso tra 0 ed 1: quanto più questo si avvicina all'estremo superiore, tanto più è elevata la probabilità che comprando un prodotto A, venga acquistato anche il prodotto B.

Indice di lift

Ma a fugare ogni dubbio, una volta identificati elevati indici di support e di confidence, è **l'indice di lift** che rappresenta una misura di normalizzazione (il risultato della stessa, infatti, è un numero puro) che permette di stabile se tra i due eventi vi sia una correlazione positiva o negativa.

Il lift è dato dal rapporto tra il confidence di $\{A \rightarrow B\}$(ovvero la probabilità che si acquistino congiuntamente le due referenze rispetto alla probabilità che si verifichi, in genere, l'acquisto della referenza A) rapportato al support di B ovvero alla probabilità che si acquisti, in genere, B.

$$\text{Lift } \{A \rightarrow B\} = \frac{\text{Confidence } \{A \rightarrow B\}}{\text{Support } (B)} = \frac{\text{Support } \{A \rightarrow B\}}{\text{Support } \{A\} \cdot \text{Support } \{B\}}$$

Anche il Lift, analogamente al support, è simmetrico e, dunque, uguale sia se analizzo la relazione da A verso B che da B verso A: se superiore ad uno evidenzia una correlazione positiva tra le due referenze, ovvero la probabilità di acquistarli congiuntamente è superiore alle probabilità di acquistare o uno o l'altro, viceversa, se inferiore ad uno, esprime una correlazione negativa.

Un esempio servirà a chiarire meglio l'importanza di tali indici e le proprietà che li caratterizzano.

Ipotizziamo di trovarci in presenza di un insieme di transazioni effettuate presso il medesimo punto vendita e di voler individuare la sussistenza di una qualche associazione tra l'acquisto di due referenze.

Utilizzando il dataset precedentemente costruito ipotizziamo che su **3.456** transazioni totali **754** contengano solo latte, **456** solo biscotti e il numero di transazioni in cui compaiono, invece, entrambi i prodotti sia pari a **309**.

Calcoliamo **l'indice di support tra LATTE e BISCOTTI**:

$$\text{Support (LATTE} \rightarrow \text{BISCOTTI)} = \frac{309}{3456} = 0,08941$$

Esso è chiaramente simmetrico per cui:

$$\text{Support (BISCOTTI} \rightarrow \text{LATTE)} = \frac{309}{3456} = 0,08941$$

Già calcolando questo indice è possibile asserire che la regola associativa che lega le due combinazioni risulta essere abbastanza significativa.

Ciò equivale a dire che nel 9% dei casi circa è presente l'acquisto combinato di (LATTE → BISCOTTI) o (BISCOTTI → LATTE) (proprietà simmetrica).

La regola appare sufficientemente rilevante per poter proseguire l'analisi con la determinazione dell'indice di confidence. La determinazione di tale indice presuppone la conoscenza del support di A e del support di B.

$$\text{Support } \{\text{LATTE}\} = \frac{309 + 754}{3456} = 0,30758$$

$$\text{Support } \{\text{BISCOTTI}\} = \frac{309 + 456}{3456} = 0,22135$$

$$\text{Confidence (LATTE} \rightarrow \text{BISCOTTI)} = \frac{309}{1063} = 0,29069$$

In questo caso si divide il support $\{A \rightarrow B\}$ di per il support di $\{A\}$ che è dato dalla somma di tutti gli acquisti in cui figura la referenza A, dunque 309+754 (N, denominatore del support, si semplifica).

Chiaramente, il **confidence** di (BISCOTTI → LATTE) sarà diverso da quello (LATTE → BISCOTTI) poiché lo stesso non gode della proprietà simmetrica.

$$\text{Confidence (BISCOTTI} \rightarrow \text{LATTE)} = \frac{309}{765} = 0,40392$$

Infine **calcoliamo il lift** di (LATTE → BISCOTTI) e (BISCOTTI → LATTE) che, per la proprietà della simmetria dovrebbe essere uguale.

$$\text{Lift (LATTE} \rightarrow \text{BISCOTTI)} = \frac{0,29069}{0,22135} = 1,31322$$

$$\text{Lift (BISCOTTI} \rightarrow \text{LATTE)} = \frac{0,40392}{0,30758} = 1,31322$$

Proviamo ora a considerare l'ipotesi di due prodotti che presentano un lift minore di uno (indicante dunque una correlazione negativa tra le due referenze), rispetto ai quali la probabilità di acquistarli disgiuntamente o meglio alternativamente è superiore a quella di acquistarli congiuntamente come ad esempio caffé ed orzo.

Ipotizziamo che su **3.456** scontrini totali in **568** vi sia la presenza della referenza caffè e in **1.064** la referenza orzo, mentre entrambe le referenze compaiono in **284** scontrini.

Partiamo con la determinazione del support.

$$\text{Support (CAFFÈ} \rightarrow \text{ORZO)} = \frac{284}{3456} = 0,08218$$

Anche in questo caso il support ORZO → CAFFÈ sarà uguale.

Procediamo con la determinazione del confidence di (CAFFÈ → ORZO)

La determinazione del confidence implica la conoscenza del support di A e del support di B:

$$\text{Support } \{\text{A}\} = \frac{568 + 284}{3456} = 0,24653$$

$$\text{Support } \{\text{B}\} = \frac{1064 + 284}{3456} = 0,39005$$

E dunque:

$$\text{Confidence (CAFFÈ} \rightarrow \text{ORZO)} = \frac{284}{852} = 0,33333$$

Il confidence ORZO → CAFFÈ, invece, sarà pari a:

$$\text{Confidence (ORZO} \rightarrow \text{CAFFÈ)} = \frac{284}{1348} = 0,21068$$

Infine calcoliamo il lift di (CAFFÈ → ORZO) e (ORZO → CAFFÈ) che, per la proprietà della simmetria dovrebbe essere uguale.

$$\text{Lift (CAFFÈ} \rightarrow \text{ORZO)} = \frac{0,33333}{0,39005} = 0,85460$$

$$\text{Lift (ORZO} \rightarrow \text{CAFFÈ)} = \frac{0,21068}{0,24653} = 0,85460$$

Il lift minore di uno evidenzia una correlazione negativa tra le due referenze inducendo dunque alla considerazione che si tratti di due prodotti che difficilmente vengono acquistati congiuntamente.