Università
Ca'Foscari
Venezia

Corso di Laurea magistrale in Marketing e Comunicazione

Tesi di Laurea

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Market basket analysis: il caso dei prodotti "white" nella realtà aziendale Pam

Relatore
Professoressa Daniela Favaretto

Laureando
Alessandra Stevanato
Matricola 830254

Anno Accademico
2013 / 2014

Indice

Indice dei grafici

1. INTRODUZIONE

I processi di marketing riguardanti le decisioni dell'impresa sono caratterizzati da un alto livello di complessità, dovuta alla presenza simultanea di numerosi obiettivi e azioni, che risultano dalla combinazione di tutte le scelte che può prendere colui che si occupa di decidere per l'azienda.

In questo, l'importanza dei modelli matematici per il marketing è andata via via crescendo, grazie allo sviluppo di sempre più efficienti database delle transazioni di vendita, che forniscono accurate informazioni su come i consumatori usano i servizi o acquistano i prodotti.

La customer relationship management (CRM), conosciuta anche come gestione delle relazioni con la clientela, è un concetto che riconduce alla recente attenzione che le aziende hanno rivolto alla propria clientela. Infatti, conoscere in maniera approfondita i clienti, capire chi sono e cosa potrebbero acquistare, e quali sono i contatti con l'azienda, permettono di conoscere al meglio i propri clienti e di conseguenza consentono di avere una più elevata probabilità di conservare nel tempo il rapporto con essi e di offrire loro servizi aggiuntivi.

La CRM è quindi un processo che comporta la comprensione delle attività del singolo cliente e il conseguente adeguamento delle procedure aziendali alle loro esigenze. Il suo vero indicatore di successo va ricercato in un aumento delle vendite, nella risposta alle promozioni, negli indici di soddisfazione della clientela.

Un aiuto alla gestione delle relazioni con la clientela è dato dalla segmentazione della clientela stessa.[1] Essa infatti ha aiutato molte aziende a dividere i buoni clienti da quelli meno buoni, ottenendo delle categorie di clienti a cui dedicare più tempo e più spese; permette quindi di evitare il marketing di massa, dal momento che oggi le aziende cercano sempre più di individuare i loro migliori clienti, cioè quelli che, spendendo di più, procurano un ampio margine di profitto per l'azienda. Individua quindi quelle nicchie di clienti che, in passato, non potevano essere distinte, poiché mancavano i dati necessari a segmentare la propria clientela; inoltre le tecnologie una volta utilizzate per la segmentazione potevano gestire solo un numero limitato di informazioni, costringendo all'esclusione di dati importanti.

[1] Ian H. Gordon, Relationship Marketing: New Strategies, Techniques and Technologies to Win the Customers You Want and Keep Them Forever, John Wiley & Sons, 1998.

Ma quello che la segmentazione non è in grado di suggerire riguarda le specifiche azioni che le aziende dovrebbero intraprendere, dal momento che ogni singola tipologia di clienti dovrebbe richiedere l'attuazione di specifiche strategie di mercato.[2]

Recentemente però le vecchie metodologie di segmentazione sono state sostituite da altre più avanzate, indicate con il termine clustering: esse permettono all'azienda di effettuare la segmentazione della clientela anche quando si è in possesso di informazioni poco dettagliate.

La segmentazione della clientela ha quindi permesso alle aziende di modificare il proprio comportamento in base ai diversi gruppi di clienti, permettendo conseguentemente un risparmio in termini di tempo e denaro e un aumento del livello di soddisfazione della clientela.

Inoltre, succede spesso che le imprese siano in grado di mettere in atto delle strategie di CRM mirate ai singoli clienti: si sta iniziando a comprendere che impiegare campagne di marketing personalizzate (one-to-one[3]) richiede molte volte la capacità di conoscere in maniera dettagliata i singoli clienti. Il marketing one-to-one sta assumendo sempre più importanza poiché i segmenti, che in precedenza erano ampi e venivano utilizzati per le campagne di marketing, stanno lasciando spazio a segmenti costituiti da un solo cliente.[4]

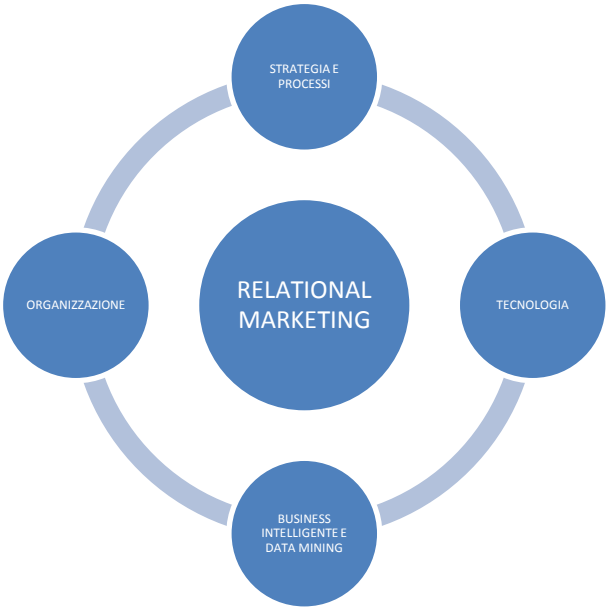
Il problema però di questo tipo di approccio sta nel fatto che spesso esso non vale gli sforzi e gli investimenti necessari ad attuarlo, soprattutto per quelle aziende che vendono prodotti di base e hanno quindi milioni di clienti. Alcuni settori, dunque, risultano più favorevoli, rispetto ad altri, per adottare strategie specifiche di CRM, mostrando un vantaggio per le imprese che per prime si sono affidate al data warehousing.

Lo scopo della strategia di marketing relazionale è quindi quello di iniziare, rafforzare, intensificare e preservare nel tempo le relazioni tra l'azienda e i suoi clienti.

[2] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.

[3] Si basa sull'idea di trattare ogni singolo cliente in maniera differente, al fine di avere una clientela soddisfatta fedele e redditizia.

[4] Peppers and Rogers, Enterprise One to One: Tools for Competing in the Interactive Age, Doubleday, 1997.



L'abilità di sfruttare le informazioni raccolte dal comportamento dei consumatori rappresenta oggi una forza competitiva per l'impresa: un'azienda, capace di raccogliere, immagazzinare, analizzare e capire l'enorme quantità di dati riguardanti i consumatori, è in grado di basare le sue azioni di marketing sulla conoscenza estratta da questi dati e di conseguenza può così raggiungere alcuni vantaggi competitivi. Le imprese scelgono di adottare queste strategie di marketing con lo scopo di trasformare contatti occasionali con i consumatori in una relazione di lungo termine: è infatti possibile aumentare la soddisfazione dei clienti e allo stesso tempo anche i profitti dell'azienda.

Le infrastrutture informatiche utilizzate comprendono il data warehouse dell'impresa, ottenuto dall'integrazione di risorse interne ed esterne e un data mart che permetta di effettuare analisi di data mining per creare un profilo di potenziali e attuali clienti; è possibile di conseguenza creare differenti segmentazioni dei consumatori, che verranno poi utilizzate per svolgere azioni di marketing mirate.

Nello specifico, la market basket analysis ha il compito di analizzare soluzioni di data mining che trovino correlazioni tra prodotti, all'interno del carrello della spesa dei clienti. I punti vendita possono di conseguenza applicare queste scoperte per rispondere in maniera più efficace ai bisogni dei consumatori.

L'esempio che più mi ha colpito e che mi ha fornito l'idea di questo elaborato è quello che riguarda la correlazione tra birra e pannolini, ormai esempio storico negli studi sulla market basket analysis.

La spiegazione a questa anomala associazione è la seguente: i neo-papà, che hanno sempre meno tempo per uscire e socializzare, compreranno una confezione da sei di birra, nella stessa esperienza di acquisto in cui avranno comprato i pannolini.[5]

Questo è un buon esempio della forte capacità di trovare relazioni, posseduta dalla market basket analysis, dal momento che è in grado di fornire una spiegazione ad una relazione a prima occhiata assolutamente poco ovvia.

I venditori possono trarre quindi utili informazioni da questa relazione, per esempio per posizionare, nei punti vendita, i due prodotti abbastanza vicini, in modo tale da agevolare i consumatori e da favorire un aumento delle vendite.

Uno strumento utile per la market basket analysis è quello delle carte fedeltà che, creando un profilo del consumatore, permettono di attuare programmi che portano i clienti alla fedeltà per quella determinata catena o punto vendita.

La market basket analysis può inoltre essere utilizzata nel web marketing, con le stesse finalità e i medesimi obiettivi, permettendo ai venditori di fornire risposte soddisfacenti ai bisogni dei propri clienti.

Si andrà quindi ad analizzare nello specifico il caso dei prodotti "white" (latte, riso, biscotti) nella realtà aziendale Pam, dopo aver affrontato i temi necessari a introdurre il concetto di market basket analysis: si parlerà infatti di data warehouse e di data mining, argomenti indispensabili per affrontare il tema obiettivo di questo elaborato.

[5] Appendice 1.

2. DATA WAREHOUSE

2.1. INTRODUZIONE AL DATA WAREHOUSE

Il presupposto dell'analisi dei dati è che gli stessi dati siano organizzati e ordinati in un database, perché, proprio da questo, viene fortemente influenzata. Oggi ci troviamo in un ambiente in cui vi è un'enorme quantità di dati, quindi risulta sempre crescente la necessità di analizzarli nella maniera più efficiente possibile.

E' per questo che diventa di grande importanza per le aziende di dimensioni medio-grandi possedere un sistema per la gestione dei dati, che consenta di ottenere una sintesi delle informazioni utili alla strategia aziendale.

Di conseguenza, risulta estremamente importante anche il processo di costruzione del database, dal momento che l'utilità delle informazioni estraibili grazie a un processo di data mining dipende dall'organizzazione dei dati stessi.

Quindi, la prima operazione da compiere per ottenere informazioni utili attraverso il data mining è possedere un valido database: si tratta della fase più onerosa dell'intero processo di data mining, sia per quanto riguarda l'allocazione delle risorse, sia per quanto riguarda i tempi di sviluppo.[6]

Sono tre gli esempi di strutture di database: il data warehouse, il data webhouse, il data mart. Le prime due sono strutture di dati dall'elevata complessità, mentre l'ultima è una base di dati così semplice da risultare pronta per l'analisi.

Il data warehouse è una sorta di contenitore di dati utili, che ha il fine di effettuare operazioni di business intelligence[7].

I dati in esso raccolti possono fornire una sequenza di livelli che partono da informazioni aggregate, che diventano sempre più dettagliate, in modo da riuscire a spiegare alcuni aspetti che durante l'analisi possono risultare critici.

Con la nascita del web ed il suo successivo impatto rivoluzionario, il data warehouse è stato leggermente accantonato, per lasciare spazio a questo nuovo ambiente nel quale operare, chiamato appunto web. Esso infatti determina un'imposizione per il data warehouse, quella di dotarsi di alcuni nuovi requisiti, rendendo così la natura del data warehouse diversa dalla precedente. Assume così, grazie a queste nuove caratteristiche, il nome di data webhouse.

---

[6] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

[7] Per business intelligence (BI) si intendono contemporaneamente sia un insieme di processi aziendali per raccogliere dati ed analizzare informazioni strategiche, sia la tecnologia utilizzata per realizzare questi processi, e infine anche le informazioni ottenute come risultato di questi processi.

Il web è un'efficientissima fonte di dati sui comportamenti di quelle persone che operano sui siti Internet: i dati che si possono ricavare da queste operazioni, anche se grezzi e molto semplici, forniscono in modo estremamente dettagliato informazioni sui movimenti compiuti durante la navigazione nel web. Tutti questi dati possono poi essere riuniti nel cosiddetto data webhouse, dove verranno analizzati e, se necessario, combinati con le altre fonti di dati, già presenti. Emerge inoltre la possibilità di rendere fruibili tutte le interfacce del data warehouse, che già esisteva, attraverso il web, tramite l'utilizzo di browser, con i quali possono essere svolte diverse operazioni.

Ne consegue che questa nuova costruzione del data warehouse debba ora tenere presenti i vari aspetti legati al web.

Un requisito fondamentale di un data webhouse è la velocità: infatti, la possibilità di raggiungere da qualsiasi parte del mondo i contenuti presenti nel web richiede al data webhouse di essere reperibile velocemente in qualsiasi momento, evitando anche solo delle brevi interruzioni nel suo funzionamento.[8]

Un data mart, invece, è un database tematico, spesso utilizzato per le attività di marketing; può infatti essere considerato un archivio aziendale, che contiene tutte le informazioni riguardanti la clientela e che aiuta a gestire i rapporti con essa. Può essere quindi considerato un data warehouse, di ridotte dimensioni e maggiore specificità.

E' questo il principale ambito operativo del data mining: infatti, la costruzione di strutture dei dati tematiche, quali sono i data mart, rappresenta il primo importantissimo passo che predisponga un ambiente informativo adatto all'attività di data mining.

Da un data warehouse è possibile estrarre tanti data mart quanti sono gli scopi della propria analisi; un data mart può essere costruito anche senza che ci sia un sistema integrato di data warehouse.

Essi condividono lo stesso quadro tecnologico: al fine di implementare le applicazioni di business intelligence, alcune aziende preferiscono progettare e sviluppare in modo incrementale una serie di data mart integrati, piuttosto che un data warehouse centrale, per ridurre i tempi di implementazione e le incertezze connesse al progetto.[9]

2.2. COS'E' UN DATA WAREHOUSE?

La definizione di data warehouse nasce negli anni Ottanta, ma nel tempo continua ad essere modificata, assumendo così più di una definizione. Il data warehouse è allo stesso tempo un

[8] Seth Godin, Permission Marketing: turning Strangers Into Friends, and Friends Into Customers, Simon & Schuster, 1999.

[9] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

contenitore di dati, un insieme di data mart più piccoli e una piattaforma hardware che consente all'utente dell'azienda di prendere decisioni.

Il data warehousing si fonda su quattro principi fondamentali:

1. Il data warehouse è solitamente un elaboratore o una piattaforma hardware, di varie dimensioni, distinta dal resto dell'apparato tecnologico.
2. I dati in esso contenuti vengono utilizzati per il supporto alle decisioni.
3. I data warehouse duplicano dati che già esistono all'interno dell'impresa.
4. Solitamente, parlando di data warehouse, ci si riferisce ad un corpo hardware, un insieme di prodotti e strumenti software e moltissimi dati.

Riassumendo questi quattro principi, il data warehouse viene comunemente definito come una raccolta di dati, che possono essere utilizzati per supportare le decisioni di management; si tratta quindi di un magazzino di dati che contiene informazioni estratte altrove nell'azienda e che vengono rese accessibili agli utenti dell'azienda stessa.[10]

E' molto importante che il data warehouse sia in grado di integrarsi perfettamente con la totalità dei dati collezionati dalle diverse applicazioni utilizzate: standard differenti devono essere ricodificati in modo univoco prima di immagazzinare le informazioni. E' poi variabile nel tempo, dal momento che l'orizzonte temporale di un data warehouse è compreso tra i 5 e i 10 anni, periodo in cui i dati vengono raccolti come una successione di precisi istanti temporali; non è volatile perché l'aggiornamento dei dati non è svolto al suo interno e quindi non subirà modifiche ad ogni aggiornamento, ma solo integrazioni.

Esistono due approcci diversi nella creazione di un data warehouse: il primo si basa sulla creazione di un solo archivio centralizzato, che raccoglie tutte le informazioni aziendali e le integra con quelle che provengono dall'esterno; il secondo invece unisce in una sola struttura interconnessa diversi database chiamati data mart, scollegati tra loro. L'approccio centralizzato ha il vantaggio di consentire un controllo costante sulla qualità dei dati che vengono immessi in esso, ma richiede una progettazione attenta che renda possibile un'espansione in futuro. Dall'altra parte, il secondo approccio appare inizialmente più semplice e risulta quindi il più diffuso attualmente, anche se sembra riscontrare alcuni problemi nel momento in cui i vari data mart vengono tra loro collegati, poiché portano a compiere un notevole sforzo per ottenere un livello di uniformità sufficiente.

E' quindi possibile sintetizzare che un data warehouse deve avere determinate caratteristiche: una sorta di magazzino di dati, ossia un archivio centralizzato; una struttura di meta-dati, che

[10] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.

sia in grado di descrivere cosa si può trovare e dove lo si può trovare all'interno del data warehouse; una serie di data mart specifici e tematici, specializzati a seconda degli obiettivi e rapidamente accessibili.

2.3. UTILIZZO DEL DATA WAREHOUSE

La maggior parte dei data warehouse oggi esistenti viene utilizzata prevalentemente per effettuare comuni funzioni operazionali, come la realizzazione di resoconti sulle entrate o di analisi delle vendite. Esse sono procedure aziendali piuttosto frequenti, ma, per quanto questi report possano essere semplici, essi consentono comunque all'utente di avere una prima impressione sui dati trattati. E' stato confermato che la tendenza di utilizzare il data warehouse soprattutto per report basilari è molto diffusa, mostrando che le ricerche dati effettuate più frequentemente sono quelle legate alle analisi finanziarie o alla redazione di report generici.[11]

Il data warehouse è però di grande importanza soprattutto per il supporto alle decisioni. Esso contempla una vasta gamma di analisi che permettono alle persone di prendere decisioni di vario genere e di diversa importanza, relative all'azienda, partendo dai dati a disposizione.

Un sistema di supporto alle decisioni è un'applicazione utilizzata attraverso il computer, interattiva, che combina dati e modelli matematici per aiutare a prendere decisioni che risolvano i problemi che si riscontrano nella gestione di affari pubblici e privati di aziende e organizzazioni.

Un sistema di supporto alle decisioni ha la capacità di trasformare i dati in informazioni e conoscenze, utili a coloro che devono prendere delle decisioni.

Infatti la differenza tra dati e informazioni è fondamentale: un data warehouse sintetizza dati che possono essere considerati informazioni solo nel momento in cui vengono trasformati in risposte o record significativi, utilizzabili nella comprensione degli eventi aziendali. Consiste proprio in questo il valore di un data warehouse, nel trasformare solo attraverso un tasto i dati in informazioni utili all'azienda.

2.4. LO SVILUPPO DEL DATA WAREHOUSE

Il data warehousing si è visto protagonista di una veloce espansione alla fine degli anni Ottanta, quando le aziende cominciarono a comprendere il valore dei dati che avevano a disposizione. Ultimamente, il valore del data warehousing è cresciuto ancora; infatti, il data

[11] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing,* Milano, Apogeo, 2000.

warehouse, con un'implementazione sempre più semplice, permette la memorizzazione di elevati volumi di dati differenti in un'unica sede, e l'economicità dei vari componenti hardware e software ne ha permesso l'espansione nel mercato, permettendo una conoscenza approfondita della clientela.

Tra gli utenti potenziali del data warehouse troviamo, e sono di notevole importanza, aziende appartenenti ad una serie di settori economici, tra cui: la vendita al dettaglio, i beni al consumo confezionati, le telecomunicazioni, i servizi finanziari, i trasporti, la sanità, il governo, i servizi pubblici, il settore manifatturiero.

2.5. GLI OBIETTIVI DEL DATA WAREHOUSING

Per l'azienda, uno degli obiettivi principali è quello di conoscere la propria clientela. Questo obiettivo ha origine però solo il secolo scorso: prima, infatti, le aziende si chiedevano continuamente chi stesse acquistando i loro prodotti. Negli anni Sessanta, invece, la domanda cambiò, e le aziende cominciarono a chiedersi cosa spingesse le persone ad acquistare i loro prodotti: questo diede inizio agli studi sulla ricerca motivazionale. Essa comprendeva una vasta gamma di pratiche, per la raccolta delle informazioni, che permettessero alle aziende di capire cosa spingesse i clienti all'acquisto. Con l'evoluzione della tecnologia e con l'aumento delle informazioni disponibili sulla clientela, le aziende hanno capito che potevano venire in possesso di queste informazioni tramite i dati ricavabili dalle operazioni aziendali eseguite ogni giorno. Potevano infatti in questo modo monitorare i comportamenti della clientela senza disturbare i clienti. Le aziende quindi iniziano a considerare non il modo in cui progettano, creano, distribuiscono e vendono, ma per chi lo fanno.[12]

Di conseguenza, gli obiettivi per cui un'azienda decide di utilizzare il data warehouse solitamente si possono ricondurre a quattro:

- fornire all'azienda una visione unica del cliente, tenendone monitorati i comportamenti;
- fornire a più utenti il maggior numero possibile di informazioni, migliorando i tempi di risposta;
- prevedere le vendite di un particolare prodotto, aumentando così la produttività globale dell'azienda;

[12] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.

- incrementare la comprensione degli eventi interni all'azienda.[13]

2.6. CATEGORIE DI ANALISI PER IL SUPPORTO ALLE DECISIONI

Le categorie di analisi per il supporto alle decisioni sono pensate in modo tale da suggerire un approccio per la realizzazione e l'utilizzo di un data warehouse di tipo aziendale. Esse sono: le query standard, l'analisi multidimensionale, la modellazione e la segmentazione, e la knowledge discovery (o scoperta della conoscenza).

Le query standard rappresentano il metodo di analisi più diffuso, perché rendono accessibile a tutti gli utenti aziendali l'informazione dettagliata e significativa relativa ai clienti, ai prodotti, al comportamento dell'impresa. Inoltre le query consentono tempi di risposta ridotti e permettono agli utenti di pensare ed agire autonomamente, in modo tale che l'azienda possa rapidamente adottare azioni dirette a controbattere le strategie di marketing dei concorrenti o a fornire risposte in tempo reale o a convincere un cliente intento ad andarsene a restare, ecc.

L'analisi multidimensionale invece rappresenta il gradino successivo alla query, per gli utenti che necessitano di tecnologie di analisi più potenti che gli consentano di studiare e confrontare l'informazione. L'analisi multidimensionale offre differenti prospettive attraverso cui guardare i dati e le informazioni, in modo tale, per esempio, da poter suddividere i clienti per aree geografiche, le vendite per città, le chiamate per periodo della giornata. Inoltre permette di raccogliere un insieme di risposte da utilizzare più volte secondo prospettive diverse.

La differenza con le query standard sta nel fatto che, mentre queste rimandano ad una grande sezione di dati differenti, l'analisi multidimensionale viene utilizzata per vedere gli stessi dati sotto diverse prospettive.

Poi, quando i dati diventano più voluminosi e dettagliati e vengono a crearsi nuove prospettive aziendali, gli utenti definiti più esperti necessitano di capacità analitiche più sofisticate, per raccogliere nuove informazioni. E' necessario quindi utilizzare strumenti di analisi specifici, che estraggono i dati dal data warehouse e li analizzano per creare una serie di modelli. La segmentazione suddivide i clienti, o altri insiemi di dati, in determinati gruppi, chiamati segmenti, che abbiano caratteristiche comuni che ne definiscono il comportamento, determinando successivamente strategie di marketing appropriate.

La knowledge discovery infine è rappresentata da algoritmi molto potenti che ricercano particolari elementi in database di grandi dimensioni: questi elementi però non vengono specificati precedentemente e viene infatti sfruttata nei casi in cui l'utente sta cercando

[13] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing,* Milano, Apogeo, 2000.

risposte a domande che non sa formulare. Sarà il data warehouse a indicare all'impresa dove si trovano gli elementi e le conseguenti principali relazioni.[14]

2.7. LE ATTIVITA' DI DATABASE MARKETING PIU' COMUNI

Ci sono numerose attività di marketing di uso comune che utilizzano la tecnologia del data warehousing e il supporto alle decisioni:

- Target marketing → Il target marketing, o marketing mirato, consiste nella commercializzazione di un determinato prodotto ad un particolare cliente o gruppo di clienti.
- Cross-selling[15] → Il cross-selling consente alle aziende di migliorare i propri rapporti con la clientela, la quale potrebbe acquistare nuovi prodotti o servizi, in un certo momento del rapporto con l'azienda, che sia esso l'acquisto di un altro prodotto, una visita al punto vendita, un reclamo telefonico.
- Analisi e previsione delle vendite → l'analisi delle vendite può risultare un'analisi soggettiva, dal momento che alcuni la considerano un'analisi delle entrate, altri invece si concentrano sulla produttività delle vendite. Si tratta comunque di un'informazione molto utile per le aziende, ai fini di modifiche di prezzo, realizzazione di campagne pubblicitarie, distribuzione di buoni sconto, ecc. L'obiettivo però non è solo quello di prevedere le entrate, ma anche di rispondere in maniera adeguata alle richieste della clientela, conoscendo i canali di vendita da utilizzare, modificando i prezzi, promuovendo particolari prodotti o servizi.
- Analisi del paniere → Si tratta sostanzialmente dell'insieme di prodotti che il cliente ha nel carrello della spesa, e può permettere all'azienda di ottenere maggiori informazioni sul cliente e sulle future possibilità di vendita. Infatti, è sempre maggiore il numero delle aziende che vuole conoscere le combinazioni di prodotti all'interno del carrello della spesa, con l'obiettivo di scoprire quali prodotti spingono all'acquisto di altri prodotti. L'analisi del paniere aiuta quindi lo studio dei trend di acquisto, oltre a contribuire alla scelta della politica dei prezzi e delle campagne da svolgere per determinati prodotti.

[14] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.
[15] Con l'espressione cross selling, che letteralmente significa "vendita incrociata", ci si riferisce ad una strategia di vendita di un prodotto o servizio in più rispetto a quanto richiesto dal cliente, in combinazione alla vendita del primo.

- Analisi delle promozioni → L'analisi delle promozioni permette alle aziende di determinare se e in che modo potrà aver successo una nuova campagna di marketing. Essa consente infatti di conoscere se si sono registrate nuove vendite o al contrario una cannibalizzazione dei prodotti, oppure se le vendite di un prodotto in promozione sono state compensate da altri acquisti.
- Mantenimento dei clienti → Con lo scopo di ottimizzare le entrate, molte aziende hanno capito che conservare un buon cliente è meno costoso che acquisirne uno nuovo. Questa analisi quindi si occupa di studiare delle costanti comportamentali manifestate da quei clienti che hanno abbandonato l'azienda e sono passati alla concorrenza, in modo tale da elaborare una strategia che aiuti ad evitare che i restanti clienti abbandonino l'azienda.
- Analisi della redditività → Viene effettuata perché le aziende preferiscono attuare strategie che consentano di aumentare la redditività della clientela esistente, invece che cercare di acquisire altri clienti.
- Valutazione del valore di un cliente → La valutazione del valore del cliente è nota anche come valutazione del valore della durata media della relazione di clientela (lifetime customer value o LCV). Il principio su cui essa si basa consiste nell'assegnare a un dato cliente un certo punteggio, che ne determina il valore per l'azienda in base a una serie di fattori per un certo periodo di tempo: indica in sostanza un valore che ogni cliente ha durante la relazione con l'azienda. Il valore di un cliente permette di determinare con precisione l'atteggiamento da assumere nei confronti di ogni singolo cliente.
- Lo studio delle tecniche di imballaggio dei prodotti → Lo scopo dell'imballaggio personalizzato dei prodotti è quello di unire più prodotti o servizi in un unico prodotto e ad un prezzo fisso, con l'obiettivo di riunire prodotti redditizi per assicurarsi degli utili, motivando il cliente a comprare proprio quell'insieme di prodotti.
- I Call Center → Il Call Center spesso è l'unico mezzo attraverso il quale il cliente può comunicare con l'azienda e ha quindi una forte influenza sulle relazioni con la clientela: sono infatti spesso l'unico canale attraverso cui le aziende possono ricevere un feedback dalla clientela.
- Analisi del contratto di vendita → Oltre a individuare prodotti e servizi che sono stati acquistati assieme, l'analisi dei contratti stipulati riesce a dare vita a programmi per il mantenimento della clientela nel momento in cui scade un contratto, quando molti clienti sono soliti passare alla concorrenza.

2.8. CONCLUSIONI SUL DATABASE MARKETING

Per concludere, l'obiettivo del database marketing è quello di vendere il prodotto giusto ai clienti giusti, attuando, al posto del mass marketing, un approccio alle vendite personalizzato in base alla clientela. Avere un'esperienza della clientela significa proprio applicare questo principio del one-to-one per trattare il cliente nel miglior modo possibile quando acquista, e prevedere anche atteggiamenti di acquisto futuri.

Il database marketing è quindi in grado di aumentare la fedeltà della clientela dell'azienda.

E' necessario conoscere quello che il cliente sta acquistando, in modo da capire il suo comportamento e prevedere anche le sue azioni future: in questo, è proprio il data warehouse a risultare di notevole aiuto.

3. IL DATA MINING

3.1. COS'E' IL DATA MINING?

Fin dal Rinascimento le persone osservavano il mondo e raccoglievano dati per spiegare i fenomeni naturali, dando origine a teorie, osservazioni, equazioni, che descrivessero il mondo naturale e le sue leggi. Egiziani e Cinesi hanno fatto studi sui triangoli e portato a quello che oggi è definito il Teorema di Pitagora; prima di essi, diversi popoli hanno osservato il movimento del sole, della luna, delle stelle, e creato il calendario.

Il cambiamento vero e proprio è avvenuto con la codifica della matematica e la creazione di macchine capaci di facilitare le misurazioni, il loro deposito e la loro analisi. Solo con l'arrivo dei nuovi computer, scienziati e ingegneri riescono a dare un senso ai dati raccolti.

Ma la storia del data mining ha inizio solo con lo sviluppo di ulteriori discipline, prima tra esse la statistica. Essa è infatti molto utile, anche se non risolve tutti i problemi di data mining. [16]

Nella società attuale, ci scontriamo ogni giorno con grandi quantità di dati relativi all'ambiente nel quale viviamo, ma questo potenziale di cui disponiamo è rimasto largamente non utilizzato.

Questo accade principalmente per due motivi: innanzi tutto questi dati sono spesso disseminati in sistemi di archiviazione sconnessi tra loro, rendendo così inefficiente l'organizzazione dei dati stessi; inoltre non si ha ancora una forte consapevolezza delle funzionalità degli strumenti statistici, utili per l'elaborazione dei dati.

A questi problemi però si contrappongono due nuove tendenze: la crescita di strumenti più efficaci e allo stesso tempo economici, e la capacità di analizzare grandi quantità di dati.

Tutto questo ha permesso la diffusione del data mining in molti contesti aziendali, come strumento di supporto alle decisioni.

Letteralmente "to mine", verbo inglese, significa "scavare per estrarre", e infatti il data mining è il processo di selezione ed esplorazione di enormi quantità di dati, con lo scopo di scoprirne relazioni per ottenere risultati utili ai fini aziendali.[17]

[16] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

[17] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

Il processo di data mining consta di varie attività:

1. Definizione degli obiettivi.

   Questi devono essere formulati in maniera chiara, dal momento che si tratta di una delle fasi più critiche del processo: è su di essa infatti che si basa tutta la successiva metodologia.

2. Selezione e organizzazione dei dati.

   Superata la prima fase, e quindi individuati gli obiettivi, bisogna scegliere i dati utili per l'analisi. Le fonti preferite nella maggior parte dei casi sono quelle interne, che derivano da esperienze proprie dell'azienda: si tratta proprio del data warehouse aziendale. Quest'ultimo permette una rappresentazione dei dati chiamata matrice dei dati, che nasce sulla base degli obiettivi iniziali. Dopo aver controllato la qualità dei dati disponibili, è preferibile fissare un campione su cui impostare l'attività di analisi: questo perché le dimensioni di un database sono spesso troppo ampie, quindi la scelta di un campione permette di ridurre i tempi di analisi ed elaborazione.

3. Analisi esplorativa dei dati.

   E' una fase necessaria perché permette di formulare i metodi statistici migliori per il raggiungimento degli obiettivi prefissati, tenendo pur sempre conto dei dati ottenuti nella fase precedente.

4. Elencazione dei metodi statistici utilizzati per l'elaborazione dei dati.

   I metodi statistici che si possono utilizzare sono numerosi: la scelta, quindi, dipenderà da un lato dal tipo di problema che si deve studiare, dall'altro lato dal tipo di dati disponibili per l'analisi.

   Esistono quattro categorie di metodi:

   -Metodi esplorativi: sono interattivi e hanno lo scopo di giungere a prime ipotetiche conclusioni sull'insieme dei dati disponibili, oltre che a fornire informazioni sulla possibilità di integrare o sostituire il database disponibile.

   -Metodi descrittivi: descrivono l'insieme dei dati, sia la sintesi delle osservazioni, sia la sintesi delle variabili.

   -Metodi previsivi: hanno l'obiettivo di spiegare ciascuna variabile in funzione delle altre, cercando di trarne regole di classificazione, che permettano di prevedere il risultato futuro di una o più variabili risposta, in funzione di quanto accade alle variabili esplicative.

   -Metodi locali: hanno l'obiettivo di individuare i tratti peculiari relativi a sottoinsiemi del database.

5. Elaborazione dei dati.

   E' necessario, a questo punto, tradurre i modelli statistici in corretti algoritmi di calcolo informatico, che siano in grado di condurre ai risultati sintetici desiderati.

6. Valutazione dei metodi utilizzati e scelta del modello finale di analisi.

   Si sceglie dunque il modello migliore di analisi dei dati, allo scopo di produrre una regola decisionale finale, sulla base di considerazioni fatte confrontando i risultati ottenuti coi diversi metodi.

7. Interpretazione del modello scelto e suo successivo utilizzo.

   Scelto il modello e verificatane la correttezza, la regola può essere applicata sull'intera popolazione di riferimento.

Inoltre, il data mining deve essere implementato correttamente nei processi aziendali: questo processo deve essere graduale, con lo scopo di raggiungere una piena integrazione del data mining con le altre attività di supporto alle decisioni, all'interno dell'impresa.

3.2. UTILIZZO DEL DATA MINING

Il data mining è l'analisi di una grande quantità di dati, che ha lo scopo di scoprire e fornire schemi e regole interessanti e significativi. Ha anche lo scopo di aiutare l'azienda a migliorare il proprio marketing, le proprie vendite, e la relazione con il cliente, analizzando e capendo il comportamento di quest'ultimo. Inoltre è da sottolineare che il data mining può essere utilizzato anche in campi che vanno dall'applicazione della legge, alla radioastronomia, alla medicina, al controllo dei processi industriali.

Quasi tutti gli algoritmi di data mining sono nati per uno scopo commerciale: in questo ambito vengono utilizzate varie tecniche che derivano da varie discipline, quali la statistica e l'informatica. La scelta di una particolare combinazione di tecniche da applicare in una determinata situazione dipende sia dalla natura dell'attività di data mining da compiere, sia dalla natura dei dati disponibili.[18]

Le attività supportate dal data mining sono: la classificazione, la stima, la previsione, il clustering e la descrizione. Alcune di queste attività meglio si rapportano con la verifica delle ipotesi, nella quale un comportamento registrato nel database in passato viene usato per verificare o confutare preconcetti, idee e sensazioni relativi a rapporti e relazioni nei dati.

Altre attività invece sono meglio collegate con la scoperta della conoscenza. Nella scoperta della conoscenza, non viene posta nessuna supposizione a priori; i dati parlano da soli. La

[18] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

scoperta della conoscenza può essere di due tipi: diretta e indiretta. La prima cerca di spiegare o classificare particolari insiemi di dati; la seconda cerca di trovare collegamenti e somiglianze tra i gruppi di record senza l'utilizzo di settori o classi predefinite.
Tutte queste attività sono comprese nella definizione di data mining.

3.3. LO SVILUPPO DEL DATA MINING

Soltanto negli ultimi anni il data mining è riuscito a farsi spazio nell'ambito commerciale. Questo è accaduto per diversi motivi, primo tra i quali è il fatto che si producono grandi quantità di dati: infatti il data mining ha importanza solo quando è presente una grande quantità di dati, dal momento che molti algoritmi di data mining ne richiedono grossi volumi, per costruire e preparare i modelli, che verranno utilizzati per compiere la classificazione, la previsione, la stima, o altre attività di data mining. Solo poche industrie da tempo erano in possesso di strumenti che permettessero un'interattiva relazione coi clienti; ma è solo oggi, con lo sviluppo di fattori quali le carte di credito e debito, cassieri automatici, scanner nei supermercati, shopping via internet, e simili, che i dati vengono prodotti e raccolti in quantità senza precedenti.[19]
I dati vengono poi immagazzinati: infatti molto spesso, questa grande quantità di dati viene estratta dalle fatture, dalle prenotazioni, dalle richieste di rimborso, dagli ordini, che entrano nel sistema e lì vengono memorizzati.
Immagazzinare i dati significa metterli insieme da fonti diverse, in uno stesso formato. Generalmente non è possibile trasformare le operazioni di data mining in un sistema operativo, ma in ogni caso i sistemi operativi immagazzinano dati in un formato adatto ad ottimizzare la performance delle attività operative.
Il potere dei calcoli inoltre risulta molto conveniente, poiché gli algoritmi di data mining solitamente richiedono numerosi passaggi, viste le grandi quantità di dati. La continua diminuzione del prezzo di dischi, memorie, processori, ha portato le tecniche, una volta costose ed utilizzate sono in rare eccezioni, nell'uso quotidiano. Questo fatto, e le sue relative conseguenze, forniscono un ambiente eccellente per un data mining di larga scala.
Di notevole importanza è anche il fatto che il data mining vive in ambienti in cui la pressione competitiva è forte, in quanto alcuni dei settori più ricchi di informazioni, come le telecomunicazioni, le assicurazioni e i servizi finanziari, stanno vivendo un forte aumento della competizione. Le compagnie di questi settori dell'economia hanno avuto per lungo

[19] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

tempo i dati e le risorse necessarie per compiere attività di data mining. Ora, per la prima volta, hanno un forte incentivo commerciale in tal senso, e le industrie, che non sono state tradizionalmente ricche di informazioni, stanno cercando di diventarlo.[20]

Diverse tendenze stanno aumentando l'importanza concorrenziale delle informazioni: un'economia basata sempre più sui servizi, l'avvento della personalizzazione di massa, la crescente importanza dell'informazione come prodotto a sé stante.

Per quanto riguarda il primo aspetto, per le compagnie appartenenti al settore dei servizi, le informazioni danno un vantaggio competitivo; le compagnie, invece, che non erano abituate a fornire determinati servizi, cominciano a pensare in maniera diversa. Inoltre, anche i comuni prodotti possono essere potenziati con dei servizi.

Per quanto riguarda invece il secondo aspetto, personalizzazione di massa significa creazione di prodotti su misura, facendo selezioni tra un grande insieme di componenti standard. In questo processo, i dati vengono raccolti secondo le preferenze dei consumatori. Inoltre, per aiutare i venditori a saperne di più sui singoli clienti, questi dati possono essere estratti per approfondimenti sul mercato nel suo complesso.

Infine, in relazione all'ultimo aspetto, spicca la figura del mediatore di informazioni, che sta creando un business sempre maggiore. Ogni compagnia che raccoglie informazioni si trova nella posizione di diventare un mediatore di informazioni.

Infine influisce il fatto che si siano resi disponibili software commerciali per il data mining, ma c'è sempre un ritardo tra il momento in cui un nuovo algoritmo appare per la prima volta nei giornali accademici, e il momento in cui i software commerciali che incorporano questi algoritmi diventano disponibili. C'è poi un altro ritardo tra la disponibilità iniziale dei primi prodotti e il momento in cui questi raggiungono un ampio consenso e, per il data mining, il periodo di una diffusa disponibilità e accettazione è solo all'inizio.

3.4. APPLICAZIONI DEL DATA MINING

Le metodologie di data mining possono essere applicate in diverse situazioni, dal controllo del marketing allo studio dei fattori di rischio nelle diagnosi mediche, dal rilevamento delle frodi alla valutazione dell'efficacia di nuovi farmaci.[21]

[20] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997

[21] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Marketing relazionale

L'applicazione del data mining in questo campo fornisce un contributo significativo alla crescita della popolarità di queste metodologie. Alcune importanti applicazioni nel marketing relazionale sono:

- l'identificazione di segmenti di consumatori più propensi a rispondere positivamente a determinate campagne di marketing;
- l'identificazione di segmenti target di consumatori per le campagne di conservazione;
- la previsione di un tasso di risposte positive a determinate campagne di marketing;
- l'interpretazione e la comprensione del comportamento di acquisto dei consumatori;
- l'analisi dei prodotti acquistati assieme dai consumatori.

Rilevamento delle frodi

Il rilevamento delle frodi è un altro campo importante per l'applicazione del data mining, dal momento che le frodi colpiscono molti settori, tra cui la telefonia, le assicurazioni, le banche.

Valutazione del rischio

Lo scopo dell'analisi del rischio è quello di stimare il rischio connesso alle decisioni future, che spesso assumono una forma dicotomica.

Text mining

Il data mining può essere applicato a tipi diversi di testi, che rappresentano dati non strutturati, con lo scopo di classificare articoli, libri, documenti, e-mail e pagine web.

Il riconoscimento delle immagini

Il trattamento e la classificazione delle immagini digitali è utile per riconoscere caratteri scritti, paragonare e identificare facce umane, applicare filtri di correzione alla fotografia e controllare comportamenti sospetti nelle video camere di sorveglianza.

Web mining

Le applicazioni di web mining sono intese come l'analisi dei clickstream, ossia le sequenze delle pagine visitate e le scelte di chi naviga nel web. Queste sono molto utili di conseguenza per l'analisi dei siti e-commerce.

Diagnosi mediche

I modelli di apprendimento sono uno strumento prezioso nel campo medico per il rilevamento delle malattie, grazie all'utilizzo dei risultati dei test clinici.

3.4.1. ALCUNI ESEMPI[22]

I federali utilizzano il data mining per rintracciare i criminali

Il governo federale degli Stati Uniti ha di recente adottato la tecnologia di data mining. Per una parte di indagini sul caso Unabomber e per altri crimini minori, l'FBI ha usato questo tipo di tecnologia per setacciare migliaia di documenti, cercando connessioni e possibili indizi.

Il Dipartimento del Tesoro usa il data mining per dare la caccia a segnali sospetti nel trasferimento di fondi internazionali, che potrebbero indicare frode o riciclaggio di denaro. I produttori di strumenti di data mining inoltre riportano che anche l'Agenzia delle Entrate ha mostrato un forte interesse per questi stessi strumenti.

Un supermercato diventa mediatore di informazioni

I supermercati si trovano nella posizione di poter avere numerose informazioni sui consumatori, ma alcuni di essi non hanno ancora le capacità tecniche per collegare i dati di acquisto presi dalle vendite con i singoli acquirenti o famiglie. Una società che invece si occupa di svolgere queste attività è la società Safeway.

Safeway, come anche altre grandi catene, è diventata un mediatore di informazioni. Il supermercato acquisisce indirizzi e dati demografici direttamente dai clienti, offrendo loro sconti, in cambio dell'utilizzo della carta "Safeway Savings Club" al momento dell'acquisto. Per avere la carta, i clienti forniscono di loro volontà informazioni personali, utili per creare dei modelli di previsione.

Così, ogni volta che il cliente presenta la carta per gli sconti, i suoi acquisti vengono memorizzati in una sorta di magazzino di dati: con ciascun momento di acquisto, il consumatore insegna al venditore qualcosa di nuovo su se stesso. Probabilmente il supermercato sarà più interessato alle aggregazioni tra i prodotti piuttosto che al comportamento del singolo consumatore, ma le informazioni raccolte sui singoli individui sono di grande interesse per coloro che dispongono i prodotti negli spazi del supermercato. Ovviamente il supermercato assicura al cliente che le informazioni saranno mantenute private. Safeway inoltre fa pagare una somma di denaro ai fornitori che vogliono un loro coupon o una particolare offerta promozionale per raggiungere solo determinate persone: è proprio nell'individuare questo particolare segmento di clienti che dà il suo contributo il data mining.

Un business basato sulla conoscenza del gruppo di consumatori

Uno dei motivi per cui imparare a conoscere il comportamento dei consumatori è dovuto al fatto di riuscire poi a generalizzare e quindi fare previsioni sul comportamento di altre, simili

[22] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

persone. Questo è quello che "Peppers and Rogers" chiamano sviluppo della conoscenza della comunità: essi citano "Firefly" come un esempio di business basato su questo tipo di conoscenza. Inoltre è un esempio di business che dipende completamente dal data mining. Firefly chiede ai suoi membri di valutare musica e film e, in base ai "mi piace" e ai "non mi piace", le persone vengono automaticamente riunite in gruppi, i cui membri abbiano la stessa opinione, in modo tale che, quando in un gruppo entra a far parte un nuovo membro, il sistema possa già sapere quali film e quale tipo di musica possa essere di suo interesse.
La bellezza del sistema sta nel fatto che più esso viene utilizzato, meglio verranno compresi i comportamenti dei membri.

Cross selling

USAA è una compagnia di assicurazione che vede tra le sue occupazioni l'interesse per il personale militare in pensione e le loro famiglie. La compagnia mantiene informazioni dettagliate sui suoi clienti e usa il data mining per prevedere dove sono in un determinato momento e di quali prodotti avranno bisogno.
Un'altra compagnia che ha usato il data mining per migliorare la propria capacità di cross selling è la "Fidelity Investments". Fidelity possiede un data warehouse, che comprende le informazioni di tutti i suoi clienti. Queste informazioni vengono utilizzate per costruire dei modelli che prevedano quali altri prodotti potrebbero interessare a ciascun cliente. Quando un ipotetico cliente chiama Fidelity, lo schermo del rappresentante telefonico mostra esattamente da dove parte la conversazione.
Inoltre, per migliorare le abilità di cross selling della compagnia, il data warehouse di Fidelity permette ai servizi finanziari di costruire modelli che conducano ad una fedeltà del cliente e che quindi aumentino di conseguenza la fidelizzazione. Questi modelli hanno costretto Fidelity a mantenere un servizio di pagamento delle fatture marginalmente profittevole perché altrimenti questo tipo di servizio sarebbe stato tagliato. Si è scoperto che le persone che hanno utilizzato il servizio erano di gran lunga meno interessate del cliente medio a legarsi ad un concorrente. Il taglio del servizio avrebbe incoraggiato un gruppo redditizio di clienti fedeli a guardarsi intorno.
Un principio fondamentale della filosofia del marketing one-to-one sta nel fatto che è molto più profittevole focalizzarsi sul "wallet share" o sul "customer share", cioè la quantità di business che si può avere con ciascun cliente, piuttosto che sulla quota di mercato (market share). Dai servizi finanziari all'industria manifatturiera, le società innovative stanno utilizzando il data mining per incrementare il valore di ciascun cliente.

Avviamento di richiesta di garanzia

Un costruttore di motori diesel riceve un flusso costante di richieste di garanzia dai rivenditori indipendenti, che hanno effettuato la manutenzione sui motori coperti dalla garanzia del produttore. Ciascuna richiesta deve essere esaminata da un esperto per determinare se la manodopera e le parti utilizzate sembrano ragionevoli e appropriate. Si ha da tempo una serie di regole utilizzate per esentare alcune richieste che sono considerate di prassi e che possono essere pagate senza un esame approfondito. La compagnia si sta ora informando sull'utilizzo del data mining per espandere il numero di richieste pagate automaticamente, scoprendo un set esteso di regole per descrivere classi di richieste sempre approvate dagli esperti. Questo avviamento di richieste automatiche ha la possibilità di far risparmiare all'azienda milioni di dollari.

Tenere stretti i buoni clienti

Il data mining viene usato per promuovere la fidelizzazione dei clienti nei settori dove i clienti sono liberi di cambiare fornitori ad un basso costo e i competitori non vedono l'ora di portarseli via. Cercando di capire chi ha intenzione di andarsene e perché, può essere sviluppato un piano di fidelizzazione che affronti i giusti problemi e si diriga verso i giusti clienti.

Il Southern California Gas è solo un esempio di un primo monopolio regolamentato che ora deve competere per i clienti. Prima della deregolamentazione la società non aveva un dipartimento di marketing; ora ha un programma di database marketing che integra i dati di fatturazione e di utilizzo con le informazioni di credito e i dati del censimento degli Stati Uniti. Applicando le tecniche di data mining a questi dati, è possibile capire chi potrebbe beneficiare maggiormente da un piano di pagamento per livelli. La direct mail[23] sulla base del modello ha prodotto tassi di risposta compresi tra il 7% e l'11%, un risultato fenomenale per la direct mail. Inoltre è stato possibile imparare che alcuni gruppi di clienti sono molto più sensibili al prezzo di altri.

Costa di più inserire un nuovo cliente che tenersi stretto uno già esistente, ma spesso l'incentivo offerto dalla fidelizzazione di un cliente è abbastanza costoso. Il data mining è la chiave per capire quali clienti possono ottenere questo incentivo e quali clienti staranno senza questo incentivo.

[23] La direct mail è una delle tecniche di marketing attraverso la quale aziende commerciali e enti di vario genere comunicano direttamente con clienti e utenti finali consentendo di raggiungere un target definito, con azioni mirate che utilizzino una serie di strumenti, ottenendo in tal modo delle risposte oggettive misurabili, quantificabili e qualificabili.

Eliminare i cattivi clienti

In molti settori, alcuni clienti costano più di quanto valgono. Questi sembrano essere quelle persone che utilizzano spesso le risorse di assistenza ai clienti, senza però comprare molto. Oppure, sembrano essere quelle persone fastidiose che hanno una carta di credito che raramente usano. Ancora peggio, potrebbero essere persone che possiedono una grande quantità di denaro, ma dichiarano la bancarotta.

Le stesse tecniche di data mining che vengono usate per riconoscere i clienti di maggior valore possono essere utilizzate anche per scegliere quelli che dovrebbero essere rifiutati per un prestito o quelli che possono rimanere in attesa per molto tempo.

Rivoluzionare un settore

Nel 1988, l'idea che il bene più prezioso di una emittente della carta di credito fossero le informazioni che essa possiede sui suoi clienti era abbastanza rivoluzionaria.

"Signet Bank Corporation" ha acquistato dati comportamentali da molte fonti e li ha utilizzati per creare modelli di previsione. Utilizzando questi modelli è stato lanciato un prodotto di grande successo che ha modificato il settore di produzione delle carte di credito. L'utilizzo delle tecniche di data mining che alimentarono questa rapida crescita sono anche responsabili di mantenere i tassi di perdita su crediti tra i più bassi del settore.

Il data mining è il fulcro della strategia di marketing delle banche relativamente alle carte di credito: esso consentirà alle società di ricavare opportunità di cross-selling per i prestiti auto, mutui e altri servizi bancari generali. L'obiettivo è inoltre quello di porre al centro del proprio approccio a tutti i servizi di banca un marketing e un metodo decisionale basato sulle informazioni.

Questi esempi hanno dato un'idea dei campi in cui è possibile applicare il data mining, ma non li ha analizzati tutti: infatti le tecniche di data mining possono essere utilizzate per fare molte delle cose che ciascun business richiede per la propria crescita e per il proprio arricchimento.

3.5. IL CIRCOLO VIRTUOSO DEL DATA MINING[24]

Le informazioni che derivano dai dati concentrano i propri sforzi sulla segmentazione della clientela, migliorano il design dei prodotti incontrando i reali bisogni dei consumatori, migliorano l'allocazione delle risorse capendo e prevedendo le preferenze del cliente.

[24] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

I dati sono il fulcro delle principali attività di molte aziende: sono generati dalle transazioni che stanno alla base di molti settori, quali il commercio al dettaglio, le telecomunicazioni, le aziende manifatturiere, i trasporti, le assicurazioni, le banche, e altri ancora. La promessa del data mining è quella di trovare relazioni interessanti nascoste nei milioni di byte di dati. Ma trovare solamente gli schemi non è sufficiente: è necessario infatti anche rispondere a queste relazioni, lavorare con esse, trasformare i dati in informazioni, le informazioni in azioni e le azioni in valore aggiunto per l'azienda. Questo è denominato circolo virtuoso del data mining. Per mantenere questa sua promessa, il data mining deve diventare un processo essenziale, incorporato al marketing, alle vendite, all'assistenza ai clienti, ecc.

Le tecniche che da sempre danno un senso a milioni di dati ora sono pronte ad affrontare un numero ancora maggiore di dati, immagazzinati nel data warehouse. Infatti ora si possiede una potenza di calcolo che permette, con complessi algoritmi, di raggiungere ottimi risultati. Ma, anche se gli algoritmi sono importanti, il data mining non è solo un insieme di tecniche e strutture di dati; queste infatti devono essere applicate nelle giuste aree e con i giusti dati. Il circolo virtuoso del data mining riconosce che il data mining è solo un passo del processo interno, che utilizza la conoscenza ottenuta dalla crescente comprensione dei clienti, del mercato, dei prodotti e dei competitori. E' quindi un processo continuo che col tempo costruisce continui risultati.

Il circolo virtuoso del data mining è costituito da quattro fasi:

1. Identificare il problema → E' la fase che si verifica durante l'organizzazione, dove le crescenti informazioni potrebbero permettere alle persone di svolgere al meglio il loro lavoro. Lo scopo è quello di identificare le aree dove le relazioni tra i dati possono procurare un valore aggiunto: queste saranno il punto iniziale della fase successiva.
2. Utilizzare le tecniche di data mining per trasformare i dati in informazioni utili → Il data mining prende i dati e produce risultati utili per la fase successiva. Importante è saper identificare la giusta fonte di dati, per ottenere risultati utili all'analisi, come anche è fondamentale saper mettere insieme i dati nel sistema di calcolo usato per l'analisi.
3. Lavorare sulle informazioni → E' la fase in cui i risultati del data mining hanno avuto effetto e vengono trasmessi alla fase successiva, ossia quella della misurazione. Il problema in questa fase è quello di incorporare le informazioni nel processo di business: le azioni da svolgere sono quindi parte fondamentale del circolo virtuoso.
4. Misurare i risultati → Le misurazioni forniscono i feedback necessari a migliorare continuamente i risultati. In questo caso, le misurazioni si riferiscono al valore del

business, che va oltre tassi di risposta, oltre la deviazione standard e la media. Queste specifiche misurazioni devono dipendere da un certo numero di fattori: le opportunità per il business, la sofisticatezza dell'organizzazione, le misurazioni passate, la disponibilità dei dati.

La fase delle misurazioni dipende fortemente dalle informazioni derivanti dalle fasi precedenti: è quindi necessario svolgere in maniera adeguata le fasi precedenti, per arrivare a possedere i giusti dati e le giuste informazioni per la fase delle misurazioni.

Il circolo virtuoso è la cornice che permette al data mining di integrarsi negli altri processi dell'azienda.

Il data mining differisce per molti aspetti dai classici processi dell'azienda.

Tipici processi aziendali	Data mining
Operazioni e rapporti nuovi, ma dati vecchi. Prevedibili e periodici flussi di lavoro, tipicamente legati al calendario. Limitato uso dei dati da parte di tutta l'azienda. Attenzione alla linea di business, non al cliente. Tempi di risposta spesso misurati in secondi/millisecondi o settimane/mesi. Sistema di registrazione dei dati. Descrittivo.	Analisi di dati storici e attuali per determinare azioni future. Imprevedibile flusso di lavoro, in base ai bisogni di marketing e a quelli dell'azienda. Più sono i dati, generalmente migliori sono i risultati. Attenzione al prodotto, al cliente, alla regione di vendita. Tempi di risposta spesso misurati in minuti o ore. Copia dei dati. Creativo.

Si deduce che: il data mining non cerca di replicare i risultati precedenti esattamente, ma è un processo creativo, i cui risultati cambiano nel tempo. Inoltre il data mining fornisce feedback per altri processi che potrebbero aver bisogno di cambiare per incorporare i risultati di data mining.

Ci sono quindi molti casi in cui il data mining è stato utilizzato per supportare le attività di marketing. In ciascun caso è stata affrontata una sfida: con l'utilizzo del data mining, sono stati costruiti dei modelli di previsione. Questi modelli sono poi stati trasferiti e utilizzati per le azioni di marketing, i cui risultati sono stati misurati e hanno rimandato il loro feedback allo stesso processo di data mining.

3.6. LE TECNICHE DI DATA MINING

Le principali tecniche di data mining, ciascuna con le proprie attività, sono otto, e sono:[25]

- La market basket analysis
- Il ragionamento basato sulla memoria (MBR)
- L'individuazione dei cluster
- L'analisi per collegamenti
- Gli alberi decisionali e la regola dell'induzione
- Le reti neurali
- Gli algoritmi genetici
- L'elaborazione analitica on-line (OLAP)

La market basket analysis

La market basket analysis è una forma di clustering, utilizzata per trovare gruppi di prodotti che sono soliti apparire assieme nelle transazioni. I modelli che costruisce mostrano la probabilità che diversi prodotti vengano acquistati assieme e possono essere espressi attraverso regole.

Lo scopo della market basket analysis è quello di ricavare dalle vendite fatte dai clienti delle informazioni utili alle future azioni di marketing. Viene infatti spesso utilizzata nel settore industriale e nelle attività di e-commerce per analizzare le vendite; oppure anche per l'analisi di acquisti effettuati con le carte di credito, di servizi attivati dai clienti su cellulari o telefoni fissi, di comportamenti d'acquisto delle famiglie. E' inoltre una tecnica molto legata al settore della vendita al dettaglio, dove le informazioni sui prodotti acquistati assieme potrebbero essere l'unico dato ricavabile sui comportamenti dei consumatori, dal momento che i dati demografici non sono ricavabili dalle transazioni, che sono anonime.

I dati utilizzati a questo scopo molto spesso si riferiscono alle vendite e ciascuna transazione consiste in una lista di oggetti acquistati: questo è ciò che viene chiamato *basket*. Le regole che ne vengono estratte possono aiutare il supporto alle decisioni, e le informazioni che ne derivano possono essere utilizzate per scopi differenti, come pianificare il layout del punto vendita, limitare le offerte speciali ad uno dei prodotti tra quelli che sembrano essere acquistati insieme, impacchettare i prodotti, offrire coupon per gli altri prodotti quando uno di questi è venduto senza gli altri, ecc. Quando le transazioni non sono anonime, la market basket analysis può essere adatta all'utilizzo su dati storici con una componente temporale.

[25] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Il ragionamento basato sulla memoria (MBR)

L'MBR è una tecnica di data mining diretto che si basa su due momenti strettamente collegati tra loro: l'individuazione di osservazioni e la ricerca nel database di tutte le osservazioni a queste simili. Utilizza quindi esempi conosciuti come modelli per fare delle previsioni riguardo casi sconosciuti: esso cerca le somiglianze negli esempi conosciuti e combina i loro valori per costruire una classificazione o una previsione; la distanza tra i dati simili fornisce una misura della correttezza del risultato.

Uno dei suoi maggiori vantaggi è la capacità di funzionare su qualsiasi fonte di dati, anche senza modificarli. I due elementi chiave dell'MBR sono la funzione della distanza, utilizzata per trovare gli elementi più vicini e la funzione di combinazione, che combina i valori per fare delle previsioni.

Un altro vantaggio è la capacità di venire a conoscenza di nuove classificazioni, soltanto introducendo nuovi esempi nel database. Una volta che le funzioni corrette sono state trovate, sono solite rimanere stabili anche come nuovi esempi per nuove categorie e vengono incorporate tra i dati conosciuti: questa facilità differenzia l'MBR da molte altre tecniche di data mining, che devono invece essere nuovamente applicate per incorporare nuove informazioni.

L'individuazione dei cluster

L'individuazione di cluster è la costruzione di modelli che si occupano di trovare record simili: questi gruppi composti da record simili sono chiamati appunto cluster. Si tratta di data mining indiretto, il cui obiettivo è quello di trovare somiglianze tra i dati ancora sconosciute. Ci sono numerose tecniche per trovare i cluster, tra cui metodi geometrici, metodi statistici, reti neurali.

L'individuazione dei cluster è un ottimo modo per cominciare ogni tipo di analisi di dati; cluster simili forniscono il punto di partenza per conoscere cosa c'è nei dati e per capire come utilizzarli al meglio.

L'analisi dei collegamenti

L'analisi dei collegamenti cerca di stabilire delle relazioni logiche tra le singole righe di un database. Si differenzia dal concetto di associazione perché essa è sostanzialmente un'analisi globale delle interrelazioni tra unità statistiche e non tra variabili. L'analisi dei collegamenti utilizza alcuni strumenti della teoria dei grafi, che favorisce un'interpretazione intuitiva delle relazioni così rappresentate. Le relazioni tra i consumatori diventano sempre più importanti, specialmente coloro che si occupano del marketing prestano molta attenzione ai consumatori, alle famiglie, ecc. Una delle principali aree in cui viene svolta l'analisi per collegamenti è

quella delle telecomunicazioni: ogni chiamata infatti collega un cliente con qualcun altro. E' inoltre utilizzata dalle forze dell'ordine per collegare crimini e cercare di risolverli. Si tratta però di un'attività poco supportata dalla tecnologia: i pochi strumenti disponibili si concentrano più sul visualizzare i collegamenti, che sull'analisi degli schemi.

Gli alberi decisionali

Gli alberi decisionali sembrano essere l'applicazione di data mining più conosciuta e utilizzata, grazie alla sua semplicità, la sua facilità di utilizzo, la velocità di calcolo, il rispetto di ogni singolo dato e la facilità di interpretazione delle regole generate. Gli alberi decisionali sono un potente modello prodotto da un insieme di tecniche che includono la classificazione, gli alberi di regressione e l'induzione automatica del "chiquadro". Gli alberi decisionali sono usati per il data mining diretto, in particolare per la classificazione. Partendo da un dataset, è possibile costruire un numero esponenziale di diversi alberi decisionali; essi dividono i record in soggetti disgiunti, ciascuno dei quali viene descritto con una semplice regola in uno o più campi.

Uno dei principali vantaggi degli alberi decisionali è che il modello spiegabile prende la forma di regole esplicite: questo permette alle persone di valutare i risultati, identificando gli attributi chiave nel processo. Inoltre esso è un metodo utile quando i dati in entrata sono di qualità incerta e falsi risultati saranno evidenti nelle regole. Le regole stesse possono essere espresse facilmente come affermazioni logiche e potranno essere direttamente applicate a nuovi record.

Le reti neurali

Le reti neurali probabilmente sono la tecnica di data mining più comune: sono destinate a simulare il comportamento di sistemi biologici composti di neuroni. Nella loro più comune accezione, prendono conoscenze dall'insieme di partenza e generalizzano gli schemi all'interno di esso; vengono utilizzate non solo per la classificazione e la previsione, ma anche per la regressione di attributi target continui. Le reti neurali possono essere applicate anche nel data mining indiretto e nelle previsioni delle serie temporali.

Una rete neurale è un grafico orientato fatto di nodi, che nell'analogia con i sistemi biologici rappresentano i neuroni, connessi da archi, che corrispondono ai dendriti e alle sinapsi. Ciascun arco è associato a un peso e ciascun nodo a una funzione di attivazione.

Uno dei principali vantaggi delle reti neurali è la loro ampia applicabilità: gli strumenti che supportano le reti neurali sono accessibili per una grande varietà di programmi. Sono inoltre interessanti perché scoprono gli schemi tra i dati in un modo analogo al pensiero umano.

Esse hanno però due importanti lati negativi: il primo riguarda la difficoltà di comprensione dei modelli che produce; il secondo è rivolto alla loro particolare sensibilità al formato dei dati in entrata.

Gli algoritmi genetici

Gli algoritmi genetici applicano la meccanica della genetica e della selezione naturale per trovare un insieme di parametri ottimale, che descriva una funzione predittiva. Essi infatti lavorano su una popolazione di individui, i quali rappresentano varie soluzioni ai problemi; al termine del processo permetteranno di individuare il modello migliore. Gli algoritmi genetici sono simili alla statistica, dal momento che il modello deve essere conosciuto in anticipo; solo i modelli più predittivi sopravvivranno di passaggio in passaggio, fino al momento in cui si arriverà alla soluzione ottimale.

L'elaborazione analitica on-line (OLAP)

L'OLAP non è nello specifico uno strumento di data mining, ma è uno degli strumenti utili ad estrarre e presentare informazioni: si tratta infatti di un modo di mostrare i dati agli utilizzatori, che ne facilita la comprensione e mette in evidenza le relazioni tra di essi.

3.7. LE ATTIVITA' DI DATA MINING[26]

Le attività che il data mining è in grado di svolgere utilizzano molto spesso le tecniche precedentemente descritte. Esse sono:

- La classificazione
- La stima
- La previsione
- Il raggruppamento per affinità
- Il clustering
- La descrizione

Ogni singolo strumento o tecnica di data mining non è ugualmente applicabile a tutte queste attività.

Nella realtà, il data mining è solitamente utilizzato in database molto grandi; questo per due motivi: innanzitutto perché è possibile trovare interessanti schemi e relazioni dal semplice

[26] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

studio dei risultati ottenuti; inoltre, molte tecniche di data mining richiedono una grande quantità di dati per generare delle regole, delle associazioni, delle previsioni, e via dicendo.[27]

Classificazione

La classificazione è la più comune attività di data mining: infatti continuamente noi classifichiamo, categorizziamo, valutiamo. Dividiamo gli esseri viventi in phylum, specie, generi; sostanze in elementi; cani in razze; e così via.

La classificazione consiste nell'esaminare le caratteristiche di un nuovo oggetto di studio e assegnarlo di conseguenza ad un predefinito insieme di elementi. A questo fine, gli oggetti, per essere classificati, sono generalmente rappresentati come record di un database, e l'atto della classificazione consiste nell'aggiornare ciascun record compilando un campo con il codice di un certo tipo di classe.

L'attività della classificazione è caratterizzata da una ben definita rappresentazione delle classi, che comprendono esempi predefiniti: essa consiste nel costruire un modello che possa essere applicato a dati che devono ancora essere classificati.

Alcuni esempi delle attività di classificazione sono: assegnare parole chiave ad articoli; determinare quali numeri di telefono corrispondono a fax; individuare crediti di assicurazione fraudolenti, ecc.

Alberi decisionali e ragionamenti basati sulla memoria sono le tecniche che meglio si adattano alla classificazione; ad essi si aggiunge, in determinate circostanze, l'analisi per collegamenti.

Stima

La stima riguarda risultati continuamente valutati. Una volta dati alcuni input, la stima è utilizzata per assegnare un valore ad alcune sconosciute variabili continue, come il reddito, l'altezza, il saldo nella carta di credito.

In sostanza, la stima viene speso utilizzata per portare a termine l'attività di classificazione.

L'attività della stima ha il vantaggio che i record individuali possono essere ordinati in ranghi.

Alcuni esempi di stima sono: il numero di figli in una famiglia, il reddito per famiglia, il valore di un cliente, ecc.

Le reti neurali sono la tecnica più utilizzata nell'attività della stima.

Previsione

La previsione assomiglia alla classificazione o alla stima, tranne che per il fatto che i record sono classificati secondo la previsione di comportamenti futuri o la stima di alcuni valori

[27] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

futuri. Nell'attività di previsione, l'unico modo per verificare la precisione della classificazione è aspettare e vedere.

Alcune delle tecniche utilizzate per la classificazione e la stima possono essere adatte anche per l'utilizzo nella previsione, usando esempi dove il valore delle variabili è già noto, assieme anche ad alcuni dati storici. I dati storici sono usati per creare un modello che spieghi il comportamento attuale osservato; quando viene applicato questo modello ad input attuali, il risultato è la previsione di un comportamento futuro.

La tecnica di market basket analysis, utilizzata per scoprire quali prodotti sono soliti essere acquistati insieme in un negozio di alimentari, può essere anche adatta per la creazione di un modello che preveda quali spese o azioni future possano svolgersi, osservando i dati attuali.

Alcuni esempi dell'attività di previsione sono: prevedere quali clienti se ne andranno nei successivi sei mesi, prevedere chi acquisterà un servizio aggiuntivo sul proprio piano telefonico, ecc.

La market basket analysis, il ragionamento basato sulla memoria, gli alberi decisionali, e le reti neurali, sono tutte metodologie applicabili nell'attività della previsione. La scelta della tecnica dipende quindi dal tipo di dati utilizzati e dal tipo di valori da prevedere.

Raggruppamento per affinità o regole di associazione

Queste attività sono utilizzate per identificare interessanti e ricorrenti associazioni tra insiemi di record di un dataset: consiste, molto semplicemente, nel determinare quali cose vengono considerate insieme. L'esempio per eccellenza è quello di determinare quali prodotti vengono acquistati assieme in un supermercato, da qui il concetto di market basket analysis. Le catene di vendita al dettaglio possono utilizzare il raggruppamento per affinità per pianificare la posizione del prodotti nel negozio o in un catalogo, e, dal momento che certi prodotti vengono acquistati assieme, verranno anche visti assieme.

Il raggruppamento per affinità può essere anche utilizzato per identificare opportunità di cross-selling e per individuare interessanti pacchetti o gruppi di prodotti e servizi.

Esso è inoltre un semplice approccio che permette di generare regole dai dati; se due prodotti appaiono frequentemente assieme, si possono generare due regole: le persone che comprano il primo prodotto comprano anche il secondo, con una probabilità $P_{(1)}$; le persone che comprano il secondo prodotto comprano anche il primo, con una probabilità $P_{(2)}$.

Clustering

Il clustering è l'attività di segmentazione di una popolazione eterogenea in più sottogruppi omogenei, chiamati cluster. Quello che distingue il clustering dalla classificazione è che il clustering non si riconduce a classi predefinite. Nella classificazione, la popolazione viene

suddivisa assegnando ciascun elemento o record ad una classe predefinita, sulla base di un modello sviluppato da esempi precedenti.
Nel clustering non ci sono classi predefinite, né esempi; i record vengono raggruppati sulla base di somiglianze, e sta a chi studia queste somiglianze capirne l'importanza e il significato. Ad esempio, cluster di sintomi indicano differenti malattie; cluster di foglie e semi indicano differenti ceppi di mais.
Spesso il clustering è precedente ad altre forme di data mining: ad esempio può essere considerato la prima fase della segmentazione del mercato, dal momento che è necessario prima di ogni attività, come potrebbe essere una promozione sulle vendite, dividere la clientela in cluster di persone con simili abitudini di acquisto e successivamente chiedersi per ogni cluster quale tipologia di promozione è preferibile attuare.
Descrizione
Alcune volte, lo scopo del data mining è semplicemente quello di descrivere cosa sta succedendo in un complicato database, in modo tale da comprendere meglio persone, prodotti, processi. Una buona descrizione del comportamento suggerirà una buona spiegazione del comportamento stesso. Al contrario del clustering e delle regole di associazione, l'analisi descrittiva non si focalizza su particolari gruppi di record del dataset, ma cerca di dare spiegazioni plausibili a relazioni nascoste tra i dati per fornire una spiegazione del fenomeno al quale i dati si riferiscono.
Un esempio di tecnica utilizzata per la descrizione è la market basket analysis.

3.7.1. L'UTILIZZO DI MODELLI NELLE ATTIVITA' DI DATA MINING

Un modello produce uno o più output, per un dato insieme di input. Analizzare i dati spesso implica la costrizione di un appropriato modello per quegli stessi dati.
Non si può però affermare che se un modello esiste, esso produce sicuramente risultati precisi, ma si può parlare di buoni e cattivi modelli: misurare i risultati di un modello è un passaggio critico del loro utilizzo e sviluppo. Un modello può essere utilizzato per il clustering, la classificazione e l'analisi delle serie temporali; i modelli forniscono un linguaggio comune per parlare di data mining.
Un modello di classificazione prende un nuovo record e assegna ad esso una classificazione esistente. Potrebbe inoltre assegnare una nuova classificazione, una probabilità di correttezza, e altre informazioni, in base alla natura del modello. Un modello predittivo è simile al modello di classificazione, ad eccezione del fatto che l'output non è limitato ad un certo numero di classi. Il modello di clustering prende più record e restituisce un numero minore di

cluster; questi cluster possono poi essere ricondotti a nuovi record, nella creazione di un modello di classificazione. Un modello di serie temporali è simile a una classificazione o ad un modello predittivo. Talvolta l'attributo target evolve nel tempo ed è quindi associato a periodi adiacenti sull'asse del tempo: in questo caso, la sequenza dei valori delle variabili target rappresenta una serie temporale. I modelli di analisi delle serie temporali lavorano su dati caratterizzati da una dinamica temporale e hanno lo scopo di prevedere il valore delle variabili target per uno o più periodi futuri.
Quando i modelli vengono creati, gli input sono di solito chiaramente specificati. In realtà, preparare i dati derivanti da diversi sistemi operativi è anche più impegnativo che creare il modello stesso. Gli input del modello possono riguardare la scelta della tecnica; l'output del modello è spesso specificato a priori, in quanto è spesso una categoria. Inoltre un modello spesso fornisce come output un grado di confidenza: questo è molto utile per determinare quando applicare i risultati del modello.

3.8. LA METODOLOGIA DI DATA MINING

Ci sono principalmente due approcci diversi al data mining: la verifica di ipotesi e la scoperta della conoscenza. Il primo è un approccio di tipo top-down[28], che cerca di verificare o confutare un'idea predeterminata. Il secondo è un approccio di tipo bottom-up[29], che parte dai dati e cerca di fornire informazioni ancora sconosciute.
Questa metodologia affronta i problemi del data mining da entrambe le direzioni, facendo avanti e indietro dall'uno all'altro approccio: da un lato vengono ideate possibili spiegazioni ai comportamenti osservati, in modo tale da lasciare che queste ipotesi dettino i dati da analizzare; dall'altro lato si lascia che siano i dati a suggerire nuove ipotesi da verificare.[30]

3.8.1. LA VERIFICA DI IPOTESI

Un'ipotesi è una proposizione la cui validità deve essere provata. La prova di validità di un'ipotesi viene fatta analizzando i dati, che vengono raccolti con l'osservazione o generati da un esperimento.
Il metodo della verifica di ipotesi si compone di più passaggi:

[28] Si tratta di una metodologia basata sul disegno complessivo del data warehouse, è più sistematica ma implica tempi di sviluppo più lunghi e maggiori rischi di non essere completato entro programma.

[29] Si basa sull'uso di prototipi e quindi le estensioni del sistema vengono eseguite mano a mano, secondo uno schema. Questo approccio di solito è più veloce, fornisce risultati più tangibili, ma manca una visione di insieme dell'intero sistema.
[30] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

1) Generare ipotesi.
2) Determinare quali dati possono permettere la verifica delle suddette ipotesi.
3) Individuare i dati.
4) Preparare i dati per l'analisi.
5) Costruire modelli basati sui dati.
6) Valutare i modelli per confermare o rifiutare le ipotesi.

3.8.2. LA SCOPERTA DELLA CONOSCENZA

La scoperta della conoscenza è il processo che permette di trovare schemi significativi nei dati, che spieghino eventi passati, in modo tale da usare gli stessi schemi per prevedere eventi futuri.
La scoperta della conoscenza può essere diretta o indiretta. Nella prima, il compito è quello di spiegare il valore di alcuni campi, rispetto a tutti gli altri: si sceglie quindi un campo e direttamente il computer ci dice come viene stimato, classificato, previsto. Nella seconda, non c'è un campo prescelto, ma viene semplicemente chiesto al computer di trovare relazioni significative tra i dati.
In altre parole si può dire che la scoperta della conoscenza indiretta viene utilizzata per riconoscere le relazioni tra i dati, mentre la scoperta della conoscenza diretta viene utilizzata per spiegare queste relazioni, una volta trovate.
La scoperta della conoscenza diretta è orientata all'obiettivo: c'è un campo specifico, il cui valore deve essere determinato, un insieme prefissato di classi da assegnare a ciascun record, o una specifica relazione da esplorare.
La scoperta della conoscenza diretta si articola in più passaggi:

1) Identificare le fonti e i dati predefiniti.
2) Preparare i dati all'analisi.
3) Costruire e preparare un modello.
4) Valutare il modello.

La scoperta della conoscenza indiretta, a differenza di quella diretta, non ha nessun campo di riferimento, ma si occupa solamente di trovare tra i dati relazioni significative. Uno dei più comuni utilizzi della scoperta della conoscenza indiretta è volto alla market basket analysis, che si occupa, per la maggior parte dei casi, di individuare quali prodotti vengono acquistati assieme.
Le fasi del procedimento della scoperta della conoscenza indiretta sono:

1) Identificare le fonti dei dati.
2) Preparare i dati per l'analisi.
3) Costruire e preparare un modello.
4) Valutare il modello.
5) Applicare il modello a nuovi dati.
6) Identificare un target potenziale per la scoperta della conoscenza diretta.
7) Generare nuove ipotesi da verificare.

3.9. MISURARE L'EFFICACIA DEL DATA MINING

Il data mining è costoso: richiede un grande sforzo nel raccogliere i dati, nel prepararli, nel formulare un problema, nel costruire un modello, nell'analisi stessa.

Il data mining può essere usato in diversi modi e per diversi scopi: i metodi da utilizzare vengono scelti in base ai risultati che si vuole raggiungere. Ciascuno dei metodi di data mining ha il suo vocabolario, la sua stima della precisione, e la sua performance.

Dove possibile, gli obiettivi più generici devono essere ripartiti in obiettivi più specifici, per monitorare più facilmente il processo che permette il raggiungimento degli stessi.

Gli obiettivi di data mining possono essere descrittivi o predittivi. Tutte le attività di data mining possono essere pensate come descrittive o predittive. Un'attività descrittiva è quella il cui obiettivo è la comprensione, la spiegazione, o la scoperta della conoscenza. Anche quando l'obiettivo del data mining è predittivo, potrebbe essere importante che il modello utilizzato sia sufficientemente descrittivo, affinché mostri chiaramente il motivo per cui è stata fatta una particolare previsione.[31]

Quando viene misurata la precisione di un modello predittivo, l'attenzione ricade sia sulla precisione del modello nel suo complesso, sia sulla precisione delle singole previsioni.

Minimum description length (MDL)

Il minimum description length per un modello è il numero di bit utilizzati per codificare sia la regola che la lista di tutte le eccezioni alla regola: meno bit richiede, migliore sarà la regola.

Alcuni strumenti di data mining utilizzano la MDL per decidere quali tegole tenere e quali eliminare.

Error rate (tasso di errore**)**

Esso è semplicemente la percentuale di record classificati scorrettamente; è utilizzato come una stima dell'errore che ci si aspetta dalla classificazione di nuovi record.

Alcune attività di data mining utilizzano questa tecnica.

[31] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Varianza

Misurare quanto i valori stimati siano lontani dalla media è il modo più semplice per descrivere la precisione di un modello di stima. E' necessario sommare i quadrati delle differenze tra i valori corretti e i valori stimati: la media di queste differenze al quadrato si chiama varianza. Minore è la varianza, più precisa è la stima. Spesso è più utile utilizzare la deviazione standard, che si ottiene facendo la radice quadrata della varianza.

Confidenza e supporto

La market basket analysis può essere usata per la stima: essa infatti studia quali prodotti sono stati venduti assieme, per prevedere quali prodotti verranno venduti assieme.

Per valutare quanto queste previsioni siano precise, vengono usati la confidenza e il supporto. La confidenza indica quanto spesso la relazione ha validità all'interno del campione. Il supporto indica quanto spesso la combinazione appare.

Distanza

Il concetto di distanza appare in molte tecniche di data mining. Quando tutte le variabili sono continue e numeriche, è possibile utilizzare la formula geometrica della distanza, cioè la radice quadrata della somma dei quadrati delle differenze su ciascun asse.

Lift

Il modo più comune di confrontare la prestazione dei modelli di classificazione è quello di utilizzare un rapporto, chiamato lift. questa tecnica può inoltre essere adatta a confrontare anche modelli creati dalle altre attività di data mining. Il lift misura il cambiamento nella concentrazione di una particolare classe, quando il modello è usato per selezionare un preciso campione da una popolazione generale.

Il lift però non è in grado di definire se un modello è valido in termini di tempo, sforzo, costo.

Lifetime value (LTV)

Il lifetime value è definito come il valore complessivo di un cliente nel tempo. Inizialmente un individuo è un cliente potenziale, che non ha ancora iniziato ad acquistare i prodotti o ad utilizzare i servizi di un'azienda: questo però diventa poi consumatore. Nella fase di maturità della relazione, si tenta di estenderne la durata e la profittabilità, si cerca cioè di massimizzare il lifetime value di ciascun cliente. L'ultima fase della relazione tra cliente e azienda è l'interruzione della relazione stessa.

4. MARKET BASKET ANALYSIS

4.1. INTRODUZIONE ALLA MARKET BASKET ANALYSIS (MBA)

La market basket analsis è una nuova ed efficace metodologia di data mining, che consente di individuare le associazioni che esistono tra i prodotti acquistati o i servizi utilizzati da un cliente.

Quello che la market basket analysis si propone di fare è quindi costruire modelli comportamentali di acquisto ricorrenti, al fine di supportare le attività di marketing strategico.

Definendo alcune cosiddette regole di associazione tra i prodotti acquistati, è possibile far emergere alcuni legami, permettendo vari nuovi tipi di indagine. In sostanza, le associazioni studiate nella market basket analysis permettono di capire e analizzare i comportamenti di acquisto dei clienti, per poi, di conseguenza, progettare determinate azioni, quali promozioni, composizioni di cataloghi, presentazione di prodotti; indaga inoltre sui prodotti che si attraggono maggiormente tra di loro, permettendo di individuare regole valide su tutta la popolazione dei clienti, che permettano di associare all'acquisto di un set di prodotti, l'acquisto di altri prodotti.

Riassumendo: l'obiettivo finale della market basket analysis è quello di conoscere i prodotti che si presentano abbinati nei panieri dei clienti e di comprendere i comportamenti d'acquisto dei clienti, per fornire loro risposte adeguate ai propri bisogni.

Nello specifico, nella grande distribuzione, le informazioni sono ottenute da lettori ottici alle casse dei supermercati, da scontrini fiscali, ecc; i dati sono di natura dicotomica, ossia l'acquisto viene contrassegnato dal numero "1", il non acquisto dal numero "0". Si rende così possibile la misurazione della vicinanza o similarità tra due clienti, tramite un coefficiente di similarità che tiene conto della presenza e dell'assenza dei prodotti.

Una volta individuate le coppie di prodotti che più si attraggono, su questi l'impresa effettuerà azioni mirate di marketing strategico, come ad esempio possono essere le promozioni sui prodotti.

4.2. COS'E' LA MARKET BASKET ANALYSIS?

Ciascun ipotetico carrello contiene un assortimento di prodotti capace di rivelarci ciò che ciascun consumatore compra in ogni singola esperienza d'acquisto: ma ancora più informazioni ci sono date dall'insieme delle spese dell'intero gruppo di consumatori.

I consumatori, infatti, non sono tutti uguali: ciascun consumatore acquista prodotti diversi, in quantità diverse, in momenti diversi. La market basket analysis utilizza le informazioni che derivano dalle molteplici esperienze di acquisto per dirci chi sono i consumatori, perché fanno

determinati acquisti, ma anche quali prodotti vengono venduti insieme e quali sono più propensi a essere soggetti a promozioni. Può quindi aiutare nella disposizione dei prodotti nel punto vendita, nel determinare i prodotti da mettere in offerta, nell'indicare quando emettere i coupon.

Sembra che i primi ad utilizzare la market basket analysis siano stati Agrawal, Imieliński, Swami; essi erano esperti informatici che avevano accesso ad un grande deposito di dati raccolti dalle varie transazioni dei clienti, e capaci di trovare regole associative tra i prodotti acquistati. Questo metodo venne subito utilizzato nell'ambito del marketing: prodotti che vengono solitamente comprati assieme risultano spesso vicini all'interno del punto vendita, per aumentare la probabilità che il consumatore compri entrambi i prodotti; i prodotti cosiddetti complementari, inoltre, vedranno abbassato il prezzo di uno dei due, in modo tale da far aumentare la domanda di entrambi.

Ci sono due principali approcci allo studio della market basket analysis, chiamati rispettivamente conoscitivo e esplicativo.

Il primo approccio, quello conoscitivo, si impegna a ricercare distinte relazioni tra diverse categorie, basandosi su esempi di prodotti acquistati. Questo approccio vale soprattutto per i metodi volti ad una rappresentazione delle associazioni simmetriche derivanti dalle tabelle incrociate delle vendite, attraverso categorie multiple; il problema che si deve affrontare deriva dal fatto che sono presenti molto poche categorie che abbiano una relazione di questo tipo.

Il secondo approccio, quello esplicativo, guarda gli effetti delle variabili di marketing mix nelle vendite, attraverso il rendiconto delle dipendenze tra categorie nell'assortimento; il problema da risolvere, in questo caso, è che talvolta il set di categorie che deve essere analizzato è piuttosto limitato.

Per concludere, entrambi gli approcci alla market basket analysis hanno dei limiti iniziali nel trovare le informazioni di cui necessitano; dall'altra parte però, ciascuno di questi approcci ha indubbiamente propri specifici ed importanti meriti.

4.3. LE REGOLE DI ASSOCIAZIONE

Uno dei punti a favore della market basket analysis è sicuramente dovuto alla chiarezza e all'utilità dei suoi risultati, che troviamo rappresentati sotto forma di regole di associazione, che rivelano quanto i prodotti sono relazionati gli uni con gli altri, e in che modo tendono ad essere raggruppati insieme.

Le regole di associazione hanno lo scopo di identificare schemi regolari e ricorrenti all'interno di un ampio insieme di transazioni. Esse sono infatti piuttosto semplici e intuitive e sono utilizzate frequentemente per indagare sulle vendite, attraverso la market basket analysis.
In molte aree di applicazione la raccolta sistematica di dati genera enormi liste di transazioni, che, una volta analizzate attraverso le regole associative, sono in grado di identificare possibili ripetizioni nei dati.
Date due proposizioni Y e Z, in termini generali è una regola un'implicazione del tipo $Y \Rightarrow Z$, che a parole significa: se Y è vera, anche Z è altrettanto vera; Y è l'antecedente della regola, mentre Z è il conseguente.
Le regole rivolte ad estrarre nuove conoscenze per un'analisi di business intelligence dovrebbero essere semplici e facilmente interpretabili, in modo tale da poter essere utili per coloro che se ne occupano e facilmente traducibili in azioni concrete.
Ci sono tre tipologie di regole di associazione: quelle utili, quelle banali e quelle inspiegabili.
Le regole di associazione utili forniscono importanti informazioni; infatti una volta delineato lo schema, non sarà difficile darne una giustificazione.
Le regole di associazione banali sono quelle regole già conosciute e per questo ritenute scontate; in questo caso infatti l'analisi produce risultati già conosciuti e compresi.
Infine, i risultati delle regole di associazione definite inspiegabili sembrano non avere una spiegazione e non suggerire un preciso modo di agire.
E' da notare che i risultati ottenuti in maggioranza sono quelli banali e quelli inspiegabili. Per quanto riguarda i primi, essi riproducono conoscenze già acquisite, rendendo inutili gli sforzi fatti nell'utilizzare sofisticati strumenti tecnici di analisi; spesso danno informazioni su azioni ormai passate, ma non forniscono indicazioni per azioni future. Le regole inspiegabili suggeriscono un'investigazione al di fuori del campo del data mining per capirle meglio.

4.4. LE REGOLE DI DISSOCIAZIONE

Una regola di dissociazione è simile ad una regola di associazione, ad eccezione del fatto che una regola di dissociazione ha anche una condizione di "and not" che si aggiunge a quelle di "and".[32]
Le regole di dissociazione possono essere create da un semplice adattamento all'algoritmo classico della market basket analysis, che consiste nell'introduzione di un nuovo insieme di prodotti che sono l'opposto di ciascuno di quelli originali. Successivamente, modifica

[32] Esempio: If A and not B then C.

ciascuna transazione in modo tale che essa contenga uno degli opposti, se e solo se non contiene l'originale.[33]

Ci sono però tre svantaggi nell'includere questi nuovi prodotti: il primo è che il numero totale dei prodotti usati nell'analisi raddoppia, rendendo ancora più complicati i calcoli da svolgere; il secondo svantaggio è dovuto al fatto che la portata di una tipica transazione cresce perché ora include anche i prodotti opposti; il terzo problema è che la frequenza dei prodotti opposti tende ad essere molto più ampia della frequenza dei prodotti originari.

A volte però è utile invertire solo i prodotti più frequenti del gruppo scelto per l'analisi. Questo diventa particolarmente importante quando la frequenza dei prodotti originali è vicina al 50%, in modo tale che le frequenze dei rispettivi opposti siano anch'esse prossime al 50%.

4.5. COME AVVIENE LA MARKET BASKET ANALYSIS?

La market basket analysis ha inizio con l'acquisto di uno o più prodotti e le relative informazioni sull'acquisto stesso.

Ci sono tre cose importanti che interessano l'utilizzo della market basket analysis: la scelta di un corretto insieme di prodotti; la generazione di regole tramite la decifrazione della matrice delle relazioni; sopraffare i limiti su cui ci si imbatte a causa della moltitudine di dati da analizzare.

4.5.1. LA SCELTA DI UN CORRETTO INSIEME DI PRODOTTI[34]

I dati utilizzati nella market basket analysis sono solitamente i dati derivanti dagli acquisti nel punto vendita. Raccogliere e usare questi dati è la parte fondamentale della market basket analysis e dipende fortemente dai prodotti scelti per l'analisi.

Ciascun prodotto ha determinati codici che lo identificano e che lo fanno rientrare in specifiche categorie: questa classificazione è chiamata tassonomia[35].

La cosa preferibile è di utilizzare, per la market basket analysis, prodotti del più alto livello della tassonomia, ad esempio "dolci freddi e semifreddi" al posto di "gelati". Dall'altra parte però, più specifici sono i prodotti, più accurati saranno i risultati. Un giusto compromesso sembrerebbe quello di utilizzare inizialmente prodotti più generici, per poi ripetere il

[33] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

[34] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

[35] Secondo la matematica, una tassonomia è una struttura ad albero di categorie appartenenti ad un dato gruppo di concetti. A capo della struttura c'è una categoria singola, il nodo radice, le cui proprietà si applicano a tutte le altre categorie della gerarchia (sotto-categorie). I nodi sottostanti a questa radice costituiscono categorie più specifiche le cui proprietà caratterizzano il sottogruppo del totale degli oggetti classificati nell'intera tassonomia.

procedimento perfezionando la ricerca con l'utilizzo di prodotti più specifici: in questo frangente dovranno essere utilizzati solo i sottoinsiemi delle transazioni che contengono quei determinati prodotti.

La complessità del procedimento è dovuta al numero di prodotti che vengono utilizzati: maggiore sarà questa quantità, più lungo sarà il procedimento di generazione di regole di associazione. Allo stesso modo, la complessità desiderata della regola determina quanto specifici o generali devono essere i prodotti.

In alcune circostanze, come potrebbe essere il minimarket o la spesa attraverso cataloghi, i consumatori non acquistano in grandi quantità; in altri casi, invece, come per esempio in un supermercato, si hanno acquisti di un numero più elevato di prodotti, quindi sono utili regole più complesse.

Salendo a livelli sempre maggiori della tassonomia, si riduce il numero dei prodotti. Centinaia di prodotti si riducono ad uno solo più generico, che spesso corrisponde ad un unico reparto o linea di prodotto: spesso usare prodotti generici ha come effetto la scoperta di relazioni interdipartimentali. Inoltre, i prodotti generici aiutano a trovare regole che abbiano un sufficiente supporto.

Ma, anche se alcuni prodotti vengono usati al livello di generici, questo non significa che tutti i prodotti devono essere utilizzati al medesimo livello. Il livello appropriato dipende dal prodotto, dalla sua importanza nel generare risultati, dalla sua frequenza nei dati ricavati dalle transazioni.

Importante da sottolineare è il fatto che, quando i prodotti appaiono all'incirca nello stesso numero nei dati derivanti dalle transazioni, la market basket analysis produce i migliori risultati. Questo affinché le regole non siano dominate dai prodotti più comuni. La tassonomia può aiutare: avvicinare i prodotti rari ai livelli più alti nella tassonomia, li rende più frequenti; mentre i prodotti più comuni non devono assolutamente subire lo stesso trattamento.

Quando si applica la market basket analysis, è utile avere una tassonomia dei prodotti che devono essere considerati nell'analisi. Si possono sostituire i prodotti nell'analisi con prodotti generici da livelli differenti della tassonomia. Scegliendo un giusto livello della tassonomia, questi prodotti generici dovrebbero verificarsi circa lo stesso numero di volte nei dati per migliorare i risultati dell'analisi.

I dati utilizzati per la market basket analysis non sono generalmente di una qualità molto elevata. I dati che derivano dal sistema operativo spesso sono sporchi, grezzi, e hanno bisogno di una ripulita, prima di diventare una buona fonte per le decisioni. Inoltre è probabile che i dati siano di diversi formati o che abbiano codici diversi; alcuni sistemi, infatti, sono più

aggiornati di altri. Sono questi alcuni dei problemi tipici dell'utilizzo dei dati per il data mining, aggravati nella market basket analysis, poiché questo tipo di analisi dipende fortemente dalle transazioni del punto vendita.

La market basket analysis ha dimostrato di poter essere utilizzata in maniera interessante in punti vendita quali supermercati, minimarket, farmacie, catene di fast food, dove la maggior parte degli acquisti vengono fatti con pagamento in contanti. I pagamenti in contanti sono anonimi: ciò significa che il punto vendita non ha alcuna conoscenza riguardo i consumatori, perché non ci sono informazioni che identifichino il consumatore di quella particolare transazione. Relativamente alle transazioni anonime, le uniche cose conosciute sulla spesa sono la data e l'ora, il luogo del punto vendita, il cassiere, i prodotti acquistati, eventuali coupon cambiati, e l'ammontare della spesa. Con la market basket analysis anche questo limitato numero di dati produce interessanti risultati.

L'uso crescente delle carte di credito e di debito sta generando transazioni sempre meno anonime, fornendo una maggiore possibilità di ottenere un maggior numero di informazioni sui consumatori e sul loro comportamento nel tempo.

4.5.2. LA GENERAZIONE DI REGOLE A PARTIRE DAI DATI[36]

Calcolare il numero di volte che una data combinazione di prodotti appare nella transazione è assolutamente corretto, ma una combinazione di prodotti non è una regola. Alcune volte si verificano combinazioni insolite e interessanti; ma in altre circostanze ha più senso trovare una regola di base comune.

Una regola ha due parti, la condizione e il risultato ed è così rappresentata: " If condition then result"[37]. Il risultato deve sempre essere uno solo.

Nello specifico: $O = \{o_1, o_2, \ldots o_n\}$ è un insieme di n oggetti; $L \subseteq O$ è un generico sottoinsieme, che contiene un certo numero di oggetti. Una transazione rappresenta un generico sottoinsieme che è stato registrato nel database in combinazione con un'attività o un ciclo di attività. Il dataset D è quindi composto da una lista di transazioni, ciascuna associata ad un unico identificatore.

Con riferimento alla market basket analysis, gli oggetti rappresentano i prodotti del punto vendita, e ciascuna transazione corrisponde ai prodotti elencati in una ricevuta di vendita.

Ciascun oggetto che appare in una transazione è associato ad un numero *(f)* , che rappresenta la frequenza con la quale l'oggetto appare nell'insieme delle transazioni; mentre la frequenza

[36] Carlo Vercellis, *Business intelligence : data mining and optimization for decision making*, Hoboken, NJ, Wiley, 2009.
[37] Letteralmente: "Se condizione allora risultato".

empirica di un certo sottoinsieme L definisce il numero di transazioni esistenti nel dataset D che contengono l'insieme L. Il rapporto tra la frequenza empirica di L *(f(L))* e il numero totale di transazioni *(m)*, mostra la probabilità del verificarsi dell'insieme L *(Pr(L))* , interpretata come la probabilità che L sia contenuto in una nuova transazione registrata nel database.

Dati due sottoinsiemi $L \subset O$ e $H \subset O$, tali che $L \cap H = \varnothing$, e una qualsiasi transazione T, una regola di associazione è un'implicazione probabilistica, del tipo $L \Rightarrow H$, che significa: se L è contenuta in T, anche H è contenuta in T, con una data probabilità *(p)*, che è la confidenza della regola nel dataset D e che viene definita come $p = \text{conf}\{L \Rightarrow H\} = [f(L \cup H)] / [f(L)]$.

La confidenza della regola indica la quota di transazioni contenenti l'insieme H, tra quelle contenenti l'insieme L, ed esprime quindi l'affidabilità della regola. Un'elevata confidenza corrisponde ad una più alta probabilità che il sottoinsieme H esista in una transazione che contiene anche il sottoinsieme L.

Inoltre, la regola $L \Rightarrow H$ ha anche un supporto *s* nel dataset D, se la quota di transazioni che contengono sia L che H è uguale a *s*, ovvero se $s = \text{supp}\{L \Rightarrow H\} = [f(L \cup H)] / m$. Il supporto di una regola esprime la quota di transazioni che contengono sia l'antecedente che il conseguente della regola stessa, e misura quindi la frequenza con la quale essi appaiono assieme nelle transazioni del dataset. Un basso supporto suggerisce che la regola si verifica occasionalmente: le regole con un basso supporto sono regole di scarso interesse.

Una volta che un dataset D contenente *m* transazioni viene assegnato, e vengono fissate delle soglie minime di supporto e confidenza, vengono determinate tutte le regole di associazione forti, caratterizzate da un supporto e una confidenza superiori a quelli minimi fissati.

Molte regole non sono forti, nel senso che non hanno supporto e confidenza superiori a quelli prefissati: è quindi appropriato individuare un metodo che sia in grado di derivare regole forti, eliminando quelle che non raggiungono il minimo supporto e la minima confidenza. Questo processo consta di due fasi: la prima, quella della generazione di insiemi frequenti, estrae tutti gli insiemi di oggetti la cui frequenza relativa è più grande del minimo supporto assegnato; la seconda, cioè la generazione di regole forti, verifica se la confidenza della regola supera la soglia minima prestabilita.

Le regole forti non sono però sempre interessanti per coloro che devono prendere delle decisioni e, molto spesso, per definire una regola efficace, non sono sufficienti il supporto e la confidenza: bisogna infatti considerare un altro indice, il lift.

Il lift è così definito: $l = \text{lift}\{L \Rightarrow H\} = [\text{conf}\{L \Rightarrow H\}] / f(H) = [f(L \cup H)] / [f(L)\, f(H)]$.

Esso infatti può essere trovato dividendo la confidenza, per la frequenza del conseguente:

$\{[f(L \cup H)] / [f(L)]\} / f(H) = \{[f(L \cup H)] / [f(L)]\} \bullet [1/ f(H)] =$

$\{[f(L \cup H)] \bullet 1\} / [f(L) \bullet f(H)] = [f(L \cup H)] / [f(L) f(H)].$

Un valore del lift maggiore di 1 indica che la regola considerata è più efficace della stima ottenuta attraverso la frequenza relativa dell'antecedente e che antecedente e conseguente della regola sono associati positivamente. Al contrario, se il lift è inferiore di 1 la regola è meno efficace della stima ottenuta attraverso la frequenza relativa dell'antecedente e l'antecedente e il conseguente sono associati negativamente. In altre parole, se la regola considerata ha un lift maggiore di 1, essa mostra il grado in cui antecedente e conseguente dipendono l'uno dall'altro, evidenziando la potenziale utilità di quella regola nella previsione di comportamenti futuri. Nel caso in cui invece il lift fosse inferiore a 1, starebbe a significare che antecedente e conseguente non sono tra loro legati e nessuna regola forte può collegarli.

4.5.3. IL SUPERAMENTO DEI LIMITI FUNZIONALI[38]

La generazione di regole di associazione è un processo composto da più passaggi:

1. La creazione della matrice delle associazioni per i singoli prodotti.
2. La creazione della matrice delle associazioni per due prodotti, che viene usata per trovare regole con due prodotti.
3. La creazione della matrice delle associazioni per tre prodotti, che viene usata per trovare regole con tre prodotti.
4. E così via.

Aumentando il numero dei prodotti nelle combinazioni, si richiedono di conseguenza maggiori calcoli; se si considerano combinazioni con più di tre o quattro prodotti, i tempi di esecuzione di questi calcoli crescono in maniera esponenziale. La soluzione è una tecnica chiamata pruning, che permette di ridurre il numero di prodotti e le combinazioni di prodotti da considerare in ogni passaggio. Quindi questa tecnica viene applicata al termine dello sviluppo dell'albero, per ridurre il numero di ramificazioni, senza peggiorare, ma con l'obiettivo di migliorare, l'accuratezza del modello finale.

Il più comune meccanismo di pruning si chiama minimum support pruning. Esso richiede che una regola si rifaccia ad un numero minimo di transazioni e questo ha un senso, perché lo

[38] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

scopo di creare queste regole è quello di svolgere determinate azioni che portino a transazioni proficue.

Il minimum support pruning quindi elimina i prodotti che non compaiono in un certo numero di transazioni. Ci sono due alternative per farlo: la prima è quella di eliminare i prodotti dall'analisi; la seconda è quella di utilizzare la tassonomia per rendere i prodotti più generici, in modo tale da incontrare la soglia desiderata.

Grazie ad esso, ad ogni passaggio della tabella delle associazioni, è possibile eliminare le combinazioni di prodotti che non incontrano la soglia, riducendone la grandezza e diminuendo il numero di combinazioni da considerare nel passaggio successivo.

La scelta migliore per il minimum support dipende dai dati e dal contesto; è quindi possibile variare il minimum support come lo sviluppo dell'algoritmo.

Poiché si usano le probabilità per creare regole di associazione, i calcoli che vengono fatti devono essere fatti per ciascuna combinazione di prodotti: questo numero di combinazioni tenderà a crescere in maniera esponenziale, se si aumenta il numero di prodotti. E anche il calcolo del supporto e della confidenza diventa sempre più difficile, man mano che il numero di prodotti nelle combinazioni aumenta; anche se i computer stanno diventando sempre più veloci ed economici, è comunque molto costoso fare questi calcoli per un numero così elevato di combinazioni: l'uso della tassonomia riduce il numero di prodotti per renderlo maneggevole.

Inoltre, anche il numero delle transazioni è molto elevato, quindi, determinando se una particolare combinazione di prodotti è presente in una determinata transazione può richiedere un po' di sforzo, che va moltiplicato per tutte le transazioni.

4.6. L'ANALISI DELLE SERIE TEMPORALI ATTRAVERSO L'UTILIZZO DELLA MARKET BASKET ANALYSIS

La market basket analysis analizza fatti che succedono in uno stesso momento, ossia quali prodotti vengono acquistati in un dato momento.

Una serie temporale, invece, è una sequenza ordinata di prodotti. Si differenzia da una transazione solo per il fatto che è ordinata. Generalmente una serie temporale contiene informazioni che identificano il cliente, dal momento che questi dati vengono utilizzati per legare assieme le diverse transazioni in una serie. E una delle tecniche utilizzate per analizzare una serie temporale è proprio la market basket analysis.[39]

[39] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Per utilizzare le serie temporali, i dati delle transazioni devono avere due ulteriori caratteristiche: il timbro dell'ora o una sequenza di informazioni che determinino come si sono verificate le transazioni; informazioni identificative, come numero di conto, documento di identità, che mostrino che differenti transazioni appartengono allo stesso consumatore o famiglia.

Le transazioni che corrispondono allo stesso consumatore sono raccolte in una serie temporale grazie alle informazioni sopracitate.

Diversamente dalle transazioni, queste serie mostrano quali prodotti vengono prima, dopo, nello stesso momento di altri; possono inoltre contenere prodotti uguali.

Un modo per utilizzare le serie temporali è analizzare le cause e gli effetti: questo infatti è un metodo per trovare le cause di un evento che accade in un particolare momento.

Si tratta quindi ora di trasformare il problema delle serie temporali in un problema di market basket analysis: ogni volta le serie temporali vengono convertite in una transazione includendo i prodotti precedenti o successivi l'evento di interesse e rimuovendo i prodotti doppi dalla transazione. Si avrà quindi un insieme di transazioni favorevoli alla market basket analysis.

Le finestre temporali sono poi un altro modo per interpretare le serie temporali. E' particolarmente utile quando ci sono pochi prodotti che si verificano in un arco di tempo. Una finestra temporale è una descrizione sintetica di tutti i prodotti che si verificano all'interno di un certo periodo di tempo. Un esempio è quello di raccogliere tutte le transazioni fatte in un mese in un'unica transazione.

4.7. PUNTI DI FORZA E PUNTI DI DEBOLEZZA DELLA MARKET BASKET ANALYSIS[40]

Punti di forza	Punti di debolezza
Produce risultati chiari e comprensibili. Funge da supporto per il data mining. Lavora con dati di lunghezza variabile. I calcoli che utilizza sono semplici da capire.	Richiede sforzi di calcolo che crescono in maniera esponenziale con il crescere della dimensione del problema. Ha un supporto limitato per gli attributi sui dati. E' difficile determinare il giusto numero di prodotti. Ignora i prodotti rari.

I risultati sono compresi in modo chiaro

I risultati della market basket analysis sono regole di associazione. L'espressione "if-then" di queste regole rende i risultati facili da interpretare e da trasformare in azioni. In alcune circostanze, soltanto l'insieme dei prodotti collegati è di interesse e le regole non hanno ancora bisogno di essere generate.

La market basket analysis è utile per il data mining

Il data mining è molto importante quando si rivolge ad una grande quantità di dati e non si sa da che parte iniziare con l'analisi. La market basket analysis è una tecnica appropriata, quando può essere applicata, per analizzare i dati e dare inizio all'analisi, e fornisce inoltre risultati chiari e comprensibili.

Lavora su dati di lunghezza variabile

La market basket analysis gestisce dati di lunghezza variabile senza aver bisogno di essere riassunti. Altre tecniche invece sono solite richiedere dati in uno stesso formato, cosa che non è una rappresentazione naturale dei dati delle transazioni. La market basket analysis riesce quindi ad occuparsi di transazioni senza perdere nessuna informazione.

Calcoli semplici

I calcoli necessari per applicare la market basket analysis sono abbastanza semplici, anche se il numero di questi calcoli cresce molto rapidamente con il numero delle transazioni e il numero dei differenti prodotti nell'analisi.

[40] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Crescita esponenziale man mano che cresce il problema

I calcoli richiesti per generare regole di associazione crescono in maniera esponenziale con il numero di prodotti e la complessità delle regole che vengono considerate. La soluzione è quella di ridurre il numero di prodotti, rendendoli generici: ma prodotti più generici sono solitamente meno corretti. Metodi di controllo dei calcoli, come il minimum support pruning, tendono ad eliminare importanti regole che invece dovrebbero essere tenute in considerazione.

Supporto limitato per gli attributi sui dati

La market basket analysis è una tecnica specializzata per i prodotti di una transazione. I prodotti devono essere identici tranne che per una caratteristica identificativa, come il tipo di prodotto. Quando è applicabile, la market basket analysis ha un forte potere, ma non tutti i problemi rientrano in questa descrizione. L'utilizzo della tassonomia e i prodotti virtuali aiutano a rendere le regole più espressive.

Determinare i giusti prodotti

Probabilmente il problema più difficile nell'applicare la market basket analysis è quello di determinare il corretto insieme di prodotti da utilizzare nell'analisi. Rendendo i prodotti generici attraverso la tassonomia, si garantisce una stessa frequenza dei prodotti utilizzati nell'analisi, ma questo processo causa la perdita di alcune informazioni, e i prodotti virtuali devono essere nuovamente inseriti nell'analisi per recuperare queste informazioni.

La market basket analysis ha delle difficoltà con i prodotti rari

La market basket analysis lavora meglio quando tutti i prodotti hanno approssimativamente la stessa frequenza nei dati. I prodotti che appaiono raramente sono presenti in un numero molto limitato di transazioni e saranno quindi tagliati dall'analisi. Modificare la soglia del minimo supporto per prendere in considerazione il valore del prodotto è l'unico modo per garantire che i prodotti costosi vengano tenuti in considerazione, anche se potrebbero essere rari nei dati. L'utilizzo delle tassonomie può garantire che i prodotti rari vengano inclusi in qualche forma nell'analisi.

4.8. APPLICAZIONI DELLA MARKET BASKET ANALYSIS

La market basket analysis viene applicata nei problemi indiretti di data mining che riguardano prodotti ben definiti, che sono raggruppati assieme in modi interessanti. Questi problemi si verificano solitamente nel commercio al dettaglio, dove le transazioni nei punti vendita sono alla base dell'analisi. Problemi simili possono essere riscontrati anche in altri settori.[41]

[41] Adriaans e Zantinge, Data Mining, Addison-Wesley, 1996.

La market basket analysis può essere applicata anche per alcuni problemi diretti di data mining in questi settori. Può essere eseguita su un sottoinsieme ben definito di transazioni. L'algoritmo di base può anche essere modificato se si vogliono considerare solamente regole che contengano un particolare prodotto, come per esempio potrebbe essere un nuovo prodotto, per creare dei modelli di vendita.[42]

L'ambito delle serie temporali, inoltre, è un'altra area dove questi metodi possono essere applicati. Alcuni problemi delle serie temporali possono essere adattati alla market basket analysis attraverso delle semplici trasformazioni nei dati delle serie temporali.

E' necessario però prestare molta attenzione durante l'utilizzo della market bastek analysis: esistono infatti tre "trappole" nelle quali si rischia di imbattersi.

La prima riguarda il fatto che i dati devono essere raccolti o conservati per tutti i possibili tipi di risultati interessanti nello studio che si sta svolgendo.

La seconda suggerisce che i dati d'archivio raccolti dalle imprese, dalle associazioni professionali, e da altre organizzazioni, sono solitamente differenti dai dati raccolti dalla ricerca accademica.

Infine, come terza trappola, c'è il fatto che imparare un nuovo strumento metodologico o una nuova tecnica statistica richiede una concreta esperienza.

[42] Robert Groth, Data Mining: A Hands-on Approach for Business Professionals, Prentice Hall, 1997.

5. IL GRUPPO "WHITE" NELLA REALTA' AZIENDALE PAM

5.1. IL GRUPPO "WHITE"

Si effettua ora un'analisi statistica svolta prendendo in considerazione Pam, azienda italiana della Grande Distribuzione, presente sul territorio nazionale con 109 supermercati a gestione diretta; si considera qui il punto vendita situato a Spinea (VE).

Il data set considerato contiene 12 variabili e più di un milione di osservazioni riferite ad un insieme di prodotti, soprannominati "white": riso, latte, biscotti.[43]

Si riporta di seguito un esempio.

VARIABILE	SIGNIFICATO	ESEMPIO
DIB_BASK_ID	Codice identificativo del basket (carrello della spesa).	144141301624908
DIB_TIME_CODE	Settimana dell'acquisto.	201349: si tratta della settimana 49 dell'anno 2013.
DIB_SHOP_DATE	Data dell'acquisto.	20131208: la data è riportata nel formato americano e si tratta quindi del giorno 8 dicembre 2013.
DIB_WEEKDAY	Giorno della settimana dell'acquisto.	1: corrisponde a lunedì.
DIB_HOUR	Ora dell'acquisto.	8: l'orario corrisponde alle ore 8 del mattino.
FLAG_PROMO	Prodotto soggetto o meno a promozione (S=sì; N=no).	N
DIB_PROD_CODE	Codice del prodotto.	8387695
DIB_PROD_DESC	Nome del prodotto.	BISCOTTO NOVARA
DIB_PROD_LEVEL10	Caratteristiche del prodotto.	10_218022: è il codice che nel dataset di Pam si riferisce ai prodotti che hanno come caratteristica quella di appartenere alla "BISCOTTERIA A PESO".
DIB_PROD_LEVEL20	Categoria di prodotto (biscotti, riso, latte).	20_218020: è il codice che nel dataset di Pam si riferisce ai prodotti appartenenti alla categoria "BISCOTTI".

[43] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

DIB_QUANTITY	Quantità acquistata.	1
DIB_SPEND	Prezzo unitario (in €).	2,99

Il gruppo white contiene tre categorie di prodotto: il latte, il riso e i biscotti, che a loro volta nel database si riferiscono ad una vasta gamma di prodotti ciascuno: esattamente si dispone di 46 prodotti appartenenti alla categoria "latte", 82 prodotti appartenenti alla categoria "riso", 224 prodotti appartenenti alla categoria "biscotti".

Riportati nella tabella sottostante vi sono alcuni esempi dei prodotti contenuti nelle tre categorie considerate.[44]

LATTE	RISO	BISCOTTI
LATT.FR. TAPPOROSSO AQ (1500 ML TP)	BLOND INSALATE GALLO (1000 GR SC)	ABBRACCI M.B. (700 GR PK)
LATTE CAPRA INTERO (1000 ML BT)	CURTIRISO AMBRA INSALA (1000 GR AS)	ALCE FROL FARRO CIOCCO (300 GR SK)
LATTE CRESCITA ESL GRA (500 ML BO)	G.RISO CARNAR.IGP DELT (1000 GR PK)	ALCE N FROLLINI KAMUT (300 GR SK)
LATTE ESL +G.INTERO GR (1500 ML BT)	GALLO BLOND RISOTTI (1 KG AS)	AMARETTI ARTIGIANALI
LATTE ESL +G.PS.GRANAR (1500 ML BT)	GALLO CHIC.PIU BETAGLU (400 GR SC)	AMARETTI BAROVERO
LATTE ESL SCREMATO GRA (1000 ML BT)	GALLO RISO ROSS.CAMARG (500 GR SC)	BACI DI DAMA
LATTE ESL.INT.BIO GRAN (1000 ML PK)	GR.RISO ARBOR.IGP DELT (1000 GR PK)	BACI DI DAMA CACAO
LATTE FERM.INTERO ORO (1000 ML BT)	GRANDI RISO ARBORIO (1000 GR SC)	BAROV.BACI PRELIBATI
LATTE FR A.Q. GRANAROL (1000 ML BT)	GRANDI RISO BASMATI (1000 GR CE)	BIS CHICCHI DI CIOCCOL (300 GR SK)
LATTE FR INT INGL.LTMI (1000 ML BT)	GRANDI RISO ORIGINARIO (1000 GR PK)	BIS KAMUT SENZA ZUCCHE (300 GR SK)
LATTE FR TRSO BIO (1000 ML BT)	GRANDI RISO RIBE (1000 GR AS)	BIS NOVELLINI 250 GENT (250 GR SC)
LATTE FR. A.Q. LACTIS (500 ML BK)	GRANDI RISO ROMA (1000 GR SC)	BIS SENZA LATTE E UOVA (400 GR SK)
LATTE FR. AQ P&P L 1 (1000 ML PT)	GRANO C`E` DI BUONO (500 GR AS)	BIS. GRANCEREALE M.B. (500 GR CO)
LATTE FR.A.Q CLMI (500 ML TP)	MIX 5 CEREALI BUONO (500 GR AS)	BIS.PAIN CROUTE INTEGR (450 GR PK)
LATTE FR.A.Q. LACTIS (1000 ML BT)	MIX RISO3CONT NATTURA (500 GR SK)	BISC COLUS G TURCHESE (400 GR PK)
LATTE FR.A.Q. P&P (1500 ML BO)	P&P BIO RISO ARBORIO (1000 GR AS)	BISC INTEGR ARTE BIANC (400 GR AS)
LATTE FR.A.Q. PET ABIT (1000 ML BT)	P&P RISO ARBORIO SV (1 KG AS)	BISC OSVEGO 250 GENTIL (250 GR SC)
LATTE FR.A.Q.PET CLTO (1000 ML BT)	P&P RISO BASMATI (1000 GR AS)	BISC. CAMPAGNOLE M.B. (1000 GR SK)
LATTE FR.A.Q.PROB ABIT (1000 ML BT)	P&P RISO BASMATI PRECO (250 GR AS)	BISC. MACINE MB (1000 GR SK)

[44] In Appendice, Tabelle 1, 2, 3, sono riportati gli elenchi completi dei prodotti considerati nelle tre categorie, con annesse quantità vendute nel periodo di riferimento.

LATTE FR.A.Q.TOP LTMI (1000 ML TP)	P&P RISO PARBOILED (2000 GR CE)	BISC. TARALLUCCI M.B. (1000 GR SK)
LATTE FR.A.Q.VETR CLTO (750 ML BT)	P&P RISO RIBE SV (1000 GR AS)	BISC.ABBRACCI M.BIANCO (350 GR SK)
LATTE FR.INT CLTO (500 ML TB)	P&P RISO.CARNAROLI.SV (1000 GR AS)	BISC.CAMPAGNOLE M.B. (700 GR SK)
LATTE FR.INT.A.Q.CLMI (1000 ML TP)	P&P RISO.ORIGINAR.SV (1000 GR AS)	BISC.CORLEGGERI GRONDO (250 GR PK)
LATTE FR.INT.TAPRS PET (1000 ML BT)	P&P RISO.PARBOILED (1 KG AS)	BISC.CUOR DI MELA MB (300 GR AS)
LATTE FR.PS.BIO GRANAR (1000 ML BK)	P&P.RISO.BASMATI. (500 GR AS)	BISC.ENERGELLI SZ (300 GR SK)
LATTE FRE.INT.ROS.ABIT (1000 ML TP)	P&P.RISO.INTEGRALE.SV (1000 GR AS)	BISC.FIOR DI LATTE MB (300 GR SK)
LATTE FRE.LEGG.GRANAR. (500 ML TB)	P&P.RISO.ROMA.SV (1 KG AS)	BISC.GALLETTI M.B. (1000 GR SK)

5.2. ANALISI DESCRITTIVA

Si utilizza per la maggior parte dell'analisi il livello di prodotto 20:

DIB_PROD_LEVEL20,DIB_PROD_LEVEL20_DESC
20_13002,RISO
20_218020,BISCOTTERIA A PESO E A NUMERO
20_57001,LATTE FRESCO INTERO
20_57003,LATTE FRESCO SCREMATO
20_57004,LATTE FRESCO MICROFILTRATO/ARRICCHITO
20_8004,BISCOTTI CONFEZIONATI

Esso tiene in considerazione le categorie di prodotto "latte", "riso", "biscotti", ma all'interno della categoria latte effettua una divisione in "latte fresco intero", "latte fresco scremato", "latte fresco microfiltrato/arricchito", e nella categoria "biscotti" considera la divisione in "biscotteria a peso e a numero", "biscotti confezionati". Ciascuna di queste divisioni contiene poi tutti i prodotti considerati nell'analisi. [45]

Si consideri infatti ora il totale delle quantità vendute (asse delle ordinate) di ciascun prodotto (asse delle ascisse), con riferimento al periodo preso in analisi, che inizia l'8 dicembre 2013 e si conclude il 24 giugno 2014.

[45] In Appendice, Tabelle 1, 2, 3.

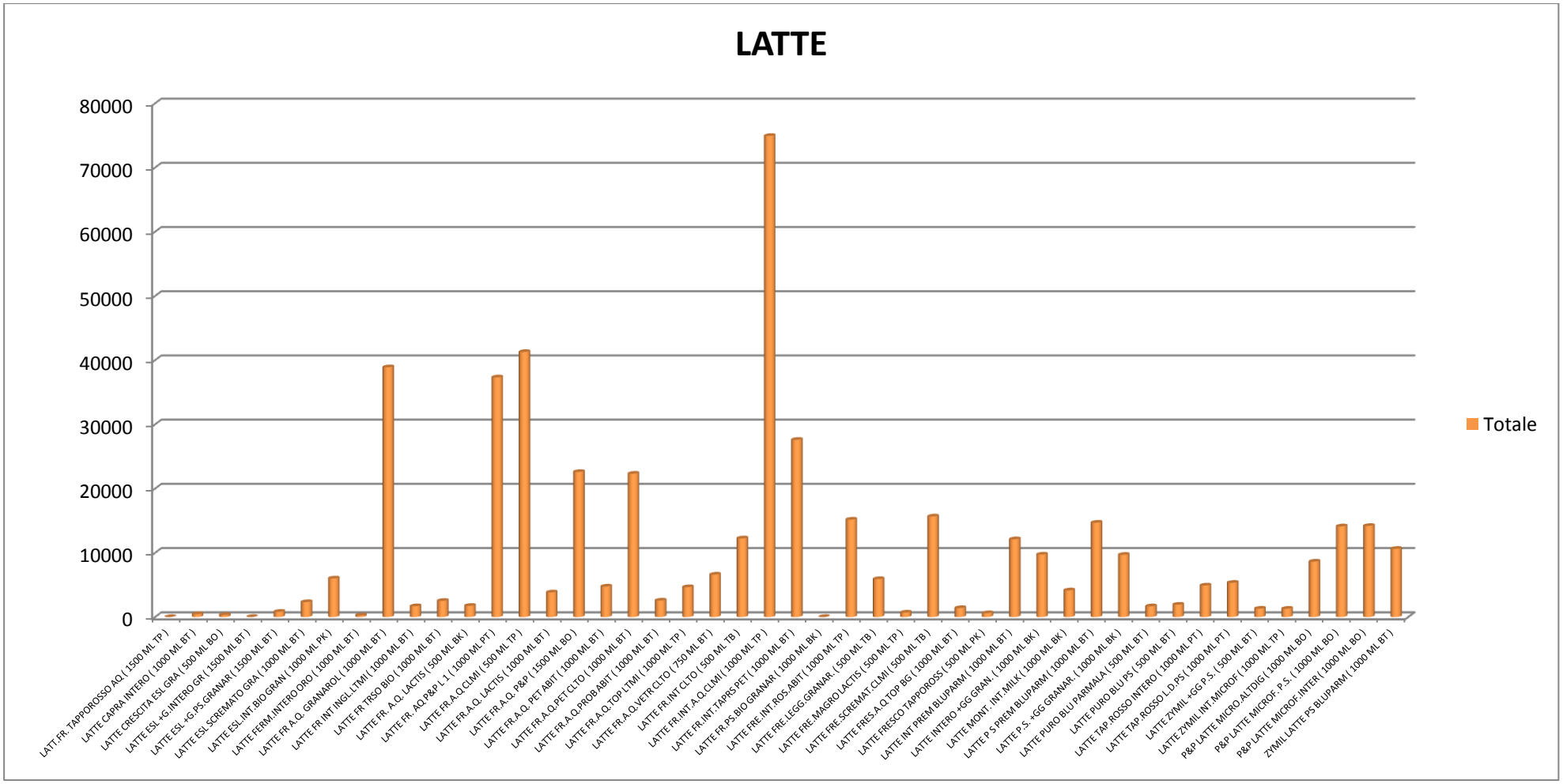


Figura 1. Quantità vendute nel periodo di riferimento dei prodotti appartenenti alla categoria “Latte”.

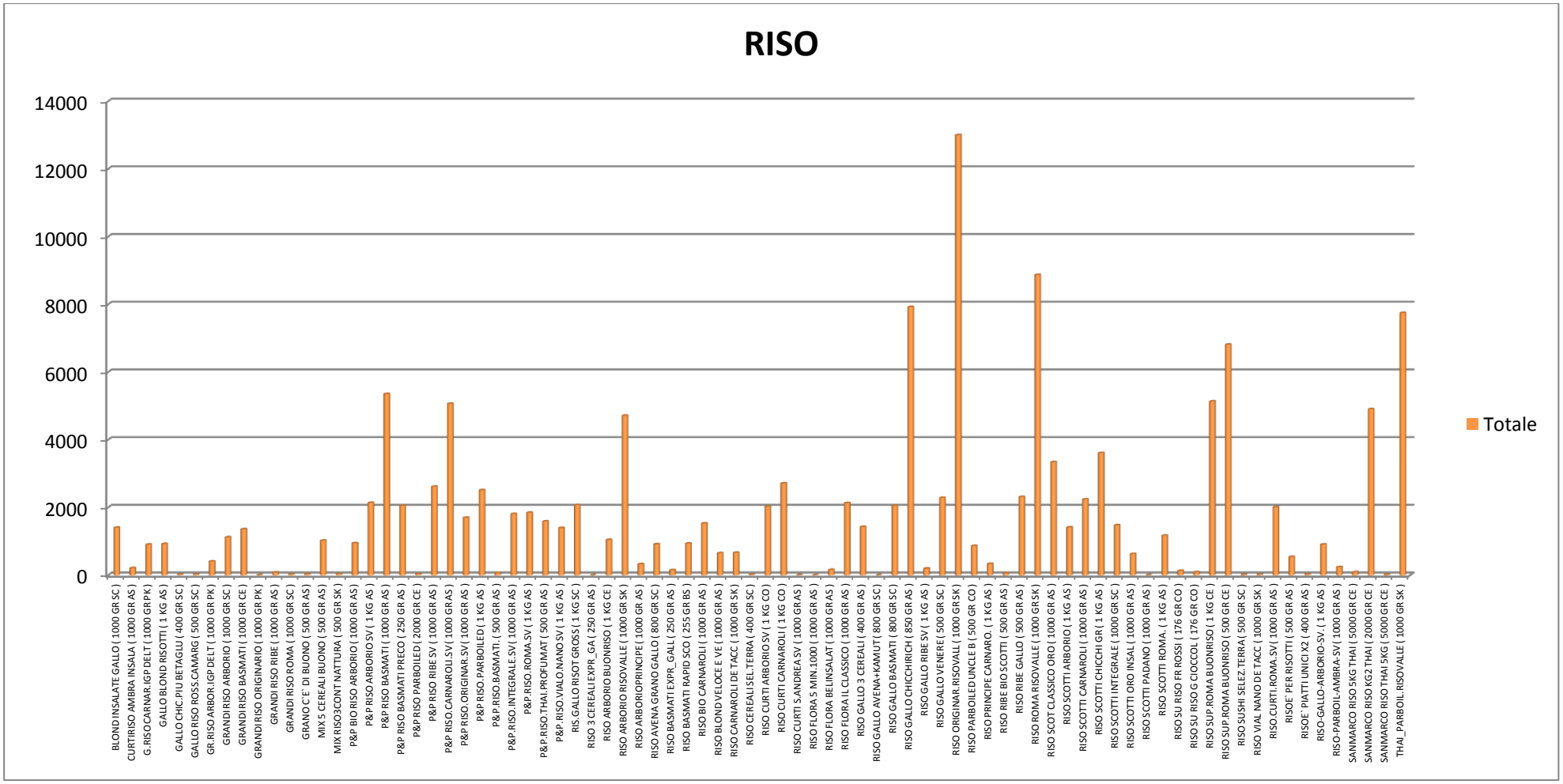


Figura 2. Quantità vendute nel periodo di riferimento dei prodotti appartenenti alla categoria “Riso”.

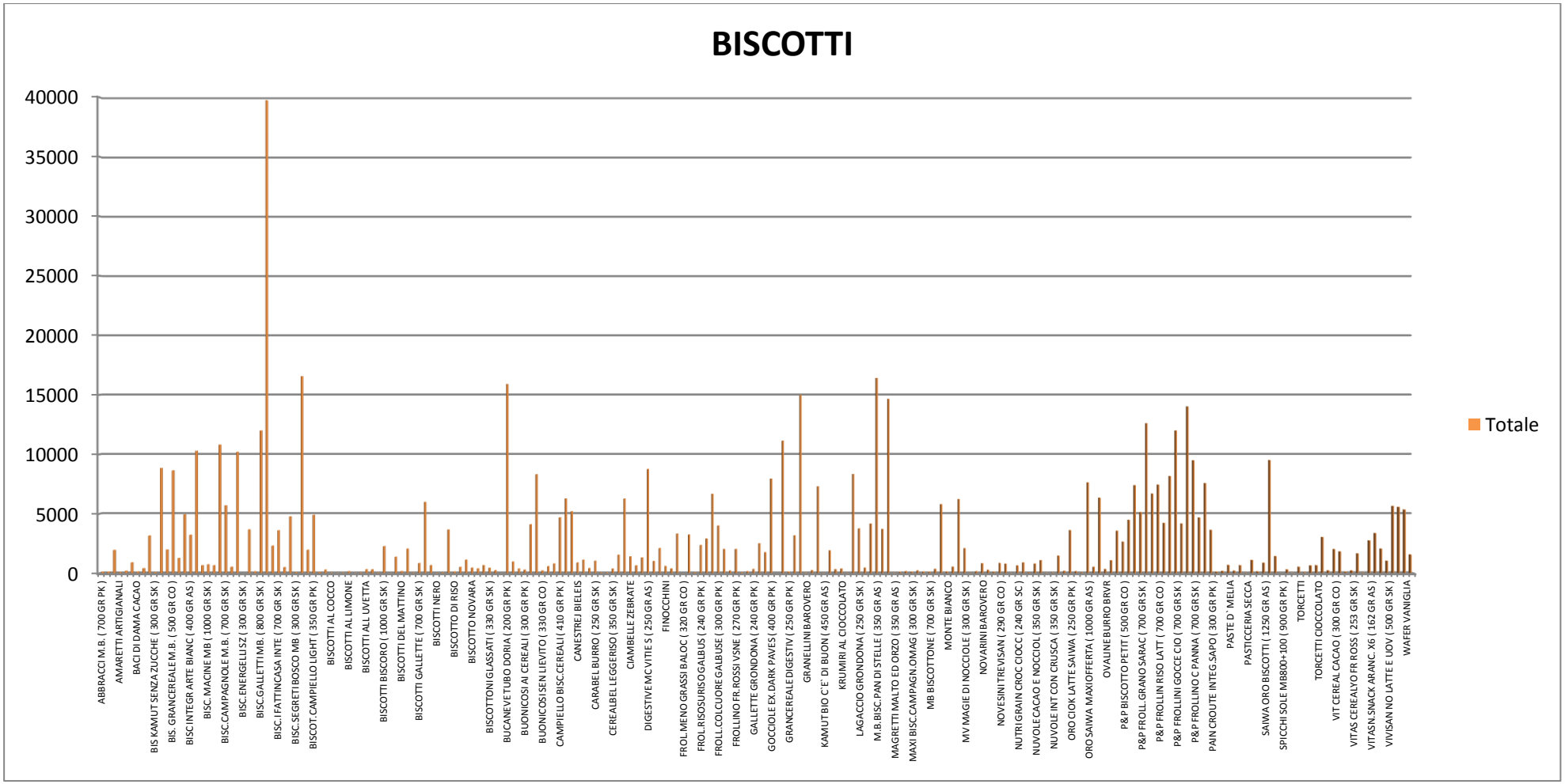


Figura 3. Quantità vendute nel periodo di riferimento dei prodotti appartenenti alla categoria “Biscotti”. Dato l’elevato numero di prodotti appartenenti alla categoria in oggetto, le 224 etichette dell’asse orizzontale non compaiono per intero nel grafico. Inoltre, quando la quantità di prodotti venduti è inferiore circa alle 500 unità, sembra non comparire nel grafico: si veda quindi in Appendice, Tabella 3, l’esatta quantità di ciascun prodotto acquistato, appartenente alla categoria “biscotti”.

E’ possibile a questo punto evidenziare l’andamento delle vendite durante il periodo preso in considerazione, che inizia l’8 dicembre 2013 (settimana numero 49 del 2013) e si conclude con il 24 giugno 2014 (settimana numero 26 del 2014).[46]

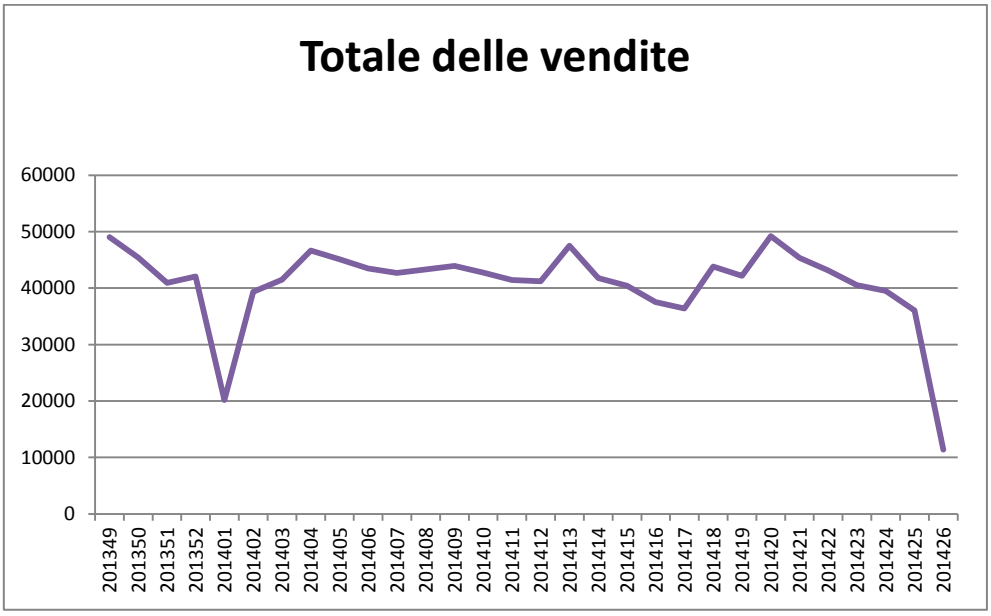


Figura 4. Andamento delle vendite, dalla settimana numero 49 del 2013 alla settimana numero 26 del 2014.

Sull’asse delle ascisse vi è l’intero periodo considerato, diviso in settimane, dalla numero 49 del 2013 alla numero 26 del 2014; l’asse delle ordinate indica invece, in migliaia, la quantità venduta dei prodotti appartenenti alle tre categorie considerate. Il tracciato indica dunque l’andamento delle vendite dei prodotti contenuti in riso, latte e biscotti, dalla settimana numero 49 del 2013, che inizia l’8 dicembre 2013, alla settimana 26 del 2014, che si conclude il 24 giugno 2014.

Si può notare un andamento delle vendite pressoché costante durante l’intero periodo, ad eccezione di due forti cali, corrispondenti alle prime settimane dell’anno e ai mesi estivi, solitamente periodi in cui le promozioni sui prodotti sono in diminuzione.

Nel dettaglio, i grafici riferiti alle settimane numero 1 e numero 20 del 2014, rispettivamente settimane di picco negativo e picco positivo delle vendite.[47]

[46] In Appendice, Tabella 4.
[47] In Appendice, Tabelle 5, 6.

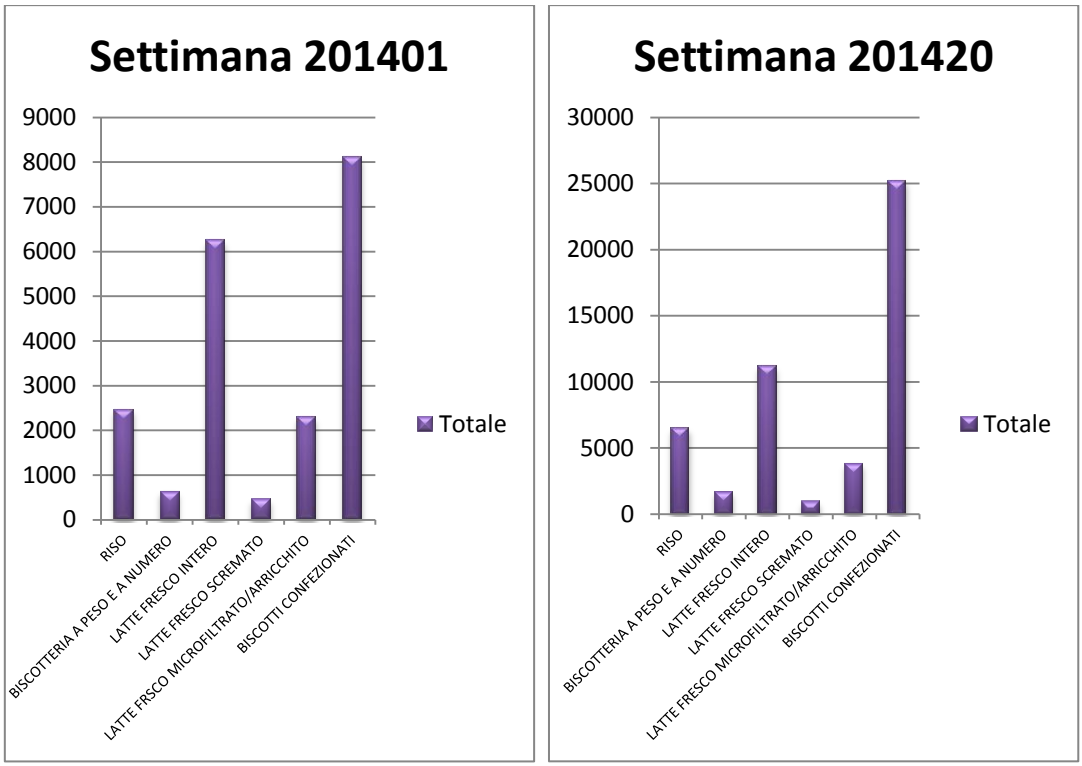


Figura 5. Quantità vendute nelle settimane 201401 e 201420.

Vediamo attraverso i grafici alcuni esempi: nella settimana numero 1 del 2014 la quantità venduta di riso è inferiore alle 3000 unità, mentre, nella settimana numero 20, vengono superate le 5000 unità; la vendita dei biscotti confezionati raggiunge solo le 8000 unità nella settimana numero 1, mentre supera le 25000 unità nella settimana numero 20; lo stesso succede poi con il latte fresco intero e il latte fresco microfiltrato arricchito, che rispettivamente toccano, nella settimana numero 1 le 6000 circa e le 2000 circa unità vendute, e nella settimana 20, le 10000 circa e le 5000 circa unità vendute.

Sono di conseguenza particolarmente interessanti i grafici relativi all'andamento delle vendite delle singole categorie di prodotto, sempre nell'arco del periodo considerato e con riferimento al livello 20.[48]

[48] In Appendice, Tabelle 7, 8, 9, 10, 11, 12.

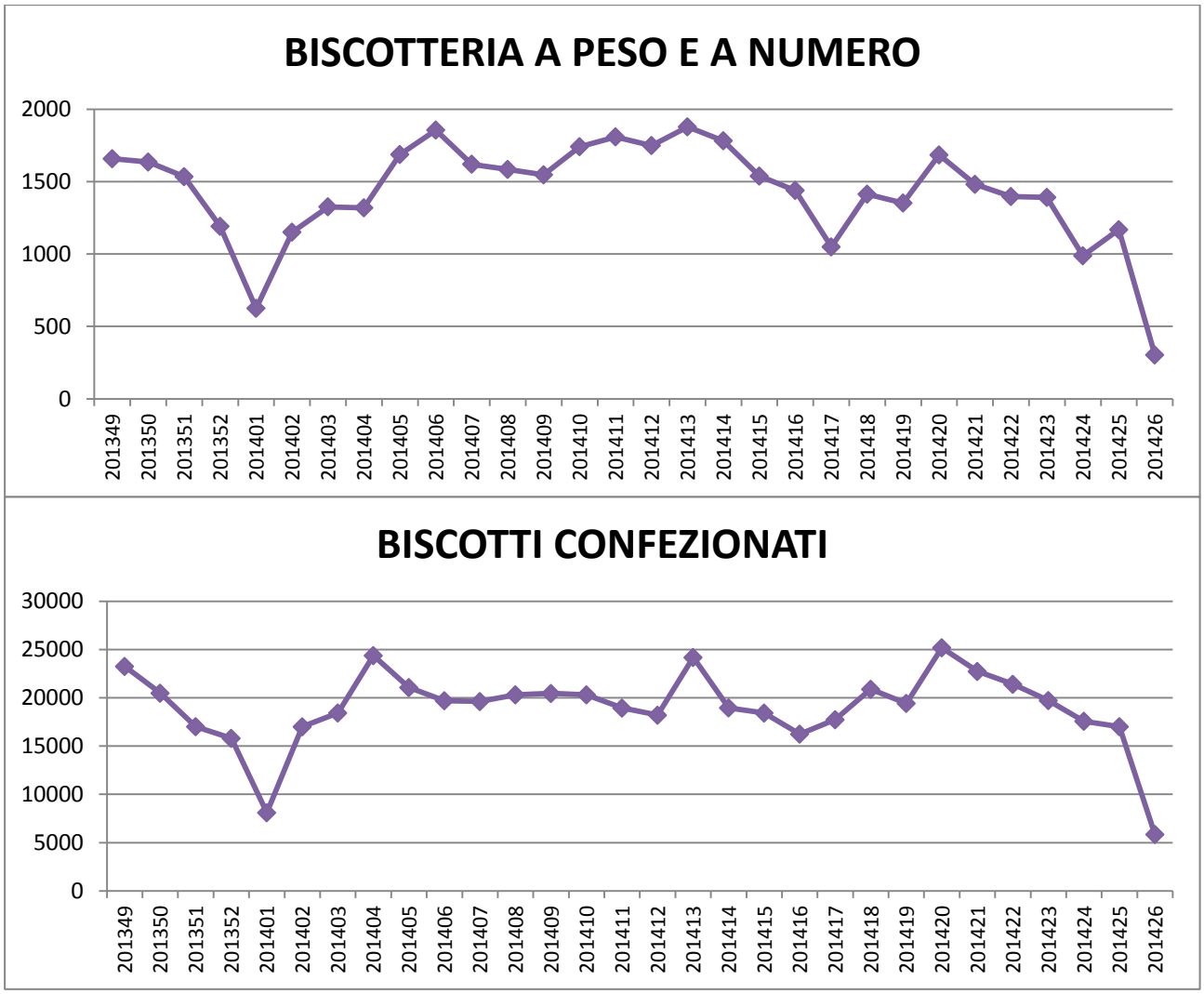


Figura 6. Andamento delle vendite della categoria “biscotti”.

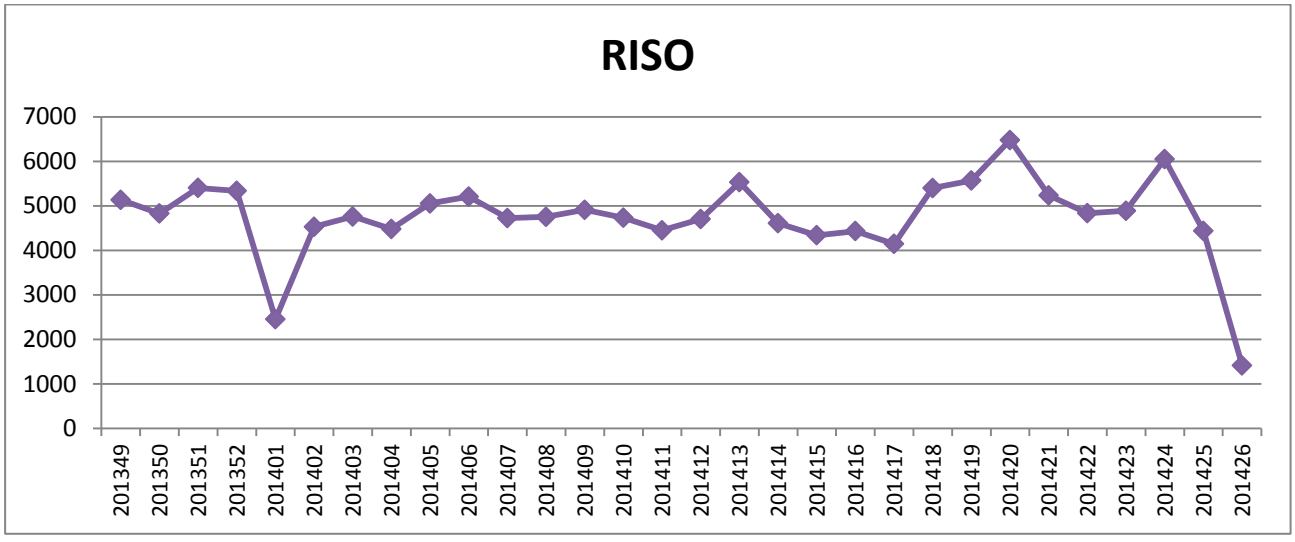


Figura 7. Andamento delle vendite della categoria “riso”.

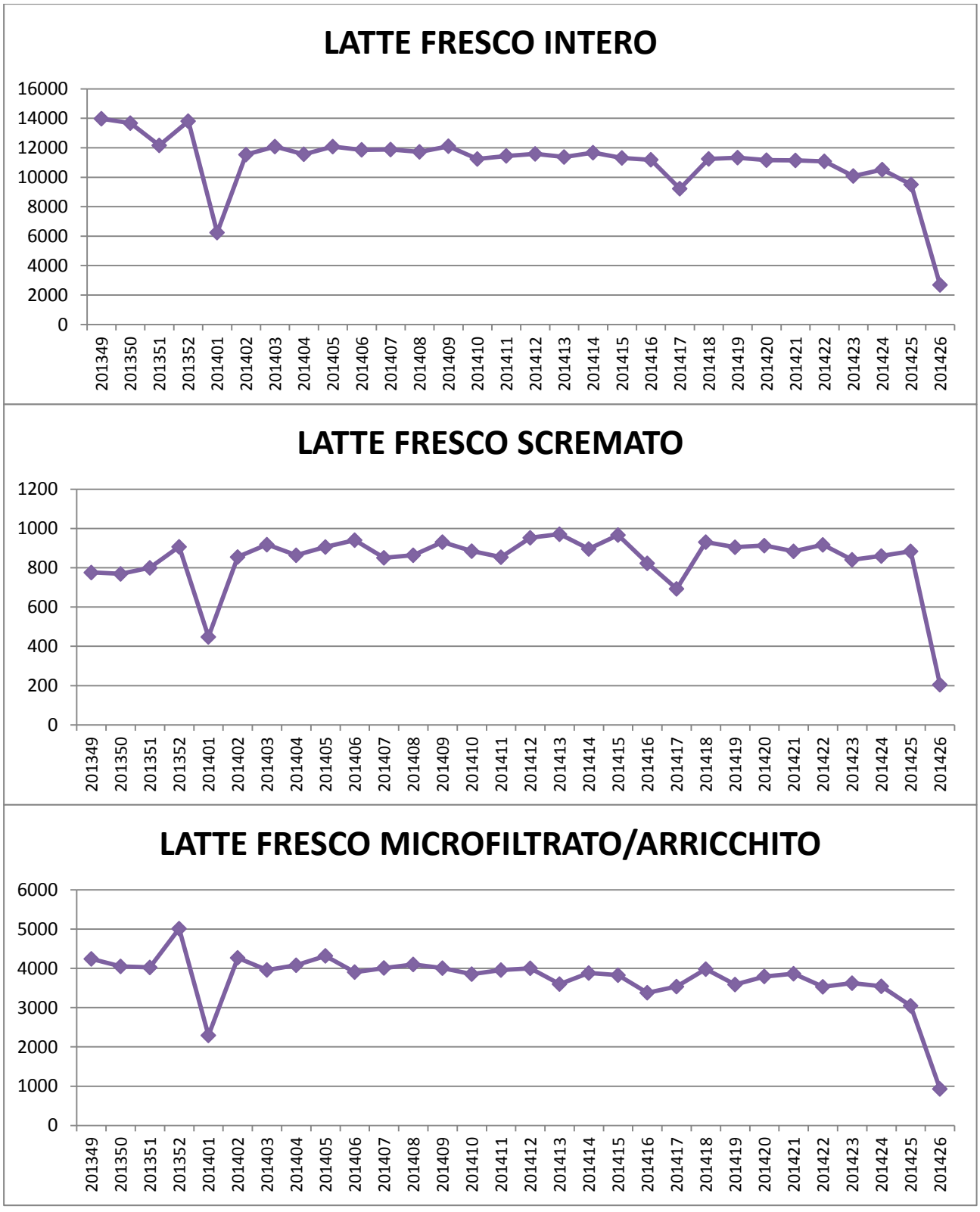


Figura 8. Andamento delle vendite della categoria "latte".

I grafici rappresentano l'andamento delle vendite delle categorie di prodotti considerate: in ascissa sono presenti le settimane del periodo di riferimento, in ordinata vi è la quantità vendute, in migliaia, dei prodotti in oggetto.

L'andamento delle vendite, anche in questo caso, sottolinea la presenza di picchi negativi in corrispondenza di periodi in cui si applica uno scarso numero di promozioni.

Si veda ora quindi un grafico che analizzi questo aspetto, cioè le promozioni attive sui prodotti appartenenti alle categorie "latte", "riso", "biscotti".[49]

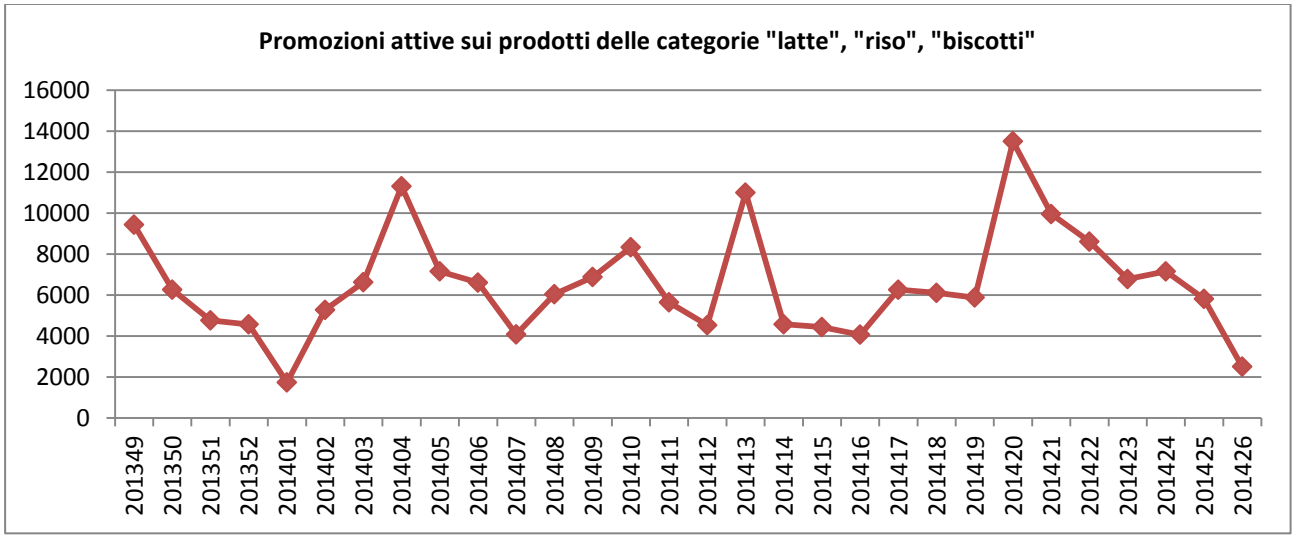


Figura 9. Andamento delle promozioni attive sui prodotti delle categorie "latte", "riso", "biscotti".

Il grafico analizza le promozioni effettuate su tutti i prodotti appartenenti alle tre categorie, in maniera congiunta: sull'asse delle ascisse ci sono infatti le settimane del periodo di riferimento, dalla numero 49 (che inizia l'8 dicembre 2013) alla numero 26 (che si conclude il 24 giugno 2014); sull'asse delle ordinate vi è invece il numero corrispondente alla quantità di promozioni attive; i nodi nel grafico mostrano dunque il livello di promozioni attive nella specifica settimana considerata in ascissa.

Osservando il grafico si possono notare proprio delle diminuzioni delle promozioni nei periodi in cui si riscontravano delle diminuzioni delle vendite, ossia i primi mesi dell'anno e i mesi estivi.

Inoltre, se per la maggior parte delle settimane, l'andamento delle promozioni resta per lo più costante, si possono comunque intravedere dei picchi in cui in numero delle promozioni aumenta: essi corrispondono alle settimane 4, 13, 20.

Risulta ora necessario quindi esaminare su quali categorie di prodotti del gruppo considerato, il gruppo "white" (latte, riso, biscotti), le promozioni hanno una maggiore applicazione.[50]

[49] In Appendice, Tabella 13.
[50] In Appendice, Tabelle 14, 15, 16, 17, 18, 19.

Ci si riferisce quindi nuovamente alle categorie di prodotto finora considerate, livello 20:

- biscotteria a peso e a numero;
- biscotti confezionati;
- riso;
- latte fresco intero;
- latte fresco scremato;
- latte fresco microfiltrato/arricchito.

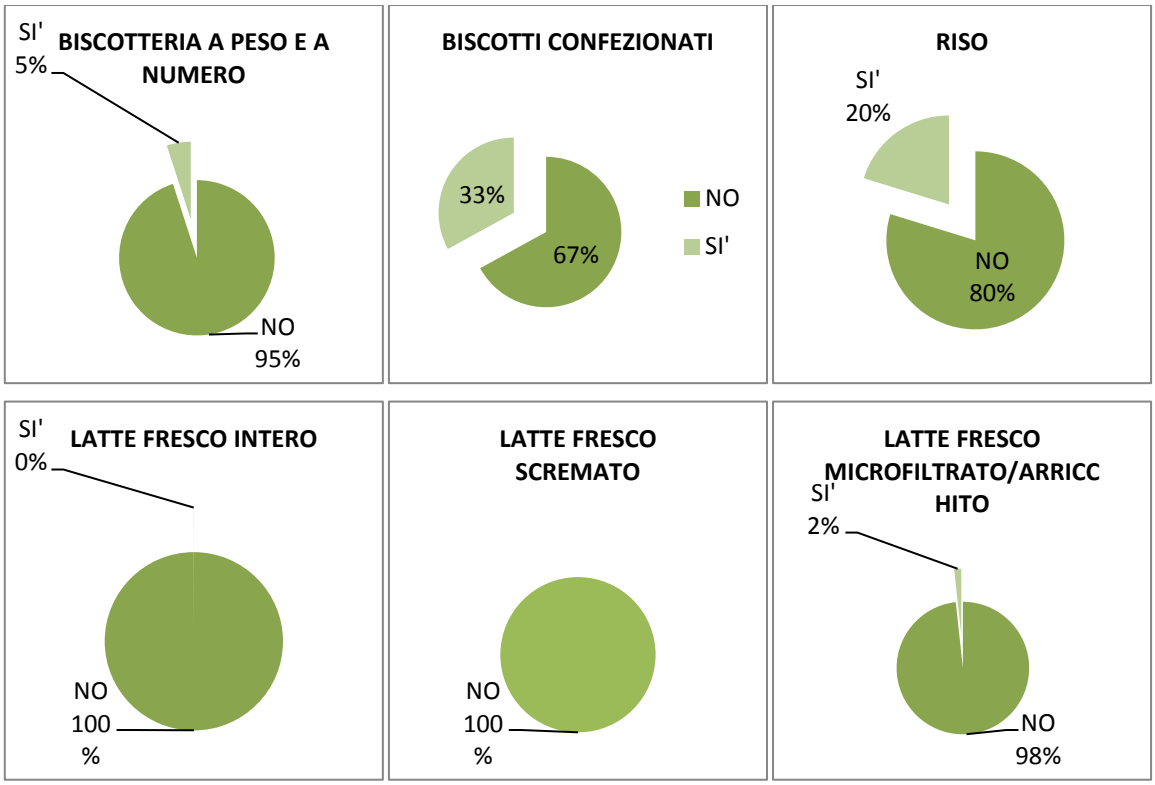


Figura 10. Applicazione delle promozioni sulle categorie di prodotto, livello 20.

I grafici mostrano la percentuale di promozioni attive con riferimento alle categorie di prodotto fin qui considerate.

Le considerazioni che si possono fare su essi sono di notevole importanza.

Il latte fresco non è quasi mai soggetto a promozioni e il suo prezzo rimane sostanzialmente stabile nell'arco dell'anno; il riso è invece un prodotto abbastanza promozionato nell'arco dell'anno, al quale si applicano tre diversi tipi di promozione: promozioni coupon, promozioni display, promozioni feature.[51]

[51] Le promozioni coupon offrono un risparmio immediatamente rimborsabile per alcuni prodotti: riducendo il prezzo di un prodotto infatti, lo rendono più desiderabile agli occhi del consumatore.

Ma più applicate ancora sono le promozioni sulla vendita dei biscotti. Si evince quindi che l'effetto di una promozione applicata ai biscotti si ripercuote sulle vendite dei prodotti a questi associati, riso e latte fresco.

5.3. ANALISI STATISTICA

Viene considerato ora il problema di trovare relazioni e regole utili da un grande dataset. Da ricordare che una relazione è ciò che ci fornisce informazioni su aspetti particolari dei dati, dandone una completa descrizione. Poi, dato un insieme di tutte le relazioni possibili, è necessario di volta in volta verificare se ciascuna di esse è significativa per l'analisi.

Quindi le regole di associazione trovano la loro principale applicazione nella market basket analysis, con la quale vengono esaminati grandi insiemi di dati, in modo tale da scoprire quali prodotti vengono venduti insieme nelle singole esperienze di acquisto: tutti i prodotti che corrispondono allo stesso ID di transazione formano una singola transazione che li contiene tutti.[52]

L'intero data set di un supermercato ha dimensioni eccessivamente elevate: è stato quindi considerato un campione che comprendesse le vendite dei prodotti delle categorie "latte", "riso", "biscotti", nel periodo compreso tra l'8 dicembre 2013 e il 24 giugno 2014. Infatti la frequenza di una regola nel campione sarà approssimativamente la stessa dell'intero dataset, quindi le relazioni scoperte nel campione producono ugualmente interessanti risultati.[53]

Si fa quindi nuovamente riferimento al livello 20, che considera le tre categorie di prodotto e non i singoli prodotti in essa contenuti, poiché questi risulterebbero una quantità troppo elevata per portare ad un concreto e visibile risultato (46 prodotti della categoria "latte", 82 della categoria "riso", 224 della categoria "biscotti").

L'esistenza di associazioni tra le tre categorie di prodotto è stata verificata utilizzando un programma statistico chiamato "R". Si tratta di un software disponibile per diversi sistemi operativi, il cui linguaggio è specifico e orientato agli oggetti. Anche se il linguaggio è fornito con un'interfaccia a riga di comando, sono disponibili diverse interfacce grafiche che consentono di integrare R con diversi pacchetti, tra cui "Arules", usato per questa analisi.

Le promozioni display consistono in aree promozionali situate all'interno dei punti vendita: i display hanno una maggiore attrattiva dal punto di vista visivo agli occhi dei consumatori rispetto al singolo prodotto sullo scaffale.

Le promozioni features, infine, vengono definite come annunci stampa e coupon distribuiti dal rivenditore: per molti di essi le feature consistono principalmente in inserti di giornale e circolari in-store.

[52] Michael Hahsler, Kurt Hornik, Bettina Grun, Christian Buchta, *Introduction to arules – A computational environment for mining association rules and frequent item sets.*

[53] David Hand, Heikki Mannila, Padhraic Smyth, *Principles of data mining,* Cambridge (Mass.) ; London : MIT press, c2001.

La ricerca di itemset frequenti e di regole di associazione è un approccio conosciuto e ben studiato, che serve per scoprire interessanti relazioni tra variabili in grandi database. Nello specifico, il pacchetto di R chiamato "arules" fornisce una struttura di base per la creazione e la manipolazione di input, forniti dai dataset, e per l'analisi degli itemset e delle regole da essi risultanti.

Una transazione nel database contiene un ID della transazione e un set di elementi; una regola in un insieme di regole di associazione contiene due set di elementi, uno come antecedente e uno come conseguente della regola stessa.

E' stato reso disponibile, grazie a Pam, un grande campione[54] del data set relativo alle quantità dei prodotti acquistati dai consumatori, appartenenti alle categorie in esame.

Il database delle transazioni può essere rappresentato come una matrice, in cui le colonne corrispondono ai prodotti e le righe agli itemset.

Infatti, posto che ogni transazione sia caratterizzata da un codice identificativo del basket di riferimento, le righe rappresentano le transazioni dello stesso basket, le colonne le categorie di prodotti considerate.

Si ricorda che esse sono "latte", "riso", "biscotti", e che comprendono rispettivamente 46, 82, 224 prodotti. L'analisi verrà svolta utilizzando le tre categorie, ma è necessario precisare e ricordare che sono sottointesi in ciascuna di esse i numerosi prodotti in esse contenuti.

Si consideri la matrice seguente[55], in cui i numeri indicano la quantità acquistata dei prodotti della categoria in oggetto, lo spazio vuoto significa invece che nessun prodotto di quella categoria è stato acquistato.

BASKET-ID	BISCOTTI	LATTE	RISO
144.141.300.000.014	1	2	1
144.141.300.000.126	1	1	
144.141.300.000.172	1	2	
144.141.300.000.176		1	1
144.141.300.000.361	2		
144.141.300.000.382		1	1
144.141.300.000.396	2		
144.141.300.000.408		2	
144.141.300.000.476	1		1
144.141.300.000.495	2		
144.141.300.000.606	1	1	
144.141.300.000.629	1	1	
144.141.300.000.637	2		

[54] Dalla settimana 49 del 2013 alla settimana 26 del 2014.
[55] Si tratta solo di una parte della matrice considerata, contenente 3560 BASKET-ID.

144.141.300.000.667	2		
144.141.300.000.673	1	1	
144.141.300.000.680		1	1
144.141.300.000.706	2		
144.141.300.000.719	2	1	
144.141.300.000.783	1	1	1
144.141.300.000.789	2		
144.141.300.000.816		2	
144.141.300.000.817	3	1	

Si trasformino ora le variabili numeriche, che indicano la quantità di prodotti acquistati, in variabili non numeriche, come suggerisce l'utilizzo di "Arules". Questo procedimento verrà svolto in Excel, grazie all'utilizzo della seguente formula: =CODICE.CARATT(Sheet5!*cella*+65).

Si ottiene così una nuova matrice:[56]

BASKET-ID	BISCOTTI	LATTE	RISO
144.141.300.000.014	B	C	B
144.141.300.000.126	B	B	A
144.141.300.000.172	B	C	A
144.141.300.000.176	A	B	B
144.141.300.000.361	C	A	A
144.141.300.000.382	A	B	B
144.141.300.000.396	C	A	A
144.141.300.000.408	A	C	A
144.141.300.000.476	B	A	B
144.141.300.000.495	C	A	A
144.141.300.000.606	B	B	A
144.141.300.000.629	B	B	A
144.141.300.000.637	C	A	A
144.141.300.000.667	C	A	A
144.141.300.000.673	B	B	A
144.141.300.000.680	A	B	B
144.141.300.000.706	C	A	A
144.141.300.000.719	C	B	A
144.141.300.000.783	B	B	B
144.141.300.000.789	C	A	A
144.141.300.000.816	A	C	A
144.141.300.000.817	D	B	A

[56] Si tratta solo di una parte della matrice considerata, contenente 3560 BASKET-ID.

Attraverso una cosiddetta “vignetta”, ossia un file pdf che aiuta nell’utilizzo dei vari pacchetti in “R”, [57] è stato possibile utilizzare “Arules”, ed è stato dato ad “R” il seguente input:[58]

```
dati<-read.csv2("matricc.csv",header=TRUE,sep=",")
Adult <- as(dati, "transactions")
itemFrequencyPlot(Adult, support = 0.1, cex.names=0.8)
rules <- apriori(Adult, parameter = list(support = 0.01, confidence = 0.6))
rules
summary(rules)
rulesIncomeSmall <- subset(rules, subset = rhs %in% "RISO=B" & lift > 1.2)
inspect(head(sort(rulesIncomeSmall, by = "confidence"), n = 13))
write(rules, file = "tutte_regole.csv", sep = ",", col.names = NA)
write(rulesIncomeSmall, file = "regole_tagliate.csv", sep = ",", col.names = NA)
```

Posto uno dei prodotti appartenenti alla categoria “riso” come elemento contenuto nel basket, sono state calcolate alcune regole che associano agli acquisti di uno dei prodotti appartenenti a questa categoria, uno o più prodotti appartenenti alle altre due categorie considerate, ossia “latte” e “biscotti”.

Si è posto inoltre un lift maggiore di 1.2, in modo tale da considerare solo quelle regole che più frequentemente si verificano.

In questo modo le regole selezionate sono le seguenti:

Regole:

```
> inspect(head(sort(rulesIncomeSmall, by = "confidence"), n = 13))

1 {RISO=1}=>{LATTE=1}
0.01285262 1.0000000 5.459679

2 {RISO=1}=>{BISCOTTI=2}
0.01746605 1.0000000 5.459679

3 {RISO=1}=>{BISCOTTI=1}
0.08448828 1.0000000 5.459679
```

Queste regole possono essere spiegate nel seguente modo:

- la regola numero 1 sta a indicare che, dato l’acquisto di 1 prodotto appartenente alla categoria “riso”, viene acquistato 1 prodotto appartenente alla categoria “latte”, con un supporto pari a 0.01285262, una confidenza pari a 1.0000000 e un lift pari a 5.459679;
- la regola numero 2 indica che, dato l’acquisto di 1 prodotto appartenente alla categoria “riso”, vengono acquistati 2 prodotti appartenenti alla categoria “biscotti”,

[57] Appendice 2.
[58] http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf

con un supporto pari a 0.01746605, una confidenza pari a 1.0000000 e un lift pari a 5.459679;

- la regola numero 3mostra che, dato l'acquisto di 1 prodotto appartenente alla categoria "riso", viene acquistato 1 prodotto appartenente alla categoria "biscotti"; con un supporto pari a 0.08448828, una confidenza pari a 1.0000000 e un lift pari a 5.459679.

In verde è possibile vedere il supporto della regola, ossia la quota di transazioni che contengono sia l'antecedente che il conseguente della regola stessa.
Si prendano ad esempio le tre regole considerate: posto il riso come antecedente di tutte e tre le regole, nella prima regola il latte è il conseguente della regola, nella seconda e nella terza lo sono i biscotti, rispettivamente con una quantità pari a 2 e a 1.
Il supporto misura quindi la frequenza con la quale questi prodotti appaiono assieme nelle transazioni del dataset. Un più alto supporto suggerisce che la regola si verifica con elevata frequenza ed è quindi una regola interessante.
In azzurro è evidenziata la confidenza della regola, cioè la quota di transazioni contenenti l'insieme dei conseguenti, tra quelle contenenti l'insieme degli antecedenti, ed esprime quindi l'affidabilità della regola.
Si prendano ad esempio le tre regole considerate: in tutte e tre le regole l'insieme degli antecedenti contiene il riso, mentre l'insieme dei conseguenti contiene, nella prima regola, il latte, nella seconda e nella terza regola, i biscotti, rispettivamente con una quantità pari a 2 e a 1.
Un'elevata confidenza corrisponde quindi ad una più alta probabilità che il sottoinsieme composto dal latte esista in una transazione che contiene anche il sottoinsieme contenente il riso.
Infine, in giallo è riportato il lift della regola, che serve a valutare la significatività di una regola.
Infatti le regole forti non sono sempre anche regole significative.
Si prendano ad esempio le tre regole considerate, tutte e tre con un lift superiore a 1: esse mostrano il grado in cui antecedente (riso) e conseguente (latte o biscotti) dipendono l'uno dall'altro, evidenziando la potenziale utilità di quella regola nella previsione di comportamenti futuri.

Si osservi un esempio riferito ai primi basket-ID della matrice utilizzata per l'analisi:

BASKET-ID	**BISCOTTI**	**LATTE**	**RISO**
144.141.300.000.014	1	1	1
144.141.300.000.126	1	1	0
144.141.300.000.172	1	1	0
144.141.300.000.176	0	1	1
144.141.300.000.361	1	0	0

Sia *I* l'insieme dei prodotti, chiamati *Items*. Sia *D* l'insieme delle transazioni, chiamato *database*. Ciascuna transazione in *D* ha un unico codice identificativo (*ID transaction*) e contiene un sottoinsieme di prodotti in *I*.

Nell'esempio, l'insieme dei prodotti considerato (*Itemset*) è biscotti e latte, e nel database 1 sta a indicare la presenza del prodotto nella transazione, 0 la sua assenza.

Per selezionare regole interessanti è necessario utilizzare supporto, confidenza e lift.

Il supporto, supp (X), di un itemset X è definito come il rapporto tra le transazioni del data set che contengono l'insieme di prodotti considerato (biscotti e latte) e il totale delle transazioni. Nell'esempio l'itemset scelto ha un supporto di 3/5=0.6, ciò significa che si verifica con una probabilità del 60%.

La confidenza di una regola è definita nel modo seguente:

$\mathrm{conf}\{X \Rightarrow Y\} = [\mathrm{supp}(X \cup Y)] / \mathrm{supp}(X)$.

Nell'esempio, la regola è $\{\text{biscotti, latte}\} \Rightarrow \{\text{riso}\}$ ha una confidenza di 0.2/0,6= 0,3circa nel database, che significa che nel 30% circa delle transazioni che contengono riso e biscotti, viene acquistato anche il latte.

Il lift di una regola è definite nel modo seguente:

$\mathrm{lift}\{X \Rightarrow Y\} = [\mathrm{supp}(X \cup Y)]/ [\mathrm{supp}(X) \bullet \mathrm{supp}(Y)]$.

La regola considerata ha un lift pari a 0.2/0.6•0.4 = 0.8 circa.

La regola, avendo un lift inferiore a 1, non è considerata utile nella comprensione dei comportamenti futuri dei consumatori.[59]

[59] Association rule learning - Wikipedia, the free encyclopedia.

Sottolineando nuovamente che la categoria "latte" sottintende 46 prodotti diversi, la categoria "riso" ne comprende 86 e quella dei "biscotti" 224, è interessante notare che in ciascuna delle regole più frequenti ed importanti i tre prodotti non vengono mai acquistati assieme, ma che l'acquisto del riso comporta con un elevata frequenza l'acquisto o del latte o dei biscotti.
Dato che l'associazione tra latte e biscotti risulta di immediata comprensione, dal momento che si tratta di prodotti che quotidianamente vengono con abbondanza utilizzati soprattutto per la colazione, si è posto come dato un prodotto appartenente alla categoria "riso", in modo tale che fosse visibile come questa si relazioni con le altre due categorie. Si è potuto riscontrare quindi, che essa spesso appare nella stessa esperienza di acquisto di latte e biscotti.
Ricordando che il latte fresco non è quasi mai soggetto a promozioni e il suo prezzo rimane sostanzialmente stabile nell'arco dell'anno, si evince che le sue vendite possano essere condizionate da quelle del riso. Esso infatti è un prodotto abbastanza promozionato nell'arco dell'anno.
Nonostante, come è già stato detto, la cosa preferibile è di utilizzare, per la market basket analysis, prodotti del più alto livello della tassonomia, un'analisi successiva, interessante da svolgere, potrebbe essere quella che comprenda un livello più ampio di prodotti, o addirittura tutti i prodotti appartenenti alle tre categorie, in modo tale da andare nello specifico delle relazioni che intercorrono tra i singoli prodotti.
Infatti, un giusto compromesso sembrerebbe quello di utilizzare inizialmente prodotti più generici, per poi ripetere il procedimento perfezionando la ricerca con l'utilizzo di prodotti più specifici.[60]
Questo lavoro richiederebbe un notevole impiego di tempo, infatti la complessità del procedimento è dovuta al numero di prodotti che vengono utilizzati: maggiore sarà questa quantità, più lungo sarà il procedimento di generazione di regole di associazione. Ma risulterebbe allo stesso tempo interessante, soprattutto per coloro che si occupano di prendere decisioni all'interno di un punto vendita, come potrebbe essere il suo layout e la disposizione dei prodotti sugli scaffali, come anche la scelta dei prodotti da sottoporre a promozione.
I consumatori potrebbero in questo modo essere attratti, e allo stesso tempo avvantaggiati, da una disposizione dei prodotti che deriva da uno studio approfondito delle scelte d'acquisto dei consumatori stessi.

[60] Michael J. A. Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

6. CONCLUSIONI

Il data mining è sempre stato un sinonimo dell'analisi statistica realizzata per sviluppare una comprensione dei dati notevolmente accurata, tale che riuscisse con facilità a trovare relazioni interessanti nell'ambito dei dati e che ne sfruttasse i risultati ottenuti.

Come già è stato detto, conoscere la propria clientela rappresenta un solido obiettivo aziendale: è necessario dunque che le aziende spostino la propria attenzione dai prodotti ai clienti.[61] Infatti, con l'evoluzione della tecnologia e con l'aumentare delle informazioni disponibili sulla clientela, le aziende hanno finalmente capito di poter monitorare e comprendere la totalità dei comportamenti della clientela, introducendo con continuità prodotti nuovi o innovativi.

La market basket analysis è un approccio metodologico che, pur nascendo nell'ambito del marketing, ha assunto recentemente una varietà di applicazioni in altri campi, quali le scienze nucleari, la geofisica, ecc. Uno dei motivi per i quali è aumentata l'adozione della tecnica di market basket analysis nei campi scientifici è data dal fatto che essa permette ai ricercatori di stimare la presenza di regole di associazione, utilizzando un approccio induttivo.[62]Inoltre, la market basket analysis permette ai ricercatori di utilizzare per le analisi dati che spesso vengono considerati inutili, ma che invece sono necessari per scoprire relazioni interessanti. Quindi, la market basket analysis ha un grande potenziale in termini di produzione di conoscenze teoriche nell'ambito del management, che sono anche in grado di dar luogo a pratiche significative e interventi organizzativi.

Infine, la market basket analysis può contribuire a ridurre il divario tra scienza e pratica, consentendo alle imprese di analizzare i dati che già possiedono e ai ricercatori di analizzare i dati che non avevano raccolto. Spesso è l'inaccessibilità di grandi insiemi di dati o costi proibitivi per la raccolta dei dati che impedisce avanzamenti nella ricerca scientifica, ma la market basket analysis è particolarmente adatta ad affrontare importanti questioni di ricerca utilizzando i dati tra sottocampi di management, quali l'ambito delle risorse umane, il comportamento organizzativo, l'imprenditorialità, il management strategico, e molti altri.

La nozione che può essere derivata dall'analisi svolta è che le scelte in una categoria influenzano le scelte in altre categorie. L'approccio presuppone che il ricercatore possa specificare la probabilità che un consumatore scelga una categoria nel proprio basket, dando in questo modo informazioni sulle scelte effettuate nelle altre categorie di analisi: in questo

[61] Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.
[62] *Journal of Management*, http://jom.sagepub.com/.

modo risulta possibile dedurre la distribuzione di mercato che spiega gli acquisti in tutte le categorie.

Dall'analisi svolta sulle tre categorie merceologiche del gruppo "white" (latte, riso, biscotti) risulta la presenza di 13 regole di associazione tra i prodotti contenuti delle categorie considerate.

Come detto inizialmente, l'associazione tra due o più prodotti può essere utilizzata per riorganizzare il layout del punto vendita e, attraverso le promozioni, per incrementare i profitti.

Una volta appurata la presenza di un'associazione tra prodotti, banale o non che sia, il primo passo da effettuare è quello di riorganizzare il punto vendita, con l'obiettivo di raggruppare i prodotti in zone omogenee, nelle quali il consumatore sia certo di non dimenticarsi di acquistare uno dei prodotti del gruppo.[63]

Una volta trovate le associazioni e individuate le promozioni più efficaci per ciascun gruppo di prodotti, si potrà realizzare un piano razionale: risulta possibile organizzare, all'interno del punto vendita considerato, un susseguirsi ciclico di promozioni, in modo tale che due o più prodotti fortemente associati non siano mai messi in promozione assieme: questo comporterà, nella maggior parte dei casi, l'acquisto da parte del consumatore del prodotto in promozione, e la presenza delle associazioni con altri prodotti spingerà il cliente a ulteriori acquisti all'interno del gruppo, in modo tale da cercare di incrementare le vendite non solo del prodotto messo in promozione, ma anche di quello, o quelli, ad esso associati.

[63] Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

APPENDICE

1. Articolo relativo all'associazione tra birra e pannolini.

FOCUS ON: Inventory Management

market-basket analysis

Market-Basket Mystery

What do beer and diapers have in common? For retailers, the answer could be powerful.

Before the dot.com boom of the 1990s, market-basket analysis was a new process that promised to turn the terabytes of data being collected at the POS into brilliant merchandising and promotional decisions.

Unfortunately, when retailers saw how much computing power was required to crunch the numbers, many lost interest in market-basket analysis, and the technology disappeared from the scene—until a few years ago.

Market-basket analysis is a term that describes data-mining solutions that find correlations between items in the customer's shopping basket. Merchandisers can apply these findings to respond to customer demand more effectively. It also helps them make more powerful planogramming decisions that consider the different kinds of items consumers are most likely to buy in the same shopping trip. The sheer number of SKUs in the store make it impossible to consider all possible correlations between items using human intellect alone. That's where market-basket analysis comes in.

The relationship between beer and diapers is an oft-cited example of a product correlation that has become data-mining urban legend. Reputedly, new fathers who suddenly have no time to go out and socialize will pick up a six-pack of beer when making a Huggies run. It's a good example of the potential findings of market-basket analysis, because it illustrates an exploitable relationship that's not obvious at first glance.

Although it's probably a stretch to expect retailers to stock diapers alongside the beer, if acted upon properly, effective market-basket analysis can bring increased sales, a stronger in-stock position and increased customer satisfaction.

According to Forrester Research analyst Lou Agosta, sometimes unprofitable, slow-selling items drive purchases in ways that can't easily be seen in the numbers. For example, if a market carries eight different types of olives but only half of them sell consistently, a manager might consider only eliminating the four poor performers. But an analysis of olive purchases might reveal that when the poor performers sell, it is usually with high-margin items. These are the kind of merchandising decisions that can help retailers become more customer-driven.

Although retailers can't be expected to stock diapers next to beer, market-basket analysis, when used properly, can increase sales and customer satisfaction.

Dedicated to the task: Market-basket analysis is especially popular among grocers. For example, Ahold USA, the domestic division of Dutch grocery giant Royal Ahold, recently replaced its legacy data warehouses with Netezza servers to process and store customer-loyalty data, and to support market-basket analysis. The data from customer-loyalty programs is fertile ground for market-basket analysis because it enables retailers to consider correlations not only between items bought on the same shopping trip, but also across multiple trips. It also permits retailers to consider purchase decisions in context with the customer's profile.

Grocery chains need lots of processing power for market-basket analysis, since they have one of the largest customer bases in retail and their SKU counts run into the tens of thousands of items. And each new item added to the sample only increases the amount of work geometrically.

According to Paula Rosenblum, director of retail research at Aberdeen Group, "It's a natural for grocery stores because of their use of loyalty programs and cards. It facilitates the cross-sell."

Market-basket analysis is being utilized by many other kinds of retailers, as well. In a recent Aberdeen Group survey, 38% of the companies polled said they used market-basket analysis and felt it had a positive effect on their business.

Chase-Pitkin Home & Garden, a division of Wegmans Food Markets, is a firm believer of market-basket analysis and has used it for the past four years to compete against home improvement giants The Home Depot and Lowe's.

"Our system started out serving one initiative but has been leveraged to be used across the entire enterprise," says CIO Chris Dorsey. "It's evolved into a complex business-intelligence tool that is being used by all levels of the organization."

One of Chase-Pitkin's most innovative applications of market-basket analysis is to aid in the creation of project cards that are distributed with the products it sells. Each card instructs the consumer on how to perform a task such as fixing a toilet or installing a rain gutter. The parts needed to complete the job are listed on the back. The cards typically are placed with the parts needed so all a customer has to do is grab the proper parts and tools, all from the same place.

The cards also help cross-sell related products, which Dorsey says has paid off. "Once you sell the core items, everything else is simple. We can certainly say it's helped our sales increase by double digits."

On the Web and beyond: Market-basket analysis is being applied to other retail channels, as well. It is proving to be very effective for e-retailers such as Amazon.com, which has become known for its powerful cross-selling. Ultimately, this could lead to multichannel transparency by creating a single source of customer data that all channels draw from. The advantage of this would be more insightful customer service, well-coordinated marketing efforts and even on-the-fly promotions that could adjust dynamically. A few companies already are exploring this last concept, though implementation is still a while away.

Market-basket analysis is especially popular among grocers, because the data from customer-loyalty programs can be used to consider correlations between items purchased on multiple shopping trips.

The bottom line, says Forrester's Agosta, is that market-basket analysis enables retailers to respond to their customers' needs more effectively. "I think we are becoming a nation of shopkeepers. Retailers and sellers of all kinds are looking for more interesting associations so they can make the right offer at the right place and the right time. It's a process in which technology really helps." RTQ

—Dennis Nishi
(pmcom@socal.rr.com)

Tabella 1. Prodotti considerati nella categoria “Latte” e relative vendite durante il periodo di riferimento.

LATTE	QUANTITA' VENDUTA
LATT.FR. TAPPOROSSO AQ (1500 ML TP)	8
LATTE CAPRA INTERO (1000 ML BT)	451
LATTE CRESCITA ESL GRA (500 ML BO)	326
LATTE ESL +G.INTERO GR (1500 ML BT)	21
LATTE ESL +G.PS.GRANAR (1500 ML BT)	824
LATTE ESL SCREMATO GRA (1000 ML BT)	2331
LATTE ESL.INT.BIO GRAN (1000 ML PK)	6009
LATTE FERM.INTERO ORO (1000 ML BT)	255
LATTE FR A.Q. GRANAROL (1000 ML BT)	38901
LATTE FR INT INGL.LTMI (1000 ML BT)	1687
LATTE FR TRSO BIO (1000 ML BT)	2521
LATTE FR. A.Q. LACTIS (500 ML BK)	1749
LATTE FR. AQ P&P L 1 (1000 ML PT)	37324
LATTE FR.A.Q CLMI (500 ML TP)	41290
LATTE FR.A.Q. LACTIS (1000 ML BT)	3840
LATTE FR.A.Q. P&P (1500 ML BO)	22598
LATTE FR.A.Q. PET ABIT (1000 ML BT)	4754
LATTE FR.A.Q.PET CLTO (1000 ML BT)	22325
LATTE FR.A.Q.PROB ABIT (1000 ML BT)	2572
LATTE FR.A.Q.TOP LTMI (1000 ML TP)	4652
LATTE FR.A.Q.VETR CLTO (750 ML BT)	6615
LATTE FR.INT CLTO (500 ML TB)	12253
LATTE FR.INT.A.Q.CLMI (1000 ML TP)	74934
LATTE FR.INT.TAPRS PET (1000 ML BT)	27600
LATTE FR.PS.BIO GRANAR (1000 ML BK)	12
LATTE FRE.INT.ROS.ABIT (1000 ML TP)	15172
LATTE FRE.LEGG.GRANAR. (500 ML TB)	5911
LATTE FRE.MAGRO LACTIS (500 ML TP)	701
LATTE FRE.SCREMAT.CLMI (500 ML TB)	15665
LATTE FRES.A.Q TOP BG (1000 ML BT)	1405
LATTE FRESCO TAPPOROSS (500 ML PK)	604
LATTE INT PREM BLUPARM (1000 ML BT)	12116
LATTE INTERO +GG GRAN. (1000 ML BK)	9733
LATTE MONT. INT.MILK (1000 ML BK)	4155
LATTE P S PREM BLUPARM (1000 ML BT)	14697
LATTE P.S. +GG GRANAR. (1000 ML BK)	9695
LATTE PURO BLU PARMALA (500 ML BT)	1677
LATTE PURO BLU PS (500 ML BT)	1927
LATTE TAP.ROSSO INTERO (1000 ML PT)	4915
LATTE TAP.ROSSO L.D.PS (1000 ML PT)	5338
LATTE ZYMIL +GG P.S. (500 ML BT)	1311
LATTE ZYMIL INT.MICROF (1000 ML TP)	1298
P&P LATTE MICRO.ALTDIG (1000 ML BO)	8634

P&P LATTE MICROF. P.S. (1000 ML BO)	14124
P&P LATTE MICROF.INTER (1000 ML BO)	14196
ZYMIL LATTE PS BLUPARM (1000 ML BT)	10628
Totale complessivo	**469754**

Tabella 2. Prodotti considerati nella categoria “Riso” e relative vendite durante il periodo di riferimento.

RISO	QUANTITA' VENDUTA
BLOND INSALATE GALLO (1000 GR SC)	1405
CURTIRISO AMBRA INSALA (1000 GR AS)	211
G.RISO CARNAR.IGP DELT (1000 GR PK)	905
GALLO BLOND RISOTTI (1 KG AS)	926
GALLO CHIC.PIU BETAGLU (400 GR SC)	17
GALLO RISO ROSS.CAMARG (500 GR SC)	26
GR.RISO ARBOR.IGP DELT (1000 GR PK)	409
GRANDI RISO ARBORIO (1000 GR SC)	1123
GRANDI RISO BASMATI (1000 GR CE)	1360
GRANDI RISO ORIGINARIO (1000 GR PK)	2
GRANDI RISO RIBE (1000 GR AS)	82
GRANDI RISO ROMA (1000 GR SC)	25
GRANO C`E` DI BUONO (500 GR AS)	33
MIX 5 CEREALI BUONO (500 GR AS)	1022
MIX RISO3CONT NATTURA (500 GR SK)	38
P&P BIO RISO ARBORIO (1000 GR AS)	947
P&P RISO ARBORIO SV (1 KG AS)	2139
P&P RISO BASMATI (1000 GR AS)	5354
P&P RISO BASMATI PRECO (250 GR AS)	2053
P&P RISO PARBOILED (2000 GR CE)	35
P&P RISO RIBE SV (1000 GR AS)	2617
P&P RISO.CARNAROLI.SV (1000 GR AS)	5066
P&P RISO.ORIGINAR.SV (1000 GR AS)	1698
P&P RISO.PARBOILED (1 KG AS)	2515
P&P.RISO.BASMATI. (500 GR AS)	67
P&P.RISO.INTEGRALE.SV (1000 GR AS)	1809
P&P.RISO.ROMA.SV (1 KG AS)	1849
P&P.RISO.THAI.PROFUMAT (500 GR AS)	1590
P&P.RISO.VIALO.NANO SV (1 KG AS)	1397
RIS.GALLO RISOT GROSS (1 KG SC)	2075
RISO 3 CEREALI EXPR_GA (250 GR AS)	1
RISO ARBORIO BUONRISO (1 KG CE)	1044
RISO ARBORIO RISOVALLE (1000 GR SK)	4714
RISO ARBORIOPRINCIPE (1000 GR AS)	327
RISO AVENA GRANO GALLO (800 GR SC)	915
RISO BASMATI EXPR_GALL (250 GR AS)	151
RISO BASMATI RAPID SCO (255 GR BS)	937
RISO BIO CARNAROLI (1000 GR AS)	1530

RISO BLOND VELOCE E VE (1000 GR AS)	652
RISO CARNAROLI DE TACC (1000 GR SK)	663
RISO CEREALI SEL.TERRA (400 GR SC)	15
RISO CURTI ARBORIO SV (1 KG CO)	2031
RISO CURTI CARNAROLI (1 KG CO)	2714
RISO CURTI S.ANDREA SV (1000 GR AS)	4
RISO FLORA 5 MIN.1000 (1000 GR AS)	1
RISO FLORA BELINSALAT (1000 GR AS)	158
RISO FLORA IL CLASSICO (1000 GR AS)	2131
RISO GALLO 3 CEREALI (400 GR AS)	1430
RISO GALLO AVENA+KAMUT (800 GR SC)	1
RISO GALLO BASMATI (800 GR SC)	2050
RISO GALLO CHICCHIRICH (850 GR AS)	7924
RISO GALLO RIBE SV (1 KG AS)	197
RISO GALLO VENERE (500 GR SC)	2287
RISO ORIGINAR.RISOVALL (1000 GR SK)	13004
RISO PARBOILED UNCLE B (500 GR CO)	866
RISO PRINCIPE CARNARO. (1 KG AS)	335
RISO RIBE BIO SCOTTI (500 GR AS)	57
RISO RIBE GALLO (500 GR AS)	2312
RISO ROMA RISOVALLE (1000 GR SK)	8873
RISO SCOT CLASSICO ORO (1000 GR AS)	3342
RISO SCOTTI ARBORIO (1 KG AS)	1414
RISO SCOTTI CARNAROLI (1000 GR AS)	2240
RISO SCOTTI CHICCHI GR (1 KG AS)	3612
RISO SCOTTI INTEGRALE (1000 GR SC)	1477
RISO SCOTTI ORO INSAL (1000 GR AS)	627
RISO SCOTTI PADANO (1000 GR AS)	3
RISO SCOTTI ROMA. (1 KG AS)	1171
RISO SU RISO FR ROSSI (176 GR CO)	130
RISO SU RISO G CIOCCOL (176 GR CO)	91
RISO SUP.ROMA BUONRISO (1 KG CE)	5134
RISO SUP.ROMA BUONRISO (500 GR CE)	6814
RISO SUSHI SELEZ.TERRA (500 GR SC)	15
RISO VIAL NANO DE TACC (1000 GR SK)	24
RISO.CURTI.ROMA.SV (1000 GR AS)	2009
RISOE` PER RISOTTI (500 GR AS)	542
RISOE` PIATTI UNICI X2 (400 GR AS)	26
RISO-GALLO-ARBORIO-SV. (1 KG AS)	908
RISO-PARBOIL-AMBRA-SV (1000 GR AS)	240
SANMARCO RISO 5KG THAI (5000 GR CE)	92
SANMARCO RISO KG2 THAI (2000 GR CE)	4907
SANMARCO RISO THAI 5KG (5000 GR CE)	12
THAI_PARBOIL.RISOVALLE (1000 GR SK)	7749
Totale complessivo	**138698**

Tabella 3. Prodotti considerati nella categoria “Biscotti” e relative vendite durante il periodo di riferimento.

BISCOTTI	QUANTITA' VENDUTA
ABBRACCI M.B. (700 GR PK)	117
ALCE FROL FARRO CIOCCO (300 GR SK)	123
ALCE N FROLLINI KAMUT (300 GR SK)	1956
AMARETTI ARTIGIANALI	93
AMARETTI BAROVERO	198
BACI DI DAMA	913
BACI DI DAMA CACAO	142
BAROV.BACI PRELIBATI	414
BIS CHICCHI DI CIOCCOL (300 GR SK)	3161
BIS KAMUT SENZA ZUCCHE (300 GR SK)	85
BIS NOVELLINI 250 GENT (250 GR SC)	8848
BIS SENZA LATTE E UOVA (400 GR SK)	1984
BIS. GRANCEREALE M.B. (500 GR CO)	8643
BIS.PAIN CROUTE INTEGR (450 GR PK)	1272
BISC COLUS G TURCHESE (400 GR PK)	4941
BISC INTEGR ARTE BIANC (400 GR AS)	3228
BISC OSVEGO 250 GENTIL (250 GR SC)	10290
BISC. CAMPAGNOLE M.B. (1000 GR SK)	665
BISC. MACINE MB (1000 GR SK)	741
BISC. TARALLUCCI M.B. (1000 GR SK)	667
BISC.ABBRACCI M.BIANCO (350 GR SK)	10814
BISC.CAMPAGNOLE M.B. (700 GR SK)	5700
BISC.CORLEGGERI GRONDO (250 GR PK)	535
BISC.CUOR DI MELA MB (300 GR AS)	10206
BISC.ENERGELLI SZ (300 GR SK)	2
BISC.FIOR DI LATTE MB (300 GR SK)	3682
BISC.GALLETTI M.B. (1000 GR SK)	155
BISC.GALLETTI MB. (800 GR SK)	11996
BISC.GOCCIOLE PAVESI (500 GR PK)	39741
BISC.I FATTI CASA EQUI (700 GR SK)	2322
BISC.I FATTINCASA INTE (700 GR SK)	3604
BISC.MAGRETTI CON GCIO (260 GR AS)	511
BISC.RIGOLI M.BIANCO (800 GR PK)	4770
BISC.SEGRETI BOSCO MB (300 GR SK)	87
BISC.TARALLUCCI M.B. (800 GR SK)	16569
BISCORO INTEGRALI (1000 GR SK)	1967
BISCOT.CAMPIELLO LIGHT (350 GR PK)	4900
BISCOTTI AI FICHI	100
BISCOTTI AL CIOCCOLATO	303
BISCOTTI AL COCCO	43
BISCOTTI AL FICO	12
BISCOTTI AL LIMONCELLO	42
BISCOTTI AL LIMONE	165

BISCOTTI ALL ANANAS	5
BISCOTTI ALL ARANCIA	65
BISCOTTI ALL UVETTA	335
BISCOTTI ALLA MANDORLA	320
BISCOTTI ALLA MELA	39
BISCOTTI BISCORO (1000 GR SK)	2264
BISCOTTI CANTUCCIONI	10
BISCOTTI CUORFELICI (350 GR SK)	1382
BISCOTTI DEL MATTINO	170
BISCOTTI FARFALLEGRE (350 GR SK)	2058
BISCOTTI FIORGOLOSI (350 GR SK)	51
BISCOTTI GALLETTE (700 GR SK)	850
BISCOTTI GIROTOTONDI (800 GR SK)	5994
BISCOTTI LAGACCIO (400 GR SK)	672
BISCOTTI NERO	33
BISCOTTI PRIMA COLAZ.	71
BISCOTTO ATENE DORIA (500 GR PK)	3672
BISCOTTO DI RISO	140
BISCOTTO GRANELLA	532
BISCOTTO KAOKAO (650 GR SK)	1132
BISCOTTO NOVARA	460
BISCOTTO NOVARINO	395
BISCOTTONE ZEBRATO (330 GR SK)	674
BISCOTTONI GLASSATI (330 GR SK)	446
BISTOCCU SARDU	252
BOCCONOTTI ALL UVA	4
BUCANEVE TUBO DORIA (200 GR PK)	15907
BUONGIORNO AL CACAO (700 GR SK)	975
BUONGIORNO ALL`UOVO (700 GR SK)	392
BUONICOSI AI CEREALI (300 GR PK)	296
BUONICOSI AL LATTE (300 GR AS)	4112
BUONICOSI FROLLINO (330 GR PK)	8321
BUONICOSI SEN LIEVITO (330 GR CO)	221
CABARET MELIGA BAROVER	583
CABARET SFOGL.MANINE	809
CAMPIELLO BISC.CEREALI (410 GR PK)	4690
CAMPIELLO BISC.SPIGARE (380 GR PK)	6278
CAMPIELLO S.ZUCCH.ACC. (350 GR PK)	5198
CANESTREJ BIELEIS	904
CANESTRELLI	1125
CANTUCCI	433
CARABEL BURRO (250 GR SK)	1038
CEREABEL CLASSICO (220 GR SK)	62
CEREABEL FRUTTA (220 GR SK)	57
CEREALBEL LEGGERISO (350 GR SK)	387
CIAMB.AGRUMI MANDORLE (200 GR SK)	1546
CIAMB.CIOCCO (200 GR SK)	6272

CIAMBELLE ZEBRATE	1406
COLUSSI BISC. OSVEGO (250 GR AS)	654
CRUSCHETTO INTEGRALE (700 GR SK)	1314
DIGESTIVE MC VITIE S (250 GR AS)	8753
DORIA BUCANEVE SACCHET (400 GR SK)	1017
FERRI DI CAVALLO	2106
FINOCCHINI	602
FIOCCHI DI NEVE TREVIS (400 GR CO)	391
FIOR DI RISO ARTE BIAN (400 GR AS)	3332
FROL.MENO GRASSI BALOC (320 GR CO)	1
FROL.PRIVOLAT MISURA (400 GR CO)	3250
FROL.RISOSURISO CACAO (220 GR PK)	74
FROL.RISOSURISO GALBUS (240 GR PK)	2370
FROLL MISURA SENZA ZUC (400 GR PK)	2911
FROLL PIU INTEGR GALB (330 GR AS)	6669
FROLL.COLCUORE GALBUSE (300 GR PK)	4008
FROLLINI CEREALI VSNEL (250 GR PK)	2027
FROLLINI FANTASIA	225
FROLLINO FR.ROSSI VSNE (270 GR PK)	2033
FRUMENTINI AVEN.LA (375 GR CO)	55
FRUMENTINI AVEN.LAZZAR (250 GR CO)	159
GALLETTE GRONDONA (240 GR PK)	351
GC SNACK NOCCIOLA (180 GR CO)	2516
GC SNACK YOGURT (180 GR CO)	1759
GOCCIOLE EX.DARK PAVES (400 GR PK)	7949
GOCCIOLOTTI BALOCCO (700 GR SK)	1
GRANCEREALE CROCCANTE (230 GR PK)	11142
GRANCEREALE DIGESTIV (250 GR PK)	127
GRANCEREALE FIBRECACAO (230 GR CO)	3178
GRANCEREALE FRUT FIBRA (250 GR PK)	14990
GRANELLINI BAROVERO	24
I RUSTICI	246
JUNIOR BISC.DOLCECOLAZ (1000 GR SK)	7306
KAMUT BIO C`E` DI BUON (450 GR AS)	17
KRUMI AL CICCOLATO	1916
KRUMIRI	333
KRUMIRI AL CIOCCOLATO	382
KRUMIRI BAROVERO	13
KRUMIRI CLASS.BISTEFAN (300 GR SK)	8331
LAGACCIO GRONDONA (250 GR SK)	3765
LANGHESINI BISC.TREVIS (400 GR CO)	468
M.B. RITORNELLI (700 GR PK)	4167
M.B.BISC.PAN DI STELLE (350 GR AS)	16413
M.B.BISCOT. BATTICUORI (350 GR PK)	3717
MACINE M.B. (800 GR SK)	14662
MAGRETTI MALTO ED ORZO (350 GR AS)	22
MANDOLINI ALBIC E MELA	109

MARGHERITE MARMELLATA	168
MAXI BISC.CAMPAGN.OMAG (300 GR SK)	114
MAXI GALLETTI M.B.OMAG (350 GR CO)	231
MAXI GOCC.DARK OMAGGIO (300 GR PK)	131
MB BISCOTTONE (700 GR SK)	123
MERINGHE RIPENE	359
MOLINETTI MULINO BIANC (800 GR SK)	5798
MONTE BIANCO	133
MUFFINS VAN & CIOCC X6	548
MULINO BIANCO GALLETTI (400 GR SK)	6234
MV MAGIE DI NOCCIOLE (300 GR SK)	2102
NOCCIOLINI PASTICC.BAR	18
NOVARA BAROVERO	165
NOVARINI BAROVERO	835
NOVARONI TREVISAN (400 GR CO)	281
NOVELLONE	124
NOVESINI TREVISAN (290 GR CO)	855
NUTRI GRAIN AL CIOCCOL (264 GR SC)	797
NUTRI GRAIN CER CROCCA (240 GR SC)	7
NUTRI GRAIN CROC CIOCC (240 GR SC)	655
NUTRI GRAIN FRUT E FIB (264 GR SC)	904
NUTRI GRAIN LATTE E CE (264 GR SC)	68
NUVOLE CACAO E NOCCIOL (350 GR SK)	805
NUVOLE CON G CIOCCOLAT (350 GR SK)	1090
NUVOLE GRANO SARACENO (350 GR SK)	32
NUVOLE INT CON CRUSCA (350 GR SK)	49
OCCHI DI BUE ALBICOCCA	1478
ORO 5 CEREALI SAIWA (400 GR PK)	197
ORO CIOK LATTE SAIWA (250 GR PK)	3619
ORO SAIWA FIBRATTIVA (400 GR PK)	172
ORO SAIWA GOCC.CIOC.X6 (300 GR PK)	70
ORO SAIWA MAXIOFFERTA (1000 GR AS)	7631
ORZO C`E` DI BUONO (500 GR AS)	537
OSVEGO BISC.5CER.GENTI (250 GR PK)	6349
OVALINE BURRO BRVR	353
OVALINE GLAS NOCC TOST	1080
P&P BISCOTTI PETIT (500 GR PK)	3572
P&P BISCOTTO PETIT (500 GR CO)	2654
P&P FROL.CEREALI S.ZUC (350 GR SK)	4486
P&P FROLL.CACAO NOCCIO (350 GR SK)	7402
P&P FROLL.GRANO SARAC (700 GR SK)	5136
P&P FROLL.LATTE MIELE (350 GR PK)	12609
P&P FROLL.PANNA CACAO (350 GR SK)	6700
P&P FROLLIN RISO LATT (700 GR CO)	7445
P&P FROLLINI AI CEREAL (500 GR SK)	4224
P&P FROLLINI GOCCE (350 GR SK)	8172
P&P FROLLINI GOCCE CIO (700 GR SK)	12002

P&P FROLLINI INTEGRALI (350 GR SK)	4181
P&P FROLLINI RIP.MELA (350 GR SK)	14023
P&P FROLLINO C PANNA (700 GR SK)	9496
P&P FROLLINO C-ZUC.GR (700 GR PK)	4685
P&P FROLLINO UOVA (700 GR SK)	7576
PAIN CROUTE INTEG.SAPO (300 GR PK)	3640
PAN DI STELLE M B (700 GR PK)	127
PAPILLONS DI SFOGLIA	185
PASTE D` MELIA	692
PASTE DI MELIGA	213
PASTE DI MELIGA (300 GR CO)	670
PASTICCERIA SECCA	38
PRIVOLAT CON G CIOCCOL (300 GR CO)	1115
SACHER	149
SAIWA ORO BISCOTTI (1250 GR AS)	883
SAIWA ORO BISCOTTI (750 GR AS)	9510
SAVOIARDI BAROVERO	1439
SPICCHI SOLE MB800+100 (900 GR PK)	10
SPUMIGLIE NOCCIOLA	314
TEGOLE VALDOST BAROVER	30
TORCETTI	541
TORCETTI AL BURRO	70
TORCETTI BURRO TREVISA (180 GR VS)	658
TORCETTI CIOCCOLATO	670
TRIANGOLINI SOIA ARTE (400 GR AS)	3043
VENTAGLI DI SFOGLIA	236
VIT CEREAL CACAO (300 GR CO)	2029
VIT CEREAL LATTE E CER (300 GR CO)	1835
VIT CEREAL NOCCIOLE (300 GR CO)	112
VITAS CEREALYO FR ROSS (253 GR SK)	231
VITASN.CEREAL YO CACAO (253 GR CO)	1666
VITASN.CEREAL YO MIELE (253 GR AS)	51
VITASN.SNACK ARANC. X6 (162 GR AS)	2755
VITASNELLA CEREAL YO (253 GR PK)	3376
VIVIS INTEG NO ZUCCHER (500 GR SK)	2069
VIVISAN NO LATTE E UOV (500 GR SK)	1050
WAFER CACAO	5657
WAFER NOCCIOLA	5573
WAFER VANIGLIA	5355
ZOO DORIA (350 GR PK)	1579
Totale complessivo	**611646**

Tabella 4. Andamento delle vendite congiuntamente di tutti i prodotti, dalla settimana numero 49 del 2013 alla settimana numero 26 del 2014.

SETTIMANE	QUANTITA' VENDUTA
201349	49034
201350	45436
201351	40938
201352	42058
201401	20162
201402	39337
201403	41474
201404	46675
201405	45106
201406	43473
201407	42699
201408	43335
201409	43952
201410	42764
201411	41459
201412	41193
201413	47533
201414	41798
201415	40401
201416	37505
201417	36382
201418	43852
201419	42170
201420	49212
201421	45336
201422	43162
201423	40546
201424	39534
201425	36054
201426	11392
Totale complessivo	**1223972**

Tabella 5. Quantità venduta nella settimana 201401.

DIB_TIME_CODE	201401

CATEGORIE DI PRODOTTO	**QUANTITA' VENDUTA**
RISO	2452
BISCOTTERIA A PESO E A NUMERO	625
LATTE FRESCO INTERO	6240
LATTE FRESCO SCREMATO	448
LATTE FRSCO MICROFILTRATO/ARRICCHITO	2292
BISCOTTI CONFEZIONATI	8105
Totale complessivo	**20162**

Tabella 6. Quantità venduta nella settimana 201420.

DIB_TIME_CODE	201420

Etichette di riga	**QUANTITA' VENDUTA**
RISO	6476
BISCOTTERIA A PESO E A NUMERO	1684
LATTE FRESCO INTERO	11155
LATTE FRESCO SCREMATO	913
LATTE FRESCO MICROFILTRATO/ARRICCHITO	3793
BISCOTTI CONFEZIONATI	25191
Totale complessivo	**49212**

Tabella 7. Quantità di “biscotti a peso e a numero” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA BISCOTTI A PESO E A NUMERO
201349	1658
201350	1635
201351	1534
201352	1191
201401	625
201402	1151
201403	1326
201404	1319
201405	1687
201406	1856
201407	1620
201408	1584
201409	1547
201410	1741
201411	1809
201412	1749
201413	1878
201414	1782
201415	1538
201416	1439
201417	1049
201418	1413
201419	1352
201420	1684
201421	1481
201422	1397
201423	1391
201424	988
201425	1168
201426	303
Totale complessivo	**42895**

Tabella 8. Quantità di “biscotti confezionati” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA BISCOTTI CONFEZIONATI
201349	23257
201350	20486
201351	17020
201352	15809
201401	8105
201402	17004
201403	18435
201404	24377
201405	21067
201406	19707
201407	19619
201408	20319
201409	20457
201410	20319
201411	18945
201412	18207
201413	24184
201414	18967
201415	18427
201416	16248
201417	17739
201418	20890
201419	19430
201420	25191
201421	22737
201422	21407
201423	19724
201424	17579
201425	17017
201426	5857
Totale complessivo	**568530**

Tabella 9. Quantità di “riso” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA RISO
201349	5133
201350	4828
201351	5403
201352	5336
201401	2452
201402	4527
201403	4758
201404	4479
201405	5052
201406	5206
201407	4724
201408	4751
201409	4909
201410	4732
201411	4449
201412	4702
201413	5531
201414	4609
201415	4339
201416	4433
201417	4145
201418	5398
201419	5569
201420	6476
201421	5234
201422	4834
201423	4887
201424	6049
201425	4436
201426	1412
Totale complessivo	**142793**

Tabella 10. Quantità di “latte fresco intero” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA LATTE FRESCO INTERO
201349	13970
201350	13669
201351	12159
201352	13805
201401	6240
201402	11533
201403	12078
201404	11558
201405	12077
201406	11859
201407	11876
201408	11718
201409	12104
201410	11235
201411	11446
201412	11581
201413	11370
201414	11664
201415	11305
201416	11183
201417	9221
201418	11241
201419	11327
201420	11155
201421	11138
201422	11078
201423	10081
201424	10516
201425	9503
201426	2686
Totale complessivo	**332376**

Tabella 11. Quantità di “latte fresco scremato” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA LATTE FRESCO SCREMATO
201349	776
201350	769
201351	800
201352	907
201401	448
201402	855
201403	918
201404	864
201405	906
201406	941
201407	851
201408	864
201409	931
201410	885
201411	854
201412	953
201413	971
201414	896
201415	967
201416	823
201417	693
201418	931
201419	905
201420	913
201421	884
201422	917
201423	841
201424	860
201425	884
201426	205
Totale complessivo	**25212**

Tabella 12. Quantità di “latte fresco microfiltrato/arricchito” venduta nel periodo considerato.

SETTIMANE	QUANTITA' VENDUTA LATTE FRESCO MICROFILTRATO/ARRICCHITO
201349	4240
201350	4049
201351	4022
201352	5010
201401	2292
201402	4267
201403	3959
201404	4078
201405	4317
201406	3904
201407	4009
201408	4099
201409	4004
201410	3852
201411	3956
201412	4001
201413	3599
201414	3880
201415	3825
201416	3379
201417	3535
201418	3979
201419	3587
201420	3793
201421	3862
201422	3529
201423	3622
201424	3542
201425	3046
201426	929
Totale complessivo	**112166**

Tabella 13. Promozioni attive nel periodo considerato.

SETTIMANE	NUMERO PROMOZIONI ATTIVE
201349	9437
201350	6265
201351	4765
201352	4571
201401	1736
201402	5276
201403	6630
201404	11312
201405	7160
201406	6609
201407	4081
201408	6043
201409	6883
201410	8336
201411	5646
201412	4529
201413	11002
201414	4574
201415	4439
201416	4070
201417	6261
201418	6109
201419	5883
201420	13507
201421	9954
201422	8610
201423	6782
201424	7155
201425	5816
201426	2507
Totale complessivo	**195948**

Tabella 14. “Biscotti a peso e a numero” soggetti a promozione.

SOGGETTI A PROMOZIONE	BISCOTTI A PESO E A NUMERO
NO	35785
SI'	1879
Totale complessivo	**37664**

Tabella 15. “Biscotti confezionati” soggetti a promozione.

SOGGETTI A PROMOZIONE	BISCOTTI CONFEZIONATI
NO	340296
SI'	167538
Totale complessivo	**507834**

Tabella 16. “Riso” soggetto a promozione.

SOGGETTI A PROMOZIONE	RISO
NO	98191
SI'	24913
Totale complessivo	**123104**

Tabella 17. “Latte fresco intero” soggetto a promozione.

SOGGETTI A PROMOZIONE	LATTE FRESCO INTERO
NO	272225
SI'	83
Totale complessivo	**272308**

Tabella 18. “Latte fresco scremato“ soggetto a promozione.

SOGGETTI A PROMOZIONE	LATTE FRESCO SCREMATO
NO	18254
Totale complessivo	**18254**

Tabella 19. “Latte fresco microfiltrato/arricchito” soggetto a promozione.

SOGGETTI A PROMOZIONE	LATTE FRESCO MICROFILTRATO/ARRICCHITO
NO	87876
SI'	1535
Totale complessivo	**89411**

Appendice 2. Vignetta per l'utilizzo di Arules.

Package 'arules'

January 7, 2015

Version 1.1-6

Date 2014-12-07

Title Mining Association Rules and Frequent Itemsets

Description Provides the infrastructure for representing, manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Also provides interfaces to C implementations of the association mining algorithms Apriori and Eclat by C. Borgelt.

Classification/ACM G.4, H.2.8, I.5.1

URL http://R-Forge.R-project.org/projects/arules/, http://lyle.smu.edu/IDA/arules/

Depends R (>= 2.14.2), Matrix (>= 1.0-0)

Imports stats, methods

Suggests pmml, XML, arulesViz, testthat

License GPL-3

Copyright The code for apriori and eclat in src/rapriori.c was obtained from http://www.borgelt.net/ and is Copyright (C) 1996-2003 Christian Borgelt. All other code is Copyright (C) Michael Hahsler, Christian Buchta, Bettina Gruen and Kurt Hornik.

Author Michael Hahsler [aut, cre, cph], Christian Buchta [aut, cph], Bettina Gruen [aut, cph], Kurt Hornik [aut, cph], Christian Borgelt [ctb, cph]

Maintainer Michael Hahsler <mhahsler@lyle.smu.edu>

NeedsCompilation yes

Repository CRAN

Date/Publication 2014-12-08 07:29:41

BIBLIOGRAFIA E SITOGRAFIA

Pieter Adriaans, Dolf Zantinge, *Data Mining*, Addison-Wesley, 1996.

Michael Berry, Gordon Linoff, *Data mining techniques : for marketing, sales, and customer support*, New York, J. Wiley, 1997.

Michael Berry, Gordon Linoff, *Mastering Data Mining*, John Wiley & Sons, 2000.

Nicola Del Ciello, Susi Dulli, Alberto Saccardi, *Metodi di data mining per il customer relationship management*, Franco Angeli, 2000.

Jill Dyché, *E-Data: come trasformare i dati in informazione con tecniche di data warehousing e data base marketing*, Milano, Apogeo, 2000.

Paolo Giudici, *Data mining : metodi statistici per le applicazioni aziendali*, Milano, McGraW-Hill, 2001.

Seth Godin, *Permission Marketing: turning Strangers Into Friends, and Friends Into Customers*, Simon & Schuster, 1999.

Ian H. Gordon, *Relationship Marketing: New Strategies, Techniques and Technologies to Win the Customers You Want and Keep Them Forever*, John Wiley & Sons, 1998.

Robert Groth, *Data Mining: A Hands-on Approach for Business Professionals*, Prentice Hall, 1997.

Jiawei Han, Micheline Kamber, *Data Mining: Concepts ans Techniques*, Morgan Kaufmann, 2001.

David Hand, Heikki Mannila, Padhraic Smyth, *Principles of data mining*, London , MIT press, 2001.

Peppers & Rogers, *Enterprise One to One: Tools for Competing in the Interactive Age*, Doubleday, 1997

Carlo Vercellis, *Business intelligence : data mining and optimization for decision making*, Hoboken, NJ, Wiley, 2009.

Riviste:

Knowledge discovery and data mining

Journal of Management: Herman Aguinis, Lura E. Forcum and Harry Joo, *Using Market Basket Analysis in Management Research*

European Journal of Operational Research: Yasemin Boztug, Thomas Reutterer, *Interfaces with Other Disciplines, A combined approach for segment-specific market basket analysis*

Advances in Consumer Research: Julien Schmitt, Loughborough University, UK, *Drawing Association Rules between Purchases and In Store Behavior: An Extension of the Market Basket Analysis.*

Springer Science+Business Media, LLC 2012: Wagner A. Kamakura, *Sequential market basket analysis.*

Journal of Retailing: Gary J. Russell, Ann Petersen, *Analysis of Cross Category Dependence in Market Basket Selection.*

Human Systems Management: Kuriakose Athappilly, Muhammad A. Razi and J. Michael Tarn, *A multi-technique data mining approach to exploring consumer behaviors.*

International Journal of Organizational Innovation: Lee-Wen Huang, Ye-In Chang, *A graph-based approach for mining closed large itemsets.*

De Gruyter: Katrin Dippold and Harald Hruschka, *A Model of Heterogeneous Multicategory Choice for Market Basket Analysis.*

Articoli:

Convenience Store Decisions: *Analyzing scan data*

Chain Store Age: *Market-Basket Mistery*

Siti Internet:

www.sciencedirect.com

http://jom.sagepub.com/

Association rule learning - Wikipedia, the free encyclopedia