UNIVERSITÀ DI PISA
DIPARTIMENTO INFORMATICA

CORSO DI LAUREA MAGISTRALE IN BUSINESS INFORMATICS

Analisi dei Pattern Individuali di acquisto nella Grande Distribuzione

Laureando
Gian Felice Meloni

Relatori
Prof. Dino Pedreschi
Dr. Diego Pennacchioli
Dott. Riccardo Guidotti

Anno Accademico 2013/2014

Ai miei genitori.

Sommario

Conoscere i propri consumatori è una delle attività più importanti di cui si devono occupare le aziende di produzione e distribuzione nei mercati attuali. Uno dei metodi principali è la segmentazione, che consiste nel dividere i propri clienti in gruppi omogenei per determinate funzioni obiettivo. Punto cruciale di tale attività è la scelta delle misure da utilizzare per approssimare il comportamento desiderato.

Il lavoro della Tesi si è concentrato su questo task, introducendo e implementando due misure di sistematicità sul comportamento di acquisto dei consumatori: il **BRI** (*Basket Regularity Index*) e lo **STRI** (*spatio-temporal Regularity Index*).

Tali misure sono state applicate ad un insieme dati di acquisto reali di utenti in possesso della *carta socio* della nota catena di vendita al dettaglio *Unicoop Tirreno*.

Le analisi effettuate hanno fornito un valido sostegno alla bontà delle misure modellate: sono state infatti individuate caratteristiche interessanti in comune per ogni sottogruppo di consumatori trovati.

Indice

Elenco delle figure

Elenco delle tabelle

Elenco degli algoritmi

Capitolo 1

Introduzione

Negli ultimi decenni, la proliferazione di prodotti da un lato e quella dei format distributivi dall'altro si sono incontrate a formare innumerevoli combinazioni che si offrono ai consumatori. Essi sono diventati più esperti, selettivi e volubili, e a ciò rispondono produttori e distributori con offerte sempre più differenziate.

Di pari passo con l'evoluzione dei comportamenti dei clienti, si sono evoluti anche i criteri di segmentazione [7]. I criteri *socio-demografici* sono andati perdendo la loro capacità di prevedere comportamenti e preferenze e sono apparsi altri approcci come la *segmentazione sui vantaggi ricercati* o quella per *stili di vita* che mal si adattano alle finalità delle aziende di *vendita al dettaglio* per gli elevati costi di misurazione.

Con l'introduzione delle *carte fedeltà* l'azienda commerciale viene in possesso non solo dei dati anagrafici dei clienti, ma anche di dati di sintesi sul comportamento di acquisto (cosa i clienti comprano, quanto spendono, etc.) e dei dati di dettaglio sulle singole transazioni. Questi dati divengono il punto di partenza di nuovi sforzi di segmentazione.

Attualmente è possibile osservare la presenza di due grandi approcci alla segmentazione dei clienti adottati dalle aziende della distribuzione al dettaglio [5]:

- uno orientato al *prodotto*: segmentazione dei clienti basata sulla creazione di profili a partire dalla composizione dettagliata dei panieri di spesa monitorati nel tempo (utilizzato ad esempio dalla catena britannica **TESCO**);
- uno orientato al *cliente*: segmentazione dei clienti basata sui comportamenti di acquisto in termini di valore, frequenza, recency, etc. (utilizzato dai principali retailers italiani).

Gli approcci elencati precedentemente hanno un limite importante: essi approssimano grandezze ma non comportamenti.

L'obiettivo della presente Tesi è, invece, quello di sintetizzare il comportamento di spesa degli individui distinguendo due componenti:

- una componente *concreta*: dove, quando, quanto e cosa comprano;
- una componente *astratta*: come comprano;

A tale proposito sono state modellate due specifiche misure di *sistematicità*: il **BRI** (*basket regularity index*) e lo **STRI** (*spatio-temporal regularity index*). La prima misura riassume il comportamento di acquisto di un cliente in termini di articoli acquistati; la seconda riassume il comportamento spazio temporale di un cliente in termini di *dove* e *quando* compie i propri acquisti.

Entrambi gli indici, considerati singolarmente (per un unico individuo), sono privi di significato. Analizzati però nella *collettività*, mostrano, invece, gruppi di consumatori per intervalli delle suddette misure.

Dopo aver introdotto e modellato le misure, sono stati realizzati due casi di studio utilizzando dati di acquisto reali forniti da *Unicoop Tirreno*.

Col primo esperimento viene fornita una strategia di segmentazione sul comportamento di acquisto basata sui due indici; col secondo vengono individuati sottogruppi di utenti che hanno debole, media o forte sistematicità.

Le analisi hanno fornito un valido sostegno alla bontà delle misure modellate. Infatti, per entrambi gli approcci, sono state individuate caratteristiche interessanti in comune per ogni sottogruppo di consumatori trovati.

1.1 Contenuto della Tesi

Viene di seguito riportata l'organizzazione della Tesi indicando le tematiche principali di ciascun capitolo.

Nel capitolo 2 verrà fornita una breve introduzione sulla disciplina del marketing soffermandosi principalmente sul *marketing relazionale* e sul *customer relationship management*.

Nel capitolo 3 verranno approfondite le tematiche del *customer profiling* e della *customer segmentation*. In particolare, saranno elencate le principali tecniche del *Data Mining* analizzando nel dettaglio gli algoritmi utilizzati nel *frequent pattern mining* e nel *clustering*. Si parlerà infine della misura dell'*entropia* nella teoria dell'informazione e dei suoi nuovi utilizzi nell'ambito dello studio del comportamento umano.

Nel capitolo 4 si presenta una descrizione del *datawarehouse* dell'*Unicoop Tirreno* sul quale sono state svolte le analisi. Seguirà una panoramica sul *datamart*, sulla tabella dei fatti e sulle tabelle dimensionali utilizzate.

Nel capitolo 5 verrà esposto il metodo realizzato. Dopo una breve introduzione sul concetto di *sistematicità* di un individuo nel comportamento

di acquisto, vengono presentate le due misure sintetizzate: il **BRI** (basket regularity index) e lo **STRI** (spatio-temporal regularity index). Per ciascuna misura vengono forniti gli pseudocodici ed esempi di applicazione.

Nel capitolo 6 si mostreranno le applicazioni delle due misure su un insieme reale di consumatori. Viene inizialmente fornita la descrizione del sottoinsieme di dati utilizzato per i test, alla quale segue il calcolo effettivo dei due indici e due casi di studio che utilizzano le misure trovate per segmentare la clientela. Per ciascun caso di studio verranno descritti i risultati ottenuti.

Infine, nel capitolo 7, si esporranno le possibili applicazioni dei due indici nell'ambito delle strategie di marketing, e si farà cenno agli sviluppi futuri.

Capitolo 2

Stato dell'arte Economico

Nel libro *'Marketing one-to-one'* l'autore Don Peppers afferma ([3]):

> *«In futuro, non sarà importante cosa conosciamo dei nostri Clienti, ma piuttosto cosa sappiamo su ciascuno di essi.»*

Questa frase, secondo gli esperti del settore, indicherebbe una delle più importanti innovazioni commerciali degli ultimi anni.

La personalizzazione del rapporto azienda-cliente, in effetti, è un concetto ha che rivoluzionato in modo significativo la concezione *classica* del marketing.

Negli anni '70 e '80, infatti, il principio prevalente che guidava la pianificazione e l'attuazione delle politiche di marketing era quello della *segmentazione del mercato*. L'obiettivo fondamentale era quello di produrre per soddisfare un *mercato di massa* e, pertanto, non si prestava attenzione nel comprendere le differenze tra le diverse esigenze dei consumatori.

Negli anni successivi, in corrispondenza di una progressiva saturazione e rallentamento della crescita dei mercati, gli esperti del marketing iniziarono a intuire la diversità delle esigenze degli acquirenti. In questo scenario si iniziò a dividere il mercato in segmenti di individui che presentassero esigenze differenti rispetto ai prodotti (design, servizi, prezzo, etc...). Ciascun segmento, attraverso opportune indagini campionarie di mercato, identificava una specifica categoria di consumatori in termini di *stile di vita*, *età*, *livello culturale*, etc. I prodotti potevano così essere modificati ed adattati alle esigenze di ciascun mercato ottenendo un triplice risultato:

- avvicinarsi maggiormente ai desideri dei clienti;
- evitare di rinunciare alle *economie di scala* della produzione industriale;
- creare una *comunicazione mirata* con i soli soggetti appartenenti ai segmenti di interesse.

Questo paradigma, fondamento delle economie industrializzate, è stato in vigore fino a metà degli anni '90.

Successivamente, i principi *classici* del marketing furono nuovamente messi in discussione dal manifestarsi di due fenomeni.

Il primo fenomeno riguarda la coscienza di una situazione di stallo dei mercati dovuta allo sviluppo della concorrenza tra le imprese. Tutte le aziende avevano utilizzato gli stessi principi di segmentazione e comunicazione nei mercati, trovandosi a competere con gli stessi strumenti per l'acquisizione di una quota di clientela. Crescevano dunque gli investimenti per guadagnare frazioni di mercato, con un rapporto sempre più discutibile tra costi e benefici. Il secondo fenomeno riguarda la diffusione di un nuovo mezzo di comunicazione dotato di caratteristiche di interattività e di personalizzazione dinamica: *internet*.

Con l'avvento di *internet* si rese necessaria una completa revisione dei canoni del marketing. Grazie a questo mezzo di comunicazione, infatti, esisteva la possibilità di poter raccogliere dei dati di mercato basati non più su campioni statistici di clientela, ma bensì su ciascun potenziale cliente. Inoltre, grazie alla comparsa della rete, diventava possibile inviare comunicazioni dirette e personalizzate a ciascun cliente, offrendogli nel contempo la possibilità di interagire e dialogare con l'azienda.

Nasce così nuova concezione di marketing, definito *one-to-one* o *marketing relazionale*, che integra la possibilità di gestire un rapporto interattivo con ciascun cliente.

Il marketing relazionale rappresenta il necessario completamento della *mass customization*. Esso fornisce un approccio di tipo individualizzato (*one-to-one*) con il cliente. L'impresa deve adottare tecnologie, processi e strutture organizzative atte a mantenere ed incentivare relazioni interattive coi clienti secondo un approccio di *Customer Relationship Management* (CRM).

Per CRM si intende la gestione del processo continuo di mantenimento e sviluppo della relazione col cliente, attraverso la permanente creazione e condivisione di valore nel tempo. Scopo del CRM è porre il cliente al centro di ogni processo di creazione di valore aziendale, focalizzando così le risorse e le competenze a disposizione per una piena soddisfazione delle esigenze del cliente stesso e una sua completa fidelizzazione.

Un adeguato sistema di CRM consente all'azienda di fornire al cliente un servizio eccellente in tempo reale, sviluppando una relazione con i clienti migliori tramite un uso efficace delle informazioni acquisite su tale cliente. Sulla base di queste informazioni, l'impresa può personalizzare la propria offerta.

Secondo i principi del CRM, il fattore fondamentale della profittabilità dell'impresa è costituito dal valore aggregato dei suoi clienti e dalle relazioni con essi.

L'impresa migliora il valore derivante dai propri clienti seguendo le seguenti strategie:

- riducendo il tasso di defezione dei clienti;
- aumentando la durata della relazione coi clienti;
- migliorando la profittabilità dei clienti poco redditizi, o terminando il rapporto;
- dedicando particolare attenzione ai clienti migliori.

Il CRM non va identificato con una nuova tecnologia, ma va inteso come una nuova strategia di business: una strategia con la quale l'azienda ottimizza il proprio profitto attraverso la soddisfazione del proprio cliente coinvolgendo processi, tecnologie e persone sia all'interno, che all'esterno dell'organizzazione.

Capitolo 3

Stato dell'arte Tecnico

3.1 Introduzione

Ai giorni nostri il mercato è caratterizzato dall'essere globale; i prodotti e i servizi offerti sono pressoché identici e vi è abbondanza di offerta. Proprio a causa delle dimensione e della complessità dei mercati, il *marketing di massa* risulta molto costoso (Figura 3.1) e il *ritorno degli investimenti* è il principale obiettivo.

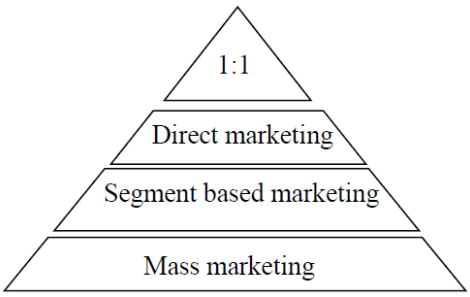


Fig. 3.1: Communication with customers, B2B, B2C [9].

Anziché puntare a tutte le opportunità o fornire le stesse scelte incentivanti a ciascuno, un'azienda può selezionare soltanto quei consumatori che soddisfano certi criteri di *redditività* (*profitability*) basati sulle necessità individuali e sullo schema di acquisto [9]. Questo viene realizzato costruendo un modello per predire il **valore di un individuo** basato su caratteristiche demografiche, stile di vita, e comportamenti precedenti. Il modello produce informazioni che si focalizzano sul mantenimento dei clienti e su programmi di reclutamento col fine di acquisire ed incrementare gli acquisti dei clienti più redditizi. Questo processo è chiamato *customer behavior modeling* (CBM) o *customer profiling*. Un *profilo di un cliente* è uno strumento che aiuta i venditori a comprendere meglio le caratteristiche dei propri clienti.

La motivazione della profilazione del cliente è quella di trasformare questa comprensione in una interazione *automatizzata* con i propri consumatori [9, 11]. Per questi compiti, il mercato odierno ha bisogno di un'ampia gamma di processi e di strumenti dell'*information technology* (IT). Questi strumenti vengono utilizzati per collezionare i dati e semplificare il processo di estrazione della conoscenza sui negozi e sulla pianificazione di campagne di marketing. Gli strumenti di *Data Mining* sono utilizzati per identificare gruppi *significativi* nei dati storici (selezione di criteri per mailing list, identificare negozi ad alto potenziale, cercare caratteristiche negli stili di vita che coincidano con i consumatori). In sostanza, gli strumenti di *Data Mining* permettono di trovare schemi interpretabili dall'uomo che **descrivano** i dati.

3.2 Customer segmentation e customer profiling

Il *customer relationship management* (CRM), include la *customer segmentation* e il *customer profiling*.

Customer segmentation è un termine utilizzato per descrivere il processo di divisione dei consumatori in gruppi omogenei sulla base di attributi condivisi o comuni (abitudini, sensazioni, etc.)

Con *Customer profiling* si intende invece la descrizione dei consumatori tramite i loro attributi, come età, reddito, stile di vita. Questo è realizzato costruendo un modello di comportamento degli utenti e stimando i rispettivi parametri. In base ai dati disponibili, possono essere trovati nuovi clienti o eliminati i clienti non desiderati. L'obiettivo è quello di predire comportamenti basati sulle informazioni che si possiedono di ciascun consumatore [11]. La profilazione viene eseguita dopo la segmentazione.

Avendo a disposizione questi due meccanismi, i venditori possono decidere quali azioni di marketing intraprendere in ciascun segmento e poter così allocare le cosiddette *risorse scarse* ai segmenti in modo da venir incontro agli specifici obiettivi di business (Figura 3.2).

3.2.1 Customer segmentation

La segmentazione viene utilizzata per avere una comunicazione mirata con i consumatori. Il processo di segmentazione descrive le caratteristiche di gruppi di utenti (chiamati *segmenti* o *clusters*) all'interno del dataset. Segmentare significa letteralmente *dividere la popolazione in segmenti* in base a caratteristiche di affinità o similarità. La *customer segmentation* è il primo passo per la classificazione di ciascun utente in base ai gruppi che sono stati definiti.

Le principali difficoltà nel produrre una buona segmentazione sono [1]:

- *rilevanza e qualità dei dati:* se un'azienda ha una quantità insufficiente o troppo elevata di dati, è possibile che le analisi risultino eccessiva-

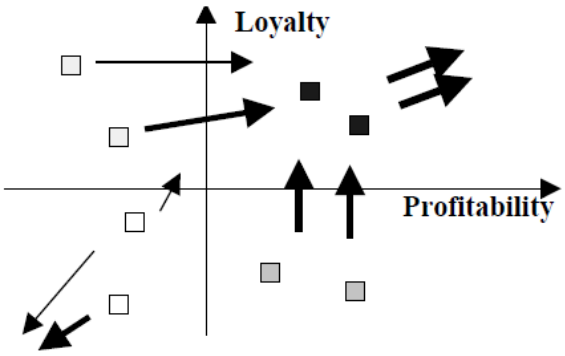


Fig. 3.2: Esempio di segmentazione della clientela sui parametri *lealtà* e *redditività*. Conoscendo il profilo di ciascun consumatore, l'azienda può trattare ciascun individuo nel modo più opportuno per incrementare il suo valore nel tempo [9].

mente complesse e dispendiose dal punto di vista del tempo. Inoltre, con dati mal organizzati (differenti formati o differenti sorgenti) è difficile estrapolare informazioni di interesse. Anche l'utilizzo di troppe variabili può portare alla creazione di segmenti inutili ai fini del marketing.

- *intuizione:* anche in presenza di dati fortemente rilevanti dal punto di vista informativo, i venditori devono sviluppare in modo continuo nuove ipotesi di segmentazione per identificare i dati *corretti* per l'analisi.

- *processi continui:* la segmentazione richiede un continuo sviluppo e aggiornamento dei dati per ogni cliente acquisito. Le strategie di segmentazione vengono infatti influenzate dai consumatori che a loro volta sono influenzati da esse; per questo è necessaria una revisione e riclassificazione degli utenti. Ad esempio, nell'ambiente dell'*e-commerce* la segmentazione richiede almeno un aggiornamento al giorno.

- *sovra-segmentazione:* un segmento potrebbe essere troppo piccolo i insufficientemente distinto per essere trattato come un segmento separato.

3.2.2 Customer profiling

Il customer profiling pone le basi ai venditori per comunicare con i clienti esistenti in modo da offrire loro i servizi migliori e fidelizzarli. Questo viene reso possibile grazie all'assemblamento delle informazioni sui consumatori. Il customer profiling viene anche utilizzato per ipotizzare i profili dei nuovi

clienti avvalendosi di dati esterni all'azienda. I dati vengono usati per dividere la base di dati in cluster di consumatori con caratteristiche di acquisto simili [4].

In base all'obiettivo prefissato, vengono selezionati i profili rilevanti al progetto. Un *profilo utente* è un file che contiene almeno i seguenti attributi: nome, indirizzo, età, città, Regione, codice postale, Nazione. All'occorrenza possono servire dei profili per specifici prodotti; in questo caso il file conterrà informazioni aggiuntive come ad esempio l'importo speso.

Gli attributi che in genere vengono utilizzati per il profiling sono [10]:

- *geografici:* Regione, Nazione, città, C.A.P.;
- *culturali :* etnia, lingue parlate;
- *condizioni economiche e potere d'acquisto:* reddito medio del nucleo familiare, quanto spende un cliente per ciascun prodotto, quante volte un cliente va a fare acquisti;
- *frequenza di shopping, preferenze, grado di soddisfazione:* attributi validi per clienti già acquisiti utili per creare profili di acquisto;
- *età:* di un gruppo di clienti target, membri del nucleo familiare;
- *valori e credenze:* l'atteggiamento di un cliente nei confronti di un prodotto o servizio offerto;
- *ciclo di vita:* regolarità di acquisto del cliente;
- *conoscenza e consapevolezza:* conoscenza dei clienti su un determinato prodotto, servizio o azienda;
- *stile di vita:* caratteristiche di un consumatore necessarie ad identificare le sue abitudini di acquisto;
- *media utilizzati:* dove il consumatore si informa, cosa legge, a quali riviste è abbonato;
- *metodo di reclutamento:* come il consumatore è stato reclutato.

3.2.3 Collezionamento e preparazione dei dati

Per quanto riguarda il collezionamento dei dati esistono diverse fonti:

- *database interno dei consumatori:* le identità possono provenire da mail dirette, programmi di acquisto, concorsi, registrazioni, ricevute e tessere soci;

- *sorgenti esterne:* esistono software o database in grado di scoprire stili di vita o informazioni geografiche utilizzando solamente il codice postale [6];
- *indagini di ricerca:* faccia a faccia, telefoniche, questionari postali, Internet.

Esistono fondamentalmente due tipi di informazioni ricavabili dai dati: *variabili di classificazione* e *variabili di tipo descrittivo*.

Le *variabili di classificazione* sono usate per classificare gli strati corrispondenti a dei segmenti. Le più comuni sono:

- *variabili demografiche:* età sesso, reddito, stato coniugale, educazione, occupazione, tipologia di residenza, etc. ;
- *variabili geografiche:* Città, Stato, codice postale, Nazione, Regione, indirizzo, densità della popolazione, clima, etc. ;
- *variabili comportamentali:* fedeltà al brand, livello di utilizzo, canali di distribuzione utilizzati, reazioni ai fattori di marketing, etc.

Le *variabili di tipo descrittivo* sono utilizzate per descrivere ciascun segmento e permettere di distinguere un gruppo dagli altri. Sono misure facilmente ottenibili dalle risorse interne e utilizzabili come variabili di classificazione.

Preparazione dei dati

Prima di utilizzare i dati all'interno di uno strumento di *data mining*, essi devono essere puliti e convertiti. I principali task riguardano:

- risoluzione di inconsistenze nei formati dei dati e nell'encoding, traduzioni geografiche, abbreviazioni, punteggiatura;
- eliminazione dei campi indesiderati e privi di significatività;
- combinare i dati provenienti da sorgenti diverse sotto una chiave comune;

3.2.4 Costruzione di un modello

Le tecniche di profiling sono un ottimo precursore per la costruzione di un modello, perché aiutano a dividere una vasta popolazione in cluster facilmente gestibili. Il modello che si otterrà andrà poi perfezionato e raffinato con informazioni comparative per uno specifico scenario di marketing. Successivamente andrà testato su un campione rappresentativo del database aziendale.

Per la costruzione di un modello è necessario:

- *identificare le variabili* da includere nel modello (prodotti acquistati, lunghezza temporale della fedeltà del consumatore, etc.);
- *costruire il modello* con i profili e i segmenti di consumatori individuati;
- *usare il modello per predire* i consumatori più propensi all'acquisto;
- *identificare le variabili maggiormente determinanti*, ad esempio variabili che aiutano a predire i gusti di un consumatore.

3.3 Data Mining e principali tecniche

Il Data Mining è un processo analitico di scoperta di relazioni, di pattern e di informazioni precedentemente sconosciute e potenzialmente utili presenti all'interno di grandi database. Un pattern indica una struttura, un modello, una rappresentazione sintetica dei dati. Per questo motivo l'informazione creata viene messa in circolo con tutti gli altri dati per poter essere utilizzata più volte, anche assieme ai dati stessi per creare altra informazione.

L'informazione ottenuta può essere tramutata in azioni commerciali allo scopo di ottenere un vantaggio di business e aumentare la redditività. Infatti la diversificazione e la globalizzazione hanno aumentato il livello di competitività e ottenere un vantaggio sulla concorrenza è diventato fondamentale per la sopravvivenza di un'azienda.

Il Data Mining serve per :

- *Classificare:* si assegna una nuova variabile ad ogni cliente che identifica l'appartenenza ad una determinata classe. Si applica un modello a dei dati grezzi che verranno poi classificati. Si può così dividere la propria clientela e dedicare gli sforzi di una campagna di Marketing ad una determinata classe. Per esempio si possono dividere i clienti secondo il reddito (basso, medio, alto). Esiste comunque un numero di classi già note e l'obiettivo è quello di inserire ogni record (cliente) in una determinata classe.
- *Stimare:* mentre la classificazione usa valori discreti, la stima usa valori continui. In base ad un determinato input, usiamo la stima per individuare il valore di una variabile continua sconosciuta (il reddito per esempio). Spesso stima e classificazione vengono utilizzate simultaneamente.
- *Fare previsioni:* può essere considerata una classificazione o una stima perché i clienti sono classificati in base ad un comportamento futuro prevedibile o stimato. Molta importanza hanno i dati storici perché servono per costruire un modello che spieghi il comportamento futuro in base a quello passato.

- *Raggruppare per affinità o regole di associazione:* l'obiettivo è di stabilire quali oggetti (in genere prodotti) possono abbinarsi tra loro. Si può utilizzare il raggruppamento per affinità per pianificare la disposizione dei prodotti sugli scaffali o nei cataloghi in modo che gli articoli che vengono acquistati insieme si trovino il più possibile vicini.

- *Clustering:* significa segmentare un gruppo di clienti eterogenei in gruppi omogenei. È diverso dalla classificazione perché non si usano classi predefinite, e i record sono raggruppati in base alle analogie. E' il ricercatore che deve stabilire il significato da dare ad ogni cluster.

- *Descrizione e visualizzazione:* una descrizione efficace di uno specifico comportamento indica da dove partire per cercare una spiegazione. La visualizzazione dei dati è una forma molto efficace di Data Mining descrittivo, ed è molto più immediato ricavare utili informazioni da dati visivi.

Le prime tre tecniche sono esempi di **Data Mining diretto**: lo scopo è di usare i dati disponibili per creare un modello che dia come output una specifica variabile obiettivo.

Le altre tre tecniche sono esempi di **Data Mining indiretto**: non esistono più variabili target ma l'obiettivo è quello di stabilire una precisa relazione tra tutte le variabili.

Il Data Mining è molto usato nel settore marketing vista la presenza di grosse quantità di dati da elaborare per ricavarne informazioni utili. Questi dati sono tutti raccolti in un database marketing e si riferiscono a tutti i potenziali clienti (*prospect*) di una campagna di mercato. Questi dati possono descrivere il comportamento del cliente già acquisito o possono contenere una serie di informazioni grezze di tipo demografico sui possibili clienti.

Il Data Mining permette all'azienda di ridurre le spese non contattando la clientela che difficilmente risponderà all'offerta.

3.3.1 Fasi del processo di Data Mining

Il processo di Data Mining è riassumibile nelle sei fasi mostrate in Figura 3.3.

- *Determinazione del problema di business:* il primo passo del processo consiste nel definire l'obiettivo dell'analisi.

- *Selezione ed organizzazione dei dati:* una volta determinato l'obiettivo di business bisogna raccogliere e selezionare i dati necessari per l'analisi.

- *Analisi esplorativa dei dati:* consiste in una prima valutazione delle variabili statistiche per una eventuale eliminazione o trasformazione.

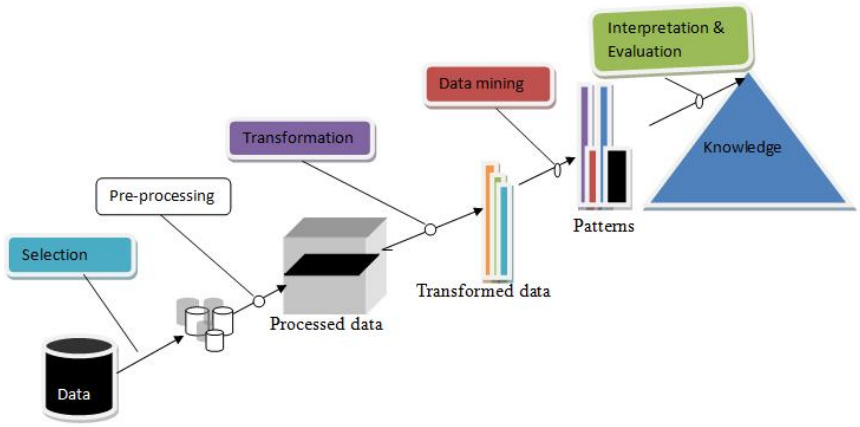


Fig. 3.3: Processo di Knowledge Discovery nei Database (KDD).

Tramite un'adeguata analisi delle variabili statistiche è possibile individuare la presenza di valori anomali che non vanno necessariamente eliminati perché possono contenere informazioni utili al raggiungimento degli obiettivi prefissati. L'analisi esplorativa può essere utile anche per individuare quali variabili sono tra loro correlate in modo da eliminare quelle che hanno la stessa informazione.

- *Data Mining:* si sceglie e si applica una o più tecniche da usare in base all'obiettivo dell'analisi e dai dati disponibili.
- *Interpretazione dei modelli identificati:* analisi e verifica dei risultati con possibile retroazione ai punti precedenti per ulteriori iterazioni al fine di migliorare l'efficacia dei modelli trovati.
- *Consolidamento della conoscenza scoperta:* integrazione della conoscenza e valutazione del sistema mettendo a confronto i risultati con l'effettivo andamento della realtà e produzione della documentazione agli utenti finali o a terze parti interessate.

3.3.2 Tecniche di Data Mining

- *Tecniche di visualizzazione dei dati:* mediante grafici multidimensionali è possibile identificare relazioni complesse per scoprire l'informazione nascosta.
- *Clustering:* tecnica di classificazione che divide una popolazione in sottogruppi. Le tecniche di clustering e le reti neurali non supervisionate consentono di effettuare operazioni di segmentazione della clientela.

- *Reti neurali:*risolvono problemi di classificazione e di previsione. Sono utilizzate in un ambiente dinamico, dove i dati cambiano sempre, data la loro capacità di generalizzare la conoscenza appresa.
- *Alberi decisionali:* rappresentazioni grafiche costruite suddividendo i dati in sottogruppi omogenei. La rappresentazione avviene in modo gerarchico ad albero. Le reti neurali supervisionate e gli alberi di classificazione, fanno uso della conoscenza acquisita per classificare nuovi oggetti o prevedere nuovi eventi.
- *Individuazione di associazioni:* tecniche esplorative per misurare l'affinità dei prodotti. Individuano gruppi di prodotti legati da analoghe abitudini di acquisto.

3.3.3 Applicazioni

- *Scoring system:* È un particolare approccio di analisi incentrato sull'assegnazione ai singoli clienti (*prospect*) della probabilità di adesione ad una campagna commerciale. La finalità è quella di classificare i clienti o gli eventuali prospect in modo tale da attuare azioni di marketing diversificate a seconda dei target individuati. L'obiettivo è quello di costruire un modello predittivo in modo da individuare una relazione tra una serie di variabili comportamentali e una variabile obiettivo che rappresenta l'oggetto di indagine. Il modello produce come risultato un punteggio (*score*) che indica la probabilità di risposta positiva alla campagna.
- *Segmentazione della clientela:* applicazione di tecniche di clustering per individuare gruppi omogenei calcolati secondo variabili comportamentali o socio-demografiche. L'individuazione delle diverse tipologie permette di effettuare campagne di marketing mirate.
- *Market basket analysis:* applicazione di tecniche di associazioni a dati di vendita per individuare quali prodotti vengono acquistati insieme. Utile per la disposizione dei prodotti sugli scaffali.

3.4 Association Analysis

Molte aziende di business accumulano enormi quantità di dati dalle proprie operazioni giornaliere; ad esempio i supermercati collezionano giornalmente gli importi sugli acquisti dei consumatori. Questo tipo di dati viene chiamato *market basket transactions* (Tabella 3.1). Ciascuna riga della tabella corrisponde ad una transazione, che contiene un identificatore univoco chiamato **TID** ed un insieme di articoli comprati da un dato consumatore. I venditori sono interessati ad analizzare i dati per acquisire informazioni

sulle abitudini di acquisto dei propri consumatori. Alcune delle informazioni ricavate possono essere utilizzate a scopi di marketing promozionale, gestione delle scorte e per il *customer relationship management*. Le problematiche principali legate al *market basket analysis* riguardano l'elevato costo computazionale per la ricerca di pattern in dataset con un numero elevato di transazioni, e la possibilità di andare incontro ad associazioni *spurie* ottenute in modo casuale ([8] Cap. 6).

TID	$Items$
1	$\{Bread, Milk\}$
2	$\{Bread, Diapers, Beer, Eggs\}$
3	$\{Milk, Diapers, Beer, Cola\}$
4	$\{Bread, Milk, Diapers, Beer\}$
5	$\{Bread, Milk, Diapers, Cola\}$

Tabella 3.1: Esempio dei carrelli di un cliente.

Itemset e Support Count: sia $I = \{i_1, i_2, ..., i_d\}$ l'insieme di tutti gli articoli in un carrello e sia $T = \{t_1, t_2, ..., t_N\}$ l'insieme di tutte le transazioni. Ciascuna transazione contiene un sottoinsieme di articoli scelti da I. Nell'*analisi associativa*, una collezione con uno o più articoli è chiamata *itemset*. Ad esempio, $\{Beer, Diapers, Milk\}$ è un esempio di 3-itemset. L'insieme nullo (o vuoto) è un insieme che non contiene articoli.

La lunghezza della transazione è definita come il numero di articoli presenti in una transazione. Una transazione t_j contiene un itemset X se X è un sottoinsieme di t_j. Una importante proprietà di un itemset è il suo *support count*: è il numero di transazioni che contengono un particolare itemset. Matematicamente il *support count* , $\sigma(X)$ è definito come:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|, \tag{3.1}$$

dove il simbolo $|.|$ denota la cardinalità dell'insieme. Il *supporto* è una misura significativa perchè un itemset con un supporto molto basso compare in modo casuale. Una regola con supporto basso è anche probabile che sia poco interessante dal punto di vista di business perché non risulta vantaggioso promuovere articoli che i clienti raramente acquistano insieme.

Itemset Frequenti: sono tutti quegli itemset che hanno un supporto maggiore o uguale ad una soglia prefissata (*minsup*).

3.4.1 Generazione itemset frequenti

In generale, un dataset che contiene k elementi può potenzialmente generare $2^k - 1$ itemset frequenti (escludendo l'insieme vuoto). Poiché k nelle applicazioni pratiche è molto grande, lo spazio di ricerca degli itemset da esplorare diventa esponenzialmente elevato. Un esempio di generazione di tutti i possibili itemset, a partire da dei itemset *candidati* è mostrato in figura (Figura 3.4). Computazionalmente si avrebbero $O(NMw)$ confronti, dove N è il numero di transazioni, M il numero di **itemset frequenti candidati**, e w è la lunghezza massima delle transazioni.

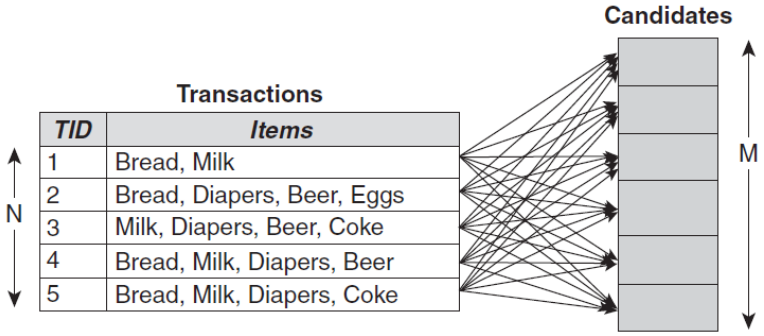


Fig. 3.4: Conteggio del supporto per ricavare gli itemset candidati.

Ci sono diverse strategie per ridurre la complessità computazionale per la generazione di itemset frequenti:

- *ridurre il numero di itemset candidati (M):* il *principio apriori* ,(Figura 3.5) , è una della strategie per eliminare alcuni degli itemset candidati senza contare il loro valore dei supporti;
- *ridurre il numero di confronti:* anziché confrontare ciascun itemset candidato con ogni transazione, si può ridurre il numero di confronti utilizzando strutture dati avanzate, sia per salvare gli itemset che per comprimere il dataset;
- *utilizzare la proprietà antimonotona del supporto:* se un itemset è infrequente, anche i suoi sovrainsiemi sono infrequenti (Figura 3.6).

3.4.2 Algoritmo Apriori

Apriori è il primo algoritmo sulla ricerca di regole associative basato sull'utilizzo del pruning incentrato sulla misura del supporto e sul controllo sistematico della crescita esponenziale degli itemset candidati. Viene di seguito mostrato lo speudocodice sulla parte della generazione degli *itemset frequenti* .

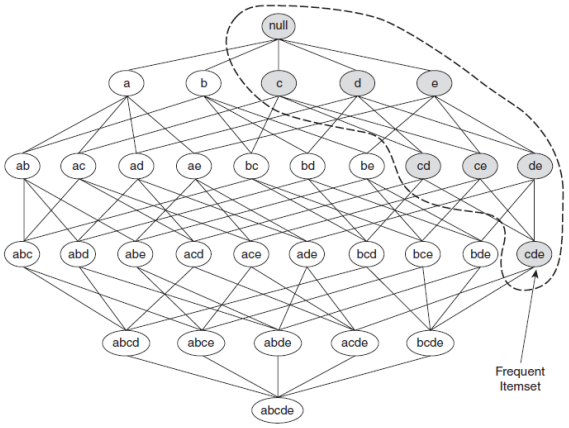


Fig. 3.5: Illustrazione del principio Apriori. Se $\{c, d, e\}$ è un itemset frequente, allora tutti i suoi sottoinsiemi sono frequenti.

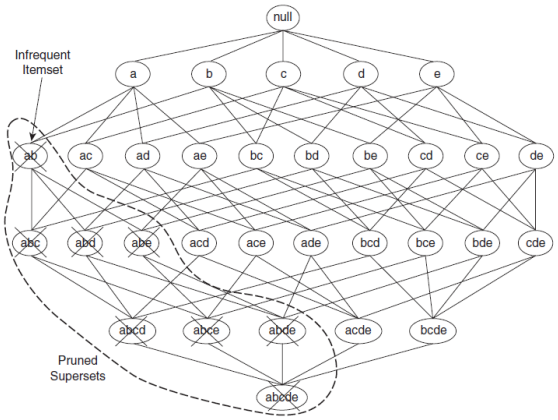


Fig. 3.6: Illustrazione della proprietà antimonotona del supporto e del pruning basato su di essa. Se $\{a, b\}$ è un itemset infrequente, allora tutti i suoi sovrainsiemi sono infrequenti.

Sia C_k l'insieme dei *k-itemset* (itemset di lunghezza k) e F_k l'insieme dei *k-itemset* **frequenti**:

- L'algoritmo inizialmente esegue uno step sul dataset per determinare il supporto di ciascun elemento. In questo modo vengono trovati tutti gli *1-itemset* frequenti (step 1 e 2);
- Successivamente l'algoritmo genera iterativamente i nuovi *k-itemset* candidati a partire dai *(k-1)-itemset* trovati nelle iterazioni precedenti (step 5). La generazione dei candidati è implementata utilizzando una funzione chiamata *apriori-gen*;
- Per il conteggio del supporto dei candidati, l'algoritmo ha bisogno di

Algoritmo 1: Generazione Itemset frequenti algoritmo Apriori

1 $k = 1$;
2 $F_k = \{i | i \in I \wedge \sigma(i) \geq N \times minsup\}$ *{Trovo gli 1-itemset frequenti}* ;
3 **repeat**
4 $k = k + 1$;
5 $C_k = apriori\text{-}gen(F_{k-1})$ *{Genero itemset dei candidati}* ;
6 **for *each*** transaction $t \in T$ **do**
7 $C_k = subset\ (C_k, t)$ *{trovo i candidati appartenenti a t}*;
8 **for *each*** candidate itemset $c \in C_t$ **do**
9 $\sigma(c) = \sigma(c) + 1$ *{Incremento il support count}* ;
end
end
10 $F_k = \{c | c \in C_k \wedge \sigma(c) \geq N \times minsup\}$ *{estraggo i k-itemset frequenti}* ;
until $F_k = \emptyset$;
11 $Result = \bigcup F_k$

un passo aggiuntivo (passi da 6 a 9). La funzione $sottoinsieme$ viene utilizzata per determinare tutti gli itemset candidati in C_k che sono contenuti in ciascuna transazione t.

- Dopo aver contato i loro supporti, l'algoritmo elimina tutti gli itemset candidati il cui supporto è minore di $minsup$ (step 10).
- L'algoritmo termina quando non ci sono più itemset frequenti generabili, ad esempio quando $F_k = \emptyset$.

Per quanto riguarda la generazione dei candidati tramite la procedura *apriori-gen*, viene utilizzato il metodo $F_{k-1} \times F_{k-1}$; esso fonde due coppie di *(k-1)-itemset* se e solo se i loro primi *(k-2)-itemset* sono identici. Formalmente, dati $A = \{a_1, a_2, ..., a_{k-1}\}$ e $B = \{b_1, b_2, ..., b_{k-1}\}$ due coppie di *(k-1)-itemset*, essi saranno fusi se soddisfano la seguente condizione:

$$a_i = b_i (for\ i\ = 1, 2, ..., k-2) \wedge a_{k-1} \neq b_{k-1} \tag{3.2}$$

3.5 Cluster Analysis

La Cluster Analysis raggruppa oggetti basandosi esclusivamente su informazioni trovate all'interno dei dati che descrivono gli oggetti e le loro relazioni. Gli oggetti all'interno di un gruppo sono simili tra loro, mentre sono diversi dagli oggetti appartenenti ad altri gruppi. Più un gruppo è

omogeneo, maggiori saranno le differenze tra i gruppi e, di conseguenza, maggiormente distinti saranno i clusters.

In molte applicazioni, la nozione di cluster non è ben definita. In generale si può affermare che la migliore definizione dei cluster dipende dalla natura dei dati e dai risultati cercati.

La Cluster Analysis è collegata ad altre tecniche come la **classificazione**; per questo ci si riferisce al clustering come **classificazione non supervisionata** ([8] Cap. 8).

3.5.1 Tipologie di Cluster

Esistono diversi tipi di cluster. I più comuni sono:

- **Well Separated:** un cluster è un insieme di oggetti ben definiti nei quali ciascun oggetto è simile agli altri oggetti all'interno del cluster, piuttosto che agli oggetti appartenenti a cluster differenti. A volte viene utilizzata una soglia per specificare che tutti gli oggetti in un cluster sono sufficientemente simili l'uno all'altro. Ovviamente questa definizione idealistica di cluster è soddisfatta solamente quando i dati contengono clusters naturali molto distanti gli uni dagli altri. La distanza tra due punti appartenenti a gruppi differenti è più grande della distanza tra qualsiasi altra coppia di punti appartenenti allo stesso gruppo (Figura 3.7 (a)).

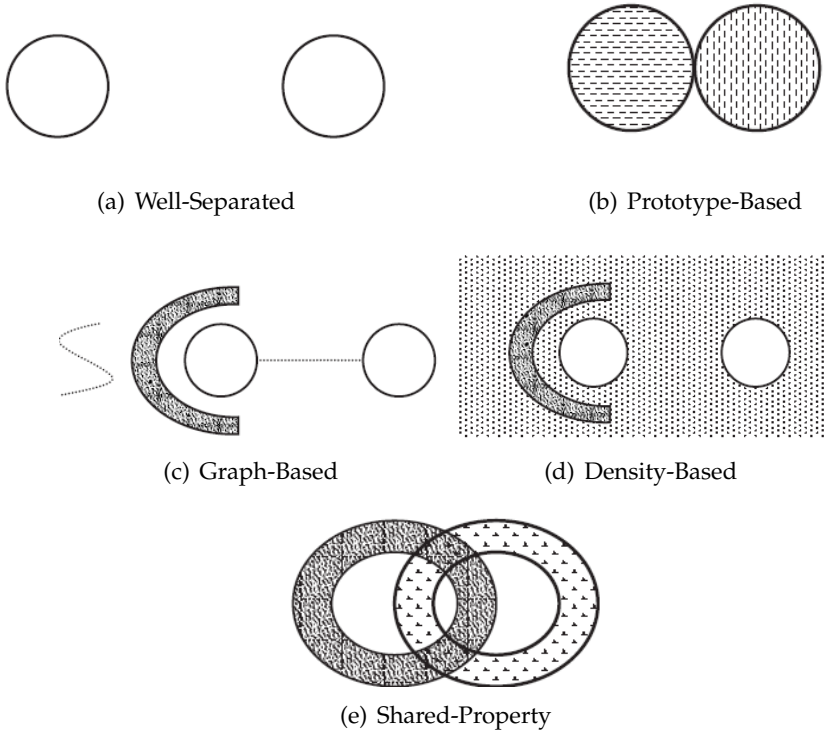
- **Prototype-Based:** un cluster è un insieme di oggetti nei quali ciascun oggetto è simile al prototipo che definisce il cluster. Per i dati con attributi continui, il prototipo è un *centroide*, ad esempio la media, di tutti i punti nel cluster. Nel caso in cui il centroide sia privo di significato, come per dati di tipo categorico, il prototipo è il *medoide*; esso è il punto più centrale di un cluster (Figura 3.7 (b)).

- **Graph-Based:** se i dati sono rappresentati mediante grafi, in cui i nodi sono oggetti e i link rappresentano le connessioni tra oggetti, allora un cluster può essere definito come componente connessa (gruppo di oggetti connessi gli uni agli altri). Un importante esempio di cluster basati sui grafi sono i cluster basati sulla contiguità, nei quali due oggetti sono connessi solamente se rientrano all'interno di una specifica distanza gli uni dagli altri (Figura 3.7 (c)).

- **Density-Based:** un cluster è una regione densa di oggetti circondati da regioni a bassa densità. La definizione di cluster basati sulla densità è impiegata su cluster irregolari e intrecciati, oppure quando sono presenti outlier e rumore (Figura 3.7 (d)).

- **Shared-Property:** un cluster può essere definito, in modo generale, come un insieme di oggetti che hanno proprietà condivise. L'algoritmo

di clustering necessita di uno specifico concetto di cluster per identificare correttamente i cluster. Il processo di calcolo di ciascun cluster è chiamato *cluster concettuale* (Figura 3.7 (e)).



(a) Well-Separated (b) Prototype-Based

(c) Graph-Based (d) Density-Based

(e) Shared-Property

Fig. 3.7: Tipologie di Cluster

3.5.2 Tecniche di Clustering

Esistono principalmente tre tecniche di clustering:

- **K-means:** è una tecnica di partizionamento *prototype-based* che cerca di trovare un numero K di clusters, con K definito dall'utente. Ciascun cluster è rappresentato dal proprio *centroide*;
- **Agglomerative Hierarchical Clustering:** è un approccio al clustering basato sulla fusione di due cluster vicini in uno singolo. In prima istanza ciascun punto rappresenta un cluster;
- **DBSCAN:** è una tecnica basata sulla densità che produce un partizionamento nel quale il numero K di cluster viene determinato in modo automatico. Le regioni a bassa intensità vengono classificate come

rumore e quindi vengono omesse. Dunque il **DBSCAN** non produce un clustering completo.

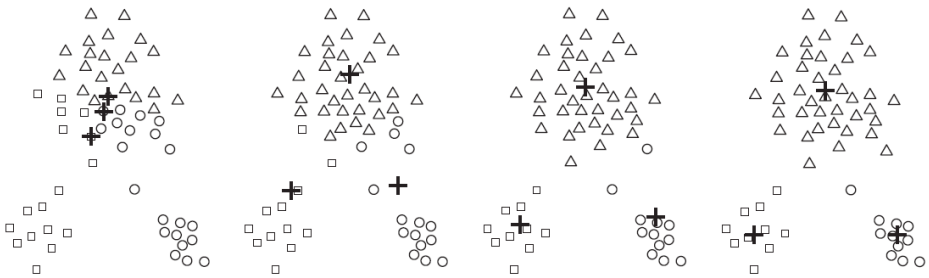
3.5.3 Simple K-MEANS

K-means definisce i cluster in termini di *centroidi*. Un centroide consiste nella *media* di un gruppo di punti applicata agli oggetti appartenenti ad uno spazio continuo n-dimensionale. Lo speudocodice dell'algoritmo è mostrato di seguito:

Algoritmo 2: Algoritmo Simple K-MEANS

```
1 Seleziono K punti come centroidi iniziali ;
2 repeat
3 |   Formo K cluster assegnando ciascun punto al cluster più vicino;
4 |   Ricalcolo il centroide di ciascun cluster;
  until I centroidi non cambiano;
```

L'algoritmo sceglie K centroidi iniziali, con K parametro specificato dall'utente che rappresenta il numero di cluster desiderato. Ciascun punto viene assegnato al centroide più vicino e ciascun gruppo di punti assegnato ad un centroide è un cluster. Il centroide di ciascun cluster viene poi aggiornato in base ai punti assegnati al cluster. La procedura di assegnamento ai cluster e la procedura di aggiornamento dei centroidi viene ripetuta finchè nessun punto cambia cluster, o equivalentemente, se nessun centroide cambia. Un esempio di esecuzione dell'algoritmo è mostrato in Figura 3.8:



(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

Fig. 3.8: Esempio di utilizzo di K-MEANS per trovare 3 clusters.

Per quanto riguarda l'assegnazione dei punti al cluster più vicino vi è bisogno di una misura di prossimità che quantifichi la nozione di *vicinanza* per i dati specifici presi in considerazione. La distanza *Euclidea* è la più utilizzata per punti appartenenti allo spazio euclideo. Le misure di similarità utilizzate in k-means sono relativamente semplici poiché l'algoritmo calcola la misura per ciascun punto di ciascun centroide.

Invece, per quanto riguarda la misura che quantifica la qualità di un cluster si utilizza una **funzione obiettivo** chiamata **SSE** (sum of the squared error). In altre parole, viene calcolato l'errore di ciascun punto, ovvero la distanza euclidea dal centroide più vicino, e si calcola la somma totale degli errori. Dati due diversi insiemi di cluster che sono stati calcolati da due diverse istanze di **K-MEANS**, si preferisce l'insieme in cui l'**SSE** risulta più piccolo.

Formalmente l'**SSE** è definito come:

$$SSE = \sum_{i=1}^{N} \sum_{x \in C_i} dist(c_i, x)^2, \tag{3.3}$$

con K il numero di cluster, x un oggetto, C_i l'i-esimo cluster e c_i il centroide del cluster C_i. La funzione *dist* rappresenta la distanza euclidea standard. Il centroide che minimizza l'**SSE** del cluster è la media. Il centroide (media) del cluster i-esimo è definito dall'equazione:

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x, \tag{3.4}$$

con m_i il numero di elementi del cluster i-esimo.

3.6 Entropia nella teoria dell'informazione

Nella teoria dell'informazione l'*entropia* misura la quantità di incertezza o informazione presente in un segnale aleatorio. Il primo studioso che si dedicò a questa tematica fu *Claude Shannon*. Il suo primo lavoro sull'argomento si trova nell'articolo *"A Mathematical Theory of Communication"* pubblicato nel 1948. Nel primo teorema di Shannon, o teorema di Shannon sulla codifica di sorgente, egli dimostrò che una sorgente casuale d'informazione non può essere rappresentata con un numero di bit inferiore alla sua entropia, cioè alla sua *autoinformazione* media.

Formalmente, l'entropia di una variabile aleatoria X è definita come il valore atteso dell'autoinformazione $I(X)$ della variabile aleatoria:

$$H(X) = \mathrm{E}[I(X)] = -\mathrm{E}[\log \cdot P(X)] \tag{3.5}$$

Se il numero di valori che può assumere X è *finito* allora il valore atteso si riduce ad una media dell'autoinformazione di ogni simbolo x_i pesata con la propria probabilità $P(x_i)$:

$$H(X) = -\sum_{i=1}^{N} P(x_i) \cdot P(x_i) \tag{3.6}$$

Se invece X è una variabile aleatoria *continua* il valore atteso dell'autoinformazione deve essere calcolato attraverso un integrale:

$$H(X) = -\int P(x_i) \cdot P(x_i) \tag{3.7}$$

L'entropia di una variabile aleatoria continua non condivide tutte le proprietà di quella per variabili discrete, ad esempio $H(X)$ può risultare negativa.

La base del logaritmo originariamente utilizzata da Shannon fu quella naturale, ma oggi è di uso comune la base 2 in quanto consente di ottenere dei risultati più chiari. Il valore di entropia ottenuto è misurato in bit e può essere interpretato come il numero di bit necessari in media a memorizzare un simbolo emesso dalla sorgente. In figura Figura 3.9 è mostrata la distribuzione probabilistica dell'entropia. Essa raggiunge il valore massimo (1.0) quando la variabile aleatoria assume probabilità 0.5.

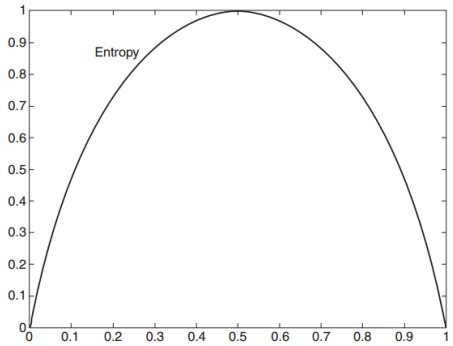


Fig. 3.9: Grafico dell'entropia nella Teoria dell'informazione

Negli ultimi anni la misura dell'entropia ha trovato applicazione anche nell'ambito del *Data Mining*. Ad esempio viene utilizzata nel campo della *classificazione* che consiste l'assegnazione di un oggetto ad un insieme di classi predefinite. Qualsiasi tecnica utilizzata (alberi di decisione, reti neurali, classificatori Bayesiani, etc.), impiega algoritmi di apprendimento per identificare un modello che meglio contraddistingua le relazioni tra un insieme di attributi e un insieme di classi a partire dai dati passati in input (sotto forma di record). L'entropia, assieme all'*indice di Gini* e all'*errore di classificazione*, è una misura di *impurità* utilizzata per scegliere il miglior

modo di suddividere i record in base alla loro distribuzione prima e dopo un *taglio* su un determinato attributo. Questa misura di impurità viene inglobata nell'**information gain** (Δ) per misurare il guadagno di informazione successivo ad uno split su un attributo. Formalmente:

$$\Delta = I(parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} \cdot I(v_j), \tag{3.8}$$

dove $I(\cdot)$ è la misura di impurità di un dato nodo, N il numero totale di record del nodo padre, k il numero di valori degli attributi e $I(v_j)$ il numero di record associati al nodo figlio v_j.

Un'altra applicazione della misura dell'entropia viene esposta da Albert-Lászlo Barabási , professore presso l'Università di Notre-Dame e direttore al Centro di Ricerca Reti Complesse (**CCNR**) presso la Northeastern University di Boston. Nel saggio "Lampi, La trama nascosta che guida la nostra vita" espone la sua tesi: *«Se si accetta che il comportamento umano è casuale, improvvisamente diventa prevedibile»* ([2] pag.289). Noi pensiamo che il nostro comportamento sia determinato da una serie di azioni che siamo abituati a considerare come eventi discreti, casuali e isolati. Secondo Barabási la realtà è molto diversa:

> *«Se tutto va bene, prima di aver finito di leggere questo libro vi avrò convinto che, nonostante tutta la spontaneità che potete mostrare, siete molto più prevedibili di quanto sareste disposti ad ammettere.»*

La popolazione umana vive attuando le proprie azioni in concentrazioni simili a *lampi* di attività frenetica, intervallati da periodi di calma, e che questo modello ricalcherebbe ciò che avviene in molti diversi sistemi naturali.

Lo studioso di sistemi complessi, quindi, premette un assunto: stabilito che gli eventi naturali sono fenomeni prevedibili, allo stesso modo sono da considerarsi i comportamenti degli esseri umani. Le azioni umane seguono dei modelli decifrabili e ripetitivi, quindi prevedibili. Le tecnologie che utilizziamo oramai nella vita di tutti i giorni (navigazione internet, spostamenti, acquisti, etc.) sono immense banche dati dei comportamenti umani in cui si possono identificare e catalogare le azioni umane per determinare modelli di comportamento.

La tesi dell'opera sostiene che, per prevedere il futuro, bisogna prima conoscere il passato. Se vogliamo sapere quanto è prevedibile una persona, come prima cosa si deve determinare la sua **entropia**. E' questo il concetto che viene ripreso dagli algoritmi di *Data Mining*. L'entropia (dal greco $\epsilon\nu$, "dentro" e $\tau\rho o\pi\eta$,"trasformazione") di ciascun individuo è fotografabile e

registrabile. L'essere umano, così come qualsiasi altro fenomeno fisico, è in possesso di una propria entropia, ed in base ad essa (ordine, disordine o trasformazione) è misurabile, profilabile, catalogabile e prevedibile. Gli algoritmi di Data Mining esistono proprio per cogliere questi aspetti per rendere come risultato la profilazione di un individuo o un gruppo di individui.

Capitolo 4

Database

Il dataset oggetto di studio riguarda la vendita al dettaglio della **Coop**, una delle più grandi compagnie di ditribuzione italiane. I dati in possesso appartengono al distretto **Unicoop Tirreno** che serve le province di Massa e Carrara, Lucca, Livorno, Grosseto, Siena, Terni, Viterbo, Roma, Latina, Frosinone, Caserta, Napoli e Avellino. L'intera base di dati contiene informazioni sui negozi, prodotti e sui consumatori coprendo un arco temporale che va dal 2 Gennaio 2007 al 10 agosto 2014.

I clienti attivi su questa finestra temporale sono 1.603.870. Un cliente è definito *attivo* se ha fatto acquisti nel periodo di riferimento; esso è *riconoscibile* se durante gli acquisti ha usato la carta socio. I 133 negozi della compagnia coprono una parte considerevole dell'Italia. La distribuzione geografica dei negozi e dei consumatori è mostrata in Figura 4.1.

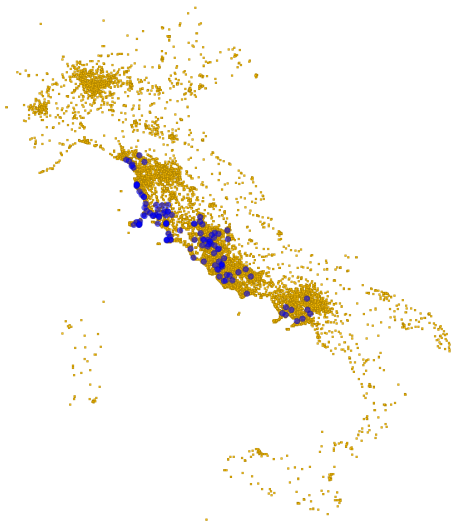


Fig. 4.1: Distribuzione dei negozi appartenenti all'Unicoop Tirreno (in blu) e dei clienti Coop (in giallo).

4.1 Data Warehouse

Lo schema concettuale del *Data Warehouse* **COOP** è mostrato in Figura 4.2. Esso rappresenta solamente una piccola parte dell'intero dataset, quella rilevante per la profilazione degli utenti in base al metodo proposto basato sul comportamento di acquisto e sul comportamento spazio-temporale.

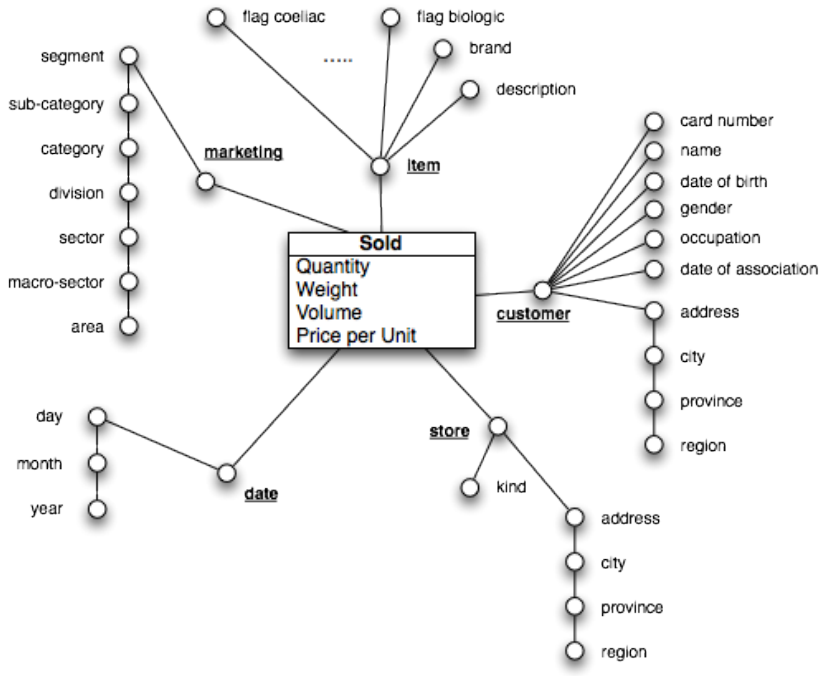


Fig. 4.2: Datamart Unicoop Tirreno

La tabella dei fatti è rappresentata dal **Venduto**, all'interno della quale ciascuna riga rappresenta un articolo venduto nella catena di supermercati. Le *misure* principali all'interno della tabella dei fatti sono:

- *quantità:* è il numero di articoli identici acquistati in una specifica spesa. Questo valore dipende da come il cassiere batte un articolo; può scansionarne uno e moltiplicarlo per il numero effettivo acquistato, oppure può scansionarli singolarmente.
- *peso:* è associato ad articoli riguardanti il reparto ortofrutticolo, salumeria, macelleria e panetteria.
- *volume:* è un valore associato ad articoli *imballati* come acqua, bibite, etc.
- *prezzo per unità:* costo effettivo di un articolo al netto di sconti, offerte, promozioni, etc.

Le principali dimensioni associate alla tabella dei fatti sono cinque.

- *Articoli:* contiene informazioni sui prodotti: una breve descrizione in linguaggio naturale, la marca e una lista di flag per i prodotti *speciali* (organici, per celiaci, prodotti di origine controllata, prodotti geograficamente protetti, etc.) La tabella contiene informazioni su 332.322 prodotti distinti.

- *Clienti:* contengono informazioni sui consumatori: numero carta socio, nome, data di nascita e informazioni sulla residenza (indirizzo e/o coordinate geografiche).

- *Negozi:* contiene informazioni sui supermercati. In particolare fornisce informazioni geografiche (indirizzo e/o coordinate geografiche) e la tipologia di negozio. Uno store può essere classificato come:

  - *Iper:* sono i negozi più gradi; hanno una struttura con un'area di vendita al dettaglio superiore ai $2.500m^2$. Vendono qualsiasi tipo di prodotto (cibo, prodotti per la pulizia, televisori, scooters, elettrodomestici, etc.).
  - *Super:* sono i negozi di medie dimensioni; hanno struttura con un'area di vendita al dettaglio che va dai $400m^2$ ai $2.500m^2$. Hanno un assortimento medio-elevato (non vendono elettrodomestici, televisori e altri prodotti costosi).
  - *GestIn:* sono i negozi più piccoli; hanno struttura con un'area di vendita al dettaglio che va dai $100m^2$ ai $400m^2$. Vendono principalmente prodotti alimentari e un sottoinsieme di prodotti non alimentari (prodotti per la pulizia della casa, per l'igiene personale, etc.).
  - *E-commerce:* negozio virtuale dove è possibile fare la spesa online con possibilità di ritiro della spesa al punto vendita o mediante consegna a domicilio.

- *Data:* contiene informazioni sul giorno, sul mese, sull'anno e sul tipo di giorno (feriale o festivo).

- *Marketing:* questa dimensione è utilizzata per classificare i prodotti; è organizzata ad albero e rappresenta una gerarchia costruita sulle tipologie di prodotto. E' stata costruita dagli esperti di marketing della compagnia. Il livello più alto della gerarchia è l'*area* di appartenenza di un articolo, che divide i prodotti in tre aree fondamentali: *alimentari*, *non alimentari* (prodotti per la persona, per la casa, bricolage, etc.) e *altri* (parafarmaci, buoni sconto, contenitori spesa, etc.). Il livello più basso della gerarchia di marketing è il *segmento*; esso contiene 4719 valori distinti. Ciascun articolo può essere raggruppato dunque su

vari livelli . Le gerarchia principalmente utilizzata è il livello *segmento*. La differenza rispetto all'*articolo* consiste essenzialmente nell'imballaggio, nelle dimensioni e nella marca. Ad esempio gli articoli distinti *Coca Cola 1,5 litri*, *Coca Cola 1 litro*, *Coca Cola 6x1,5 litri* e *Pepsi 1,5 litri* appartengono allo stesso segmento *Cola Drinks*.

Capitolo 5

Metodo

Sin dalla nascita della disciplina, il Data Mining trova una naturale applicazione nel campo delle vendite al dettaglio. Il sogno di ogni venditore è quello di capire i criteri che i clienti utilizzano per la scelta dei beni da acquistare. La sintesi di questi criteri può essere utilizzata per trasformare il *marketing di massa* in *marketing diretto*. La personalizzazione della relazione con ciascun consumatore in termini di specifiche offerte, promozioni o coinvolgimenti a eventi speciali potrebbe realmente migliorare l'efficienza delle sezioni aziendali che si occupano del *customer relationship management* (**CRM**). Lo scopo è il raggiungimento di uno o più obiettivi di business (mantenere la fedeltà del consumatore, incrementare i propri consumatori, prevedere i consumatori che abbandoneranno lo store) anche per le aziende con risorse limitate in termini di tempo e denaro da investire.

Per scoprire informazioni sui consumatori sono state applicate in passato numerose tecniche:

- *Market Basket Analysis:* utilizzata per ricercare pattern comuni su diversi carrelli di spesa;
- *Churn and Retention Analysis:* utilizzata per identificare gli utenti che stanno per abbandonare uno store;
- *Customer Profiling:* utilizzata per dividere gli utenti in gruppi omogenei in base ad attributi prefissati. Un esempio è l'analisi **RFM** (*regency, frequency and monetary value*) che calcola il valore dei consumatori in base a parametri come la frequenza di acquisto, la redditività di un cliente, etc.

Il metodo proposto vuole introdurre una nuova misura che descriva quanto un consumatore sia sistematico nel proprio comportamento di acquisto in maniera tale da profilare i consumatori e successivamente ricercare i comportamenti distintivi basandosi esclusivamente sulle spese effettuate nel passato.

Attualmente il comportamento di acquisto è stato studiato secondo due prospettive:

- una basata sulla composizione di un *carrello di spesa*: tipicamente un consumatore *sistematico* tende ad acquistare in modo frequente insiemi omogenei di prodotti;
- una basata sul *dove* e *quando* un consumatore va nei supermercati: un consumatore *sistematico* tende ad andare nello stesso negozio gli stessi giorni della settimana e in una fascia oraria definita.

Nelle sezioni successive verrà descritta nel dettaglio l'implementazione del metodo focalizzandosi sulla:

- sintesi dell'*indice di sistematicità del comportamento di spesa*;
- sintesi dell'*indice di sistematicità spazio-temporale*;

5.1 Indice di sistematicità del comportamento di spesa

Lo scopo dell'indice è la valutazione della composizione di ciascun carrello di spesa di uno specifico consumatore, in maniera tale da poter misurare se un consumatore tende ad acquistare sempre gli stessi prodotti. Un cliente può avere diverse abitudini di acquisto: ad esempio durante le vacanze potrebbe andare a fare la spesa più spesso comprando pochi prodotti, mentre durante i periodi lavorativi potrebbe fare la spesa soltanto nei weekend e comprare invece molti prodotti per coprire tutta la settimana. Per queste motivazioni è necessario estrarre i cosiddetti *carrelli tipici* che descrivano tutti i possibili comportamenti di un cliente durante gli acquisti.

Come prima approssimazione, i passi che esegue il metodo per il calcolo dell'indice (su un singolo cliente) sono i seguenti:

- esecuzione del *Frequent Pattern Mining* sulle spese e successiva estrazione dei gruppi di prodotti acquistati assieme (*itemset frequenti*);
- ordinamento dei *pattern* trovati rispetto alla loro lunghezza e supporto;
- inizializzazione della *frequenza* di ciascun pattern a zero;
- per ciascuna spesa viene cercato nell'insieme di pattern, precedentemente ordinato, quello che sia un sottoinsieme della spesa (oppure abbia un match completo), e:
  - se viene trovato un pattern corrispondente ai criteri, la sua frequenza viene incrementata di 1;

  - se non viene trovato un pattern corrispondente ai criteri, la spesa viene inserita come pattern con frequenza pari ad 1;

- le frequenze vengono utilizzate come probabilità per il calcolo dell'*entropia normalizzata* sull'insieme di pattern.

L'entropia, o *Basket Regularity Index* (**BRI**) ci darà una misura compresa nel dominio $[0, 1]$. Un indice vicino allo zero indicherà un consumatore *molto sistematico*, mentre un indice tendente ad uno indicherà un consumatore *non sistematico* o *casuale* (poiché tutte le sue spese sono diverse le une dalle altre).

5.1.1 Pseudocodice del Basket Regularity Index

Viene di seguito riportato lo speudocodice dell'algoritmo per calcolo dell'entropia sul comportamento di acquisto di un insieme di clienti:

Algoritmo 3: BRI()

Data: $clients\_list$,$support$
begin
1 $Result$ = {};
2 **for** $client$ **in** $client\_list$ **do**
3 $start\_time = time()$;
4 $baskets = load\_baskets(client)$;
5 $frequent\_item\_sets = compute\_apriori(baskets, support)$;
6 $order\_frequent\_item\_sets\_by\_length()$;
7 $tag\_item\_sets(baskets, frequent\_item\_sets)$;
8 $entropy = calculate\_entropy(baskets, frequent\_item\_sets)$;
9 $weighted\_avg = calculate\_avg(baskets, frequent\_item\_sets)$;
10 $computation\_time = time() - start\_time$;
11 $Result.add(client, entropy, weighted\_avg, computation\_time)$;
end
12 **return** $Result$
end

L'algoritmo prende come input una *lista di clienti* e il *supporto minimo* da utilizzare per la ricerca dei pattern frequenti. Per ciascun cliente vengono caricate le spese effettuate (step 4). I carrelli della spesa risiedono in un file (uno per cliente) nel quale ciascuna riga rappresenta un singolo carrello: un insieme di articoli *distinti*. Sull'insieme dei carrelli viene eseguito l'algoritmo *Apriori* con il supporto minimo passato come parametro (step 5). I pattern trovati vengono ordinati in base alla loro lunghezza (step 6) e, su ognuno di essi, viene calcolata la frequenza di "matching" su ciascun carrello di spesa mediante la funzione *tag_item_sets* (step 7 descritta in seguito). Infine viene calcolata l'entropia normalizzata sul numero totale di

spese (step 8) e il numero medio di articoli distinti presenti nei pattern (step 8). Come risultato finale si otterrà un unico file *csv* nel quale ciascuna riga rappresenterà un cliente. Gli attributi raccolti su ogni consumatore sono: *cliente_id* (identificativo univoco del cliente), *entropia* (indice di regolarità dei carrelli), *weighted_avg* (numero medio di articoli presenti nei pattern) e *computation_time* (tempo di elaborazione in millisecondi per l'analisi del cliente). L'ultimo parametro viene utilizzato per valutare la *performance* dell'algoritmo e per una corretta scelta del parametro *supporto* passato in input. Ci si aspetta infatti che, utilizzando un supporto troppo basso, la procedura *Apriori* impieghi più tempo per calcolare tutti i possibili pattern frequenti.

Lo speudocodice della funzione *tag_item_sets*, che si occupa dell'associazione di un itemset frequente ad un carrello della spesa, è mostrato di seguito:

Algoritmo 4: tag_item_sets()

```
   Data: baskets, frequent_item_sets
1  begin
2      candidates = {};
3      for basket in baskets do
4          while exists item_set in item_sets such that
           item_set ⊆ basket do
5              candidates.add(item_set);
6              candidates = candidates.select_best_fitting_subset();
           end
7          if candidates is null then
8              candidates.add(basket);
           end
9          update_frequency_set(frequent_item_sets);
       end
   end
```

L'algoritmo prende come input un insieme di spese ed un insieme di itemset frequenti di uno specifico cliente, restituendo come risultato il sottoinsieme di pattern frequenti che rappresentano i *carrelli tipici* di un consumatore. A partire da questo sottoinsieme e dalla frequenza di ogni pattern, verrà poi calcolata l'entropia normalizzata o **BRI**. In prima istanza viene inizializzato un insieme vuoto rappresentante i carrelli tipici "candidati" (step 2). Successivamente, per ciascun carrello di spesa, viene ricercato tra i pattern frequenti quello che meglio rappresenta il carrello stesso (step 3-8). Tutti i pattern che sono un sottoinsieme del carrello preso in analisi, vengono inseriti nell'insieme dei candidati (step 5); qualora non esistesse un pattern idoneo, il carrello stesso verrà inserito nell'insieme dei pattern candidati (step 7-8). Il cuore dell'algoritmo è rappresentato dal metodo

select_best_fitting_subset() (step 6): esso si occupa di *eleggere* tra i pattern candidati quello maggiormente rappresentativo nel comportamento di spesa. Per far ciò utilizza i seguenti criteri di ordinamento sull'insieme dei pattern:

- predilige la lunghezza dell'itemset a prescindere dal supporto ad esso associato;
- a parità di lunghezza dell'itemset viene scelto quello con il supporto più elevato;
- a parità di lunghezza e supporto viene scelto l'itemset avente il *coefficiente di lift* più basso. Ad esempio, dato un carrello di spesa $basket = \{a, b, c, d\}$ e dati due pattern $pattern_1 = \{a, b, c\}$ e $pattern_2\{a, b, d\}$ aventi stessa lunghezza e stesso supporto, viene calcolato il coefficiente di lift di ciascun itemset come:

$$Lift(pattern_i) = \prod_{j=1}^{K} support(item_j) \tag{5.1}$$

  Successivamente viene selezionato il pattern avente il valore più basso, perchè maggiormente significativo. Supponendo dunque che i supporti dei singoli items siano $a = 0.3$, $b = 0.4$, $c = 0.5$, $d = 0.6$ si avrà che $lift_{pattern_1} = 0.06$ e $lift_{pattern_2} = 0.07$. L'itemset prescelto sarà il primo.

- a parità di lunghezza, supporto e coefficiente di lift, a ciascun pattern viene assegnata frequenza $1/N$ (con N il numero di pattern che hanno medesima lunghezza, supporto e lift).

5.1.2 Approfondimento Formule

Il **BRI** (*basket regularity index*) di un cliente x corrisponde all'entropia normalizzata calcolata su tutti i pattern frequenti supportati dai propri carrelli di spesa, e tale supporto costituirà la probabilità dell'evento.

Il calcolo del **BRI** viene ottenuto mediante la seguente formula:

$$BRI(x) = \frac{-\sum_{i=1}^{N}(\frac{frequency_i}{M} \cdot log_2(\frac{frequency_i}{M}))}{log_2(N)}, \tag{5.2}$$

dove x è un utente, M è il numero di spese e N è il numero di pattern frequenti trovati dall'algoritmo di *Frequent Pattern Mining*.

Il calcolo della media ponderata del numero di item all'interno di un pattern frequente viene ottenuto mediante la seguente formula:

$$Wavg(x) = \frac{\sum_{i=1}^{N} frequency_i \cdot length(pattern_i)}{N}, \tag{5.3}$$

dove x è un utente, N è il numero di pattern frequenti trovati dall'algoritmo di *Frequent Pattern Mining* e $length$ la lunghezza del pattern in termini di $items$ che lo compongono.

Nel caso in cui *Apriori* non trovasse pattern frequenti, le formule precedenti vengono sostituite dalle seguenti:

$$BRI(x) = 1.0, \tag{5.4}$$

$$Wavg(x) = \frac{\sum_{i=1}^{N} 1.0 \cdot length(pattern_i)}{N}, \tag{5.5}$$

dove N non rappresenta il numero di pattern frequenti ma il numero di spese effettuate dal cliente x.

5.1.3 Esempi di Applicazione del Basket Regularity Index

Verranno di seguito riportati alcuni esempi di applicazione del **BRI**. Il primo riguarda il calcolo dell'indice su un utente classificabile come *sistematico*.

$Basket$	$Items$
1	$\{a, b, c\}$
2	$\{a, b, c\}$
3	$\{a, b, c\}$
4	$\{a, b, c\}$
5	$\{a, b, c\}$

$Freq.Pattern$	$Supp.$
$\{a\}$	1.0
$\{b\}$	1.0
$\{c\}$	1.0
$\{a, b\}$	1.0
$\{a, c\}$	1.0
$\{b, c\}$	1.0
$\{a, b, c\}$	1.0

Tabella 5.1: Carrelli di un cliente sistematico e i pattern frequenti trovati.

Nella tabella 5.1 sono mostrate a sinistra le spese effettuate dal cliente. Ciascuna riga rappresenta un carrello con un insieme di prodotti. A destra sono mostrati gli itemset frequenti trovati dall'algoritmo di pattern mining; è stato scelto un supporto minimo il 50%. I pattern vengono poi ordinati secondo le politiche descritte precedentemente e infine vengono estratti gli itemset maggiormente rappresentativi (Tabella 5.2). In questo esempio

$SortedFreq.Pattern$	$Supp.$
$\{a, b, c\}$	1.0
$\{a, b\}$	1.0
$\{a, c\}$	1.0
$\{b, c\}$	1.0
$\{a\}$	1.0
$\{b\}$	1.0
$\{c\}$	1.0

$Frequentpatterns$	$Frequency.$
$\{a, b, c\}$	5.0

Tabella 5.2: Pattern frequenti riordinati e pattern frequenti candidati .

viene scelto il pattern più grande, dato che i supporti delle altre regole sono identici. Il carrello *tipico* dell'utente sarà dunque $\{a, b, c\}$. Il **BRI** risultante sarà:

$$BRI(x) = \frac{5}{5} \cdot \log_2(\frac{5}{5}) = 0.0 \tag{5.6}$$

Il consumatore in questione presenta entropia minima e dunque massima sistematicità: il sottoinsieme di prodotti acquistati in ciascuna spesa è sempre lo stesso.

Il secondo esempio riguarda il calcolo dell'indice su un utente classificabile come *casuale*.

$Basket$	$Items$
1	$\{a, b, c\}$
2	$\{d, e, f\}$
3	$\{g, h, i\}$
4	$\{j, k, l\}$
5	$\{m, n, o\}$

$Freq.Pattern$	$Supp.$
$\{a, b, c\}$	0.2
$\{d, e, f\}$	0.2
$\{g, h, i\}$	0.2
$\{j, k, l\}$	0.2
$\{m, n, o\}$	0.2

Tabella 5.3: Carrelli di un cliente casuale e i pattern frequenti trovati.

Nella tabella 5.3 sono mostrate a sinistra le spese effettuate dal cliente. Ciascuna riga rappresenta un carrello con un insieme di prodotti. Supponendo di selezionare come supporto minimo il 50%, l'algoritmo di pattern mining non trova itemset frequenti. Ciascun carrello di spesa verrà inserito dunque come pattern avente frequenza $\frac{1}{N}$ (tabella 5.3 a destra). La politica di ordinamento lascia i pattern invariati, così come l'insieme dei candidati (tabella 5.4). In questo esempio non esiste un carrello tipico, dato che i supporti delle regole sono identici. Il **BRI** risultante sarà:

$Freq.Pattern$	$Supp.$
$\{a, b, c\}$	0.2
$\{d, e, f\}$	0.2
$\{g, h, i\}$	0.2
$\{j, k, l\}$	0.2
$\{m, n, o\}$	0.2

$Freq.Pattern$	$Frequency$
$\{a, b, c\}$	1.0
$\{d, e, f\}$	1.0
$\{g, h, i\}$	1.0
$\{j, k, l\}$	1.0
$\{m, n, o\}$	1.0

Tabella 5.4: Pattern frequenti riordinati e pattern frequenti candidati .

$$BRI(x) = \frac{\frac{5}{5} \cdot \log_2(\frac{1}{5})}{log_2(5)} = 0.999 \tag{5.7}$$

Il consumatore in questione presenta entropia massima e dunque minima sistematicità: non esiste un sottoinsieme di prodotti acquistati caratteristico.

Un ultimo esempio riguarda il calcolo dell'indice su un utente classificabile come *standard*.

$Basket$	$Items$
1	$\{a, b, c, d, e\}$
2	$\{a, b, c\}$
3	$\{a, b, c, d\}$
4	$\{f, g, h\}$
5	$\{g, h, i\}$
6	$\{g, h, k\}$
7	$\{a, b, e\}$
8	$\{b, e, h\}$
9	$\{k, m, n\}$
10	$\{o\}$

$Freq.Pattern$	$Supp.$
$\{g, h\}$	0.3
$\{g\}$	0.3
$\{b, e\}$	0.3
$\{e\}$	0.3
$\{a, b, c\}$	0.3
$\{a, c\}$	0.3
$\{b, c\}$	0.3
$\{c\}$	0.3
$\{h\}$	0.4
$\{a, b\}$	0.4
$\{a\}$	0.4
$\{b\}$	0.5

Tabella 5.5: Carrelli di un cliente standard e i pattern frequenti trovati.

Nella tabella 5.5 sono mostrate a sinistra le spese effettuate dal cliente. Ciascuna riga rappresenta un carrello con un insieme di prodotti. A destra sono mostrati gli itemset frequenti trovati dall'algoritmo di pattern mining; è stato scelto un supporto minimo il 30%. I pattern vengono poi ordinati secondo le politiche descritte precedentemente e infine vengono estratti gli itemset maggiormente rappresentativi (Tabella 5.6). Per ciascuna spesa, l'algoritmo cerca il pattern col massimo fitting. Ad esempio il carrello

$\{a, b, c, d, e\}$ incrementerà la frequenza di 1 del solo pattern $\{a, b, c\}$. Lo stesso varrà per i carrelli $\{a, b, c\}$ e $\{a, b, c, d\}$. Il carrello $\{f, g, h\}$ incrementerà il supporto del pattern $\{g, h\}$ e così via. Va notato invece come il carrello $\{k, m, n\}$ non abbia un riscontro con nessun pattern della lista; per questa motivazione il carrello stesso verrà inserito come pattern avente frequenza $\frac{1}{10}$.

$Freq.Pattern$	$Supp.$
$\{a, b, c\}$	0.3
$\{a, b\}$	0.4
$\{g, h\}$	0.3
$\{b, e\}$	0.3
$\{a, c\}$	0.3
$\{b, c\}$	0.3
$\{b\}$	0.5
$\{h\}$	0.4
$\{a\}$	0.4
$\{g\}$	0.3
$\{e\}$	0.3
$\{c\}$	0.4

$Freq.Pattern$	$Frequency.$
$\{a, b, c\}$	3.0
$\{a, b\}$	1.0
$\{g, h\}$	3.0
$\{m, k, n\}$	1.0
$\{m, k, n\}$	1.0
$\{o\}$	1.0

Tabella 5.6: Pattern frequenti riordinati e pattern frequenti candidati .

Dopo aver trasformato il supporto in probabilità, viene calcolato il **BRI** che equivale a :

$$BRI(x) = \frac{(\frac{6}{10} \cdot \log_2(\frac{3}{10})) + (\frac{4}{10} \cdot \log_2(\frac{1}{10})}{log_2(6)} = 0.9172 \tag{5.8}$$

Dal valore si può capire come il cliente preso in esame non sia ritenibile un consumatore sistematico. Ciò nonostante esso è rappresentato da due carrelli tipici aventi un supporto del 30% ($\{a, b, c\}$ e $\{g, h\}$).

Va comunque sottolineato il fatto che il valore assoluto del **BRI**, preso singolarmente, non ha alcun tipo di significato. Esso deve essere utilizzato solo per capire la relazione di un cliente rispetto al comportamento collettivo.

5.2 Indice di sistematicità del comportamento spazio-temporale

Lo scopo dell'indice è la sintesi del comportamento di uno specifico consumatore da una prospettiva *spazio-temporale*. Ci si potrebbe aspettare

che un utente possa andare a fare acquisti in più luoghi comprando articoli differenti in ciascuno di essi. Ad esempio un consumatore potrebbe recarsi in un *Iper store* per fare le spese settimanali o mensili e comprare gli articoli maggiormente ingombranti (acqua, prodotti per la pulizia della casa, etc.) sfruttando un mezzo di locomozione; per fare le spese giornaliere (pane, affettati, frutta e verdura, etc.) potrebbe invece recarsi in un *GestIn store* che può raggiungere a piedi in breve tempo. Gli *Iper store* potrebbero inoltre essere frequentati il fine settimana nel primo pomeriggio, mentre i *GestIn store* potrebbero essere frequentati infrasettimanalmente e dopo l'orario lavorativo. Per tutte queste motivazioni occorre un indice che sintetizzi il comportamento spazio-temporale di un consumatore.

Per misurare la sistematicità di un individuo ci si è basati su tre variabili, una spaziale e due temporali:

- i *negozi* nei quali va a fare la spesa;
- i *giorni della settimana* oppure la *tipologia dei giorni* in cui effettua gli acquisti (giorni feriali o giorni festivi);
- lo *slot temporale* nel quale va a fare la spesa: può comprendere una o più *ore specifiche* della giornata, oppure una o più *fasce orarie*.

Ciascuna spesa effettuata da un cliente può essere associata ad ogni possibile combinazione delle variabili descritte precedentemente. La sintesi dell'indice di *sistematicità spazio-temporale* (**STRI** *spatio-temporal regularity index*) risulta molto più semplice da calcolare rispetto al **BRI** perché le classi sono già note: 2 valori per la tipologia di giorno (feriale o festivo), 7 valori per il giorno della settimana, 12 valori per l'ora ($[8 - 20]$) e 5 valori per la fascia oraria.

Come prima approssimazione, i passi che esegue l'algoritmo sono i seguenti:

- reperimento dal database delle seguenti informazioni (per ciascun utente):

Scontrino_ID	*Ora*	*Giorno*	*Negozio*
1	10 : 00	Lunedì	x
2	12 : 00	Martedì	y
...	...	...	...
n	18 : 00	Sabato	x

Tabella 5.7: Dati di uno specifico utente da passare in input all'algoritmo

- calcolo delle entropie su tutte le permutazioni di tutte le variabili spazio-temporali:

Entropia Singola	*Entropia Coppie*	*Entropia Triple*
Ora	*Ora-Giorno*	*FasciaOraria-TipoGiorno-Negozio*
Giorno	*FasciaOraria-Giorno*	*FasciaOraria-Giorno-Negozio*
Negozio	*Ora-TipoGiorno*	*Ora-Giorno-Negozio*
Tipo Giorno	*FasciaOraria-TipoGiorno*	
Fascia Oraria		

Tabella 5.8: Entropie spaziali e temporali calcolate dall'algoritmo

5.2.1 Pseudocodice del Spatio-Temporal Regularity Index

Viene di seguito riportato lo speudocodice dell'algoritmo per calcolo dell'entropia sul comportamento spazio-temporale di un insieme di clienti:

Algoritmo 5: STRI()

```
   Data: clients_list
   begin
1      Result = {};
2      for client in client_list do
3          baskets = load_st_baskets_info(client);
4          entropies_list = calculate_entropies(baskets);
5          Result.add(client, entropies_list);
       end
6      return Result
   end
```

L'algoritmo prende come input un insieme di clienti e restituisce per ciascuno di essi le 12 entropie sopra discusse. Per ciascun consumatore viene reperito un file all'interno del quale sono contenute, per ogni spesa, le variabili $\{ora, giorno, negozio\}$ (step 2). Successivamente vengono calcolate le entropie (step 3). Tutti i valori calcolati per ogni cliente verranno poi raccolti in un unico file per le analisi successive (step 6).

5.2.2 Approfondimento formule

Il calcolo dello **STRI** viene ottenuto mediante la seguente formula:

$$STRI(x) = \frac{-\sum_{i=1}^{N}(\frac{frequency_i}{M} \cdot log_2(\frac{frequency_i}{M}))}{log_2(N)}, \tag{5.9}$$

dove x è un utente, M è il numero di spese, N è numero di tuple e $frequency_i$ è la frequenza della tupla i-esima. Il numero di tuple N varia al

variare della combinazione di attributi utilizzati per il calcolo dell'entropia (uno dei 12 valori presenti in tabella 5.8). Ad esempio, se si utilizza come attributo il *giorno* e si trovano come tuple *{Lunedì}*, *{Martedì}*, *{Giovedì}* e *{Sabato}* allora $N = 4$ (anziché 7, numero dei giorni della settimana); se invece si utilizzano come attributi *fasciaOraria* e *tipoGiorno* e si trovano come tuple *{fascia1,feriale}* e *{fascia2,feriale}*, allora $N = 2$ (anziché 10 che é il prodotto tra le fasce orarie e le tipologie di giorno).

5.2.3 Esempi di Applicazione del Spatio-Temporal Regularity Index

Verranno di seguito riportati alcuni esempi di applicazione del **STRI**. Il primo riguarda il calcolo dell'indice su un utente classificabile come *sistematico*.

$FasciaOra$	$TipoGiorno$	$Neg.$
2	*Feriale*	*x*
2	*Feriale*	*x*
2	*Feriale*	*x*
2	*Feriale*	*x*
2	*Feriale*	*x*

$Triple$	$Frequency$
$\{2, Feriale, x\}$	5.0

Tabella 5.9: Rappresentazione tabellare della fasciaOraria, tipoGiorno e Negozio in cui un cliente sistematico compie i propri acquisti.

Nella tabella 5.9 sono mostrate a sinistra le informazioni spazio-temporali su ciascuna spesa effettuate da un cliente; nella tabella a destra è mostrata invece l'unica tripla presente con la rispettiva frequenza. Lo **STRI** risultante sarà:

$$STRI(x) = \frac{5}{5} \cdot \log_2(\frac{5}{5}) = 0.0 \tag{5.10}$$

Il consumatore in questione presenta entropia minima e dunque massima sistematicità: ciascuna spesa viene effettuata la stessa tipologia di giorno (feriale), nella stessa fascia oraria (10-12)e nello stesso negozio.

Il secondo esempio riguarda il calcolo dell'indice su un utente classificabile come come *casuale*.

Nella tabella 5.10 sono mostrate a sinistra le informazioni spazio-temporali su ciascuna spesa effettuate da un cliente; nella tabella a destra sono mostrate invece le triple con le rispettive frequenze (in questo caso il numero di triple e il numero di spese coincide). Lo **STRI** risultante sarà:

$FasciaOra$	$TipoGiorno$	$Neg.$
1	*Feriale*	x
2	*Feriale*	x
3	*Feriale*	x
4	*Feriale*	x
5	*Feriale*	x

$Triple$	$Frequency$
$\{1, Feriale, x\}$	1.0
$\{2, Feriale, x\}$	1.0
$\{3, Feriale, x\}$	1.0
$\{4, Feriale, x\}$	1.0
$\{5, Feriale, x\}$	1.0

Tabella 5.10: Rappresentazione tabellare della fasciaOraria, tipoGiorno e Negozio in cui un cliente casuale compie i propri acquisti.

$$STRI(x) = \frac{\frac{5}{5} \cdot \log_2(\frac{1}{5})}{log_2(5)} = 0.999 \tag{5.11}$$

Il consumatore in questione presenta entropia massima e dunque minima sistematicità: nonostante il cliente utilizzi sempre lo stesso negozio sempre nei giorni *feriali*, varia sempre la fascia oraria. Il suo comportamento di acquisto spazio-temporale è considerato pertanto casuale.

Un ultimo esempio riguarda il calcolo dell'indice su un utente classificabile come *standard*.

$FasciaOra$	$TipoGiorno$	$Neg.$
2	*Feriale*	x
2	*Feriale*	x
4	*Feriale*	x
2	*Feriale*	x
2	*Feriale*	x
4	*Feriale*	x
4	*Festivo*	x
2	*Feriale*	x
2	*Feriale*	x
2	*Feriale*	x

$Pattern$	$Frequency$
$\{2, Feriale, x\}$	7.0
$\{4, Feriale, x\}$	2.0
$\{4, Festivo, x\}$	1.0

Tabella 5.11: Rappresentazione tabellare della fasciaOraria, tipoGiorno e Negozio in cui un cliente standard compie i propri acquisti.

Come è visibile in tabella 5.11 le triple del consumatore sono 3. La frequenza maggiore corrisponde alla tripla $\{2, Feriale, x\}$ che riassume il comportamento spazio-temporale del soggetto preso in analisi. Lo **STRI** risultante sarà:

$$STRI(x) = \frac{\frac{7}{10} \cdot \log_2(\frac{7}{10}) + \frac{2}{10} \cdot \log_2(\frac{2}{10}) + \frac{1}{10} \cdot \log_2(\frac{1}{10})}{log_2(3)} = 0.7293 \tag{5.12}$$

5.3 Ipotesi sulle distribuzioni degli indici BRI e STRI

Fino ad ora abbiamo discusso su come calcolare l'*indice di sistematicità degli acquisti* (**BRI**) e l'*indice di sistematicità spazio-temporale* (**STRI**) per ciascun cliente. Entrambi gli indici possono essere misurati con differenti livelli di dettaglio. Ad esempio il **BRI** può essere calcolato a livello di singolo articolo acquistato in ciascuna spesa, oppure a livello di codice di marketing (la marca o la categoria a cui appartiene un prodotto). Anche per il **STRI** possono essere svolte diverse analisi: solo spaziali (negozio), solo temporali (giorno della settimana, tipo di giorno, ora, fascia oraria), oppure loro aggregati.

Essendo la nostra un'analisi di tipo sperimentale, verranno analizzate tutte le entropie al fine di trovare correlazioni tra di esse e scartare così le misure ridondanti. L'intento finale rimane quello di poter descrivere ciascun utente secondo due aspetti: il comportamento di acquisto e il comportamento spazio-temporale. Di tutte le misure calcolate, pertanto, verranno utilizzate quelle meglio distribuite: una misura per tipologia di comportamento.

Ci si aspetta che le distribuzioni dei clienti su un valore entropico specifico assumano un andamento di tipo *lognormal*. Dagli esempi sul calcolo del **BRI** e del **STRI** appare evidente che valori prossimi allo 0.0 o all'1.0 (estremi dell'intervallo) apparterranno a un numero piccolo di utenti, mentre valori prossimi alla media o mediana apparterranno alla stragrande maggioranza della popolazione. Si suppone inoltre che il valore medio assunto dai consumatori sia alto perché l'intervallo temporale di riferimento che si intente considerare è di un anno. In tale arco temporale ci potranno infatti essere influenze stagionali, sia sul comportamento di spesa che sul comportamento spazio-temporale che potrebbero innalzare il valore delle entropie. In figura 5.1 è mostrato l'andamento probabilistico atteso dell'entropia in base alle osservazioni precedenti.

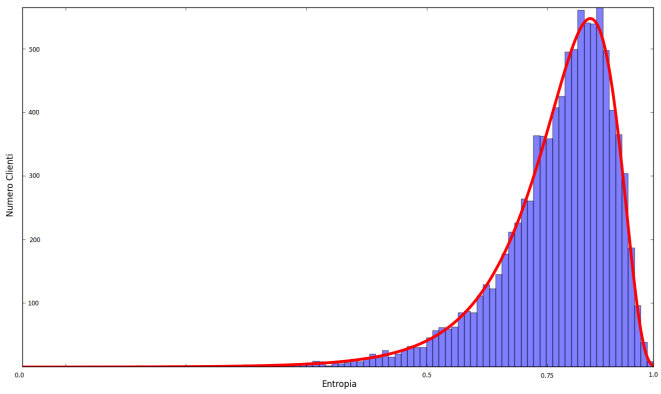


Fig. 5.1: Distribuzione ipotetica dell'entropia su un insieme di clienti

Una volta calcolate tutte le entropie, il passo successivo consiste nella segmentazione dei consumatori in gruppi omogenei che abbiano un profilo collettivo. Nel capitolo successivo verranno mostrati i risultati dell'applicazione del metodo. In particolare verranno discussi due tipologie di analisi:

- una *descrittiva* basata sul *clustering* con successivi riscontri economici per ciascun cluster trovato ;
- una *applicativa* basata sullo studio delle *distribuzioni probabilistiche* delle entropie con successiva estrazione degli articoli tipici acquistati da gruppi omogenei di consumatori.

Capitolo 6

Risultati

Una volta definiti i metodi per ricavare le misure che descrivono il comportamento di spesa dei clienti, questi sono stati applicati ad un dataset reale di vendite al dettaglio. Il *datawarehouse* in questione è quello dell'**UNICOOP TIRRENO** descritto nel capitolo 4.

Sono stati svolti due tipi di esperimenti:

- una *segmentazione* della clientela mediante clustering (sulle distribuzioni del **BRI** e dello **STRI**) col fine di trovare le caratteristiche distintive di ciascun gruppo di consumatori in termini di parametri monetari;
- una *segmentazione* della clientela mediante tagli sulle distribuzioni del **BRI** e dello **STRI** col fine di individuare gli articoli distintivi acquistati da gruppi omogenei di consumatori.

6.1 Sottoinsieme di dati utilizzato

Il *datawarehouse* presenta una granularità fine. La tabella dei fatti possiede un record per ogni articolo acquistato. Il numero totale di record è 3.355.323.532. Nonostante la base di dati sia correttamente indicizzata, anche le interrogazioni elementari risultano dispendiose in termini di tempo. Ad esempio, per sapere il contenuto di una singola spesa di un singolo cliente è necessario eseguire un raggruppamento degli articoli che appartengono allo stesso scontrino. Applicando questa interrogazione a tutti i clienti di tutti i negozi della **COOP** i risultati arriverebbero dopo giorni.

Per semplificare la sintesi del metodo dello studio del comportamento di spesa dei clienti si è pertanto deciso di studiare un sottoinsieme dei dati in un arco di tempo ristretto.

Più precisamente l'analisi ha riguardato:

- il **Venduto** dell'anno 2012;
- i **Negozi** appartenenti alla Provincia di Livorno;

- gli **Articoli** provvisti di *codice di marketing*.

E' stata pertanto creata una tabella, con i filtri sopra discussi, mediante la seguente interrogazione:

```
create table PROVINCIA_LI_2012 as
  SELECT * FROM VENDUTO
  WHERE DATA_ID BETWEEN 2557 AND 2922
  AND CLIENTE_ID  > 0
  AND IMPORTO > 0
  AND COD_MKT_ID IS NOT NULL;
```

Infine, per quanto riguarda i clienti, sono stati filtrati quelli che hanno effettuato meno di una spesa al mese nel periodo di riferimento. Anche questa scelta è stata dettata dal fatto che lo studio è incentrato sulle spese dei consumatori. E' stata così creata una tabella ausiliaria che raccoglie gli *id* dei clienti considerati idonei:

```
create table IDONEI as
  select cliente_id, count(*)
  from(
      select  prov.cliente_id, d.MESE_N
      from  DATA d, PROVINCIA_LI_2012 prov
      where prov.DATA_ID = d.DATA_ID
      group by prov.cliente_id, d.MESE_N
   )
  group by cliente_id
  having count(*) = 12
   order by cliente_id;
```

Il sottoinsieme di dati ha una dimensionalità notevolmente ridotta:

- 71.172.672 record totali;
- 56.448 utenti attivi;
- 84.362 articoli distinti;
- 23 negozi dei quali un *Iper*, 9 *Super*, e 13 *GestIn*.

Sull'insieme di dati target sono state svolte alcune analisi preliminari. La prima ha riguardato il numero totale di scontrini emessi nei 23 negozi della Provincia di Livorno. Come è visibile nella Figura 6.1, l'unico *Iper Store* presente arriva a quasi un milione di scontrini battuti. I *Super Store* hanno una media di circa 458.200 scontrini ed infine i *GestIn Store* hanno una media di 83.000 scontrini.

Questi risultati confermano solamente che il numero di spese effettuate nei negozi è direttamente proporzionale alle loro dimensioni e probabilmente al diverso assortimento di prodotti presenti in ciascuno di essi.

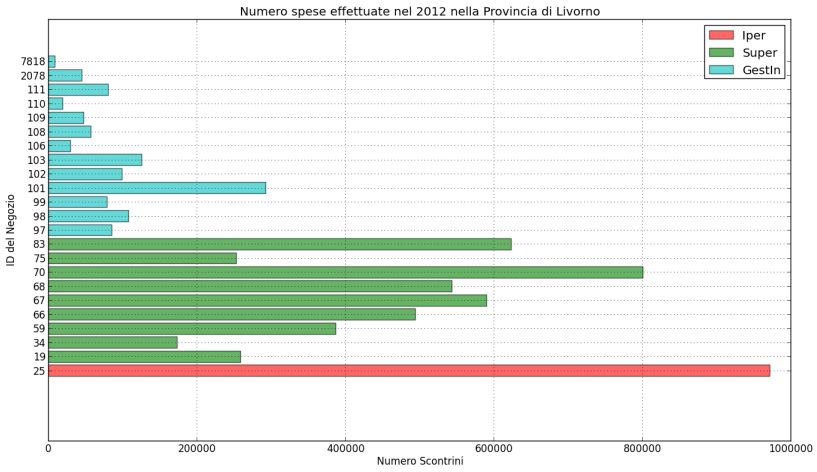


Fig. 6.1: Istogramma del numero di spese effettuate nel 2012 negli Store della Provincia di Livorno

Altre due analisi hanno invece riguardato il comportamento di acquisto dei clienti, in termini di numero di spese totali effettuate nel 2012 e le composizioni dei singoli carrelli di spesa, in termini di articoli distinti acquistati. Le rispettive distribuzioni sono mostrate in Figura 6.2 e in Figura 6.3.

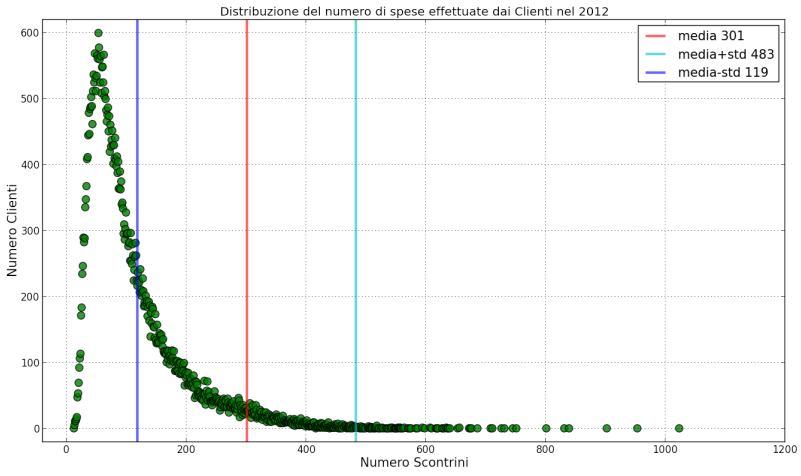


Fig. 6.2: Distribuzione del numero di spese effettuate dai clienti nel 2012.

Entrambe le distribuzioni ricordano l'andamento della distribuzione probabilistica *Log Normale*. In Figura 6.2 è evidenziata in rosso la media delle spese effettuate dai clienti che è pari a 301. Inoltre, la maggior parte dei clienti

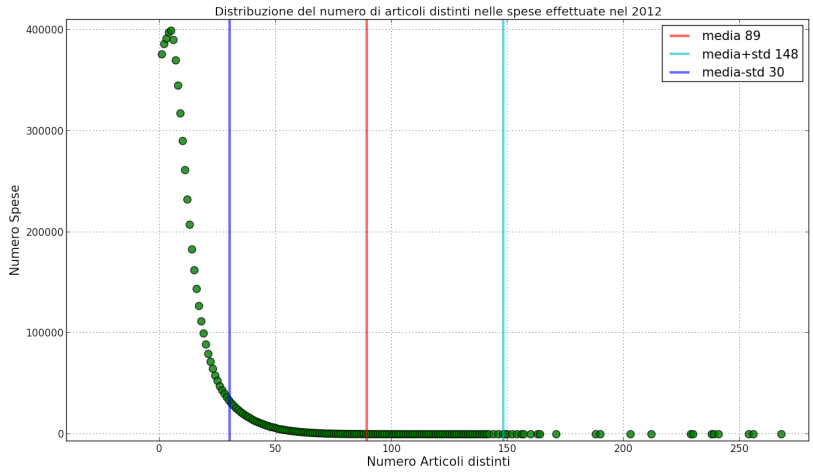


Fig. 6.3: Distribuzione del numero di articoli distinti per scontrino.

è situata nell'intervallo compreso tra 119 (in blu) e 482 (in celeste), valori ricavati sommando/sottraendo la deviazione standard al valore medio.

Per quanto riguarda la distribuzione del numero distinto di articoli nelle singole spese (Figura 6.3) si osserva come il numero medio sia 89, mentre il range identificato dalla deviazione standard è compreso tra 30 e 148 articoli distinti. Questo tipo di informazione è poco interessante perché il numero di scontrini appartenenti all'intervallo è un numero piuttosto piccolo. La distribuzione risulta sbilanciata perché esiste un numero ristretto di spese con un elevato numero di articoli acquistati.

Molto più interessante è la campana della funzione che mostra come la maggior parte delle spese abbia un numero di articoli distinti che va da un minimo di 5 ad un massimo di 20.

Un'ultima analisi ha riguardato la dimensione temporale delle spese effettuate nel 2012 in tutti i negozi aderenti all'**Unicoop Tirreno**. In particolare si sono analizzate:

- la distribuzione delle spese nell'arco di una giornata per identificare le fasce orarie all'interno delle quali avvengono i picchi di spesa;
- la distribuzione delle spese nell'arco della settimana per identificare i giorni nei quali avviene il maggior numero di spese.

Come è evidente dalla figura 6.4, si notano cinque fasce orarie ben definite: due picchi di spesa si verificano nelle fasce (10-12) e (16-18), mentre negli intervalli (7-10), (12-14) e (18-20:30) vi è un calo di acquisti.

Per quanto riguarda la distribuzione settimanale (6.5), si nota come l'andamento generale delle vendite sia lineare dal Lunedì al Martedì, sia

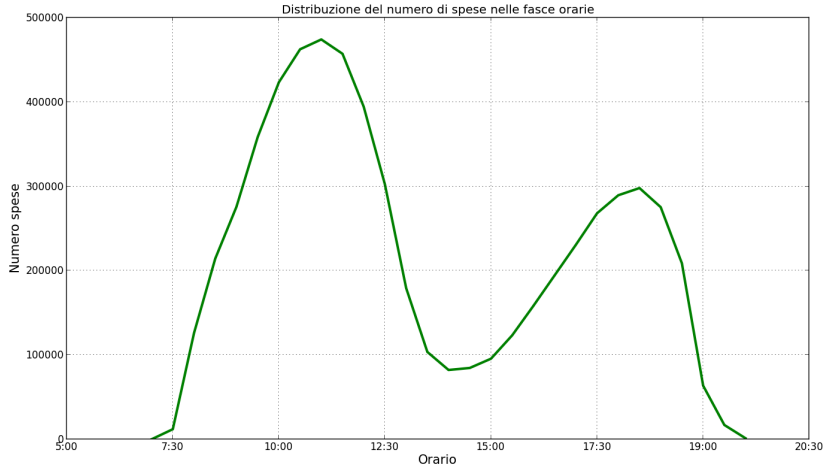
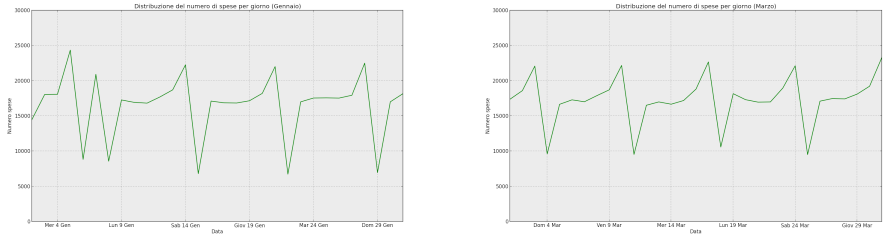


Fig. 6.4: Distribuzione delle spese nelle fasce orarie.



(a) Gennaio (b) Marzo

Fig. 6.5: Distribuzione delle spese nei giorni della settimana (per mese).

crescente dal Mercoledì al Sabato, mentre decresce la Domenica. Questo tipo di andamento è costante in tutti i mesi; cambia solo il numero di spese effettuate. E' da sottolineare come il picco di vendite, sia positivo che negativo, dipenda dalla prossimità delle festività. Sono stati riportati, per brevità, soltanto i grafici del mese di Gennaio e Marzo.

6.2 Calcolo indice BRI

Il calcolo del *Basket Regularity Index*, per la sintesi del comportamento di acquisto di un cliente, è stato sviluppato a due distinti livelli di granularità:

- a livello *articolo*: un carrello tipico è descritto dai singoli articoli ad esso appartenenti;

- a livello *segmento*: un carrello tipico è descritto da gruppi di articoli appartenenti allo stesso segmento (ad esempio *Coca Cola* e *Pepsi* risultano un unico prodotto appartenente al segmento *Cola Drinks*).

Il primo problema affrontato è stato la scelta del supporto minimo da passare come input all'algoritmo. Ci sono infatti dei pro e dei contro per ogni possibile valore assegnato. All'aumentare del supporto diminuisce la probabilità di trovare pattern frequenti significativi, ma i tempi di computazione si riducono notevolmente. Viceversa, al diminuire progressivo del supporto, aumenta la probabilità di trovare pattern significativi, ma aumenta esponenzialmente sia il numero di pattern frequenti banali (di lunghezza 1), sia il tempo di computazione.

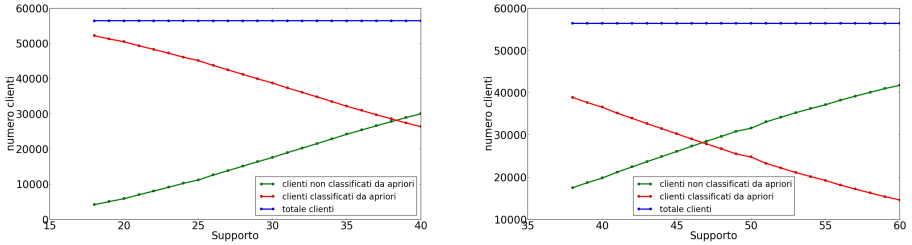
Sono state pertanto lanciate N istanze dell'algoritmo facendo variare il supporto al fine di determinare il valore ottimale. Per quanto riguarda il livello *articolo* il supporto è stato fatto variare nel range [18 - 40], mentre per il livello *segmento* è stato fatto variare nel range [38 - 60]. I parametri sui quali ci si è basati sono stati:

- numero di clienti per i quali *Apriori* non trova pattern frequenti;
- tempo di esecuzione totale per tutti i clienti;
- tempo di elaborazione per singolo cliente;
- distribuzione del **BRI**;
- numero medio di articoli distinti per pattern frequente;

La scelta di un supporto minimo nel range si è resa necessaria per motivi di complessità computazionale. Esiste infatti un sottoinsieme di clienti che, nell'arco temporale di un anno, compie un numero modesto di spese ma, per ciascuna di esse acquista un numero elevato di articoli tutti distinti tra loro da una spesa all'altra. Per questa tipologia di consumatori *Apriori* trova un numero elevato di pattern frequenti (milioni di insiemi) che successivamente devono essere associati alle singole spese per determinare i *carrelli tipici*. L'operazione in questione ($select\_best\_fitting\_subset()$, discussa nel capitolo 5) richiede una quantità di risorse eccessive in termini di memoria e di tempo rispetto all'architettura utilizzata per eseguire i test.

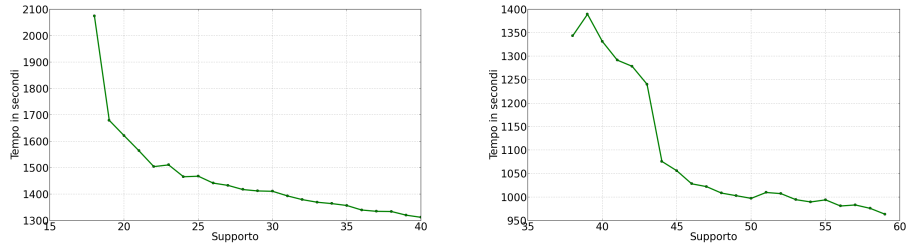
In figura 6.6 è mostrata la variazione del numero di utenti per i quali l'algoritmo *Apriori* trova pattern frequenti al variare del supporto. Sia a livello *articolo* che al livello *segmento* il numero di utenti diminuisce all'aumentare del supporto. Nel primo caso (figura 6.6(a)) il numero di clienti che non hanno pattern frequenti decresce lentamente, mentre nel secondo caso (figura 6.6(b)) decresce velocemente.

In figura 6.7 è mostrata la variazione dei tempi di esecuzione dell'algoritmo al variare del supporto. Mentre a livello *articolo* (figura 6.7(a)) il



(a) articolo (b) segmento

Fig. 6.6: Numero di utenti classificati da Apriori al variare del supporto.

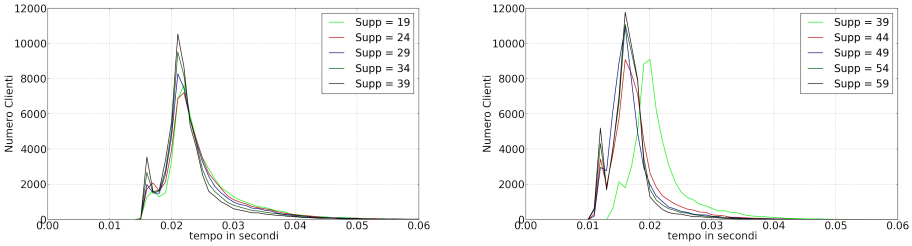


(a) articolo (b) segmento

Fig. 6.7: Variazione dei tempi di esecuzione dell'algoritmo al variare del supporto.

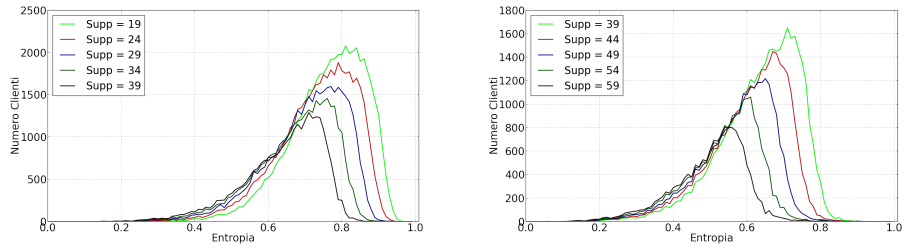
tempo di esecuzione diminuisce notevolmente a partire dal supporto 24, a livello *segmento* (figura 6.7(b)) il tempo è sempre ottimale per qualsiasi supporto si utilizzi. Per quanto riguarda invece la variazione dei tempi di esecuzione per cliente (figura 6.8), a livello *articolo* il tempo medio è sempre 0.022 a prescindere dal supporto utilizzato (figura 6.8(a)), mentre a livello *segmento* il tempo medio è sensibilmente più basso (0.015) eccetto che per il supporto minimo, in corrispondenza del quale è pari a 0.02 (figura 6.8(b)).

Le distribuzioni del *Basket Regularity Index*, al variare del supporto, sono mostrate in figura 6.9. Apparentemente sembrerebbe che la distribuzione del **BRI** a livello *segmento* sia migliore rispetto che al livello *articolo* perchè la maggior parte dei valori assunti dall'entropia sono lontani dall'aleatorietà (il limite destro è 0.9). In realtà questa tendenza è dettata dal supporto utilizzato. I valori delle entropie assunte da ciascun consumatore risultano più bassi rispetto al livello *articolo* perché i pattern frequenti trovati hanno un supporto più elevato. Per la maggior parte dei clienti, infatti, non esistono *carrelli tipici* e la loro entropia sarà quella massima (1.0). Pertanto la distribuzione maggiormente rappresentativa risulta quella in figura 6.9(a). Con la granularità a livello *articolo*, il numero di clienti per cui *Apriori* non trova pattern frequenti è sensibilmente inferiore.



(a) articolo (b) segmento

Fig. 6.8: Variazione del tempo di esecuzione dell'algoritmo per cliente al variare del supporto.



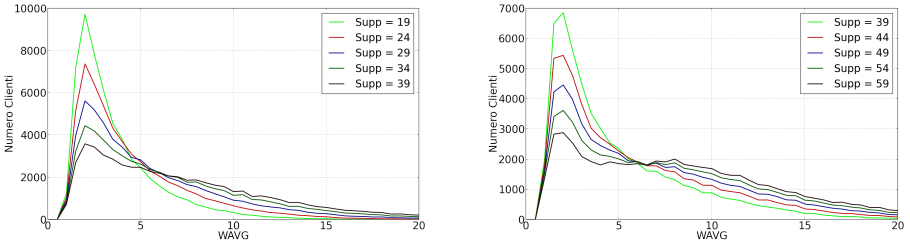
(a) articolo (b) segmento

Fig. 6.9: Variazione delle distribuzioni del BRI al variare del supporto.

In figura 6.10 viene invece mostrata la distribuzione del numero medio di articoli presenti nei pattern frequenti al variare del supporto fornito all'algoritmo. Si nota come a livello *articolo* la maggior parte dei clienti abbia *carrelli tipici* di lunghezza compresa nel range $[2, 5]$, mentre a livello *segmento* è compresa nel range $[2, 10]$. Anche in questo caso valgono le considerazioni fatte sulla distribuzione del **BRI**. In figura 6.10(b) si trovano pattern frequenti mediamente più lunghi, ma su un numero ristretto di utenti.

Per tutte le osservazioni sopra discusse si è scelta come distribuzione più rappresentativa quella riguardante il calcolo del **BRI** utilizzando come granularità il livello *articolo* e come supporto 24. Il tempo di elaborazione totale, per tutti i clienti, è di circa 20 minuti e la percentuale degli utenti per i quali *Apriori* non trova pattern frequenti è del 12%. La distribuzione prescelta del *Basket Regularity Index* è mostrata in figura 6.11.

E' da notare come l'indice di sintesi del comportamento di acquisto realizzato dia una maggiore informazione rispetto all'utilizzo convenzionale dell'entropia. Eliminando infatti la dimensione del *carrello di spesa* e considerando ogni *articolo* (o *segmento*) come distinta classe dell'entropia, l'indice assumerebbe la seguente forma:



(a) articolo (b) segmento

Fig. 6.10: Media articoli per pattern frequente al variare del supporto.

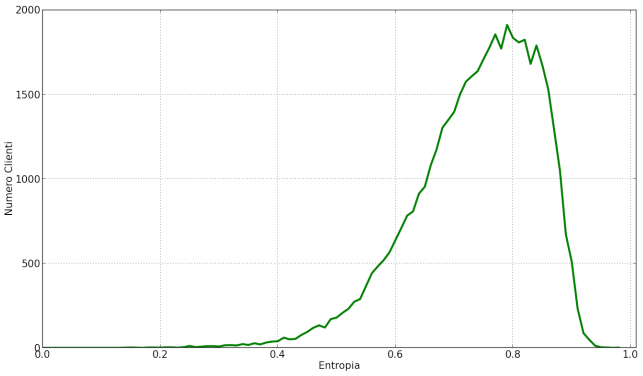


Fig. 6.11: Distribuzione del Basket Regularity Index con supporto 24

$$Entropy(x) = \frac{-\sum_{i=1}^{N}(\frac{frequency_i}{M} \cdot log_2(\frac{frequency_i}{M}))}{log_2(N)}, \tag{6.1}$$

dove x è un utente, M è il numero totale di articoli acquistati, N è il numero di articoli distinti (o classi) e $frequency_i$ è il numero di volte che l'articolo i è stato acquistato in un anno. La distribuzione dell'entropia è mostrata in figura 6.12:

I valori assunti dalla quasi totalità dei consumatori sono tutti compresi nell'intervallo $[0.8, 0.98]$. Tale intervallo è troppo piccolo per poter effettuare una segmentazione della clientela. L'indice da noi sintetizzato, invece, ha un range molto più grande ($[0.2, 0.99]$) nel quale operare una distinzione dei comportamenti di acquisto tipici di gruppi di consumatori.

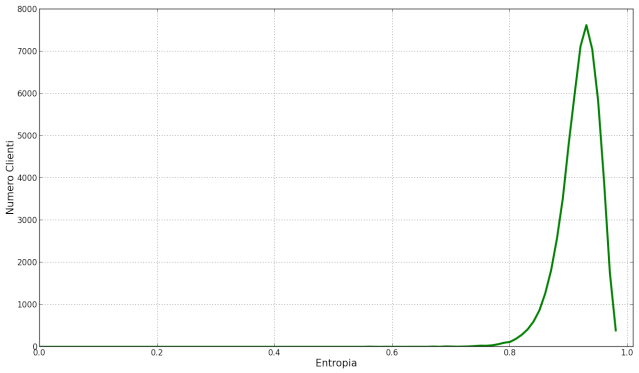


Fig. 6.12: Distribuzione entropia di acquisto rispetto agli articoli.

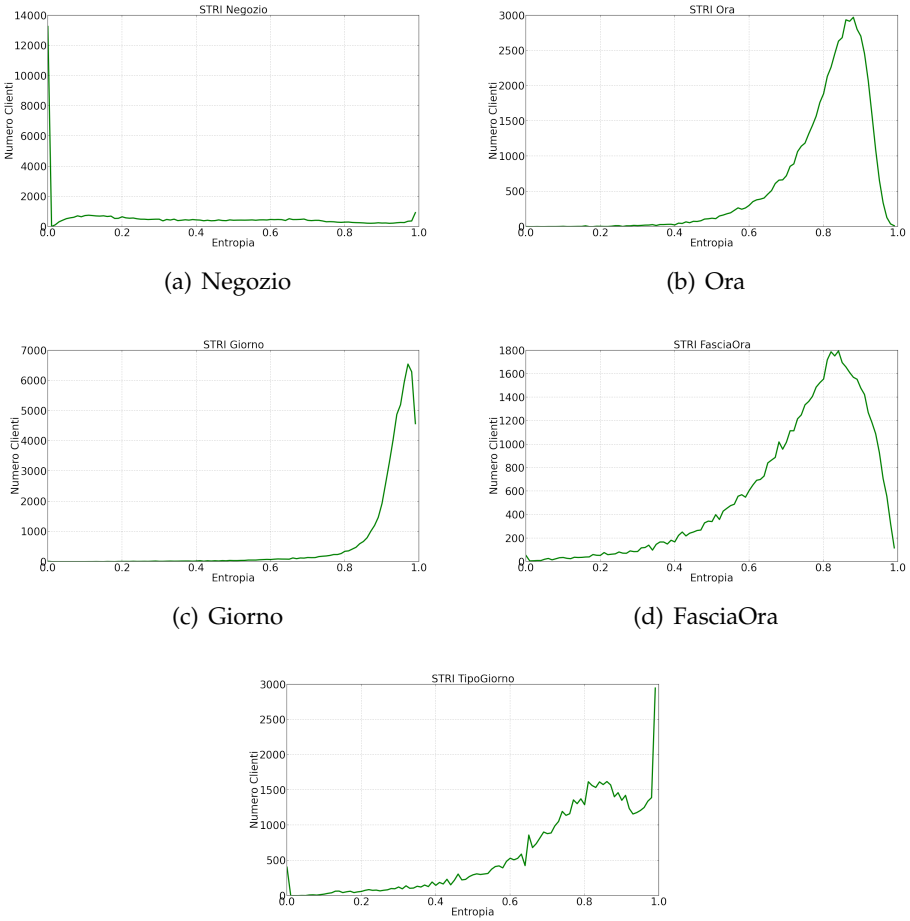
6.3 Calcolo indice STRI

Il calcolo dello *Spatio-Temporal Regularity Index*, per la sintesi del comportamento di acquisto spazio-temporale di un cliente, è stato sviluppato studiando inizialmente le distribuzioni delle entropie riguardanti le dimensioni spaziali e temporali singole. Esse sono: *Negozio*, *Ora*, *Giorno*, *FasciaOra* e *TipoGiorno* (figura 6.13).

Si nota come le distribuzioni più promettenti, dal punto di vista temporale, siano quelle riguardanti le dimensioni *Ora*, *FasciaOra* e *TipoGiorno*. La distribuzione riguardante il *Giorno* della settimana non è particolarmente significativa. Questo risultato indica che non esiste un numero rilevante di consumatori che vanno a fare la spesa sempre lo stesso giorno della settimana. Di conseguenza, un indice di sistematicità basato su questa dimensione, non offrirà alcun risultato rilevante.

Per quanto riguarda invece la dimensione spaziale *Negozio* (figura 6.13(a)), si nota dal grafico come la maggior parte dei consumatori abbia un valore prossimo allo zero. Questo risultato mostra come gli utenti della **COOP** siano legati ad un singolo punto vendita.

Poiché il nostro intento è quello di creare un unico indice di sistematicità spazio-temporale per ciascun consumatore, si è calcolato lo *Spatio-Temporal Regularity Index* sulle dimensioni *FasciaOra*, *TipoGiorno* e *Negozio*. E' stata pertanto esclusa dalle analisi la dimensione *Ora*, perchè meno significativa rispetto alla dimensione *FasciaOra*. La distribuzione dell'indice è mostrata in figura 6.14.



(a) Negozio (b) Ora

(c) Giorno (d) FasciaOra

(e) TipoGiorno

Fig. 6.13: Distribuzioni delle entropie singole

6.4 Segmentazione mediante Clustering (K-Means)

La prima analisi sugli indici di sistematicità sul *comportamento di acquisto* (**BRI**) e sul *comportamento spazio-temporale* (**STRI**) ha previsto due fasi:

- il *clustering* bi-dimensionale sugli indici sopra citati col fine di dividere i consumatori in gruppi omogenei;
- un'analisi di tipo *descrittivo* per comprendere il comportamento di spesa di ciascun gruppo in termini di parametri economico-monetari.

La tecnica del clustering si è resa indispensabile per trovare un numero di gruppi di consumatori che andasse oltre i preconcetti umani; poiché le misure usate sono entropie con distribuzione nota (distribuzione *lognormal*), verrebbe intuitivo dire che i gruppi di consumatori siano banalmente tre:

Fig. 6.14: Distribuzione STRI sulle dimensioni FasciaOra, TipoGiorno, Negozio

- *sistematici:* hanno un comportamento di acquisto ripetitivo sia sul profilo degli articoli acquistati che su quello spazio-temporale (stesso luogo, stessa fascia oraria, stessa tipologia di giorno);
- *casuali:* hanno un comportamento imprevedibile sotto ogni aspetto; per loro non è possibile prevedere *dove*, *quando* oppure *cosa* acquisteranno nel futuro;
- *standard:* hanno un comportamento non classificabile ne come *sistematico*, ne come *casuale*.

In realtà, potrebbero comparire anche altri gruppi di consumatori non osservabili in modo diretto. Ad esempio potrebbero esistere persone sistematiche nel solo comportamento spazio-temporale, ma casuali nel comportamento di acquisto (o viceversa).

Il clustering è stato effettuato mediante l'algoritmo **K-means**. Come input è stata passata una tabella contenente, per ciascun utente, i due indici di sistematicità, e il parametro k rappresentante il numero di cluster da trovare. La scelta del parametro k ottimale viene effettuata mediante lo studio dell' **SSE** (*sum of squared error*):

- vengono eseguite n istanze di **K-means**, con $k = \{1, \ldots, n\}$;
- viene costruita la funzione dell'**SSE** al variare di k;
- viene calcolata la derivata sulla funzione dell'**SSE** in corrispondenza di ogni k e viene scelto come valore ottimale quello la cui pendenza risulta maggiore.

In tabella 6.1 sono mostrati i centroidi del cluster trovati dall'algoritmo **K-MEANS** utilizzando come parametro ottimale $k = 5$; la figura 6.15 mostra invece la funzione dell'**SSE** al variare di k nell'intervallo $[1, \ldots, 12]$.

Attribute	*Full Data*	0	1	2	3	4
	56448	11040	9775	19510	12244	3879
BRI	0.7846	0.7866	0.9943	0.8039	0.6169	0.6824
stdDev	0.1353	0.0538	0.0222	0.0539	0.077	0.1078
STRI	0.7791	0.7085	0.7989	0.8548	0.7907	0.5134
stdDev	0.1105	0.0556	0.0865	0.043	0.0659	0.1022

Tabella 6.1: Centroidi dei Cluster con $k = 5$ sulle dimensioni BRI e STRI

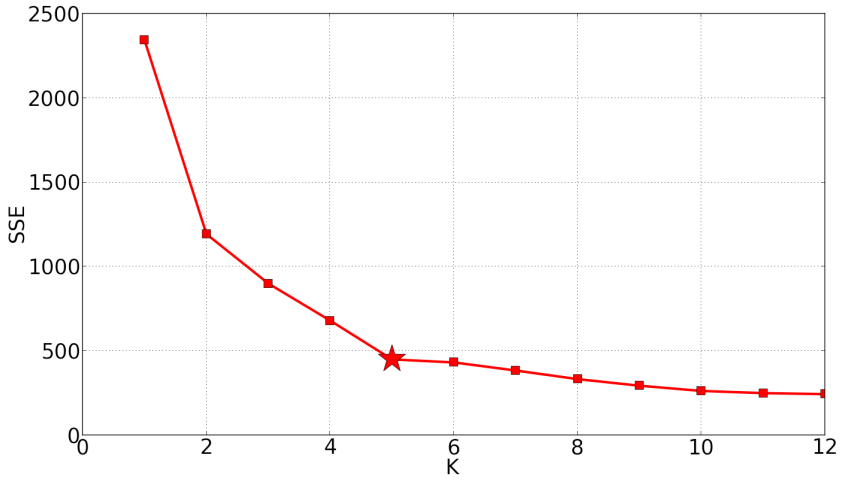


Fig. 6.15: Funzione dell'SSE al variare del parametro k

Si nota come, per $k = [2, \ldots, 5]$, l'**SSE** diminuisce in maniera costante mentre da $k = [6, \ldots, 12]$ l'errore non diminuisce in modo significativo. Con $k = 5$ si ha l'ultimo decremento positivo per la riduzione dell'**SSE**.

In figura 6.16 viene mostrato l'esito del clustering. Ciascun colore rappresenta uno dei 5 cluster trovati. Ciascun punto rappresenta un utente in termini di **BRI** e **STRI**. Lo scatter plot rispecchia gli andamenti delle singole distribuzioni sul *comportamento di acquisto* (figura 6.11) e sul *comportamento spazio temporale* (figura 6.14). Infatti la maggior parte dei consumatori si concentra dopo il valore 0.6 per entrambe le coordinate cartesiane.

Sebbene i cluster siano schiacciati verso valori alti dell'entropia, essi risultano ben definiti. Inoltre, il cluster con la maggior percentuale di utenti (il blu col 34.6 % degli utenti) risulta molto denso: tutti i consumatori risiedono in un'area molto piccola. Al contrario, il cluster col minor numero di utenti al suo interno (l'arancione col 6.9% degli utenti) è sparso: ci sono pochi utenti su un'area più vasta.

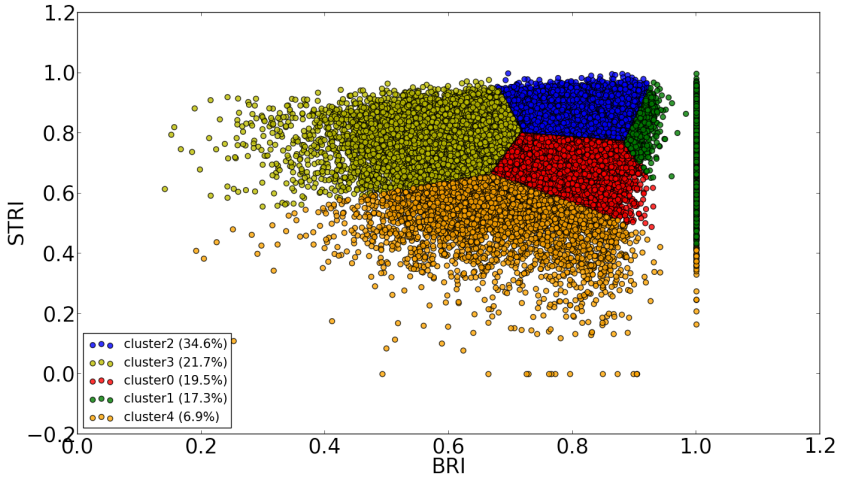


Fig. 6.16: Scatter plot della distribuzione degli utenti in ciascun Cluster

Per avere una migliore interpretazione di ciascun gruppo trovato, è stato creato uno scatter plot che mostra i centroidi normalizzati di ciascun cluster.

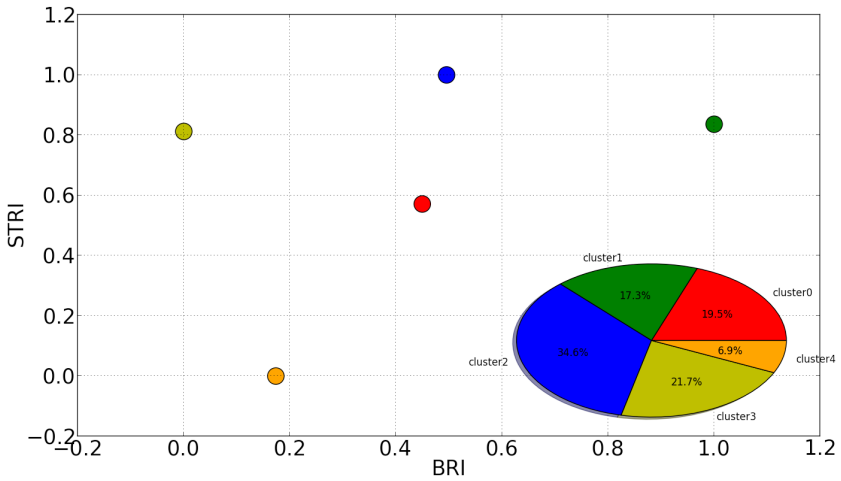


Fig. 6.17: Scatter plot dei centroidi (normalizzati) di ciascun Cluster

Come è visibile dalla figura 6.17 possiamo dividere la clientela nel seguente modo:

- *bi-sistematici:* (cluster arancione) costituiscono il 6.9% dei consumatori. Essi sono ripetitivi sia per quanto riguarda gli articoli acquistati, che per i criteri spazio-temporali che adotta (luogo, fascia oraria, tipologia di giorno);

- *sistematici sugli acquisti:* (cluster giallo) costituiscono il 21.7% dei consumatori. Essi risultano ripetitivi solamente sugli articoli acquistati, ma non mostrano sistematicità sulla dimensione spazio-temporale;
- *bi-casuali:* (cluster verde) costituisco il 17.3% dei consumatori. Essi non mostrano alcun criterio nei loro acquisti sia per la composizione dei carrelli, sia per la fascia oraria, il giorno, e il negozio in cui si recano a fare la spesa;
- *casuali sugli acquisti:* (cluster blu) costituiscono il 34.6% dei consumatori. Essi mostrano aleatorietà sulla dimensione spazio-temporale, mentre presentano un qualche criterio di acquisto;
- *standard:* (cluster rosso) costituiscono il 19.5% dei consumatori. Sono coloro che mostrano dei criteri di spesa sia su ciò che comprano che sul dove e quando. Il loro valore dell'entropia è troppo elevato affinché possano essere classificati come sistematici.

La precedente segmentazione della clientela è stata visualizzata anche mediante i box plot mostrati in figura 6.18. Ciascun rettangolo mostra i valori delle entropie assunti dai cluster nell'intervallo tra il primo e il terzo quartile, con evidenziata all'interno la mediana.

Non sono state applicate altre tecniche di clustering con k determinabile in modo automatico (come nel **DBSCAN**) perché molti utenti sarebbero stati classificati come *rumore* e pertanto esclusi dalle analisi successive. Il nostro scopo è invece inserire ciascun consumatore all'interno di un gruppo senza tralasciare nessun utente.

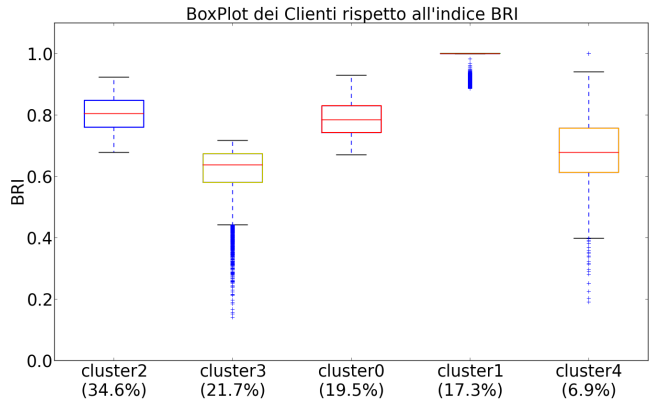
Per quanto riguarda invece l'analisi di tipo *descrittivo* mediante parametri economico-monetari, su ciascun cluster sono stati calcolati i seguenti valori:

- *totale importo per cluster:* è la somma totale degli importi annui spesi da tutti gli utenti appartenenti ad uno specifico cluster. Formalmente:

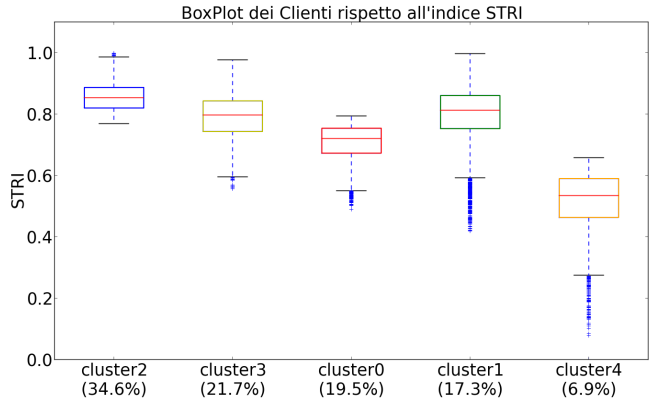
$$tot\_amount(x) = \sum_{i=1}^{N} annual\_amount(i); \tag{6.2}$$

- *totale spese per cluster:* è la somma totale delle spese annue effettuate da tutti gli utenti appartenenti ad uno specifico cluster. Formalmente:

$$tot\_baskets(x) = \sum_{i=1}^{N} annual\_baskets(i); \tag{6.3}$$



(a) BRI



(b) STRI

Fig. 6.18: Box Plot dei cluster rispetto agli indici BRI e STRI

- un *indice di redditività media* di un cliente appartenente ad un cluster. Rappresenta l'importo medio annuo speso da un cliente ed è ottenuto dalla seguente formula:

$$profitability(x) = \frac{\sum_{i=1}^{N} annual\_amount(i)}{N}; \tag{6.4}$$

- un *indice di acquisto medio* di un cliente appartenente ad un cluster. Rappresenta il numero medio di spese effettuate in un anno ed è ottenuto dalla seguente formula:

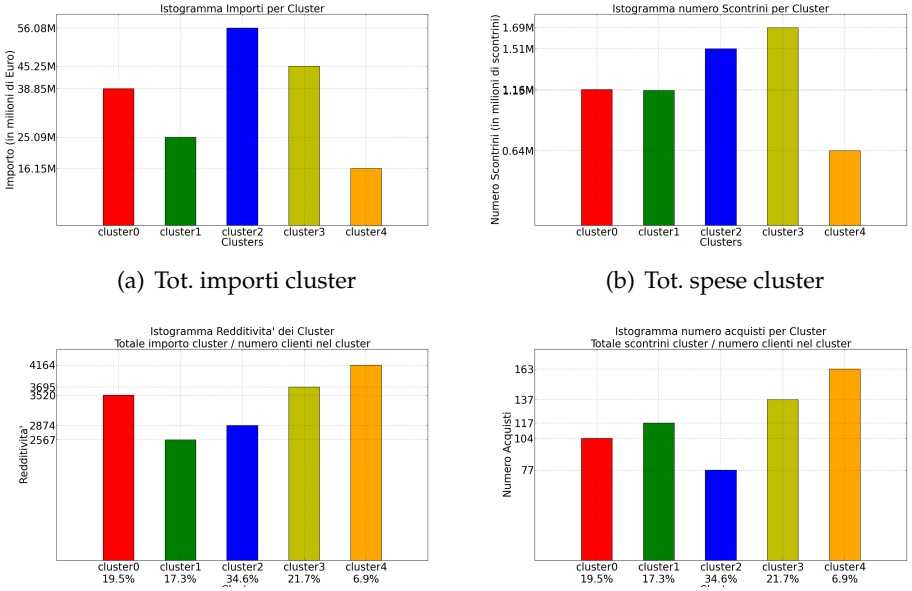
$$mean\_baskets(x) = \frac{\sum_{i=1}^{N} annual\_baskets(i)}{N}, \tag{6.5}$$

con x il cluster in analisi, N il numero di consumatori nel cluster, i il cliente i-esimo, $annual\_amount(i)$ l'importo annuo del cliente i-esimo e $annual\_baskets(i)$ il numero di spese annue effettuate dal cliente i-esimo.

Gli ultimi due indici sono indipendenti dalla cardinalità dei cluster trovati. I risultati dell'applicazione degli indici è riportato negli istogrammi in figura 6.19.

Il grafico 6.19(a) mostra come la somma totale degli importi degli utenti appartenenti a ciascun cluster sia direttamente proporzionale alla cardinalità di ciascun gruppo. La somma totale degli importi oscilla da un valore minimo di 16 milioni di euro per il cluster più piccolo, ad un massimo di 56 milioni di euro per il cluster più grande.

La stessa proporzione non viene mantenuta per quanto riguarda il numero di spese annue di ciascun cluster (figura 6.19(b)). Ad esempio il $cluster3$, corrispondente ai *sistematici sugli acquisti*, effettua spese più che proporzionali rispetto al numero di clienti al suo interno. Il numero totale di spese annue oscilla da un minimo di 640.000 ad un massimo di 1.69 milioni.



(a) Tot. importi cluster

(b) Tot. spese cluster

(c) Importi medi cluster

(d) Acquisti medi cluster

Fig. 6.19: Istogrammi sul *comportamento economico* dei cluster

Un primo risultato interessante è presentato in figura 6.19(c). L'istogramma mostra la redditività media annua di un cliente appartenente ad uno specifico cluster. I cluster che producono più "ricchezza" non sono i cluster più popolati come ci si potrebbe aspettare. Quelli con l'indice più

elevato sono invece il $cluster4$ e il $cluster3$ appartenenti rispettivamente ai *bi-sistematici* ed ai *sistematici sugli acquisti*. I *bi-sistematici*, in particolare, spendono in media 644 euro in più rispetto al cluster degli utenti classificati come *standard*.

Un secondo aspetto non banale riguarda l'indice degli acquisti medi per cluster, mostrato in figura 6.19(d). Anche in questo caso sono i cluster dei *bi-sistematici* e dei *sistematici sugli acquisti* ad effettuare un numero di spese annue maggiore rispetto a tutti gli altri cluster. Essi eseguono rispettivamente 59 e 33 spese annue in più rispetto agli utenti classificati come *standard*.

In tabella 6.2 è fornita una descrizione di ciascun cluster in base agli indici precedentemente analizzati:

Nome Cluster	%	Descrizione
bi-sistematici	6.9%	Fanno il più alto numero di spese e hanno l'importo annuo maggiore rispetto a tutti gli altri cluster
sistematici sugli acquisti	21.7%	Fanno un numero di spese medio alto e hanno un importo annuo medio alto
bi-casuali	17.3%	Fanno un numero di spese medio e hanno un importo annuo basso
casuali sugli acquisti	34.6%	Fanno un numero di spese basso e hanno un importo annuo medio basso
standard	19.5%	Fanno un numero di spese medio basso e hanno un importo annuo medio

Tabella 6.2: Descrizione dei cluster trovati

6.5 Segmentazione mediante analisi delle distribuzioni del BRI e dello STRI

Il secondo tipo di analisi consiste nello studio delle distribuzioni del **BRI** e dello **STRI**.

A differenza dell'analisi svolta con la tecnica del *clustering*, il cui scopo era una segmentazione della clientela in base a caratteristiche simili rispetto alle due misure considerate contemporaneamente, ora si vogliono identi-

ficare, per ogni misura, i clienti che cadono agli estremi dei valori assunti dagli indici, e calcolarne successivamente le intersezioni.

In particolare si vuole dividere la clientela in tre gruppi distinti:

- *sistematici:* rientrano in questa categoria i consumatori considerati ripetitivi sia negli articoli acquistati in ciascuna spesa, sia per i criteri spazio-temporali che adotta per i suoi acquisti (luogo, fascia oraria, tipologia di giorno);
- *casuali:* rientrano in questa categoria i consumatori che si comportano in maniera totalmente aleatoria; essi non presentano alcun criterio ne sulla tipologia di articoli acquistati nelle singole spese, ne sul negozio, fascia oraria o giorno in cui si recano negli store;
- *standard:* sono quei consumatori che non rientrano in nessuna delle due categorizzazioni precedenti. Hanno un valore *centrale* sia per il **BRI** che per lo **STRI**.

Una rappresentazione grafica del processo di *segmentazione* è mostrato in figura 6.20.

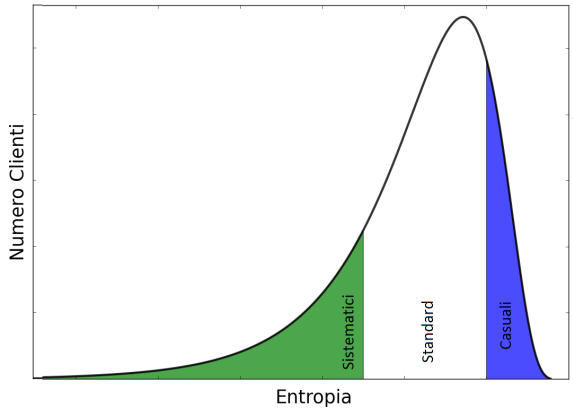


Fig. 6.20: Segmentazione dei clienti in tre gruppi: *sistematici*, *standard* e *casuali*

Poiché le distribuzioni sono di tipo *lognormal* viene intuitivo classificare come *standard* i consumatori appartenenti alla campana della funzione. In corrispondenza di tale area compaiono quei clienti i cui valori delle entropie sono più vicini alla *moda*. Detto in altri termini, nella campana risiedono i consumatori con un comportamento omogeneo e dunque meno interessanti per le analisi che si vogliono svolgere in questa tesi. I clienti maggiormente interessanti, sono invece quelli che si discostano significativamente dalla *massa*; essi sono collocati al di fuori della campana. Più precisamente, alla sinistra della campana risiederanno gli utenti *sistematici* che avranno valori entropici piuttosto bassi; alla destra della campana risiederanno invece gli utenti considerati *casuali* che avranno invece valori entropici alti.

Trattandosi di due segmentazioni realizzate basandosi su due diverse misure, vengono poi identificati gruppi di clienti significativi mettendo insieme sottogruppi comuni delle partizioni sopra elencate.

Come caso di studio, una volta identificati tali gruppi di clienti, viene proposta la ricerca di prodotti, il cui acquisto sia significativo, comuni ai clienti appartenenti agli stessi gruppi. Maggiori dettagli verranno forniti in seguito.

Il metodo di selezione è il seguente:

- si definiscono due tagli sulle distribuzioni del **BRI** e dello **STRI** per isolare le "code" della funzione. Si otterranno così, per ciascuna delle due misure, tre insiemi distinti: l'insieme dei *sistematici*, quello dei *casuali* e infine quello degli utenti *standard*;
- su ciascun insieme, verranno calcolate le intersezioni insiemistiche degli utenti appartenenti a ciascuna categoria. Si otterranno così tre nuovi insiemi contenenti gli utenti *sistematici*, *casuali* e *standard* per entrambi gli indici. Chiameremo i tre insiemi *bi-sistematici*, *bi-casuali* e *bi-standard*. Si noti che ciascuno dei tre insiemi costituisce un sottoinsieme dei clienti;
- sull'insieme dei *bi-sistematici* verranno calcolati i sottoinsiemi di articoli comprati da almeno il 60% degli utenti con una frequenza di acquisto minima x;
- si calcolerà il valore di significatività degli articoli identificati precedentemente negli altri sottogruppi utilizzando una misura di interesse.

6.5.1 Metodi di taglio sulle distribuzioni BRI e STRI

Per individuare i tagli sulle distribuzioni probabilistiche delle entropie sono state utilizzate tre tecniche:

- una basata sull'individuazione del *valore medio* e della *deviazione standard* di una specifica distribuzione. La coda sinistra, in cui risiedono gli utenti *sistematici*, apparterrà all'intervallo $[0.0, media - stdDev]$; la coda destra, in cui risiedono gli utenti *casuali*, apparterrà all'intervallo $[media + stdDev, 1.0]$. Infine gli utenti considerati *standard* apparterranno all'intervallo $(media - stdDev, media + stdDev)$.
- una basata sui *percentili*: l'insieme delle entropie dei clienti viene ordinato in modo non decrescente e, successivamente, viene diviso in tre gruppi in base ad una percentuale passata come parametro. Ad esempio, se si sceglie come variabile il 10%, l'insieme dei *sistematici* sarà il 10% degli utenti col valore dell'entropia più piccolo; viceversa, l'insieme dei *casuali* sarà rappresentato dal 10% dei consumatori aventi

il valore dell'entropia più elevato. L'80% dei clienti rimanenti saranno classificati come *standard*.

- una basata sull'*interquartile range* [12]: è la differenza tra il terzo (75%) e il primo quartile (25%), ovvero l'ampiezza della fascia di valori che contiene la metà "centrale" dei valori osservati. Formalmente lo scarto interquartile è dato dalla formula $IQR = Q3 - Q1$. La coda sinistra dei *sistematici* sarà compresa nell'intervallo $[0.0, Q1 - 1.5 \cdot IQR]$, mentre la coda destra dei *casuali* sarà compresa nell'intervallo $[Q3 + 1.5 \cdot IQR, 1.0]$

Le tabelle sottostanti mostrano l'applicazione delle tre tecniche di taglio sulle distribuzioni del **BRI** e dello **STRI**.

Metodo	Range sistematici	Range standard	Range casuali
$\mu \pm \sigma$	$[0.0, 0.64]$	$(0.64, 0.85)$	$[0.85, 1.0]$
15-percentile	$[0.0, 0.6328]$	$(0.6328, 0.85)$	$[0.85, 1.0]$
IQR	$[0.0, 0.46]$	$(0.46, 1.0)$	$[1.0, 1.03]$

Tabella 6.3: Intervalli di segmentazione della clientela per l'indice BRI

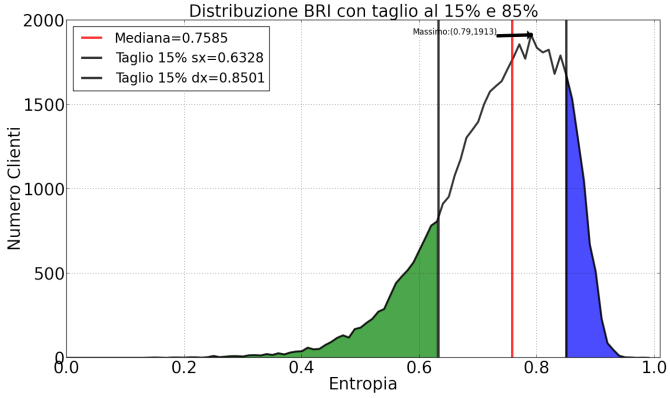
Dalle tabelle 6.3 e 6.4 si nota come i primi due metodi di taglio ($\mu \pm \sigma$ e $15 - percentile$) suddividano i valori delle entropie in intervalli molto simili tra loro. Per quanto riguarda il **BRI** gli intervalli di segmentazione sono identici, mentre per lo **STRI** cambiano gli intervalli degli utenti *standard* e *casuali* (di 0.01).

Per quanto riguarda il metodo di taglio basato sull'*interquartile range*, esso non è applicabile sulle distribuzioni in esame. Infatti, sia per l'indice **BRI** che per l'indice **STRI** uno dei tagli supera il valore massimo 1.0.

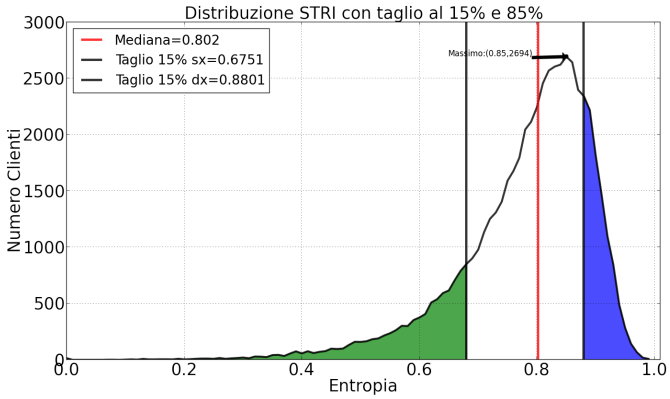
Metodo	Range sistematici	Range standard	Range casuali
$\mu \pm \sigma$	$[0.0, 0.67]$	$(0.67, 0.89)$	$[0.89, 1.0]$
15-percentile	$[0.0, 0.6758]$	$(0.6758, 0.88)$	$[0.88, 1.0]$
IQR	$[0.0, 0.53]$	$(0.53, 1.0)$	$[1.0, 1.05]$

Tabella 6.4: Intervalli di segmentazione della clientela per l'indice STRI

Come metodo di taglio per le successive analisi verrà utilizzata la tecnica del *15-percentile* perché, oltre ad avere quasi lo stesso andamento del taglio $\mu \pm \sigma$, restituisce gli insiemi dei *sistematici* e dei *casuali* equipopolati. Con gli insiemi più interessanti aventi la stessa cardinalità, si potranno apprezzare meglio le differenze nel loro comportamento di acquisto in termini di frequenza di articoli distinti comprati.



(a) BRI



(b) STRI

Fig. 6.21: Taglio al 15-percentile applicato alle distribuzioni del BRI e del STRI

In figura 6.21 sono mostrati i tagli applicati alle distribuzioni **BRI** (6.21(a)) e **STRI** (6.21(b)). Sono evidenziati in verde i consumatori *sistematici*, in bianco i consumatori *standard* ed in blu i consumatori *casuali*.

6.5.2 Cardinalità dei segmenti trovati

E' stato applicato il taglio del *15-percentile* sulle distribuzioni del **BRI** e dello **STRI**. I risultati sono mostrati in tabella 6.5. Per identificare il segmento dei *sistematici* è stato isolato il 15% degli utenti aventi il valore dell'entropia più basso; per identificare il segmento dei *casuali* è stato isolato il 15% degli utenti aventi il valore dell'entropia più alto. I rimanenti sono stati inclusi nell'insieme degli *standard*.

Successivamente sono state calcolate le intersezioni insiemistiche su ciascuno dei segmenti trovati precedentemente, col fine di trovare tre nuovi

BRI	*Cardinalità*	%
Sistematici	*8467*	15%
Standard	*39514*	70%
Casuali	*8467*	15%

STRI	*Cardinalità*	%
Sistematici	*8467*	15%
Standard	*39514*	70%
Casuali	*8467*	15%

Tabella 6.5: Cardinalità degli insiemi di ciascun segmento

gruppi, caratterizzati da una doppia sistematicità: i *bi-sistematici*, i *bi-standard* e i *bi-casuali*. La cardinalità di ciascun insieme è mostrato in tabella 6.6.

Osservando le percentuali di ciascun gruppo, si nota come l'insieme dei *bi-standard* racchiuda circa il 60% degli utenti che hanno i valori "centrali" dell'entropia sul *comportamento di acquisto* e dell'entropia sul *comportamento spazio-temporale*. Questo risultato da un'importante informazione: gli utenti *standard* tendono ad avere un comportamento "centrale" sia sugli articoli acquistati, che sul *dove* e *quando* effettuano le loro spese.

Segmento	**Cardinalità**	**% per segmento**	**% sul totale**
Bi-Sistematici	*1885*	22.3%	3.4%
Bi-Standard	*23424*	59.3%	41.5%
Bi-Casuali	*1645*	19.4%	2%

Tabella 6.6: Cardinalità dell'intersezione di ciascun segmento

Lo stesso risultato non si trova, invece, per gli altri due gruppi. I *bi-sistematici* e i *bi-casuali* sono rispettivamente il 22.3% e il 19.4% degli insiemi singoli (*sistematici* e *casuali*). Questo significa che la maggior parte degli utenti *sistematici* è ripetitiva solo in uno dei due aspetti: comportamento di acquisto o comportamento spazio-temporale. Allo stesso modo, gli utenti *casuali*, all'80%, hanno un comportamento aleatorio solamente per uno dei due criteri di ripetitività. Questi risultati si riallacciano al metodo di segmentazione basato sul *clustering* mediante **K-MEANS**, nel quale venivano identificati 5 cluster che potrebbero essere mappati nelle seguenti categorie; *bi-sistematici*, *sistematici sugli acquisti*, *bi-casuali* , *casuali sugli acquisti* e *standard*.

6.5.3 Articoli acquistati dai segmenti

La prima analisi effettuata sui clienti appartenenti al segmento dei *bi-sistematici* ha riguardato la ricerca di pattern di acquisto comuni. Detto in altri termini, si vuole sapere se i consumatori classificati come ripetitivi sul *comportamento di acquisto* e sul *comportamento spazio-temporale* abbiano anche gli stessi carrelli tipici.

Per raggiungere questo obiettivo, su ciascun cliente è stato eseguito l'algoritmo *Apriori* per ricavare i pattern frequenti di acquisto. Successivamente, si è calcolata l'intersezione di ciascun pattern per trovare i carrelli tipici comuni. I risultati sono mostrati in tabella 6.7.

Articolo	$\%Clienti$
Banane @@@	6.5%
Baguettino @@@ 120 *gr.*	9.47%
Baguettino Integrale @@@ 150 *gr.*	6.35%
Schiacciata @@@	6.43%
Baguette @@@ 250 *gr.*	4.16%
Acqua Minerale Naturale @@@ $1,5$*L.*	4.58%
Latte UHT @@@ Brick 1*L.*	4.68%
Latte Microfiltrato @@@ Brick 1*L.*	3.85%
Latte UHT @@@ Brick $1,5$*L.*	3.22%

Tabella 6.7: Carrelli tipici comuni del segmento dei bi-sistematici

Dalle percentuali sopra riportate si nota come i consumatori *bi-sistematici* non abbiano carrelli tipici in comune. Questo risultato non giunge inaspettato proprio per la natura della misura sintetizzata. Infatti, un valore basso dell'entropia sul *comportamento di acquisto* , per un sottoinsieme di persone, non implica necessariamente che queste comprino gli stessi articoli.

La seconda analisi effettuata ha riguardato la ricerca di articoli comuni acquistati da almeno il 60% dei clienti appartenenti al segmento dei *bi-sistematici*. Anziché ricercare i carrelli tipici di ciascun individuo, si è calcolato il sottoinsieme di articoli distinti che ognuno di essi compra ripetutamente. L'elenco dei prodotti è riportato in tabella 6.8.

Articolo	$\%Clienti$
Prezzemolo 70 *gr.*	62.71%
Uova IT BIO M	64.03%
Zucchine IT Scure	67.8%
Zucchine IT Chiare	67.37%
Finocchi IT	69.07%
Pomodoro Rosso Grappolo	74.22%
Banane @@@	82.44%
Zucchero Semolato IT	72.04%

Tabella 6.8: Articoli comuni nel segmento dei bi-sistematici

Le percentuali di consumatori che adottano il sottoinsieme di articoli varia da un minimo del 62.71% per il *prezzemolo* ad un massimo del 82.44%

per le *banane*.

Una volta individuati gli articoli acquistati dal segmento dei *bi-sistematici* sono state calcolate le percentuali di adozione anche negli altri due segmenti. I risultati sono riportati in tabella 6.9.

Articolo	**% bi-sistematici**	**% bi-casuali**	**% bi-standard**
Prezzemolo 70 gr.	62.71%	21.82%	36.25%
Uova IT BIO M	64.03%	28.57%	45.47%
Zucchine IT Scure	67.8%	30.21%	45.93%
Zucchine IT Chiare	67.37%	23.51%	39.16%
Finocchi IT	69.07%	28.14%	44.81%
Pomodoro Rosso Grappolo	74.22%	25.41%	45.04%
Banane @@@	82.44%	46.26%	65.06%
Zucchero Semolato IT	72.04%	34.83%	49.36%

Tabella 6.9: Percentuale di adozione degli articoli nei segmenti trovati.

Si nota come il segmento dei *bi-sistematici* abbia una percentuale di adozione nettamente superiore per ciascun prodotto rispetto agli altri due. Le misure **BRI** e **STRI**, oltre a riassumere, con un unico valore, il *comportamento di acquisto* e il *comportamento spazio-temporale* di un consumatore, trovano un riscontro anche sulle percentuali degli articoli acquistati da ciascun segmento.

L'ultima analisi sulla bontà degli indici **BRI** e **STRI** ha previsto il calcolo del *coefficiente di lift* per ciascuna coppia $\{cliente, articolo\}$ rispetto a tutti gli articoli acquistati nel periodo di riferimento. Tramite questa misura è possibile infatti determinare l'*importanza* di uno specifico articolo, per uno specifico cliente, rispetto a tutti gli altri articoli acquistati. Valori del *lift* maggiori di 1 indicheranno una significatività dell'acquisto del prodotto rispetto al cliente, mentre valori inferiori o uguali ad 1 indicheranno un'assenza di significatività.

Formalmente il $lift$ è espresso dalla seguente formula:

$$LIFT_{CP} = \frac{W_{CP} \cdot W}{W_C \cdot W_P}, \tag{6.6}$$

con W_{CP} il numero di volte che il cliente C ha acquistato l'articolo P, W il numero totale di articoli acquistati, W_C il numero di articoli acquistati da C e W_P il numero di volte che l'articolo P è stato acquistato. Il $lift$ calcola il rapporto dell'acquisto rispetto allo stesso acquisto fatto sotto ipotesi di indipendenza statistica.

In tabella 6.10 vengono riportate, per ciascun segmento, le percentuali dei clienti aventi $lift > 1$.

Articolo	**% bi-sistematici** $lift > 1$	**% bi-casuali** $lift > 1$	**% bi-standard** $lift > 1$
Prezzemolo 70 gr.	41.27%	19.51%	26.74%
Uova IT BIO M	36.18%	25.77%	33.3%
Zucchine IT Scure	39.68%	26.68%	32.81%
Zucchine IT Chiare	40.95%	20.85%	27.89%
Finocchi IT	39.25%	24.31%	31.86%
Pomodoro Rosso Grappolo	43.81%	21.09%	30.63%
Banane @@@	42.28%	29.96%	34.3%
Zucchero Semolato IT	40.53%	30.15%	34.47%

Tabella 6.10: Percentuale utenti con $lift > 1$ per ciascun segmento.

Dai valori in tabella si nota che, per ogni prodotto, il segmento dei *bi-sistematici* ha una percentuale di clienti con $lift > 1$ superiore agli altri segmenti. Ad esempio, per il *pomodoro*, le *zucchine chiare* e il *prezzemolo*, la percentuale dei *bi-sistematici* è superiore a quella dei *bi-standard* del 13% per i primi due articoli e del 15% per il terzo articolo. Per gli altri prodotti, fatta eccezione per le *uova*, la percentuale dei *bi-sistematici* supera quella dei *bi-casuali* del 7% e dell'8%.

Questo risultato mostra come le misure **BRI** e **STRI** catturino effettivamente il *comportamento di acquisto* dei consumatori. Rispetto agli altri segmenti, quello dei *bi-sistematici* ha una percentuale significativamente superiore di utenti che comprano un determinato insieme di prodotti. Viceversa, nel segmento dei *bi-casuali*, le percentuali di utenti *legati* al sottoinsieme di articoli preso in esame, è sensibilmente inferiore. In questo caso le misure catturano l'*aleatorietà* negli acquisti.

E' da notare, inoltre, che le analisi sono state effettuate a livello *articolo*, cioè al livello gerarchico più basso. Nei punti vendita della Provincia di Livorno sono presenti circa 85000 articoli distinti. Eseguendo le stesse analisi al livello gerarchico superiore (livello *segmento*), i risultati sarebbero stati ancora più significativi.

Capitolo 7

Conclusioni

Il metodo proposto definisce un approccio innovativo per la segmentazione dei clienti, misurando quanto un consumatore sia sistematico nei suoi acquisti. Le misure generate, il *basket regularity index* (**BRI**) e lo *spatio-temporal regularity index* (**STRI**), stimano il comportamento dei consumatori sotto due aspetti: uno legato agli articoli acquistati in ciascuna spesa e uno legato al luogo, giorno e ora in cui vengono fatte.

Dagli esperimenti eseguiti sul *datawarehouse* della **Coop** emergono due importanti risultati:

- un cliente *sistematico* spende in media un importo annuo maggiore rispetto ad un cliente *non sistematico* a prescindere dagli articoli acquistati;
- per un sottoinsieme di prodotti di prima necessità (frutta e verdura), un cliente *sistematico* presenta una maggiore significatività di acquisto rispetto agli altri gruppi di consumatori.

Un'importante caratteristica aggiuntiva del metodo consiste nella possibilità di estrarre i *carrelli tipici* oppure le *triple stazio-temporali* che descrivono il comportamento di acquisto di un consumatore.

In conclusione, le misure proposte riescono a sintetizzare, per ciascun individuo, la componente concreta (dove, quando, quanto e cosa compra) e astratta (come compra) del comportamento di acquisto.

7.1 Sviluppi futuri

In aggiunta alle analisi effettuate, sarebbe interessante estendere i casi di studio precedentemente discussi ad un livello superiore di astrazione. In particolare, il metodo di segmentazione basato sui tagli delle distribuzioni degli indici potrebbe essere applicato al livello *segmento di marketing* anziché al livello *articolo*.

Ulteriori casi di studio potrebbero riguardare l'evoluzione delle due misure considerando un arco temporale di riferimento più ampio, per capire se **BRI** e **STRI** assumono valori costanti nella vita del cliente o se sono soggetti a variazioni. Ad esempio, sarebbe interessante sapere se le due misure sono influenzate dalle componenti stagionali o dalle festività.

Per quanto riguarda le possibili applicazioni, il metodo permetterebbe ai reparti di marketing di effettuare offerte mirate ai clienti al fine di individuare la possibile esistenza di prodotti che spingano i consumatori ad essere maggiormente sistematici.

Ringraziamenti

Ringrazio il Professor Dino Pedreschi per avermi dato l'opportunità di contribuire ad un importante progetto di ricerca.

Un sentito riconoscimento a Diego Pennacchioli e Riccardo Guidotti per il loro costante supporto tecnico e per i preziosi consigli dati durante tutta la fase di realizzazione del metodo e della tesi.

Un doveroso ringraziamento alla mia famiglia, agli amici e a tutti coloro che mi hanno sostenuto e incoraggiato in questa splendida avventura.

Bibliografia

[1] plc. Abraham, Hawks. *Market Segmentation: can you really divide and conquer?* 1998.

[2] Albert-László Barabási. *Lampi, la trama nascosta che guida la nostra vita*. Einaudi, 2011.

[3] Bob Dorf Don Peppers, Martha Rogers. *Marketing One to One*. Il Sole 24 Ore, 2006.

[4] P. Holmes. *Customer Profiling and Modeling in Direct Marketing Association*. PhD thesis.

[5] Hunt T. Humby C. *Scoring Points: How Tesco is winning Customer Loyalty*. Kogan Page, 2003.

[6] Spatial Insight Inc. Mosaic lifestyle segmentation database. *http://www.spatialinsights.com/Lybrary/data/attribute/mosaic/*.

[7] Jean-Jacques Lambin. *Marketing strategico e operativo*. McGraw-Hill, 2004.

[8] Vipin Kumar Pang-Ning Tan, Michael Steinbach. *Introduction to Data Mining*.

[9] PriceWaterHouseCoopers. *The CRM Handbook: from group to multiindividual*. PriceWaterHouseCoopers, 1999.

[10] R.F.Wilson. *Preparing a Customer Profile for your Internet Marketing*. 2000.

[11] K. Thearling. *Data Mining and Customer Relationship*.

[12] Kokoska S. Zwillinger, D. *CRC Standard Probability and Statistics Tables and Formulae*. 2000.