Università degli Studi di Pisa

Facoltà di Lettere e Filosofia

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Specialistica in Informatica Umanistica

Tesi di Laurea

Un caso di studio nella profilazione dei nuovi clienti nell'ambito dei servizi fiscali con tecniche di Data Mining

RELATORE:

Prof. Salvatore Ruggieri

CANDIDATO:

Patrizia Vergassola

Anno accademico 2008-2009

Università degli Studi di Pisa

Facoltà di Lettere e Filosofia

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Specialistica in Informatica Umanistica

Curriculum MANAGEMENT DELLA CONOSCENZA

UN CASO DI STUDIO NELLA PROFILAZIONE DEI NUOVI CLIENTI NELL'AMBITO DEI SERVIZI FISCALI CON TECNICHE DI DATA MINING

Tesi di

Patrizia Vergassola

Relatore:

Prof. Salvatore Ruggieri .....

Candidato:

Patrizia Vergassola .....

Anno accademico 2008-2009

Ai miei genitori

Indice

Introduzione

Oggigiorno, uno dei problemi fondamentali delle organizzazioni che utilizzano database di grandi dimensioni è costituito dal non sapere come agire per estrarre l'informazione presente nell'enorme quantità di dati memorizzata in quegli stessi database. Infatti, l'informazione in questione è sì presente, ma risulta altresì nascosta tra i dati e l'operazione di trasformazione della medesima informazione in conoscenza utile al supporto alle decisioni e ad azioni di business non è affatto semplice né scontata.
Questo è proprio il caso del CAAF-CISL: un ente che opera sul territorio per fornire a lavoratori e pensionati assistenza e consulenza nel campo fiscale e delle agevolazioni sociali e che ha manifestato un profondo interesse nel voler migliorare la propria conoscenza riguardo alle particolarità caratterizzanti i clienti che per la prima volta decidono di avvalersi del servizio di assistenza per la compilazione della dichiarazione dei redditi.
L'obiettivo della tesi è dunque lo studio delle eventuali caratteristiche fiscali o demografiche in grado di identificare e differenziare quei clienti che per la prima volta si avvalgono del servizio di assistenza fiscale e che, per questa ragione, possono essere definiti nuovi.
L'obiettivo appena presentato viene raggiunto attraverso un processo di analisi consistente in sei passi, ognuno dei quali viene concretamente esemplificato da ogni capitolo di questo lavoro.
In un primo momento, lo scopo è quello di cercare di avere un quadro ge-

nerale ed esauriente riguardo lo stato dell'arte del *Data Mining*[1] e le regole associative in generale e, in un secondo momento, attraverso un approfondimento delle regole di classificazione in particolare, dal momento che proprio questa tecnica è stata scelta come quella in grado di portare a termine l'obiettivo sopra proposto.

Viene poi presentato uno studio approfondito delle regole di classificazione, studio che viene effettuato con lo scopo di cercare di comprendere al meglio la loro natura, anche in un'ottica successiva di estrazione e di selezione di queste ultime. La prima parte di questo lavoro prosegue presentando le misure di interesse - oggettive, soggettive e semantiche - proposte in letteratura per valutare le suddette regole di classificazione, di modo da poter poi effettuare la selezione automatica di quelle stesse regole che, in base alla definizione rappresentata dalla misura scelta, risultano essere le più interessanti.

Una volta portata a termine la prima fase riguardante lo studio dello stato dell'arte del Data Mining in generale e della tecnica scelta e delle misure di interesse in particolare, abbiamo dato il via alla parte più propriamente concreta e sperimentale di questo lavoro.

La descrizione di questa seconda fase parte dal secondo capitolo ed arriva al sesto; nell'arco di questi capitoli in primo luogo spieghiamo in che modo siamo arrivati alla definizione finale dell'insieme dei dati da utilizzare in fase di modellazione e, in secondo luogo, proprio nel capitolo dedicato alla fase di modellazione, descriviamo esaurientemente come la tecnica scelta sarà implementata grazie all'utilizzo di uno specifico tool, DCUBE. La descrizione del lavoro che porta alla selezione delle regole più interessanti prosegue presentando le due misure di interesse oggettivo scelte per poter far emergere le

[1] Il Data Mining è la disciplina a cui si può fare appello quando si vogliono estrarre dai dati delle informazioni utili che sono però nascoste e non possono essere svelate con gli strumenti tradizionali di analisi dei dati. Una definizione esauriente di esso può essere trovata nel capitolo 1.

regole più interessanti. Le misure scelte - ovvero l'*extended lift* (*elift*) e il *4th quantifier of founded double implication* - e implementate in DCUBE sono quelle che più si avvicinano a ciò che noi vogliamo fare emergere dai dati, vale a dire caratteristiche in grado di differenziare, in particolari contesti, i nuovi clienti dai vecchi.
Una volta ottenuti i risultati, rappresentati sotto forma di particolari regole di classificazione individuate sulla base dei valori delle misure scelte, questa tesi si conclude con la presentazione di questi stessi risultati agli esperti del dominio del CAAF-CISL - presentazione descritta nel penultimo capitolo di questo lavoro - e, per tutto il settimo capitolo, con la descrizione e l'individuazione di alcuni media tramite i quali potrebbero essere effettuate delle campagne promozionali valorizzanti, appunto, i risultati emersi.

In sintesi, la tesi appena introdotta si articola nei seguenti capitoli:

- **cap. 1**: si introduce la disciplina del Data Mining descrivendo le fasi che compongono un processo di analisi dei dati. Il capitolo prosegue presentando la tecnica scelta per risolvere l'obiettivo della tesi: le regole di classificazione. Di queste ultime viene proposta un'ampia rassegna: si parla infatti delle diverse metodologie di estrazione delle regole, ma anche di selezione di queste ultime grazie all'utilizzo delle misure di interesse (oggettive, soggettive e semantiche);

- **cap. 2**: viene descritta la situazione iniziale, unitamente al dominio in cui si è operato. Vengono elencati gli obiettivi di business ed i criteri di successo considerati determinanti. Il capitolo si conclude con l'analisi della situazione e delle risorse a disposizione per le successive fasi di pre-processing e di modellazione;

- **cap. 3**: questo capitolo contiene una descrizione dettagliata della struttura dell'insieme dei dati iniziale, già sviluppato pronto per le analisi grazie al lavoro di [2];

- **cap. 4**: in questa parte della tesi vengono motivate le scelte che hanno portato ad una riduzione del numero degli attributi facenti parte dell'insieme iniziale dei dati: si parla cioè dello studio delle distribuzioni e delle correlazioni, così come dell'opinione dataci dagli esperti del dominio riguardo il livello di rilevanza di ogni attributo;

- **cap. 5**: in questo capitolo viene descritto il processo di modellazione: si parte dalla selezione della tecnica di data mining, cioè la tecnica delle regole di classificazione, e si arriva alla descrizione dell'applicazione di quest'ultima mediante l'illustrazione del tool di analisi utilizzato - DCUBE - e alla selezione delle misure di interesse scelte per scoprire quali regole si sarebbero rivelate potenzialmente interessanti;

- **cap. 6**: questo capitolo contiene una descrizione dettagliata di tutti i risultati emersi nella fase precedente, divisi per provincia di analisi e per misura di interesse e validati insieme agli esperti del dominio;

- **cap. 7**: in quest'ultima parte della tesi vengono trattati i possibili utilizzi dei risultati ottenuti nella sezione precedente. Il tutto è integrato in un'ampia sezione relativa allo stato dell'arte del marketing mix in generale e del promotion mix in particolare.

Capitolo 1

Data Mining: stato dell'arte

Quando si vogliono estrarre dai dati delle informazioni utili che sono però nascoste e non possono essere svelate con gli strumenti tradizionali di analisi dei dati, si fa ricorso a tecniche più particolari di analisi note con il nome di *data mining*. Una definizione che può essere data del data mining è la seguente [3]:

Il data mining è un'attività di analisi semi-automatica, esplorativa, basata sull'apprendimento e finalizzata all'estrazione di conoscenza sotto forma di modelli utili ai decisori per formulare ipotesi di azioni.

Il *data mining* cerca quindi di estrarre dai dati conoscenza, ovvero informazione interessante non immaginabile a priori, rappresentata in una forma opportuna, detta *modello* o *pattern*. Un modello può essere *descrittivo* o *predittivo*. Nel primo caso esso fornisce informazioni utili solo sui dati usati per costruirlo, mentre nel secondo caso può essere usato anche per fare previsioni su un valore sconosciuto di un attributo di nuovi dati.
Il *data mining* non è quindi, per fare un esempio banale, la ricerca di un cognome in un elenco telefonico, ma, al contrario, è la scoperta che certi cognomi risultano più diffusi in certe regioni italiane rispetto ad altre: è il

portare a galla una informazione non scontata.

Il *data mining* rappresenta la confluenza di varie discipline, quali quelle

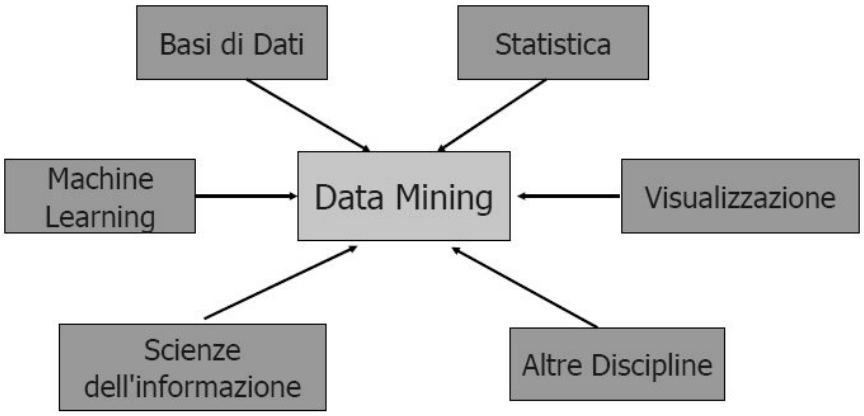


Figura 1.1: Data Mining: la confluenza di varie discipline

rappresentate nella Figura 1.

Il *machine learning* (apprendimento automatico) è un filone di studi, collegato all'informatica e all'intelligenza artificiale, che si occupa di ricavare delle regolarità a partire dai dati. In questo senso, l'apprendimento automatico è dunque una delle basi tecniche del *data mining*. I metodi del *data mining* differiscono però da quelli dell'apprendimento automatico puro in quanto:

- focalizzati ad estrarre informazione dai database;
- si occupano tipicamente dell'analisi di grandi moli di dati e sono quindi di interesse per il *data mining* soltanto gli algoritmi di *machine learning* scalabili.

La statistica si è sempre occupata di metodologie per l'analisi dei dati: recentemente molti statistici si stanno interessando al *data mining*. Dunque

anche la stessa statistica fornisce basi tecniche al *data mining* sia per il processo di costruzione di pattern, che per il processo di verifica della validità di questi ultimi. I metodi di data mining differiscono da quelli puramente statistici perché:

- focalizzati ad estrarre informazione dai database;
- si occupano tipicamente dell'analisi di grandi moli di dati;
- i database su cui operano contengono spesso dei dati che non sono stati raccolti appositamente per lo scopo dell'analisi in corso.

La ricerca nell'ambito del *data mining* è stata vista per lungo con sospetto dagli statistici, i quali hanno coniato i termini dotati di un'accezione negativa quali *data fishing*, *data dredging* e *data snooping* (in riferimento cioè ad una analisi esplorativa, senza ipotesi a priori). Queste sono le critiche che vi venivano mosse:

- nel data mining non vi è un unico modello di riferimento, ma numerosi modelli in competizione. È sempre possibile trovare un modello complesso che si adatti bene ai dati;
- l'abbondanza di dati può portare a pattern in realtà inesistenti.

Tuttavia:

- le tecniche moderne pongono molta attenzione alla generalizzabilità dei pattern, preferendo modelli più semplici;
- molti risultati di interesse per un'applicazione non sono noti a priori, mentre i metodi statistici hanno di solito bisogno di una ipotesi di ricerca data a priori.

L'uso del *data mining* pone inoltre una serie implicazioni etiche. Quando viene infatti applicato a persone, il *data mining* è usato per discriminare:

- chi ottiene il prestito? Certe discriminazioni (ad esempio in base a sesso e razza) sono eticamente poco corrette o anche illegali;
- tuttavia, molto dipende dall'applicazione.

Questo avviene perché alcuni attributi possono essere correlati ad informazioni problematiche: ad esempio, il luogo di residenza può essere correlato al gruppo etnico.
Sempre in riferimento alle implicazioni etiche, sorgono inoltre domande importanti che nascono nelle applicazioni:

- chi ha il diritto di accedere ai dati?
- per che scopo erano stati raccolti i dati?

Tutte le analisi dovrebbero ovviamente avvenire con il consenso esplicito delle persone coinvolte.

1.1 Il processo di *Knowledge Discovery in Databases* (KDD)

Il *data mining* costituisce la parte più importante di un processo più ampio, noto come *Knowledge Discovery in Databases* (KDD). Il KDD è stato così definito [4]:

processo interattivo ed iterativo, costituito dall'insieme dei metodi e delle tecniche utilizzate allo scopo di identificazione di relazioni tra dati, che siano valide, nuove, potenzialmente utili e comprensibili.

Il processo KDD prevede come input dati grezzi e fornisce come output informazioni utili ottenute attraverso le seguenti cinque fasi. Esso è in sostanza il processo che permette all'informazione di potersi eventualmente

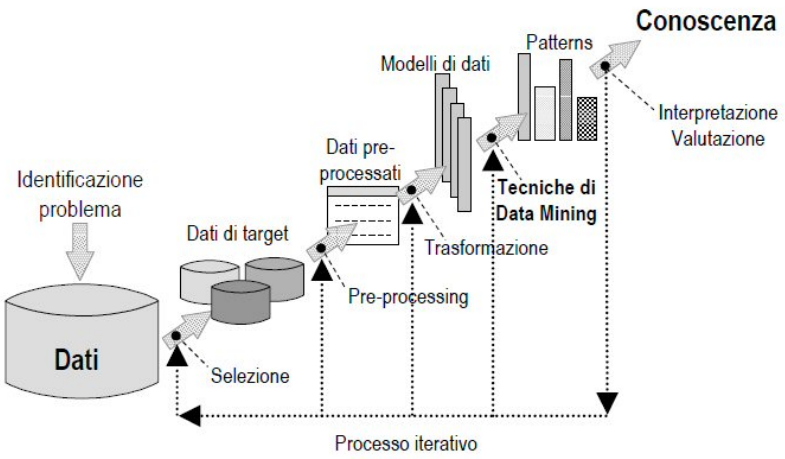


Figura 1.2: Le fasi del processo di KDD [4].

trasformare in conoscenza utile all'utente.

1. *Selezione*: è l'estrazione di parte dei dati secondo alcuni criteri, i quali dipendono dall'obiettivo preposto all'analisi. Una volta effettuata questa selezione dei dati si perviene ad un sottoinsieme di questi ultimi, i quali rappresenteranno i nostri *target data* o *dati obiettivo*. È abbastanza chiaro che in un database possono essere contenute diverse informazioni, molte delle quali possono risultare inutili per il problema in esame; per spiegarla con un esempio, se l'obiettivo dell'analisi è lo studio delle associazioni tra i prodotti acquistati dai clienti di una catena di distribuzione al dettaglio, non ha senso conservare i dati relativi al titolo di studio dei clienti; è invece sbagliato non considerare questa variabile nel caso in cui lo scopo fosse quello di segmentare la clientela in base alle sue caratteristiche socio-demografiche;

2. *Pre-elaborazione*: è la 'pulizia' dei dati (*data cleaning*) da certe informazioni ritenute inutili e che possono rallentare le future interrogazioni. In questa fase i dati possono essere trasformati per evitare

eventuali inconsistenze dovute al fatto che dati simili possono provenire da sorgenti diverse e quindi con metadati leggermente diversi (ad esempio in un database il sesso di una persona può essere salvato come 'm' o 'f' e in un altro come 0 o 1). È sempre propria di questa fase, inoltre, la decisione dei meccanismi di comportamento da adottare in caso di dati mancanti;

3. *Trasformazione* : i dati non sono semplicemente trasferiti da un archivio ad uno nuovo, ma sono trasformati in modo tale che sia possibile anche aggiungere informazione a questi, come per esempio informazioni demografiche comunemente usate nella ricerca di mercato. I dati vengono quindi resi 'usabili e navigabili': si possono convertire alcuni tipi di dati in altri o definirne di nuovi, attraverso l'uso di operazioni matematiche e logiche sulle variabili;

4. *Data mining*: ai dati trasformati vengono applicate una serie di tecniche in modo da poter ricavare delle informazioni che siano non banali o scontate ma utili e non predicibili a priori. Per fare ciò vengono eseguiti i seguenti passi, di solito più volte con lo scopo di arrivare ad una soluzione soddisfacente:

   - scelta della strategia di analisi;
   - scelta della tecnica per applicare la strategia;
   - scelta dell'algoritmo per applicare la strategia;
   - interpretazione e valutazione dei risultati.

5. *Interpretazione e valutazione*: la fase di *data mining* crea dei *pattern*, ovvero dei modelli, che possono costituire un valido supporto alle decisioni. Occorre però valutare ed interpretare questi modelli cosicché la conoscenza che se ne acquisice possa essere di supporto alle decisioni, quali ad esempio la previsione, la classificazione di elementi, il

riassunto dei contenuti di un database o la spiegazione dei fenomeni osservati.

1.2 Strategie e tecniche di *data mining*

Una volta iniziato il processo *data mining*, la prima decisione da prendere riguarda quale tipo di strategia usare per risolvere il problema in esame. La strategia definisce il tipo di modelli di analisi che vogliamo estrarre dai dati. Le strategie si suddividono in:

1. *Strategia supervisionata*

   L'obiettivo è quello di costruire un modello dei dati per spiegare il valore di un attributo, detto *attributo* (*variabile*) *dipendente* (*target*), in funzione dei valori di altri, detti *attributi* (*variabili*) *indipendenti* (*esplicative*). La strategia è detta *supervisionata* perché il modello si costruisce a partire da casi di cui è nota la classificazione, ovvero il valore dell'attributo dipendente (*apprendimento con un insegnante*). Il modello trovato è detto *predittivo*, *di previsione* quando è adatto non soltanto per descrivere i dati attuali, ma anche per prevedere il valore sconosciuto dell'attributo dipendente in nuovi dati. La strategia supervisionata è adatta per rispondere a domande di questo tipo: *perché si verifica questo fatto?*

2. *Strategia non supervisionata*

   L'obiettivo è quello di costruire un modello dei dati senza fare ricorso ad un apprendimento eseguito su casi di cui sappiamo il valore dell'attributo dipendente (*apprendimento senza un insegnante*). La strategia non supervisionata è adatta per rispondere a domande di questo tipo: *c'è qualche informazione interessante nascosta nei miei dati?*

Una volta scelta la strategia di *data mining* appropriata per risolvere il problema in esame occorre decidere di quale tecnica avvalersi per applicare ai

dati la strategia selezionata e quale algoritmo usare per implementare la tecnica utilizzata.
La tecnica definisce in che modo viene applicata ai dati la strategia selezionata con lo scopo di creare i modelli attraverso cui verrà rappresentata l'informazione trovata. Esistono diverse tecniche in base al tipo di strategia (supervisionata e non supervisionata) e in base al tipo di modello che vogliamo ottenere dai dati (predittivo o descrittivo). In ogni caso, quelle più comuni sono le seguenti:

- classificazione [predittiva e descrittiva];
- regressione [predittiva];
- raggruppamento [descrittiva];
- regole associative [descrittiva];
- scoperta di pattern sequenziali [descrittiva].

I modelli di data mining sono delle rappresentazioni matematiche che evidenziano le relazioni tra gli attributi che descrivono i dati. I più comuni sono:

- alberi decisionali;
- classificatori bayesiani;
- reti neurali;
- cluster;
- regole associative;
- pattern sequenziali;
- pattern frequenti.

Più in generale, comunque, come già visto, i modelli possono essere divisi in due categorie principali: i modelli descrittivi e i modelli predittivi.
Una volta scelta la tecnica da applicare ai dati, è nostro compito scegliere l'algoritmo giusto atto ad implementarla. I più comuni sono:

- C4.5 per la *classificazione*;
- Apriori ed FP-growth per la *generazione di regole associative*;
- K-means per il *raggruppamento*.

1.2.1 Classificazione e predizione

La classificazione può essere definita come la costruzione di un modello di classificazione (*classificatore*) e può avere sia una valenza descrittiva che una predittiva. Nel primo caso il *classificatore* può servire come strumento esplicativo per distinguere tra oggetti appartenenti a classi differenti. Per esempio, in campo medico, può essere utile sapere quali sono i sintomi che possono portare a dire che un paziente ha contratto la varicella. Nel secondo caso, invece, il *classificatore* può essere usato per predire il valore dell'attributo dipendente in dati sconosciuti, diversi da quelli di addestramento sui quali il modello è stato costruito.
Quando si parla di classificazione il tipo dell'attributo dipendente deve essere discreto, infatti, quando si ha a che fare con attributi continui si parla di *regressione*.
Il tipo di *classificatore* ottenuto può essere rappresentato in varie forme. Quelle più comuni sono: *alberi decisionali* o *di classificazione*, *regole di classificazione*, *reti neuronali*, *regressione statistica* e *classificatore bayesiano*.

1.2.2 Regole associative

L'obiettivo è quello di costruire un modello di rappresentazione dei dati composto da regole del tipo *se-allora* che descrivano la presenza di certe

relazioni tra determinati valori presenti nei dati.
L'esempio più tipicamente esplicativo riferito a questa tecnica è l'analisi del carrello della spesa di un supermercato (*market basket analysis*), cioè l'insieme di articoli comprati da un cliente durante una sua visita al supermercato, per stabilire quali prodotti hanno una maggiore probabilità di essere acquistati insieme. La scoperta che esistono certe tendenze di acquisto da parte dei clienti permette a chi deve prendere decisioni di organizzare in maniera differente e più adatta i prodotti sugli scaffali, o nel catalogo, o in eventuali campagne pubblicitarie, in modo da fare risaltare e far cadere l'attenzione del cliente sui prodotti acquistati insieme.
Dal momento che più grande è la dimensione dei dati analizzati e più grande è il numero delle regole che possono essere generate, uno degli scopi principali di questo tipo di analisi è quello di evitare di estrarre delle regole che rappresentino delle relazioni banali e scontate e che non producano nuova informazione.

1.2.3 Raggruppamento

La tecnica *di raggruppamento* (*clustering*) ha come scopo quello di raggruppare i dati in gruppi (*cluster*) non definiti a priori sulla base di un criterio di similitudine, ma con la caratteristica che tutti i dati facenti parte di uno specifico gruppo siano omogenei al loro interno ed eteregonei rispetto ai dati di un altro gruppo.
Alcune volte nei dati possono trovarsi dei casi atipici (*outlier*) che se non fanno parte di gruppi a sé stanti possono intaccare la corretta formazione di altri gruppi. Gli *outlier* non vanno necessariamente eliminati né vanno considerati errori poiché in alcune situazioni la loro scoperta è l'obiettivo dell'analisi *di raggruppamento*. Per esempio, nella valutazione degli acquisti effettuati con la carta di credito, un *outlier* potrebbe essere un probabile esempio di utilizzo fraudolento della carta.

Per esempio, il raggruppamento è utilizzato per segmentare la clientela di un'azienda o per selezionare le aree della Terra che hanno un impatto simile sul clima del pianeta.

1.3 Un modello standard per il *data mining*: il CRISP-DM

Il *CRISP-DM* (*CR*oss *I*ndustry *S*tandard *P*rocess for *Data Mining*) è un progetto finanziato dalla Commissione Europea il cui obiettivo è quello di definire un approccio standard ai progetti di *data mining*. Il CRISP-DM affronta la necessità di tutti gli utenti coinvolti nella diffusione di tecnologie di *data mining* per la soluzione di problemi aziendali. Scopo del progetto è definire e convalidare uno schema d'approccio indipendente dalla tipologia di business.
Come possiamo vedere dalla Figura 1.3 il ciclo di vita di un progetto di *data mining* consiste di sei fasi la cui sequenza non è rigida:

1. Definizione del problema di business;
2. Definizione della struttura dei dati;
3. Preparazione dei dati;
4. Modellazione;
5. Valutazione dei risultati;
6. Utilizzo dei risultati.

La creazione di un modello di *data mining* è un processo dinamico e iterativo. Dopo aver esplorato i dati, è possibile scoprire che questi non sono sufficienti per la creazione di modelli di data mining appropriati e che pertanto è necessario cercare altri dati. Analogamente, è possibile creare vari

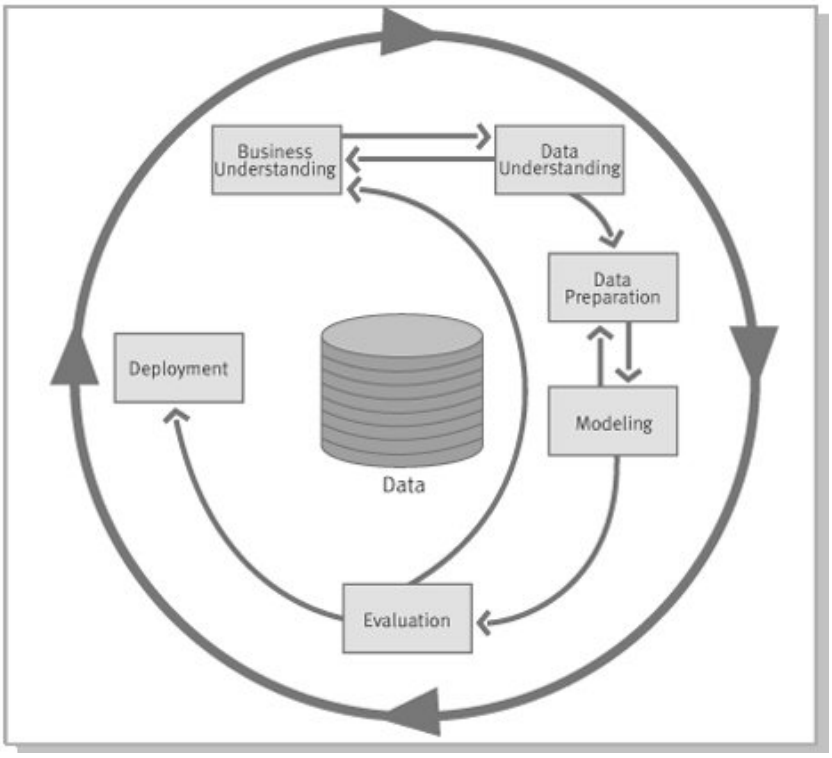


Figura 1.3: [5]Le fasi del processo CRISP-DM

modelli e scoprire che questi non consentono di risolvere il problema definito in precedenza e che pertanto è necessario ridefinire il problema. Potrebbe inoltre essere necessario aggiornare i modelli dopo la relativa distribuzione poiché, a questo punto, sono disponibili altri dati. È quindi importante comprendere che la creazione di un modello di *data mining* è un processo in cui ogni passaggio può essere ripetuto il numero di volte necessario per ottenere risultati soddisfacenti.

Le fasi del CRISP-DM appena elencate sembrano ricalcare nella sostanza le fasi del più generale processo di estrazione di conoscenza dai database (KDD). In primo luogo possiamo infatti osservare che la metodologia del CRISP-DM incorpora i passi del processo di KDD: la fase di *definizione del problema di business* può essere identificata con la fase di *identificazione del*

problema; la fase di *definizione della struttura dei dati* può essere identificata come la combinazione delle fasi di *selezione* e di *pre-elaborazione* e la fase di *preparazione dei dati* può essere identificata con la corrispondente fase di *trasformazione*; infine, la fase di *modellazione* può essere identificata con la fase relativa al *data mining* e la fase di valutazione dei risultati può essere identificata con la fase di *interpretazione e valutazione*. In realtà, però, questo progetto di *data mining* ingloba al suo interno le fasi del processo di KDD e dimostra quanto già affermato in precedenza riguardo al fatto che il processo di KDD non è altro che il flusso che l'informazione attraversa prima di potersi trasformare in conoscenza utile all'utente e KDD non è sinonimo di *data mining*.

Definizione del problema di business

Questa prima fase si focalizza nel tentare di comprendere gli obiettivi e le richieste del progetto guardando a questi ultimi da una prospettiva 'di business', cercando di convertirli in uno o più problemi di *data mining* e quindi procedendo alla stesura preliminare di un piano per raggiungere gli obiettivi stabiliti.
Un 'problema di business' è un problema per cui il cliente richiede una soluzione, soluzione che si sospetta sia nascosta nell'informazione presente nei dati e lo fa dichiarando gli obiettivi del progetto usando una terminologia aziendale.
Un 'problema di *data mining*' enuncia invece gli obiettivi del progetto facendo uso di una terminologia tecnica, specialistica.
Un 'problema di business' potrebbe essere questo: 'Aumentare le vendite attraverso i cataloghi tra i clienti già acquisiti'; un 'problema di data mining' potrebbe invece essere: 'predire quanti oggetti comprerà un cliente, sulla base dei suoi acquisti nei tre anni precedenti all'analisi, date le sue informazioni demografiche (età, salario, città di residenza, ecc.) e il prezzo

dell'oggetto'.

Definizione della struttura dei dati

Lo scopo di questa fase è quello di acquisire i dati che serviranno per lo svolgimento del processo di *data mining* in questione e, nel caso in cui le sorgenti dei dati siano multiple, scopo di questa fase è anche quello di integrare i dati in un'unica base di dati. Vanno elencati i dati che verranno utilizzati, cioè il dataset o i dataset, così come le loro sorgenti, i metodi utilizzati per acquisirli e qualsiasi problema riscontrato nel farlo.
In questa fase devono inoltre essere descritti i dati utilizzati: i tipi dei dati, la quantità dei dati - cioè il numero di record e di attributi per ciascun dataset - e il contenuto dei dati - cioè la lista dei file di dati, delle tabelle e dei campi che contengono.
In questa fase è anche prevista una serie di operazioni preliminari sui dati così da acquisire maggiore familiarità con essi e così da identificare eventuali problemi nella qualità di questi ultimi. Queste prime operazioni possono riguardare: il calcolo della distribuzione dei valori di un attributo chiave - per esempio di un attributo dipendente, operazioni di aggregazione e semplici analisi statistiche (medie, indici di variabilità, ecc.).
È quindi chiaro come queste due prime fasi siano collegate, dal momento che entrambe rappresentano l'individuazione dei fini e dei mezzi di un progetto di *data mining*.

Preparazione dei dati

Questa terza fase riguarda l'estrazione e la preparazione dei dati che popoleranno il dataset finale. È importante sottolineare che il dataset ottenuto nella fase precedente costituisce soltanto la struttura nella quale i dati verranno memorizzati, ma per adesso è ancora vuoto. In questa fase questa struttura viene popolata in modo da poter poi essere utilizzata nella

fasi successive.
Per arrivare a 'riempire' il dataset finale è necessario:

- *Selezionare i dati* da utilizzare per le operazioni di analisi successive nel processo di *data mining* (con ciò si fa riferimento sia alla selezione degli attributi sia alla selezione dei record). I criteri per svolgere questo compito includono: la rilevanza, la completezza e l'esattezza dei dati in base alle specifiche di progetto, la qualità dei dati e vincoli tecnici quali quelli che fanno riferimento a delle limitazioni sul volume dei dati o sui tipi di questi ultimi;

- *Pulire i dati.* In questa fase la qualità dei dati deve essere portata al livello richiesto dalle successive operazioni di *data mining* che verranno effettuate; in particolare la qualità dei dati deve essere quella richiesta dalla tecnica di *data mining* scelta;

- *Costruire i dati.* Questa fase comprende operazioni di preparazione dei dati costruttive come la produzione di attributi derivati, di interi nuovi record o la trasformazione di valori in attributi già esistenti;

- *Integrare i dati.* Questa fase fa riferimento all'operazione di *join* tra più tabelle recanti diversi tipi di informazione facente riferimento allo stesso oggetto di indagine, così come all'operazione di aggregazione che permette di computare nuovi valori sommando insieme informazione presente in più record e/o tabelle;

- *Portare i dati nel formato richiesto* dalla successiva funzione di mining che si intende adottare. Le trasformazioni che si effettuano in questa fase sono di tipo sintattico, cioè sono modifiche che non cambiano il significato dei dati bensì soltanto la loro forma.

Modellazione

Dopo aver definito gli obiettivi di business, la struttura dell'insieme dei dati ed aver preprocessato le informazioni provenienti dalle sorgenti dati, un processo di *data mining* comprende la fase cardine in cui viene scelta la strategia che meglio si adatta all'obiettivo dell'analisi. Questa scelta non riguarda soltanto la tecnica o la combinazione di tecniche da utilizzare, ma comprende anche i modi in cui tali tecniche dovranno essere implementate ed applicate ai dati.
Questa fase è particolarmente delicata poiché, mentre in alcuni casi la scelta della tecnica da utilizzare può risultare piuttosto scontata, in molti altri casi è invece difficile scegliere a priori una tecnica rispetto ad un'altra e quindi spesso conviene provare diverse alternative al fine di costruire il miglior modello possibile per il problema in esame.

Valutazione dei risultati

Prima di procedere all'impiego del modello o dei modelli costruiti, è molto importante valutare il modello e i passi eseguiti per costruirlo, accertarsi che attraverso tale modello si possano veramente raggiungere obiettivi di business e capire se qualcosa di importante non è stato sufficientemente considerato nella costruzione del modello.
Dal momento che i risultati ottenuti da ogni processo di *data mining* possono generare una mole di informazioni che a volte risulta molto difficile da interpretare, è molto importante che durante questa fase, a supporto dell'esperto di *data mining*, intervenga anche un esperto di business in grado di tradurre i risultati del mining nei termini del contesto di business. Poiché è improbabile che l'esperto di business sia anche un esperto di *data mining*, è fondamentale che i risultati vengano presentati in una forma chiara e facilmente comprensibile.

Utilizzo dei risultati

È la fase finale che prevede l'utilizzo nella pratica del modello o dei modelli creati e valutati che possono permettere il raggiungimento dei fini desiderati. Questa fase è particolarmente cruciale poiché l'utilizzo che si farà del modello o dei modelli darà modo di sfruttare molto o poco il potenziale dei risultati proposti dall'analisi di *data mining* condotta.

1.4 Itemset frequenti e regole associative: stato dell'arte

Molte aziende accumulano un'enorme quantità di dati a partire dalle loro operazioni giornaliere. Per esempio, nelle casse dei negozi della piccola e della grande distribuzione sono raccolti giornalmente una grande quantità di dati riguardanti gli acquisti dei clienti. Questi dati sono comunemente noti come *market basket transactions*. Ciascuna riga in una tabella di questo tipo corrisponde ad una transazione, la quale contiene un unico identificatore chiamato *TID* e un insieme di item comprati da uno specifico cliente. I rivenditori sono interessati nell'analizzare i dati per imparare qualcosa in più circa il comportamento d'acquisto dei loro clienti. Tale conoscenza può poi essere usata a supporto di una grande varietà di applicazioni di business quali le promozioni, l'amministrazione del rapporto con i clienti e la disposizione dei prodotti sugli scaffali.
L'*analisi associativa* è quindi utile per scoprire relazioni interessanti ma nascoste in data set di grandi dimensioni. Le relazioni scoperte possono essere rappresentate nella forma di *regole associative* o di insiemi di item frequenti. Per esempio, una regola di questo tipo può essere estratta da un qualsiasi data set transazionale:

$$\{Pannolini\} \rightarrow \{Birra\}.$$

La regola suggerisce che esiste una forte relazione tra la vendita dei pannolini e la birra poiché molti clienti che comprano pannolini comprano anche la birra. I rivenditori possono usare questo tipo di regole per identificare nuove opportunità per operare con tecniche di *cross-selling* sui loro prodotti.
Esistono due problemi chiave che necessitano di essere affrontati quando l'estrazione delle regole associative viene applicata a database transazionali. Per prima cosa, la scoperta di pattern da un data set di grandi dimensioni può essere computazionalmente dispendiosa. Secondo, alcuni dei pattern scoperti possono essere potenzialmente spuri perché essi possono accadere semplicemente per caso.

1.4.1 Definizione del problema

I dati transazionali possono essere rappresentati in formato binario; in questo caso ogni riga corrisponde ad una transazione ed ogni colonna corrisponde ad un item. Un item può essere trattato come una variabile binaria il cui valore è uno se l'item in questione è presente in una transazione, oppure zero se non lo è. Poiché la presenza di un item in una transazione è solitamente considerata più importante della sua assenza, un item è una variabile binaria *asimmetrica*.
Sia $I = \{i_1, i_2, ..., i_m\}$ l'insieme di tutti gli *item* presenti in un data set transazionale e sia $D = \{t_1, t_2, ...t_N\}$ l'insieme di tutte le transazioni. Ciascuna transazione t_i contiene un sottoinsieme di item scelti da I. Nell'analisi associativa una collezione di zero o più item è detta *itemset*. Se un itemset contiene k item, allora è detto k-itemset. Per esempio, {Birra, Pannolini, Latte} è un esempio di un 3-itemset.
L'ampiezza di una transazione è definita come il numero di item presenti in quella stessa transazione. Una transazione t_j è detta contenere un itemset X se X è un sottoinsieme di t_j.
Definizione 1: *Per un itemset $X \subseteq I$ il supporto di X è il numero di*

transazioni di D contenenti X.

Definizione 2: *Una regola associativa r è un costrutto $X \rightarrow Y$, dove X e Y sono insiemi disgiunti, cioè $X \cap Y = 0$.*

La forza di una regola associativa può essere misurata in termini di *supporto* e *confidenza*. Il supporto determina quanto spesso una regola è applicabile ad uno specifico data set, mentre la confidenza determina quanto frequentemente gli item in Y appaiano nelle transazioni che contengono X. La definizione formale di queste metriche è la seguente:

$$supporto(r) = \frac{supporto(X \cap Y)}{N}$$
$$confidenza(r) = \frac{supporto(X \cap Y)}{supporto(X)}$$

Il supporto è una misura molto importante poiché una regola che ha un basso supporto può occorrere semplicemente per caso. Una regola con supporto basso è anche probabile che sia non interessante da una prospettiva di business perché può non essere profittevole costruire delle promozioni su degli item che i clienti raramente comprano insieme. Per queste ragioni, il supporto è spesso usato per eliminare le regole non interessanti.

La confidenza, d'altro canto, misura l'affidabilità della deduzione fatta attraverso una regola. Per una data regola $X \rightarrow Y$, più alta è la confidenza, più è probabile che Y sia presente nelle transazioni che contengono X. La confidenza fornisce inoltre una stima della probabilità condizionale di Y dato X.

I risultati dell'analisi associativa dovrebbero essere comunque interpretati con cautela. La deduzione espressa tramite una regola associativa non implica necessariamente causalità. Al contrario suggerisce una forte relazione di co-occorrenza tra gli item presenti nell'antecedente e nel conseguente della regola.

Il problema riguardante l'estrazione delle regole associative, introdotto nel 1993 da Agrawal, Imieilinski e Swami, tre ricercatori dell'IBM, può quindi essere formulato come segue:

INPUT: è formato da una quadrupla (I, D, σ, γ) con $1 \leq \sigma \leq |D|$ e $0 \leq \gamma \leq 1$ dove I è l'insieme di tutti gli item, D è la collezione di transazioni su I, mentre σ e γ sono rispettivamente il supporto e la confidenza minimi che le regole che verranno generate dovranno rispettare.

OUTPUT: è costituito da tutte le regole r che rispetto alla collezione D hanno $supporto(r) \geq \sigma$ e $confidenza(r) \geq \gamma$. Per quanto riguarda le regole generate si possono anche porre delle condizioni sintattiche addizionali da rispettare riguardanti gli item che possono comparire. Per esempio potremmo essere interessati solo a regole che hanno un determinato item solo nella parte destra (o sinistra) dell'implicazione.

L'esempio che i tre ricercatori utilizzano per illustrare l'origine del problema è un caso di *Market Basket Analysis*: abbiamo cioè a disposizione un grande database di acquisti fatti dai clienti di un negozio (o di una catena di negozi). L'obiettivo è quello di ricavare dal database tutte le regole associative 'importanti' tra insiemi di articoli venduti nel negozio stesso.

Quindi una regola del tipo $X \rightarrow Y$ sta a significare che clienti che comprano gli articoli dell'insieme X con alta probabilità acquistano anche gli articoli dell'insieme Y.

Il *supporto* di una determinata regola dà misura della significatività statistica del campione di transazioni preso in esame, mentre la *confidenza* è una misura della verosimilità della regola, ovvero più alta la confidenza più alta è la probabilità che la regola venga verificata. Si noti che la motivazione per la condizione del supporto minimo che le regole devono avere deriva da ragioni commerciali. Se il supporto non è sufficientemente elevato, significa che la regola non ha un'importanza tale da essere presa in considerazione.

1.4.2 Le principali strategie per l'estrazione di regole associative: Apriori e FP-growth

Il problema dell'estrazione delle regole associative è stato così formulato:

trovare in una collezione di transazioni D tutte le regole r aventi $supporto(r) \geq \sigma$ *e* $confidenza(r) \geq \gamma$.

Un approccio poco intelligente per risolvere questo problema potrebbe essere quello di calcolare il supporto e la confidenza per ogni possibile regola. Questo approccio è altamente dispendioso poiché esiste un numero esponenziale di regole che può essere estratto da un data set. Per evitare di condurre calcoli non necessari, sarebbe quindi utile scartare le regole prima di calcolare i loro valori di supporto e confidenza. Un primo passo in avanti per migliorare le performance degli algoritmi per l'estrazione delle regole associative è quello costituito dal dividere in due sottoproblemi distinti il problema del calcolo del supporto e della confidenza di una regola. Quindi, gli algoritmi noti per l'estrazione di regole associative sono solitamente articolati in due passi:

1. Generazione di tutti gli itemset frequenti: *un itemset X è detto frequente se il* $supporto(X) \geq \sigma$;

2. $\forall$ itemset frequente X, generazione di tutte le regole $r = (X - Y \Rightarrow Y)$, con $Y \subset X$, e tali che $confidenza(r) \geq \gamma$.

Agrawal e Srikant nel 1994 notarono però che le regole associative forti (cioè quelli aventi i valori di supporto e confidenza maggiori o uguale alle soglie di supporto minimo e confidenza minima, rispettivamente σ e γ) sono composte da *itemset frequenti* sia nella proposizione antecedente, sia nella conseguente, nella struttura delle regole. Il seguente enunciato diventa quindi determinante per generare in modo computazionalmente efficiente itemset frequenti.

Principio Apriori. Se un k-itemset (un itemset composto da k item) è frequente, allora qualsiasi suo sottoinsieme (w-itemset, con $w < k$) è, a sua volta, frequente.

Il principio Apriori definisce un criterio di valutazione degli itemset frequenti: se un insieme di cardinalità k ha un supporto superiore alla soglia minima σ, ne consegue che tutti gli itemset di cardinalità inferiore, estraibili dal k-itemset frequente, hanno implicitamente garantito il rispetto del vincolo di supporto minimo. Se invece un k-itemset non è frequente, nulla si può inferire circa la frequenza degli itemset di cardinalità inferiore da esso generabili.

Se il principio Apriori garantisce la relazione logica:

k-itemset I è frequente $\rightarrow$ qualsiasi w-itemset, contenuto in I, con $w \leq k$, è frequente.

è possibile invertire tale relazione, negando antecedente e conseguente, formulando la seguente:

w-itemset I, con $w \leq k$, non è frequente $\rightarrow$ qualsiasi k-itemset, contenente I, non è frequente.

Dato un itemset I_w non frequente di cardinalità w, qualsiasi k-itemset I_k ottenibile da I_w, aggiungendo elementi a quelli presenti in $I_w (k > w)$, è non frequente, in quanto:

$$supporto(I_k) = \frac{K(I_k)}{m} < supporto(I_w) = \frac{K(I_w)}{m} < \sigma$$

Ciò è conseguenza dell'evidenza empirica: la probabilità di trovare, nelle m transazioni contenute in un dataset, un itemset di cardinalità $k > w$ è inferiore o uguale alla probabilità di trovare un w-itemset di dimensione inferiore a k.

L'algoritmo derivato dal principio Apriori permette di eliminare automaticamente, 'a priori', tutti gli itemset di cardinalità superiore a quella di ciascun

itemset non frequente, senza doverli esplicitamente generare e senza doverne calcolare il supporto.
L'algoritmo Apriori rappresenta una tecnica efficiente, dal punto di vista elaborativo, per la generazione di regole associative forti. Esso è composto da due fasi, strettamente sequenziali:

1. *fase 1*: generazione degli itemset frequenti;

2. *fase 2*: generazione delle regole associative forti.

La *fase 1* dell'algoritmo si occupa della generazione di tutti gli itemset frequenti da un insieme di m transazioni contenenti al massimo N prodotti, cioè dei k-itemset, $k = 1, ..., N$, il cui supporto è superiore alla soglia minima σ, definita in fase di parametrizzazione dell'algoritmo.
A partire da tutti gli 1-itemset frequenti, l'algoritmo genera, sulla base del principio Apriori, tutti i possibili 2-itemset derivabili da ciascun 1-itemset; per ciascuno di essi viene calcolato il supporto. I 2-itemset non frequenti vengono eliminati; da quelli frequenti di cardinalità 2, si procede quindi alla generazione ed alla valutazione dei 3-itemset frequenti. Si prosegue in modo iterativo, arrestandosi non appena si giunge alla k-esima iterazione, in corrispondenza della quale tutti i k-itemset generati dai $(k-1)$-itemset frequenti risultano non frequenti.
La *fase 2* ha l'obiettivo di generare le regole associative forti, partendo dagli itemset frequenti determinati durante la *fase 1*, preventivamente inseriti in una lista L. Per ciascun k-itemset frequente, si procede alla generazione di tutte le possibili regole da esso estraibili, mediante combinazioni di oggetti che compongono l'antecedente ed il conseguente della regola. Da un k-itemset possono essere estratte sino a 2^{k-2} regole associative.
Per ciascuna regola generata, i due itemset che compongono antecedente e conseguente sono frequenti, così come l'itemset che si ottiene dall'unione di antecedente e conseguente.

La *fase 2* dell'algoritmo Apriori calcola la confidenza per ciascuna regola associativa generata; solo le regole che soddisfano la condizione di confidenza minima γ vengono selezionate e inserite nella lista delle *regole forti*.
In molti casi, l'algoritmo Apriori riduce significativamente il numero degli itemset frequenti candidati facendo uso del principio da cui prende il nome. Su di esso possono comunque pesare due costi di non poco conto: può generare un insieme di itemset candidati molto grande e scorrere ripetute volte il database (in particolare sono necessarie $(n + 1)$ scansioni, dove n è la lunghezza del pattern più lungo). Per queste due ragioni Han e altri (2000) hanno proposto una strategia alternativa, radicalmente diversa da quella di Apriori, in grado di determinare gli itemset frequenti senza generare alcun candidato. Tale strategia prende il nome di *FP-Growth*, cioè *Frequent Pattern Growth*.
FP-Growth lavora con una metodologia *divide and conquer*. La prima scansione del database genera una lista di item frequenti in cui gli item sono ordinati in ordine di frequenza discendente. In base a questa lista discendente di frequenze, il database viene compresso in un *albero dei pattern frequenti*, o *FP-tree*, il quale memorizza tutte le informazioni riguardo gli itemset frequenti. L'*FP-tree* è estratto partendo da ciascun pattern di lunghezza 1 (come se fosse un suffisso), costruendo la sua *conditional pattern base* (un 'sottodatabase' che consiste di un insieme di cammini di prefissi nell'*FP-tree* che occorrono insieme ai suffissi), costruendo poi il suo corrispondente *FP-tree condizionale*, e, infine, eseguendo ricorsivamente l'algoritmo su tale albero. I pattern frequenti sono ottenuti attraverso la concatenazione dei suffissi con i pattern frequenti generati dall'*FP-tree condizionale*.
L'algoritmo *FP-Growth* trasforma il problema di trovare itemset frequenti lunghi in quello di cercare itemset frequenti più corti in maniera ricorsiva e poi concatenando i suffissi. Esso usa gli itemset meno frequenti come se fossero un suffisso, offrendo così buona selettività. Studi effettuati sulle

performance dell'algoritmo dimostrano che questo metodo riduce sostanzialmente i tempi di ricerca.

1.4.3 Estrarre gli itemset frequenti chiusi e gli itemset frequenti massimali

Una delle controindicazioni nell'estrarre gli itemset frequenti con supporto minimo molto basso a partire da un data set di grandi dimensioni è il fatto che tale estrazione spesso genera un notevole numero di itemset che soddisfano la soglia minima di supporto. Questo accade perché se un itemset è frequente, lo sono anche tutti i suoi sotto-itemset. Un itemset numeroso conterrà così un numero esponenziale di sotto-itemset più piccoli, ma ugualmente frequenti. Per superare questo problema è stata proposta *l'estrazione di itemset frequenti chiusi* e *l'estrazione di pattern frequenti massimali.*
Un itemset X è un *itemset frequente massimale* in un dataset D se X è frequente e non esiste alcun sovra-itemset Y di X tale che $X \subset Y$ con Y frequente in D. Un itemset X è invece un *itemset frequente chiuso* in un dataset D se X è frequente in D e non esiste alcun sovra-itemset Y di X tale che Y ha lo stesso supporto di X (se consideriamo soltanto questa seconda condizione diremo che un itemset è *chiuso*).

[6]Per illustrare il primo di questi due concetti prendiamo in considerazione il reticolo degli itemset mostrato dala figura 1.3. Gli itemset nel reticolo sono divisi in due gruppi: quelli frequenti e quelli non frequenti. Il confine che separa gli itemset frequenti da quelli non frequenti è rappresentato dalla riga tratteggiata. Ciascun itemset localizzato sopra il confine è frequente, mentre quelli localizzati sotto il confine (i nodi colorati di grigio) sono non frequenti. Tra gli itemset che risiedono vicino al confine, $\{a, d\}$, $\{a, c, e\}$ e $\{b, c, d, e\}$ sono considerati itemset frequenti massimali poiché i loro immediati sovrainsiemi sono non frequenti. Un itemset come {a, d} è frequente massimale

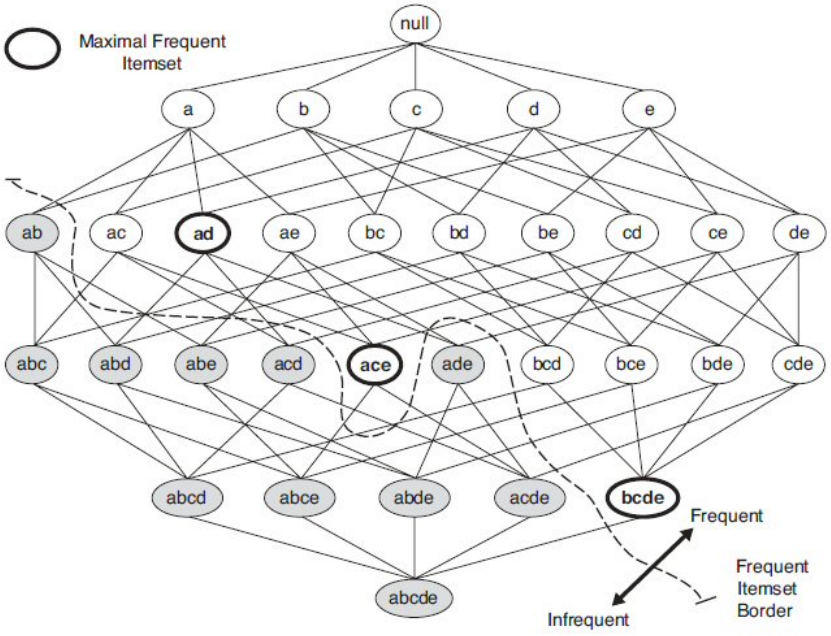


Figura 1.4: [6]Gli itemset frequenti massimali.

perché tutti i suoi immediati sovrainsiemi, $\{a, b, d\}$, $\{a, c, d\}$ e $\{a, d, e\}$, sono non frequenti. Al contrario, $\{a, c\}$ non è massimale perché uno dei suoi immediati sovrainsiemi, $\{a, c, e\}$, è frequente.

Gli itemset frequenti massimali producono effettivamente una rappresentazione compatta degli itemset frequenti. In altre parole, essi formano il più piccolo insieme di itemset a partire dai quali tutti gli itemset frequenti possono essere derivati. Per esempio, gli itemset frequenti mostrati nella figura 1.3 possono essere divisi in due gruppi:

- gli itemset frequenti che cominciano con l'item a e che possono contenere gli item c, d o e. Questo gruppo comprende itemset come $\{a\}$, $\{a, c\}$, $\{a, d\}$, $\{a, e\}$ e $\{a, c, e\}$;

- gli itemset frequenti che cominciano con gli item b, c, d o e. Questo gruppo comprende itemset come $\{b\}$, $\{b, c\}$, $\{c, d\}$, $\{b, c, d, e\}$, eccetera.

Gli itemset frequenti che rientrano nel primo gruppo possono essere sottoinsiemi sia di $\{a, c, e\}$ che di $\{a, d\}$, mentre quelli che rientrano nel secondo gruppo sono sottoinsiemi di $\{b, c, d, e\}$.

Nonostante producano una rappresentazione compatta, gli itemset frequenti massimali non contengono l'informazione relativa al supporto dei loro sottoinsiemi. Per esempio, il supporto degli itemset frequenti massimali $\{a, c, e\}$, $\{a, d\}$ e $\{b, c, d, e\}$ non fornisce alcun suggerimento circa il supporto proprio dei loro sottoinsiemi. È quindi necessaria un'ulteriore scansione del data set per determinare il supporto degli itemset frequenti non massimali.

[6]Per meglio illustrare il secondo dei due concetti espressi sopra, facciamo

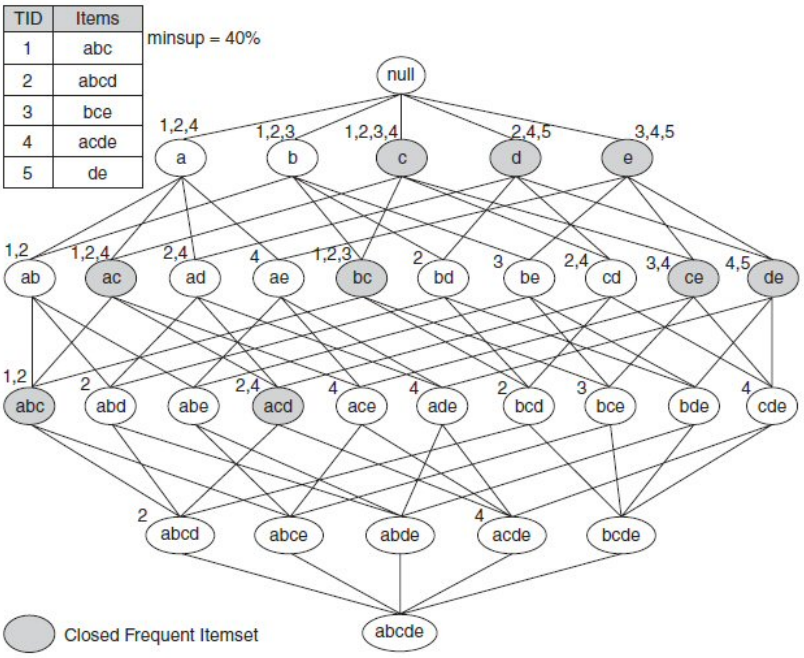


Figura 1.5: [6]Un esempio di itemset frequenti chiusi (con una soglia di supporto minimo uguale al 40%).

riferimento alla figura 1.4. In questa, per rendere più chiaro il supporto di ciascun itemset, abbiamo associato ad ogni nodo (itemset) nel reticolo una

lista degli *ID* delle transazioni che gli corrispondono. Per esempio, dal momento che il nodo $\{b, c\}$ è associato con gli *ID* delle transazioni 1, 2 e 3, il suo supporto è uguale a tre. Se assumiamo che la soglia minima di supporto è del 40%, $\{b, c\}$ è un itemset frequente chiuso poiché il suo supporto è pari al 60%. Il resto degli itemset frequenti chiusi sono indicati dai nodi colorati di grigio.

Gli itemset frequenti chiusi possono essere usati per determinare il supporto per quelli frequenti non chiusi. Per esempio, consideriamo l'itemset frequente $\{a, d\}$ mostrato nella figura 1.4. Dal momento che questo itemset non è chiuso, il suo supporto deve essere identico a quello di uno dei suoi immediati sovrainsiemi. La chiave è di determinare quale sovrainsieme (tra $\{a, b, d\}$, $\{a, c, d\}$, o $\{a, d, e\}$) ha esattamente lo stesso supporto di $\{a, d\}$. Il principio di Apriori stabilisce che qualsiasi transazione che contenga il sovrainsieme di $\{a, d\}$ deve anche contenere $\{a, d\}$. Comunque, qualsiasi transazione che contenga $\{a, d\}$ non deve contenere i sovrainsiemi di $\{a, d\}$. Per questa ragione, il supporto di $\{a, d\}$ deve essere uguale al supporto più grande che esiste tra quelli dei suoi sovrainsiemi. Dal momento che $\{a, c, d\}$ ha un supporto più grande rispetto sia a $\{a, b, d\}$ che a $\{a, d, e\}$, il supporto di $\{a, d\}$ deve essere identico al supporto di $\{a, c, d\}$. Usando questa metodologia, può essere sviluppato un algoritmo per calcolare il supporto per gli itemset frequenti non chiusi.

L'estrazione degli itemset frequenti a partire dagli itemset frequenti chiusi è stata proposta nel 1999 [7] ed è stato presentato un algoritmo basato sulla proprietà di Apriori chiamato *A-Close*. Altri algoritmi per l'estrazione di itemset frequenti chiusi sono: [8]*CLOSET*, [9]*CHARM*, [10]*CLOSET+*, [11]*FPClose* e [12]*AFOPT*.

L'estrazione degli itemset frequenti chiusi fornisce quindi un'alternativa valida ed interessante all'estrazione degli itemset frequenti, dal momento che ne eredita lo stesso potere analitico ma genera un numero notevolmente minore

di risultati. Con l'estrazione degli itemset frequenti chiusi è anche garantita una maggiore scalabilità ed una maggiore interpretabilità dei risultati.

Questo punto può essere facilmente spiegato con un esempio [6]. Con-

TID	a_1	a_2	a_3	a_4	a_5	b_1	b_2	b_3	b_4	b_5	c_1	c_2	c_3	c_4	c_5
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
5	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
6	0	0	0	0	0	1	1	1	1	1	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1

Figura 1.6: [6]Un data set transazionale per estrarre gli itemset chiusi.

sideriamo la tabella della figura 1.5, la quale contiene dieci transazioni e quindici item. Gli item possono essere divisi in tre gruppi: (1) Gruppo A, che contiene gli item che vanno da a_1 a a_5; (2) Gruppo B, che contiene gli item che vanno da b_1 a b_5; e (3) Gruppo C, il quale contiene gli item che vanno da c_1 a c_5. Va notato che gli item contenuti in ciascun gruppo sono in perfetta associazione l'uno con l'altro e non appaiono insieme ad item facenti parte degli altri gruppi. Assumendo che la soglia minima di supporto è del 20%, il numero totali degli itemset frequenti è $3x(2^5 - 1) = 93$. Comunque, ci sono sono tre itemset frequenti chiusi nei dati: ($\{a_1, a_2, a_3, a_4, a_5\}$, $\{b_1, b_2, b_3, b_4, b_5\}$, e $\{c_1, c_2, c_3, c_4, c_5\}$). Spesso è sufficiente e preferibile presentare solo gli itemset frequenti chiusi agli analisti al posto dell'intero insieme degli itemset frequenti, poiché i risultati sono più chiari e facili da interpretare. L'estrazione degli itemset frequenti massimali è stata studiata da *MaxMiner* [13], un metodo di ricerca basato su Apriori, *level-wise*, *breadth-first*, creato per trovare gli itemset frequenti massimali eseguendo il *pruning* dei sovrainsiemi frequenti e dei sottoinsiemi

non frequenti, riducendo così lo spazio di ricerca. Un altro metodo efficiente per trovare gli itemset frequenti massimali è *MAFIA* [14].

1.4.4 Le regole associative multilivello e multidimensionali

Supponiamo che, invece di usare un database transazionale, le vendite e le relative informazioni siano memorizzate in un data warehouse, il quale è, per definizione, *multidimensionale*. Per esempio, oltre a tenere traccia degli item presenti nelle transazioni, un data warehouse può memorizzare altri attributi associati agli item, quali la quantità o il prezzo dei prodotti acquistati. Possono essere inoltre memorizzate altre informazioni che riguardano il cliente che ha fatto gli acquisti, quali la sua età, la sua occupazione, il suo reddito ed il suo indirizzo. Se consideriamo ciascuna dimensione del data warehouse come un predicato, possiamo estrarre regole associative che contengono più di un predicato come quella che segue:

$$\{Età \in [20, 29) \wedge Occupazione = Studente\} \rightarrow \{Compra = laptop\}$$

Le regole associative che coinvolgono due o più dimensioni o predicati sono dette *regole associative multidimensionali*. La regola vista sopra contiene tre predicati (*età*, *occupazione* e *compra*), ciascuno dei quali compare solo una volta nella regola.

Bisogna ricordare che gli attributi dei database possono essere *categorici* o *quantitativi*. I primi hanno un numero finito di possibili valori, senza che questi valori siano ordinati. Gli attributi categorici sono anche detti *nominali* poiché i loro valori sono effettivamente i 'nomi delle cose'. Gli attributi quantitativi sono numerici e i loro valori sono implicitamente ordinati. Le tecniche per estrarre le regole associative multidimensionali possono essere divise in due categorie riguardo il trattamento da riservare agli attributi quantitativi.

Nel primo approccio, gli attributi quantitativi sono discretizzati usando delle gerarchie di concetti predefinite.

Nel secondo approccio, gli attributi quantitativi sono discretizzati o raggruppati basandosi sulla distribuzione di questi ultimi nei dati.
Nelle applicazioni dove gli item formano una gerarchia può essere difficile trovare regole associative forti a bassi livelli di astrazione a causa della sparsità dei dati in uno spazio multidimensionale. Le regole associative forti possono solitamente essere trovate ai livelli più alti della gerarchia, ma esse spesso rappresentano una conoscenza di tipo generale, se non già nota. Per esempio, con riferimento alla figura 1.6, la regola $milk \rightarrow bread$ è probabile che abbia un supporto molto alto, ma è banale. Allo stesso tempo, la regola $skim\_milk \rightarrow large\_white\_bread$ può essere utile, ma può avere un supporto basso. Un algoritmo dovrebbe essere in grado di generare le regole

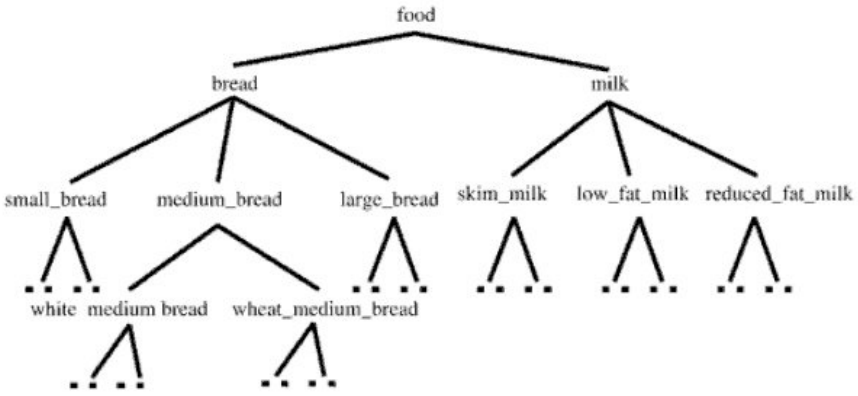


Figura 1.7: [15]Un esempio di una gerarchia di concetti riguardante il cibo.

e di traversare tra esse a differenti livelli di astrazione. Le *regole associative multilivello* sono generate eseguendo una ricerca dall'alto verso il basso (*top-down*) e di approfondimento iterativo (*iterative deepening*). Per parlare in maniera più semplice, prima troviamo le regole associative forti ai livelli più alti della gerarchia e, in un secondo momento, cerchiamo le regole meno forti ai livelli più bassi. Per esempio, prima generiamo la regola $milk \rightarrow bread$ e poi ci concentriamo nel trovare le regole che riguardano il pane (*bread*) di

diverse dimensioni (*small*, *medium* e *large*) e il latte (*milk*) a diverso contenuto di grasso (*skim*, *low_fat* e *reduced_fat*).
Esistono due famiglie di metodi per estrarre le regole associative multilivello:

- i metodi basati sul *supporto uniforme*, dove è usata la stessa soglia minima di supporto per generare le regole a tutti i livelli di astrazione. In questo caso la ricerca delle regole è semplificata, dal momento che possiamo assumere con certezza che gli itemset che contengono item i cui antenati nella gerarchia non soddisfano la soglia minima di supporto sono ugualmente non frequenti. Allo stesso tempo, è altamente improbabile che gli item ai livelli più bassi di astrazione occorrano tanto frequentemente quanto quelli ai livelli più alti. Conseguentemente, se la soglia minima di supporto è troppo alta, l'algoritmo non troverà associazioni potenzialmente utili ai livelli più bassi di astrazione. D'altro canto, se la soglia è troppo bassa, potrebbero essere estratte ai livelli più alti un grande numero di associazioni potenzialmente non interessanti;

- i metodi basati sul *supporto ridotto*, un approccio che evidenzia le controindicazioni nell'uso del supporto uniforme. In questo caso, ciascun livello di astrazione è provvisto della sua propria soglia minima di supporto. Più basso è il livello di astrazione, minore è la corrispondente soglia di supporto.

Una volta che le regole associative multilivello sono state generate, alcune di esse possono essere ridondanti a causa della relazione antenati, genitori e figli tra gli item. Una regola $X \rightarrow Y$ è ridondante se esiste una regola più generale $\widehat{X} \rightarrow \widehat{Y}$, dove $\widehat{X}$ è un antenato di X, $\widehat{Y}$ è un antenato di Y, ed entrambe le regole hanno una confidenza molto simile. Le regole che coinvolgono gli item appartenenti ai livelli più bassi della gerarchia sono considerate ridondanti poiché esse possono essere riassunte da delle regole

che coinvolgono gli item antenati. Fortunatamente, data la conoscenza della gerarchia, è facile eliminare tali itemset ridondanti durante la generazione degli itemset frequenti.

1.4.5 Le regole associative quantitative

La maggior parte delle ricerche nell'ambito del *data mining* si sono focalizzate sulla scoperta delle regole associative *booleane*, ovvero, per esempio:

Se un cliente compra uno o più spazzolini, allora comprerà anche il dentifricio.

Alla quale comunemente ci si riferisce con:

$$Spazzolino \rightarrow Dentifricio.$$

Le regole associative booleane non prendono però in considerazione la quantità degli item acquistati in una transazione. Molti autori hanno indicato che le *regole quantitative* sono importanti per alcune applicazioni, e che il processo di estrazione di tali regole necessita ancora di molto lavoro. Un esempio di una regola quantitativa è la seguente:

Se un cliente compra due, tre, o quattro spazzolini, allora comprerà tra i tre ed i sei tubetti di dentifricio.

La quale può essere espressa in questa forma:

$$Spazzolino : [2, 4] \rightarrow Dentifricio : [3, 6],$$

ed è chiamata *regola associativa quantitativa* [16]. Le regole associative contengono attributi continui ed esistono tre metodi per trattare questi ultimi: (1) i metodi basati sulla discretizzazione, (2) i metodi statistici e (3) i metodi non basati sulla discretizzazione. Le regole associative che vengono estratte usando ognuno di questi metodi differiscono abbastanza l'una dall'altra.

I metodi basati sulla discretizzazione

La discretizzazione costituisce l'approccio più comune per trattare gli attributi continui. Questo approccio raggruppa i valori adiacenti di un attributo continuo in un numero finito di intervalli. Per esempio, se abbiamo a che fare con un attributo *Età* che comprende valori che vanno dai 12 ai 60 anni, questo può essere diviso nei seguenti intervalli:

$$Età \in [12, 16),\ Età \in [16, 20),\ Età \in [20, 24),\ Età \in [56, 60),$$

dove $[a, b)$ rappresenta un intervallo che include a, ma non b. La discretizzazione può essere effettuata utilizzando diverse tecniche quali l'*equal interval width* [6], l'*equal frequency* [6], l'*entropy-based* [6], o il *clustering*. Gli intervalli discreti sono poi mappati su attributi binari asimmetrici cosicché possano essere applicati gli esistenti algoritmi di analisi associativa.
Un parametro chiave da tenere in considerazione nella discretizzazione di un attributo è il numero di intervalli usati per partizionarlo. Questo parametro è tipicamente fornito dagli utenti e può essere espresso in termini di ampiezza dell'intervallo (per l'approccio *equal interval width* [6]), di numero medio di transazioni per intervallo (per l'approccio *equal frequency* [6]), o di numero di gruppi desiderato (per l'approccio basato sul *clustering*).
Uno dei possibili problemi che si incontrano durante la discretizzazione di un attributo continuo è quello riguardante la definizione dell'ampiezza di ogni intervallo. Questo può portare ad altri problemi, perché:

1. Se l'intervallo è troppo ampio, allora potremmo perdere alcuni pattern interessanti semplicimente perché questi non hanno un valore di confidenza abbastanza alto;

2. Se l'intervallo è troppo stretto, allora potremmo perdere alcuni pattern a causa della loro mancanza di supporto.

Un modo per evitare questo genere di problemi è quello di considerare ogni possibile raggruppamento di intervalli adiacenti. Questo approccio permet-

te ovviamente di trovare tutte le regole associative quantitative forti ed interessanti, ma porta ai seguenti problemi computazionali:

1. **La computazione diventa estratamente dispendiosa**. Se la serie di valori è inizialmente divisa in k intervalli, allora devono essere generati $k(k-1)/2$ item binari per rappresentare tutti i possibili intervalli [6]. Inoltre, se un item corrispondente all'intervallo $[a, b)$ è frequente, allora tutti gli altri item che corrispondono agli intervalli che sottointendono $[a, b)$ devono essere ugualmente frequenti. Questo approccio può quindi generare troppi candidati e itemset frequenti. Per risolvere questi problemi, può essere usata una soglia massima di supporto in modo da prevenire la creazione di item che corrispondono ad intervalli troppo ampi e in modo da ridurre il numero degli itemset;

2. **Sono estratte molte regole ridondanti**. Se i valori della confidenza delle due regole estratte sono i medesimi, allora dovrebbe essere mantenuta la regola più generale, poiché copre molti esempi, inclusi quelli coperti dalla regola più specialistica.

I metodi statistici

Le regole associative quantitative possono essere usate per dedurre le proprietà statistiche di una popolazione. Per esempio potremmo voler estrarre da un data set una regola associativa quantitativa come questa:

$$\{RedditoAnnuale > 100K, Acquisti\ Online = S\grave{i}\} \rightarrow Età : Media = 38\ [6].$$

La regola afferma che l'età media degli utenti Internet il cui reddito annuale supera i $\$100K$ e che comprano abitualmente online è di 38 anni.
Per generare le regole associative quantitative statistiche deve essere specificato l'attributo target usato per caratterizzare i segmenti interessanti della popolazione. Escludento l'attributo target, i rimanenti attributi categorici

e continui presenti nei dati sono trasformati nel formato binario. A questo punto degli algoritmi già esistenti quali Apriori o FP-growth possono essere applicati per estrarre gli itemset frequenti a partire dai dati binarizzati. Ciascun itemset frequente identifica un segmento interessante della popolazione. La distribuzione dell'attributo target in ciascun segmento può essere riassunta usando tecniche proprie della statistica descrittiva, quali la media, la mediana, la varianza, o la deviazione assoluta. Per esempio, la regola scritta sopra è ottenuta facendo la media dell'età degli utenti Internet che supportano l'itemset frequente $\{RedditoAnnuale > 100K, Acquisti\ Online = S\grave{i}\}$. Il numero delle regole associative scoperto usando questo metodo è pari al numero di itemset frequenti trovati. A causa del modo in cui le regole associative quantitative sono definite, la nozione di confidenza non è applicabile a tali regole per cui è necessario trovare un metodo alternativo per validarle. Una regola associativa quantitativa è interessante solo se le statistiche calcolate a partire dalle transazione coperte dalle regola sono differenti rispetto a quelle calcolate a partire dalle transazioni non coperte dalla regola. Per esempio, la regola vista sopra è interessante solo se l'età media degli utenti Internet che non supporta l'itemset frequente $\{RedditoAnnuale > 100K, Acquisti\ Online = S\grave{i}\}$ è significativamente più alta o più bassa di 38 anni. Per determinare se la differenza nella media delle età è statisticamente significativa, devono essere applicati metodi di test di ipotesi statistica [6].

I metodi non basati sulla discretizzazione

Esistono talune applicazioni in cui gli analisti sono più interessati nel trovare le associazioni tra gli attributi continui, rispetto alle associazioni tra intervalli discreti degli attributi continui. Per esempio, consideriamo il problema di trovare le associazioni tra le parole presenti nei documenti di testo. Consideriamo di avere a che fare con una matrice, in cui ciascuna riga fa riferimento a un documento e in cui ciascuna colonna fa riferimento ad

ogni parola presente in quei documenti; in questa stessa matrice ciascuna voce rappresenta la frequenza normalizzata di una specifica parola presente in un dato documento. I dati sono normalizzati dividendo la frequenza di ciascuna parola per la somma della frequenza di quella stessa parola in tutti i documenti. Una delle ragioni di questa normalizzazione è che così siamo sicuri che il valore del supporto risultante sia un numero compreso tra 0 ed 1.

Nell'ambito del *text mining*, gli analisti sono più interessati nel trovare le associazioni tra le parole (per esempio, *data* e *mining*) rispetto alle associazioni tra gamme di frequenze di parole (per esempio, $data \in [1, 4]$ e $mining \in [2, 3]$. Un modo per fare questo è quello di trasformare i dati in una matrice 0/1, dove una voce è 1 se la frequenza normalizzata eccede una qualche soglia t, e 0 altrimenti. Mentre questo approccio permette agli analisti di applicare gli esistenti algoritmi di estrazione degli itemset frequenti, trovare la giusta soglia per la binarizzazione può essere piuttosto insidioso. Se la soglia è troppo alta, è possibile escludere delle associazioni interessanti. Al contrario, se la soglia è troppo bassa, c'è la possibilità di generare un ampio numero di associazioni spurie.

Un metodo alternativo per trovare le associazioni tra parole è conosciuto come l'algoritmo min-*Apriori* [6].

1.5 Le regole classificative

La classificazione associativa (AC - Associative Classification) integra due compiti ben noti del *data mining*: la scoperta delle regole associative e la classificazione, con lo scopo di costruire un modello (un classificatore) predittivo.

La classificazione associativa costituisce un caso speciale di scoperta delle regole associative in cui solo l'attributo che riguarda la classe viene considerato come facente parte del lato destra della regola (il conseguente); per

esempio in una regola come $X \rightarrow Y$, Y deve assere un attributo di classe. Uno dei principali vantaggi derivanti dall'uso di una classificazione basata sulle regole associative al posto di un classico approccio classificativo è che l'output di un algoritmo di classificazione associativa è rappresentato da un semplice insieme di regole della forma *se-allora*, le quali rendono più facile per l'utente finale capire e interpretare il risultato.
Definiamo quindi il problema di classificazione associativa: abbiamo un data set di addestramento T, il quale ha m attributi distinti $A_1, A_2, ..., A_m$ e C è una lista di classi. Il numero delle righe in T è dato da $|T|$. Gli attributi possono essere categorici (ciò sta a significare che il loro valore fa parte di un insieme finito di possibili valori) o continui (cioè reali o interi). Nel caso si tratti di attributi categorici, tutti i loro possibili valori sono mappati su un insieme di interi positivi. Per gli attributi continui si procede invece alla discretizzazione di quest'ultimi.

Definizione 1 Una riga o un oggetto di addestramento in T può essere descritto come una combinazione di nomi di attributi A_i e valori a_{ij}, più una classe rappresentata da c_j.
Definizione 2 Un item può essere descritto come un nome di attributo A_i e un valore a_j, rappresentato da $\langle(A_j, a_j)\rangle$.
Definizione 3 Un itemset può essere descritto come un insieme di valori di attributi disgiunti contenuti in un oggetto di addestramento, rappresentato da $\langle(A_{i1}, a_{j1}),..., (A_{ik}, a_{jk})\rangle$.
Definizione 4 Un *ruleitem* r è nella forma $\langle itemset, c\rangle$ dove $c \in C$ è la classe.
Definizione 5 L'effettiva occorrenza (*actoccr*) di un *ruleitem* r in T è il numero di righe in T che combaciano con l'itemset di r.
Definizione 6 Il *supporto count* (*suppcount*) di un *ruleitem* r è il numero di righe in T che combaciano con l'itemset di r e appartengono alla classe c

di r.

Definizione 7 L'occorrenza di un itemset i (*occitm*) in T è il numero di righe in T che combaciano con i.

Definizione 8 Una regola associativa classificativa (*CAR - Classificative association rule*) è nella forma $(A_{i1}, a_{i1}) \wedge ... \wedge (A_{jk}, a_{jk}) \rightarrow c$ dove l'antecedente della regola è un itemset e il conseguente è una classe.

Lo scopo principale della classificazione associativa è quello di costruire un insieme di regole (un modello) che sia in grado di predire le classi di dati mai visti prima, conosciuti come data set di test, il più accuratamente possibile.

1.5.1 Uno schema risolutivo del problema

Lo scopo della classificazione associativa è bensì diverso da quello della scoperta di regole associative. La differenza più scontata tra le due è che la classificazione associativa considera soltanto la classe come conseguente delle regole. Inoltre, la prevenzione dell'*overfitting* è un problema essenziale nella classificazione associativa, mentre non lo è nella scoperta delle regole associative dal momento che la classificazione associativa usa un sottoinsieme delle regole scoperte per predire le classi di cui nuovi oggetti faranno parte. L'*overfitting* spesso occorre quando le regole scoperte si comportano bene sul data set di addestramento, ma male sul data set di test. Questo può essere dovuto a ragioni diverse quali la piccola quantità di oggetti di addestramento o il rumore.

Il problema di costruire un classificatore usando la classificazione associativa può essere diviso in quattro passi principali:

1. **Passo 1**: la scoperta di tutti i *ruleitem* frequenti;
2. **Passo 2**: la produzione di tutte le regole associative classificative aventi i valori di confidenza sopra la soglia minima di confidenza *minconf*

a partire dai *ruleitem* frequenti estratti al Passo 1;

3. **Passo 3**: la selezione di un sottoinsieme di regole associative classificative per formare il classificatore a partire dalla regole generate al Passo 2;

4. **Passo 4**: misurare la qualità del classificatore sugli oggetti di test.

1.5.2 Le tecniche per la scoperta dei ruleitem frequenti

Il passo che si occupa di trovare i *ruleitem* frequenti è particolarmente duro poiché necessita di una grande quantità di calcoli. Diversi differenti approcci per trovare i ruleitem frequenti sono stati 'rubati' dalla scoperta delle regole associative. Per esempio, alcuni metodi fanno uso del metodo di generazione dei candidati caratteristico di Apriori, altri adottano l'approccio di FP-growth, altri ancora estendono invece i metodi di intersezione delle liste di *tid* propri dei dati organizzati nel formato verticale.

La generazione dei candidati

L'algoritmo *CBA* è stato uno dei primi algoritmi di classificazione associativa ad usare la modalità di generazione dei candidati di Apriori per trovare le regole. Il collo di bottiglia di Apriori è però rappresentato dal trovare gli itemset frequenti a partire da tutti i possibili itemset candidati ad ogni livello. Le tecniche di classificazione associativa che utilizzano la procedura di generazione dei candidati di Apriori per trovare i *ruleitem* frequenti hanno generalmente un buon riscontro quando la dimensione dei *ruleitem* candidati è piccola. In circostanze che coinvolgono la classificazione di data set con molti attributi altamente correlati, e una soglia minima di supporto molto bassa, il numero potenziale dei *ruleitem* candidati ad ogni livello può diventare enorme e questi algoritmi possono consumare una notevole quantità di tempo e di memoria della CPU.

Per esempio, CBA richiede una completa scansione del data set di addestramento ad ogni livello per trovare i *ruleitem* candidati e quindi, per trovare i *ruleitem* candidati ad ogni livello, una fusione di tutte le possibili combinazioni di *ruleitem* frequenti trovati al livello precedente ed una totale scansione del data set di addestramento per aggiornare le frequenze dei *ruleitem* candidati. Questo processo di continua scansione del database è costoso riguardo ai tempi di esecuzione dell'algoritmo.
Inoltre, gli algoritmi di classificazione associativa sperimentano spesso una crescita esponenziale riguardo al numero delle regole. Questo è dovuto all'approccio usato per la scoperta delle regole, il quale esplora tutte le possibili associazioni tra i valori degli attributi presenti in un database. Questo problema è stato reso noto da varie ricerche, le quali rendono tutte chiaro che il numero così accresciuto delle regole può causare seri problemi quali l'*overfitting* del data set di addestramento e può portare ad errori di classificazione (che avvengono quando più regole presenti in un classificatore aventi diverse etichette di classe coprono un singolo oggetto presente nel database) così come ad alti requisiti per quanto riguarda la memoria e la CPU.
Per migliorare l'efficienza del passo di generazione dei candidati di Apriori, almeno tre metodi di classificazione associativa, *CMAR*, L^3 e L^3G, usano approcci basati sul metodo FP-growth per trovare le regole. Il metodo FP-growth costruisce un albero molto denso per il data set di addestramento, dove ciascun oggetto di addestramento è rappresentato da al più un cammino nell'albero. Come risultato, la lunghezza di ciascun cammino è uguale al numero di item frequenti nella transazione che rappresenta proprio quel cammino. Questo tipo di rappresentazione è molto utile per le seguenti ragioni. (1) Tutti gli itemset frequenti presenti in ciascuna transazione che fa parte del database originale sono rappresentati nell'FP-tree, e, dal momento che tra gli item frequenti vi è molto in comune, l'FP-tree è di dimensione più piccola rispetto al database originale. (2) La costruzione dell'FP-tree

richiede soltanto due scansioni del database; nella prima sono trovati gli itemset frequenti così come il loro supporto per ogni transazione e nella seconda, invece, viene costruito l'FP-tree.

Una volta che l'FP-tree è stato costruito, viene utilizzato un metodo *pattern growth* per trovare le regole utilizzando i pattern di lunghezza uno presenti nell'FP-tree. Per ciascun pattern frequente sono generati e memorizzati in un FP-tree condizionale tutti gli altri possibili pattern frequenti che co-occorrono con quest'ultimo nell'FP-tree. Il processo di estrazione delle regole avviene concatenando i pattern con quelli prodotti dall'FP-tree condizionale. Questo processo usato dall'algoritmo FP-growth non è come quello di Apriori poiché non vi è nessuna generazione delle regole candidate. Una delle principali debolezze del metodo FP-growth è che non c'è nessuna garanzia che l'FP-tree starà sempre nella memoria principale, specialmente nei casi in cui il database in esame è molto grande.

Gli algoritmi *CMAR*, L^3 e L^3G memorizzano le regole in un albero dei prefissi chiamato CR-tree. Il CR-tree è usato per memorizzare le regole in un ordine discedente secondo la frequenza dei valori degli attributi che appaiono nell'antecedente. Una volta che una regola è generata, sarà inserita nel CR-tree come un cammino che parte dal nodo radice, e il suo supporto, la sua confidenza e la sua classe di appartenenza saranno memorizzati nell'ultimo nodo del cammino. Quando una regola candidata ha delle caratteristiche in comune con una regola che è presente nell'albero, il cammino della regola già esistente è esteso per riflettere l'aggiunta della nuova regola.

Gli algoritmi che usano il CR-tree considerano i valori comuni degli attributi contenuti nelle regole, il che porta a consumare meno memoria in confronto all'algoritmo CBA. In più, le regole possono essere reperite efficientemente dal momento che il CR-tree le indicizza.

L'approccio greedy

Per ciascuna classe presente nel data set di addestramento, l'algoritmo *FOIL* costruisce euristicamente le regole a partire dagli item di addestramento, utilizzando il metodo *FOIL-gain*. Il metodo *FOIL-gain* misura l'informazione derivata dall'aggiungere una condizione alla regola in esame. Assumiamo che nel dataset di addestramento siano presenti $|P'|$ oggetti positivi e $|N'|$ oggetti negativi associati alla classe c. Gli oggetti positivi per c sono quegli oggetti di addestramento che contengono c, mentre gli oggetti negativi per c sono quegli oggetti di addestramento dove c non appare mai. *FOIL* comincia costruendo una regola per ciascuna classe (c) aggiungendo un item (i) al suo antecedente. Dopo aver aggiunto i, ci saranno $|P|$ oggetti di addestramento positivi e $|N|$ oggetti di addestramento negativi che combaciano con la regola in esame, cioè con $i \rightarrow c$:

$$\text{FOIL - gain}(i) = |P| \, (\log \frac{|P|}{|P|+|N|} - \log \frac{|P'|}{|P'|+|N'|})$$

L'algoritmo FOIL cerca l'item che massimizza il FOIL-gain per una particolare classe del data set di addestramento. Una volta che quell'item è stato identificato, tutti gli oggetti di addestramento ad esso associati vengono scartati, e il processo è ripetuto sino a che sono coperti tutti gli oggetti positivi per la classe in questione. A quel punto viene selezionata un'altra classe e viene ripetuto il medesimo processo per quest'ultima, e così via.
È stata proposta una tecnica di classificazione associativa alternativa a FOIL, che si chiama *CPAR* e migliora la strategia di FOIL per generare le regole. CPAR differisce da FOIL nel senso che non rimuove tutti gli oggetti associati con un item una volta che questo è stato determinato; al contrario, i pesi degli oggetti associati con quell'item sono ridotti di un fattore moltiplicante e il processo è ripetuto sino a che non sono coperti tutti gli oggetti positivi per la classe in esame. Questa versione pesata di FOIL estrae più regole, dal momento che ad un oggetto di addestramento è per-

messo di essere coperto da più di una singola regola, similmente agli approcci per la scoperta delle regole associative.

La confidenza delle regole

La soglia minima di supporto è la chiave che porta al successo sia nella scoperta delle regole associative sia nella classificazione associativa. Comunque, per certe applicazioni, alcune regole ad alta confidenza vengono ignorate semplicemente perché non hanno abbastanza supporto. Gli algoritmi classici di classificazione associativa utilizzano un'unica soglia di supporto per controllare il numero delle regole estratte e possono quindi non essere in grado di catturare le regole ad alta confidenza che hanno però un supporto basso. Per esplorare uno spazio di ricerca ampio e per estrarre il maggior numero possibile di regole ad alta confidenza, tali algoritmi tendono a impostare una soglia minima di supporto molto bassa, la quale può sollevare problemi quali l'*overfitting*, la generazione di regole che hanno statisticamente un basso supporto e un grande numero di regole candidate con alti requisiti per quanto riguarda il tempo e la memoria della CPU.
In risposta a questo problema, è stato proposto un approccio che sospende la soglia di supporto e usa soltanto la soglia di confidenza per scoprire le regole. Questo approccio basato sulla confidenza mira ad estrarre tutte le regole che in un data set superano la soglia minima di confidenza *minfconf*. Senza la soglia minima di supporto, il passo di generazione dei candidati non è più applicabile e la proprietà *downward-closure* utilizzata da Apriori non è ugualmente valida. È quindi necessario introdurre una proprietà analoga a quest'ultima, per fare in modo di mantenere il processo efficiente e scalabile. Così, è stata introdotta una nuova proprietà chiamata '*existential upward-closure*' nell'ambito dell'approccio di classificazione associativa basato su un albero di decisione chiamato *ADT*. Il modo migliore per spiegare questa proprietà è quello di usare un esempio. Consideriamo allora le seguenti tre

regole:

R_1, reddito alto *allora* carta di credito = sì;
R_2, reddito alto ed età > 55 *allora* carta di credito = sì;
R_3, reddito alto ed età ≤ 55 *allora* carta di credito = sì.

Assumiamo che R_1 abbia una confidenza del 70%. Una regola $\langle (X, x), (A_i, a_i) \rangle \rightarrow c$ è una A_i-specializzazione di $(X, x) \rightarrow c$ se a_i è un valore di A_i. Dal momento che R_2 e R_3 sono specializzazioni di R_1 e sono vicendevolmente esclusive, allora se una condizione implica un valore di confidenza negativo, l'altra condizione deve implicare un valore di confidenza positivo. Così, almeno una regola tra R_2 e R_3 ha un valore di confidenza più alto rispetto a R_1. In altre parole, se un attributo A_i non è in una regola $R_i : x \rightarrow c$ e R_i supera la soglia minima di confidenza, lo stesso accade per qualche A_i-specializzazione di R_i.

L'approccio multi-supporto

In alcuni dati, le etichette di classe possono essere distribuite non equamente, causando così la generazione di molte regole per le classi dominanti e di poche regole, e in alcuni casi nessuna regola, per le classi minoritarie. Usare un'unica soglia globale di supporto può essere inefficiente, soprattutto se si lavora su data set con una distribuzione impari delle frequenze delle classi; questo accade perché quando l'utente imposta il valore della soglia minima di supporto *minsupp* al di sopra dei valori delle frequenze di alcune classi è ovvio che non ci saranno regole generate per tali classi, ed alcune regole ad alta confidenza verrano scartate.
Per trattare tale difetto sono state sviluppate delle estensioni per alcuni degli approcci di classificazione associativa già esistenti, quali CBA e CMAR. Queste estensioni hanno dato come risultato un nuovo approccio che considera i valori di distribuzione delle frequenze di tutte le classi esistenti in un

data set ed assegna ad ognuna di queste classi una soglia minima di supporto diversa da quella di tutte le altre.
Un altro approccio iterativo multi-supporto è stato proposto nel 2002; esso analizza il risultato del ciclo di estrazione delle regole e aggiorna la soglia di supporto per ciascuna classe. Questo approccio iterativo inizialmente assegna un'unica soglia di supporto a ciascuna classe e poi, analizzando il numero di regole generate durante uno specifico ciclo di generazione delle regole, diminuisce quella soglia di un fattore moltiplicante per il ciclo successivo per quelle classi che sono coperte da un numero basso di regole. Di conseguenza questo assicura che saranno prodotte un numero sufficiente di regole per ciascuna classe.
Queste tecniche di classificazione associativa che utilizzano il multi-supporto usano sia il metodo di Apriori sia il metodo di FP-growth per trovare i *ruleitem* frequenti e sono abbastanza simili a un algoritmo per trovare le regole associative già sviluppato noto come *MSapriori*. Infatti Msapriori è stato uno dei primi algoritmi per la scoperta delle regole associative ad aver introdotto l'idea di usare il multi-supporto per risolvere il problema di eliminare gli item rari presenti nei database.

L'intersezione delle posizioni degli oggetti di addestramento

Con la rappresentazione verticale dei dati, il supporto degli itemset frequenti è calcolato attraverso la semplice intersezione dei tid, cioè degli identificativi delle transazioni. Per esempio, il supporto degli itemset candidati di dimensione k può essere facilmente ricavando intersecando le liste dei tid di qualsiasi due sottoinsiemi $k-1$. Le liste dei tid che mantengono tutte le informazioni relative agli item presenti nel database sono una struttura relativamente semplice e facile da mantenere, e, in questo modo, non c'è bisogno di scansionare il database durante ciascuna iterazione per calcolare il supporto di nuovi itemset candidati, dando così modo di risparmiare tem-

po di *I/O*.

La maggior parte degli algoritmi di classificazione associativa in circolazione adotta il formato orizzontale per rappresentare i dati; il primo algoritmo di classificazione associativa ad utilizzare il formato verticale per eseguire semplici intersezioni tra liste di tid di itemset frequenti è *MCAR*.

Il metodo per trovare gli itemset frequenti impiegato da MCAR scansiona il data set di addestramento per contare le occorrenze degli item di grandezza uno, a partire dai quali determina quelli che hanno abbastanza supporto. Durante la scansione vengono determinati gli itemset di grandezza uno frequenti, e le loro occorrenze nel data set di addestramento (i tid) vengono memorizzate in un *array*. Qualsiasi itemset di grandezza uno che non supera la soglia minima di supporto viene scartato. Il risultato di una semplice intersezione tra le liste di tid di due itemset dà un insieme in cui vengono memorizzati i tid in cui i due itemset occorrono insieme nel data set di addestramento. Questo insieme assieme all'array delle classi, il quale è stato creato durante la prima scansione e memorizza le frequenze delle classi, può essere usato per calcolare il supporto e la confidenza del nuovo *ruleitem* risultato dell'intersezione.

1.6 L'estrazione degli itemset frequenti interessanti

Nonostante siano stati sviluppati numerosi metodi scalabili per trovare gli itemset frequenti e gli itemset frequenti chiusi e massimali, tali metodi spesso generano un numero troppo grande di pattern frequenti. Le persone che si trovano a dover usare i risultati di queste analisi vorrebbero avere a che fare soltanto con quelli interessanti. Che cosa sono allora i pattern interessanti e come possono essere estratti efficientemente? Per rispondere a queste domande, ci vengono in aiuto diversi studi recenti che hanno contribuito al-

la ricerca dei pattern frequenti e delle regole interessanti, compresi i metodi di ricerca basati su un vincolo (*constraint-based*), l'estrazione dei pattern incompleti o compressi, le misure di interesse e le analisi di correlazione.

1.6.1 La procedura di estrazione *constraint-based*

Sebbene un processo di *data mining* possa portare alla luce centinaia di pattern a partire da un determinato insieme di dati, un particolare utente può essere interessato soltanto ad un piccolo sottoinsieme di questi pattern, sottoinsieme che soddisfa alcuni vincoli pre-specificati dall'utente in questione. L'estrazione efficiente di questi pattern è chiamata *estrazione basata su un vincolo* (*constraint-based*).
Le ricerche hanno dimostrato che i vincoli possono essere divisi in diverse categorie secondo il loro grado di interazione con il processo di estrazione. Per esempio, all'inizio del processo di *mining*, durante la fase iniziale di selezione dei dati, possono essere inseriti dei vincoli *succinti*; quando invece ci si trova nel pieno del processo di *mining* possono essere applicati dei vincoli *anti-monotonici* per controllare la crescita del numero di pattern trovati; infine, possono essere verificati dei vincoli *monotonici*, e una volta che questi sono stati soddisfatti, i vincoli in questione possono essere non più verificati fino alla fine del processo di *mining*.
Dal momento che i vincoli più usati appartengono ad almeno una delle categorie viste in precedenza, essi possono essere usati durante il processo di *mining*. Sono stati proposti diversi algoritmi per implementare l'applicazione di questi vincoli durante il processo di estrazione dei pattern frequenti e, uno di questi, chiamato *ExAnte* [17], è utile per ridurre ulteriormente lo spazio di ricerca con vincoli di tipo *monotonico* imposti dall'utente.

1.6.2 L'estrazione di pattern compressi o approssimati

Le ricerche più recenti si sono focalizzate sull'estrazione di un insieme di pattern frequenti compresso o approssimato, per ridurre così l'enorme insieme di pattern frequenti generati in un processo di *data mining* senza però incidere sulla qualità di questi ultimi. In generale, la compressione dei pattern può essere divisa in due categorie: la compressione senza perdita e la compressione con perdita, in termini dell'informazione che l'insieme finale contiene, se confrontato con l'intero insieme di pattern frequenti.

L'estrazione dei pattern frequenti chiusi è una compressione senza perdita dei pattern frequenti, poiché da questi ultimi è possibile derivare l'intero insieme degli itemset frequenti; l'estrazione dei pattern frequenti massimali è invece un esempio di compressione con perdita, dal momento che da questi ultimi non è possibile derivare l'intero insieme degli itemset frequenti. Altri esempi di compressione con perdita sono: l'estrazione dei primi k pattern chiusi più frequenti, l'estrazione di k itemset riassuntivi e la compressione basata sul raggruppamento.

Per estrarre i primi k pattern chiusi più frequenti, è stato proposto un algoritmo, chiamato *TFP* [18], il quale scopre i primi k itemset chiusi più frequenti di lunghezza non minore ad una soglia minima min_l. *TFP* alza gradualmente la soglia minima di supporto durante il processo di *mining* e riduce l'ampiezza dell'FP-tree sia durante, sia dopo, la fase di costruzione dell'albero.

A causa della distribuzione di frequenza ineguale tra gli itemset, i primi k pattern chiusi più frequenti solitamente non rappresentano i k pattern più significativi. Per questa ragione, un'altra parte del lavoro di compressione fa suo un approccio 'riassuntivo', il cui scopo è quello di derivare i k itemset più rappresentativi che coprano l'intero insieme di itemset frequenti chiusi. Questi k itemset più rappresentativi forniscono una rappresentazione compatta su tutta la collezione di pattern frequenti, rendendo più facile il

compito di interpretarli e di usarli.

1.6.3 Dall'estrazione dei pattern frequenti interessanti all'estrazione delle regole associative interessanti

L'estrazione degli itemset frequenti porta naturalmente alla scoperta di associazioni e correlazioni tra gli itemset presenti in un ampio data set transazionale, cioè alla scoperta delle regole associative. In particolare, abbiamo già visto, che una regola associativa è considerata interessante solo se soddisfa sia una soglia minima di supporto che una soglia minima di confidenza.

Basandosi sulla definizione di regola associativa, la maggior parte degli studi in materia considera il processo di estrazione dei pattern frequenti come il primo ed essenziale passo nel processo di estrazione delle regole associative. Comunque, non tutte le regole associative generate sono interessanti, specialmente quando il processo di *mining* viene portato avanti con una soglia minima di supporto bassa o quando si cercano pattern piuttosto lunghi. All'interno dell'insieme delle regole associative trovate possono anche essere presenti delle regole associative *ridondanti*, cioè delle regole che portano la stessa informazione già portata da altre regole.

Gli approcci che cercano di arginare questo problema possono essere classificati in tre gruppi principali. Gli approcci che fanno parte del primo gruppo forniscono dei meccanismi per filtrare le regole associative estratte. Gli altri due approcci 'estendono' invece la definizione di regola associativa per non estrarre regole associative 'simili'.

Gli approcci del primo gruppo permettono agli analisti di definire dei *template* o degli operatori *booleani* oppure, ancora, degli operatori simili a quelli dell'SQL per permettere all'utente di selezionare le regole secondo le sue proprie preferenze. Selezionando così un sottoinsieme delle regole associative

estratte, questi approcci riducono il numero di regole da analizzare durante la visualizzazione, ma non sono comunque soppresse le regole associative ridondanti.

Nel secondo gruppo, alcuni approcci utilizzano una tassonomia degli item per estrarre regole associative generali, cioè regole associative tra insiemi di item che appartengono a livelli diversi della tassonomia. Alcuni approcci fanno invece use di misure statistiche oggettive per misurare la precisione di una regola, come, per esempio, la correlazione di Paerson o il χ^2 test, tralasciando invece la confidenza. Altri approcci presenti in questo gruppo permettono inoltre di estrarre solo le regole con antecedenti massimali tra quelli che hanno lo stesso supporto e gli stessi conseguenti. Accade cioè che una regola r verrà scartata se esiste un'altra regola r' che ha lo stesso conseguente di r, ma ha un antecedente che costituisce un sovrainsieme dell'antecedente di r. Infine, gli ultimi due approcci identificano le regole migliori o sulla base di alcune misure oggettive di interesse (confidenza, convinzione, lift, Laplace, ecc.) o sulla base di alcune misure soggettive di interesse, le quali prendono in considerazione la conoscenza dell'utente e i suoi obiettivi in riferimento all'analisi che sta compiendo.

Gli approcci appartenenti all'ultimo gruppo fanno inoltre uso della *connessione di Galois* per estrarre delle *basi* per le regole associative. Una base è un insieme non ridondante che è minimo secondo alcune proprietà matematiche e, a partire dal quale, possono essere dedotte tutte le regole associative, con i relativi valori di supporto e confidenza, senza più accedere al data set iniziale.

1.6.4 Le misure oggettive di interesse

Dal momento che il numero di possibili regole associative estratte a partire da un data set transazionale cresce esponenzialmente all'aumentare del numero di item presenti nel data set, il processo di selezione delle regole più

interessanti diventa essenziale. È quindi necessario misurare il livello di interesse di una determinata regola, e validare le regole veramente interessanti rispetto a questa specifica misura.
Una misura oggettiva di interesse è un approccio guidato a partire dai dati per valutare la qualità dei pattern associativi. Dal momento che una misura oggettiva si basa soltanto sui dati per trarre le proprie conclusioni, essa, riguardo alle nozioni che caratterizzano l'interesse di un pattern [19], fa sue quelle di *concisione*, *generalità*, *affidabilità*, *peculiarità* e *diversità*. In particolare:

- *concisione*: un pattern è *conciso* se contiene relativamente poche coppie attributo-valore, mentre un insieme di pattern è *conciso* se contiene relativamente pochi pattern. Un pattern conciso o un insieme conciso di pattern è relativamente facile da capire e da ricordare e, così, è aggiunto più facilmente all'insieme di conoscenze proprie dell'utente;

- *generalità/copertura*: un pattern è *generale* se copre un sottoinsieme relativamente grande del data set iniziale. La generalità misura la copertura di un pattern, cioè, la frazione di tutti i record presenti nel data set che combaciano con il pattern in questione. Se un pattern caratterizza un numero maggiore di informazione del data set, esso tende ad essere più interessante;

- *affidabilità*: un pattern è *affidabile* se la relazione che esso descrive occorre in un'alta percentuale di casi pratici. Per esempio, una regola di classificazione è affidabile se le sue predizioni sono altamente accurate, e una regola associativa è affidabile se ha un alto valore di confidenza;

- *peculiarità*: un pattern è *peculiare* se esso è molto distante rispetto agli altri pattern già scoperti, il tutto secondo qualche misura di distanza. I pattern peculiari sono generati a partire da dati peculiari (o *outlier*), i quale sono relativamente pochi e sono significativamente differenti

rispetto al resto dei dati. I pattern peculiari possono essere sconosciuti all'utente, ma nonostante questo possono essere molto interessanti;

- *diversità*: un pattern è *diverso* se i suoi elementi differiscono in maniera significativa l'uno dall'altro, mentre un insieme di pattern è diverso se i pattern presenti nell'insieme differiscono significativamente l'uno dall'altro. La diversità non è comunque ancora un fattore comune per misurare l'interesse sia delle regole associative che delle regole classificative.

Una misura oggettiva è inoltre indipendente dal dominio e richiede che gli utenti specifichino soltanto una soglia minima per filtrare i pattern a bassa qualità. Una misura oggettiva è solitamente calcolata sulla base del conteggio delle frequenze espresso in una *tabella di contingenza*.

Tabella 1.1: Tabella di contingenza per $A \rightarrow B$

	$\boldsymbol{B}$	$\neg B$	
$\boldsymbol{A}$	a	b	n_1
$\neg A$	c	d	n_2
	m_1	m_2	N

La Tabella 1.1 mostra un esempio di una tabella di contingenza per una coppia di variabili binarie, A e B. Viene indicata la notazione $\neg A(\neg B)$ per indicare che $A(B)$ è assente in una transazione. Ciascuna voce in questa tabella $2x2$ denota un conteggio di frequenza. Per esempio, a è il numero delle volte in cui A e B appaiono insieme nella stessa transazione, mentre c è il numero di transazioni che contengono B ma non A. La riga di somma n_1 rappresenta il supporto per A, mentre la colonna di somma m_1 rappresenta il supporto per B.
Gli algoritmi per l'estrazione delle regole associative come *Apriori* e *FP-Growth*, che abbiamo visto nella sezione 1.4, utilizzano l'approccio basato sul supporto e sulla confidenza per trovare tutte le regole associative a partire

da un data set iniziale. In particolare, questi algoritmi trovano gli itemset il cui supporto eccede la soglia minima di supporto definita dall'utente e, successivamente, a partire da questi ultimi, estraggono le regole associative i cui valori di confidenza superano la soglia minima di confidenza definita dall'utente.

La proprietà antimonotonica del supporto fa sì che l'approccio basato sul supporto e sulla confidenza per l'estrazione delle regole sia abbastanza attraente. L'utilità propria di questo approccio è comunque discutibile, nonostante il vero significato del supporto e della confidenza sia tradotto in misure molto facili da comprendere.

Per prima cosa gli algoritmi che si basano su questo approccio generano un numero veramente molto grande di regole, molte delle quali sono poco interessanti. Inoltre, la condizione che si basa sul supporto, posizionata al cuore del processo di estrazione, ha un incoveniente: scarta le regole che hanno un basso valore di supporto sebbene alcune di queste possano avere un alto valore di confidenza, alto valore che le rende così veramente interessanti, situazione che è molto comune nel *marketing* (sono le cosiddette 'pepite' del *data mining*). Se, però, per porre rimedio a questo problema, la soglia di supporto viene abbassata, vengono prodotte ancora più regole, bloccando così gli algoritmi per l'estrazione di queste ultime.

Infine, le sole condizioni di supporto e confidenza non assicurano che le regole estratte siano realmente interessanti. Anche la confidenza ha infatti un inconviente, proprio come il supporto. Il problema che ha la confidenza è più sottile, e per questa ragione, è più facilmente dimostrabile con un esempio [6].

Supponiamo, quindi, che siamo interessati nell'analizzare la relazione che intercorre tra le persone che bevono tè e le persone che bevono caffè. Possiamo raccogliere le informazioni necessarie all'analisi intervistando un gruppo di persone e possiamo riassumere le loro risposte in una tabella di contingenza

come quella della Tabella 1.2.

Tabella 1.2: Le preferenze riguardo alle bevande in un gruppo costituito da 1000 persone.

	Caffè	$\neg Caffè$	
Tè	150	50	200
$\neg Tè$	650	150	800
	800	200	1000

L'informazione riportata in questa tabella può essere usata per valutare la regola associativa $\{Tè\} \rightarrow \{Caffè\}$. Ad un primo sguardo può sembrare che le persone che bevono tè, tendono anche a bere caffè poiché i valori del supporto e della confidenza di questa regola (rispettivamente il 15% ed il 75%) sono ragionevolmente alti. Questa conclusione sarebbe stata accettabile se avessimo escluso il fatto che la percentuale di persone che bevono caffè, indipendentemente che questi siano anche consumatori di tè, è dell'80%, mentre la percentuale di consumatori di tè che però bevono anche caffè è solo del 75%. Così, conoscere che una persona è un consumatore di tè diminuisce le sue probabilità di essere un consumatore di caffè dall'80% al 75%! La regola $\{Tè\} \rightarrow \{Caffè\}$ è quindi fuorviante, nonostante sia supportata da un alto valore di confidenza.

Il trabocchetto riguardo alla confidenza può essere rintracciato nel fatto che essa ignora il supporto dell'itemset che costituisce il conseguente della regola. Così, se fosse preso in considerazione il supporto dei consumatori di caffè, di sicuro non ci sorprenderemmo nello scoprire che molte delle persone che bevono tè, tendono a bere anche caffè. Ciò che è più sorprendente è che la percentuale di consumatori di tè e caffè è più piccola rispetto alla percentuale totale di persone che bevono caffè, fatto che sottolinea una relazione inversa tra le persone che bevono tè e le persone che bevono caffè.

A causa delle limitazioni appena viste proprie dell'approccio che si basa sul supporto e sulla confidenza, devono quindi essere esaminate delle misure

che vadano oltre queste ultime, per fare in modo di selezionare le regole realmente interessanti.

Tabella 1.3: Lista delle misure oggettive di interesse.

Misura	Formula
Support	a
Confidence	$\frac{a}{n_1}$
Centred Confidence	$\frac{a-n_1m_1}{n_1}$
Symmetric Confidence	$\max(\frac{a}{n_1}, \frac{n_1}{a})$
All-confidence	$\min(\frac{a}{n_1}, \frac{a}{m_1})$
Ganascia	$2\frac{a}{n_1} - 1$
Coverage	n_1
Prevalence	m_1
Recall	$\frac{n_1}{a-n_1m_1}$
Specificity	$\frac{d}{n_2}$
Accuracy	$\frac{a}{n_1} + n_2m_2$
Weighted relative accuracy	$n_1(\frac{a}{n_1} - m_1)$
Lift	$\frac{\frac{a}{n_1}}{\frac{m_1}{N}}$
Extended Lift (Elift)	$\frac{\frac{a}{n_1}}{\frac{(a+c)}{(n_1+n_2)}}$
Selection Lift (Slift)	$\frac{\frac{a}{n_1}}{\frac{c}{n_2}}$
Olift	$\frac{\frac{a}{n_1}(1-\frac{c}{n_2}}{\frac{c}{n_2}(1-\frac{a}{n_1}}$
Interest factor	$\frac{a}{n_1m_1}$
Correlation Analysis	$\frac{ad-bc}{\sqrt{N}}$
Leverage	$a - n_1m_1$
Added Value / Change of support	$\frac{a}{n_1} - \frac{m_1}{N}$
Relative risk	$\frac{\frac{a}{n_1}}{\frac{a}{n_2}}$
Jaccard	$\frac{a}{n_1+m_1-a}$

Tabella 1.3: continua nella prossima pagina

Tabella 1.3: continua dalla pagina precedente

Misura	Formula
Certainty factor	$\frac{\frac{a}{n_1}-m_1}{(1-m_1)}$
Odds ratio	$\frac{a*d}{b*c}$
Yule's Q	$\frac{a*d-b*c}{a*d+b*c}$
Yule's Y	$\frac{\sqrt{a*d}-\sqrt{b*c}}{\sqrt{a*d}+\sqrt{b*c}}$
Cohen's	$\frac{a+d-n_1m_1-n_2m_2}{1-n_1m_1-n_2m_2}$
Klosgen	$\sqrt{a}(\frac{a}{n_1}-\frac{m_1}{N}), \sqrt{a}\max \frac{a}{n_1}-\frac{n_1}{N}, \frac{n_1}{a-n_1m_1}-\frac{n_1}{N}$
Convinction	$\frac{n_1m_2}{b}$
Collective strength	$\frac{a+(\frac{m_2}{n_2})}{n_1m_1+n_2m_2} * \frac{1-n_1m_1-n_2m_2}{1-a-(\frac{m_2}{n_2})}$
Laplace correction	$\frac{a+1}{n_1+2}$
Gini index	$n_1 * (\frac{m_1}{n_1})^2 + (\frac{m_2}{n_1})^2 + n_2 * (\frac{m_1}{n_2})^2 + (\frac{m_2}{n_2})^2 - (m_1)^2 - (m_2)^2$
J-Measure	$a\log(\frac{\frac{m_1}{n_1}}{m_1}) + b\log(\frac{\frac{m_2}{n_1}}{m_2})$
One-Way Support	$\frac{a}{n_1} * \log_2 \frac{a}{n_1m_1}$
Two-Way Support	$a * \log_2 \frac{a}{n_1m_1}$
Two-Way Support Variation	$a * \log_2 \frac{a}{n_1m_1} + b * \log_2 \frac{b}{n_1m_{+2}} + c*\log_2 \frac{c}{n_2m_1} + Nd_1 * \log_2 \frac{d}{n_2m_2}$
ϕ-coefficient	$\frac{a-n_1m_1}{\sqrt{n_1m_1n_2m_2}}$
Piatetsky-Shapiro	$a - n_1m_1$
Cosine	$\frac{a}{\sqrt{n_1m_1}}$
Loevinger	$1 - \frac{n_1m_2}{b}$
Information Gain	$\log \frac{a}{n_1m_1}$
Sebag-Schoenauer	$\frac{a}{b}$
Least Contradiction	$\frac{a-b}{m_1}$

Tabella 1.3: continua nella prossima pagina

Tabella 1.3: continua dalla pagina precedente

Misura	Formula
Example and Counterexample Rate	$1 - \frac{b}{a}$
Zhang	$\frac{a - n_1 m_1}{\max(am_2, m_1 b)}$
Bayes factor	$\frac{am_2}{m_1 b}$
Convinction	$\frac{n_1 m_2}{b}$
Examples and counter examples rate	$\frac{a-b}{a} = 1 - \frac{1}{\frac{n_1}{b} - 1}$
Entropic intensity of implication[1]	$[(1 - h_1(\frac{b}{N})^2) * (1 - h_2(\frac{b}{N})^2)]^{\frac{1}{4}} INTIMP^{\frac{1}{2}}$
Implication index (IMPIND)	$\frac{n_1 m_1 - a}{\sqrt{b}}$
Intensity of implication (INTIMP)	$P[N(0,1) >= IMPIND]$
Kappa Coefficient	$2\frac{a - n_1 m_1}{n_1 m_1 - 2 n_1 m_1}$
Probabilistic discriminant index[2]	$P[\mathcal{N}(0,1) > IMPIND^{\mathcal{CR}/\mathcal{B}}]$
Paerson's correlation coefficient	$\frac{a - n_1 m_1}{\sqrt{n_1 m_1 n_2 m_2}}$
Gray and Orlowska[3]	$((\frac{a}{n_1 m_1})^k - 1) * (a)^m$
Chi-Square Test	$\sum_i \sum_j \frac{(N_i j - E_i j)^2}{E_i j}$
4ft Quantifier of founded double implication	$\frac{a}{n_1 + c}$
4ft Quantifier of lower critical implication[4]	$\sum_{i=a}^{n_1} p^i (1-p)^{n_1 - i}$

Tabella 1.3: continua nella prossima pagina

[1] $h_1(t) = -(1 - \frac{n*t}{n_a}) \log_2(1 - \frac{n*t}{n_a}) - \frac{n*t}{n_a} \log_2(\frac{n*t}{n_a})$ se $t \in [0, n_a/(2n)]$; altrimenti $h_1(t) = 1$. $h_2(t) = -(1 - \frac{n*t}{n_{\bar{b}}}) \log_2(1 - \frac{n*t}{n_{\bar{b}}}) - \frac{n*t}{n_{\bar{b}}} \log_2(\frac{n*t}{n_{\bar{b}}})$ se $t \in [0, n_{\bar{b}}/(2n)]$; altrimenti $h_2(t) = 1$.

[2] $\mathcal{N}(0,1)$ sta per la normale funzione di ripartizione centrata e ridotta. $IMPIND^{\mathcal{CR}/\mathcal{B}}$ corrisponde a IMPIND, centrato e ridotto (CR) per un insieme di regole $\mathcal{B}$.

[3] k ed m sono coefficienti di dipendenza e di generalità, i quali pesano rispettivamente la relativa importanza dei due fattori.

[4] Con $0 < p \leq 1$

Tabella 1.3: continua dalla pagina precedente

Misura	Formula
4ft Quantifier of lower critical double implication[4]	$\sum_{i=a}^{n_1+c}(\frac{n_1+c}{i}))p^i(1-p)^{n_1+c-i}$
4ft Quantifier of founded equivalence	$\frac{a+d}{N}$
4ft Quantifier of lower critical equivalence[4]	$\sum_{i=a+d}^{N}(\frac{N}{i}))p^i(1-p)^N$
Fisher's Test	$\sum_{i=a}^{\min n_1,m_1}(\frac{(\frac{m_1}{i})(\frac{N-m_1}{n_1-i})}{\frac{n_1}{N}})$
Fisher's Exact Test	$\sum_{i=0}^{\min c,b}\frac{(n_1)!(n_2)!(m_1)!(m_2)!}{N!(a+i)!(c-i)!(b-i)!(d+i)!}$
Class Correlation Ratio	$\frac{a*m_2}{c*m_1}$
4ft Quantifier of E-equivalence	$\max(\frac{b}{n_1},\frac{c}{n_2})$
4ft Quantifier of above average dependence	$\frac{\frac{a}{n_1}}{\frac{c}{N}}$

Tabella 1.3: si conclude dalla pagina precedente

Ecco una breve descrizione delle misure elencate:

Support: il supporto è spesso utilizzato per rappresentare la significatività e la generalità di un pattern associativo. Esso è inoltre utile da un punto di vista computazionale poiché gode di una proprietà di chiusura di tipo *downward* che permette di ridurre lo spazio di ricerca esponenziale dei pattern candidati. Misure analoghe al supporto, utilizzate cioè per misurare il livello di generalità di un pattern associativo, sono la **coverage** e la **prevalence**;

Confidence, **added value**, **Laplace correction**, **convinction**, **centred confidence**, **symmetric confidence**, **Ganascia**, ***accuracy***, **all-confidence** e **recall**: la confidenza, così come l'*added value*, viene usata per misurare l'accuratezza e l'affidabilità di una data regola. Comunque, essa può spesso produrre dei risultati fuorvianti, specialmente quando il valore del supporto del conseguente della regola è più alto rispetto al

valore della confidenza dell'intera regola. Le altre varianti della confidenza comprendono la correzione di Laplace, la convinction, la confidenza centrata, la confidenza simmetrica, Ganascia, l'accuratezza e l'*all-confidence*. La confidenza simmetrica è stata introdotta come variante simmetrica della confidenza, dal momento che la confidenza è una misura asimmetrica (si veda il paragrafo 1.6.5 per ulteriori chiarimenti);

Come già detto sopra, molte delle misure qui elencate sono trasformazioni lineari della confidenza: esse cercano soprattutto di migliorarla permettendo dei confronti con il supporto del conseguente della regola in esame. Questa trasformazione è generalmente effettuata centrando la confidenza sul supporto del conseguente, cioè su m_1, usando coefficienti a differenti scale (**centred confidence**, **Piatetsky-Shapiro**, **Loevinger**, **Zhang**, l'**implication index** e la **least contradiction**). Altre misure, come quella di **Sebag e Schoenauer** o la **example and counterexample rate**, sono trasformazioni della confidenza che crescono monotonicamente, mentre l'**information gain** è una trasformazione crescente monotonicamente del *lift*. Altre misure si focalizzano sui controesempi, come la *convinction* e il già citato *implication index*. Quest'ultima misura in particolare è la base di altre differenti misure probabilistiche come il **probabilistic discriminant index**, l'**intensity of implication**, o la sua versione entropica, cioè l'**entropic intensity of implication**, la quale prende in considerazione un coefficiente d'entropia, migliorando il potere discriminante dell'*intensity of implication*. Queste ultime due misure sono state adattate per permettere che avessero la proprietà di rimanere costanti con la presenza di un'ipotesi nulla. Il **Bayes factor**, chiamato anche *odd multiplier* o *sufficiency*, è una specie di *odds ratio*, basato sul confronto della probabilità di A e B su B piuttosto che sulla probabilità di A e $\neg A$ su B. Infine, la misura di**Laplace** è una variante della confidenza, che però prende in considerazione l'intero numero di record N;

Lift: uno dei problemi della confidenza è che essa ignora il supporto dell'itemset conseguente della regola. Un modo per risolvere questo problema è quello di utilizzare una metrica nota come *lift*. Il *lift* calcola il rapporto tra la confidenza della regola e il supporto dell'itemset conseguente;

Elift, Slift e Olift: queste tre misure sono nate all'interno dello studio di potenziali eventi discriminatori presenti e nascosti nei dati.
Una regola del tipo $sesso = donna, macchina = si \rightarrow richiestadiprestito = rifiutata$ con un valore di elift pari a 3 significa che essere una donna aumenta di 3 volte la probabilità

di trovarsi ad avere rifiutata la richiesto di un prestito rispetto alla confidenza di persone che possiedono una macchina. L'elift quindi non è altro che il rapporto tra la proporzione del gruppo svantaggiato A nel contesto B ottenente il beneficio C sulla proporzione totale di A in B.
Il selection lift considera invece il rapporto tra le persone aventi una determinata caratteristica (per esempio l'essere una donna) che ottengono un determinato beneficio C rispetto alle persone che non hanno quella determinata caratteristica (gli uomini) e che anch'esse ottengono il beneficio C.
L'Olift si basa invece sulla *odds ratio*. Nella terminologia delle scommesse, 2/3 (2 a 3) significa che per ogni 2 casi in cui un evento può accadere, esistono altri 3 casi in cui lo stesso evento può non accadere. Se ripetiamo il medesimo concetto in termini di probabilità p dell'evento, l'odds ratio è: $\frac{p}{1-p}$. Nella letteratura riguardante la discriminazione presente nelle assunzioni, l"evento' modellato è la promozione o il licenziamento di una persona. Le probabilità di $A, B \to C$ possono essere definite come:

$$odds(A, B \to C) = \frac{conf(A,B \to C)}{1-conf(A,B \to C)} \equiv \frac{conf(A,B \to C)}{1-conf(A,B \to \neg C)}$$

La odds ratio nell'assunzione dei dipendenti è il rapporto tra le probabilità di assunzione di una persona appartenente ad una minoranza rispetto alle probabilità di assunzione di una persone non appartenente a quella stessa minoranza. Se estendiamo lo stesso concetto alle regole otteniamo l'olift:

$$Olift(A, B \to C) = \frac{odds(A,B \to C)}{odds(\neg A,B \to C)}$$

Interest factor: in riferimento alle variabili binarie, il *lift* è equivalente ad un'altra misura oggettiva chiamata *interest factor* (I). Questa misura è usata abbastanza ampiamente nel *data mining* ed essa mette a confronto la frequenza di un pattern (cioè il suo supporto, presente al numeratore) con un denominatore calcolato sotto l'assunzione di indipendenza statistica (se due variabili sono indipendenti è infatti valida questa equazione: $a = \frac{n_1 m_1}{N}$). Usando questa equazione e quella dell'*interest factor* arriviamo ad interpretare la misura come segue: $I(A, B) = 1$, se A e B sono indipendenti, $I(A, B) > 1$, se A e B sono positivamente correlate e $I(A, B) < 1$, se A e B sono negativamente correlate.
L'*interest factor* conosce però un limite, il quale è facilmente illustrabile con un esempio appartenente al dominio del *text mining* [6]. Nel *text mining* è ragionevole assumere che l'associazione tra una coppia di parole dipende dal numero di documenti che contengono entrambe le parole. Per esempio, a causa della loro forte associazione, con riferimento ad una collezione di articoli informatici, ci aspettiamo che le parole *data* e *mining* appaiano

insieme più frequentemente delle parole *compilatore* e mining.

La Tabella 1.4 mostra la frequenza delle occorrenze tra due coppie di parole, $\{p, q\}$ e

Tabella 1.4: Le tabelle di contingenza per le coppie di parole $\{p, q\}$ e $\{r, s\}$

	$\boldsymbol{p}$	$\neg p$	
$\boldsymbol{q}$	880	50	930
$\neg q$	50	20	70
	930	70	1000

	$\boldsymbol{r}$	$\neg r$	
$\boldsymbol{s}$	20	50	70
$\neg s$	50	880	930
	70	930	1000

$\{r, s\}$. Usando la formula dell'*interest factor*, abbiamo che il valore di I per $\{p, q\}$ è 1.02 e 4.0 per $\{r, s\}$. Questi risultati sono in qualche modo problematici per le seguenti ragioni. Sebbene p e q appaiano insieme nell'88% dei documenti, il loro valore di I è vicino ad 1, il che significa che p e q sono statisticamente indipendenti. Dall'altro lato abbiamo che il valore di IS per $\{r, s\}$ è più alto di quello di $\{p, q\}$, sebbene p e q raramente appaiano insieme nello stesso documento. In questo caso, forse, la confidenza risulterebbe la scelta migliore, dal momento che considera l'associazione tra p e q (94.6%) essere molto più forte rispetto a quella tra r ed s (28.6%).
Dal momento che l'*interest factor* è sensibile al supporto degli item (n_1 e m_1), DuMouchel ha recentemente proposto una correzione statistica all'I per i campioni di piccole dimensioni, usando la tecnica empirica di *Bayes*. Altre varianti di questa misura, le quali cercano tutte di combinare la proprietà di generabilità con quella di affidabilità, sono *Piatetsky-Shapiro*, il **certainty factor**, la **collective strength**, il **one-way support**, il **two-way support**, la **two-way support variation**, *Jaccard*, *Gray and Orlowska* e *Klosgen*;

Specificity: la *specificity* è la probabilità condizionale della falsità di B, sapendo che A è ugualmente falsa. Nei problemi di classificazione binaria, essa è uguale al *recall* dei casi negativi nell'*information retrieval*, cioè $Specificity = \frac{TN}{TN+FP} = \frac{TN}{N} = \frac{d}{n_2}$, dove TN sta per *true negative* ed occorre quando il risultato della classificazione è negativo così come lo è il vero valore mentre FP sta per *false positive*, caso che invece occorre quando il risultato della classificazione è positivo, mentre il vero valore è negativo;

Correlation analysis: la *correlation analysis* è una tecnica basata sulla statistica per analizzare le relazioni presenti tra coppie di variabili binarie. Per le variabili continue, la correlazione è definita usando il coefficiente di **correlazione di Paerson**. I valori della correlazione spaziano da -1 (correlazione perfettamente negativa) a $+1$ (correlazione

perfettamente positiva). Se le variabili sono statisticamente indipendenti, allora il valore della correlazione sarà uguale a 0. Per esempio, la correlazione tra le persone che bevono tè e le persone che bevono caffè presentate nella Tabella 1.2 è di -0.0625.
La controindicazione presente nell'utilizzo della correlazione può rintracciarsi nell'esempio riguardante l'associazione di parole presentato nella Tabella 1.4. Nonostante le parole p e q appaiano insieme più frequentemente rispetto alle parole r e s, i loro coefficienti di correlazione sono identici. Questo accade perché il coefficiente di correlazione dà uguale importanza sia alla co-presenza che alla co-assenza degli item in una transazione ed è quindi più adatto ad analizzare le variabili binarie simmetriche. Un'altra importante limitazione di questa misura è che non rimane costante quando avvengono dei cambiamenti proporzionali nella dimensione del campione. Questo problema verrà discusso più in dettaglio nel paragrafo 1.6.5;

Leverage: la l*everage* misura la differenza tra il supporto dell'itemset AB e ciò che ci aspetterebbe se A e B fossero statisticamente indipendenti. Per esempio, facendo riferimento alla grande distribuzione e ad A e B come a due prodotti distinti, essa dà modo di scoprire quante più volte A e B verrebbero venduti insieme rispetto a quante volte essi sarebbero venduti singolarmente. L'utilizzo di una soglia minima di *leverage* dà inoltre la possibilità di includere nei dati un vincolo implicito di frequenza. Per esempio, per impostare una soglia minima di *leverage* allo 0.01% (il che corrisponde a 10 occorrenze in un data set composto da 100.000 transazioni), si può usare prima un algoritmo per trovare tutti gli itemset che superano una soglia minima di supporto dello 0.01% e poi filtrare gli itemset che sono stati trovati usando il vincolo di *leverage*;

Convinction: dal momento che l'*interest factor* è una misura simmetrica (vedi paragrafo 1.6.5), è stata proposta [20] un'altra misura, sua evoluzione, per superare questo problema. La *convinction* rappresenta questa evoluzione, ma, anch'essa, è simmetrica nel complemento. Essa dà infatti lo stesso valore per le regole associative $A \rightarrow B$ e $\neg B \rightarrow \neg A$, il che costituisce una seria controindicazione al suo utilizzo;

Ahmed-El-Makky-Taha's: viste le controindicazioni insite nell'utilizzo del supporto, della confidenza, dell'*interest factor* e della *convinction*, è stata proposta [21] un'altra misura per valutare l'affidabilità di una regola associativa, misura che non dà lo stesso risultato per $A \rightarrow B$ e $B \rightarrow A$. Essa è la differenza tra la probabilità condizionata di B dato A e la probabilità incondizionata di B. Misura l'effetto dell'informazione disponibile su A sulla probabilità di B. Più grande è questa differenza (assoluta), più forte

è l'associazione di $A \rightarrow B$. Una correlazione positiva è indicata da un risultato maggiore o uguale a zero, mentre una correlazione negativa è indicata da un valore inferiore allo zero;

Cosine(**IS**): questa misura può essere derivata dalla *correlation analysis* ed è stata proposta per gestire le variabili binarie asimmetriche. È la media geometrica tra l'*interest factor* e il *supporto*. Bisogna notare che il valore di *IS* è ampio quando il valore di *interest factor* e di supporto del pattern in questione sono ugualmente ampi. Per esempio, in riferimento alla Tabella 1.4, i valori di *IS* per la coppia di parole $\{p, q\}$ e $\{r, s\}$ sono rispettivamente 0.946 e 0.286. Contrariamente ai risultati proposti dall'*interest factor* e dalla *correlation analysis*, il *cosine* suggerisce che la correlazione tra $\{p, q\}$ è più forte rispetto a quella tra $\{r, s\}$, cosa che coincide con ciò che noi ci aspetteremmo da un'analisi associativa delle parole presenti in quei documenti.
Anche il *cosine* conosce però delle limitazioni. Il valore di *IS* per una coppia di itemset indipendenti, A e B, in riferimento alla Tabella 1.1, è dato da: $IS_indep(A, B) = \frac{a}{\sqrt{n_1 m_1}} = \frac{n_1 m_1}{\sqrt{n_1 m_1}} = \sqrt{n_1 m_1}$. Dal momento che il valore dipende dal supporto di $A(n_1)$ e dal supporto di $B(m_1)$, l'*IS* condivide un problema simile alla confidenza, cioè che il valore della misura può essere abbastanza grande, anche per i pattern non correlati o negativamente correlati. Per esempio, nonostante l'ampio valore di *IS* tra gli item p e $q(0.889)$, questo stesso valore è sempre più piccolo rispetto al valore che ci aspetteremmo quando gli item sono statisticamente indipendenti ($IS_indep = 0.9$);

Piatetsky-Shapiro(***rule-interest function***, ***RI***): questa misura è utilizzata per quantificare la correlazione tra gli attributi in una *semplice regola classificativa*. Una *semplice regola classificativa* è quella dove il lato destro e sinistro dell'implicazione logica $A \rightarrow B$ corrispondono ad un singolo attributo. In particolare a è il numero di tuple che soddisfa il supporto dell'itemset (AB) e $n_1 m_1$ è il numero di tuple che ci si aspetterebbe se A e B fossero indipendenti. Quando RI è uguale a 0, allora A e B sono statisticamente indipendenti e la regola non è interessante. Quando $RI > 0$ ($RI < 0$), allora A è positivamente (negativamente) correlato con B;

Jaccard e **Klosgen**: la misura di *Jaccard* è ampiamente usata nell'area dell'*information retrieval* per quantificare il livello di similitudine tra i documenti, mentre la misura di *Klosgen* è stata usata dal sistema di *discovery knowledge Explora* [22];

Goodman and Kruskal($\lambda - coefficient$): Il $\lambda - coefficient$, altrimenti conosciuto come l'indice di associazione predittiva, era stato inizialmente proposto da *Goodman* e

Kruskal [23]. L'intuizione dietro questa misura sta nel fatto che se due variabili sono altamente dipendenti l'una dall'altra, allora l'errore nel predire una delle due sarebbe piccolo qualora fosse conosciuto il valore dell'altra variabile. λ è usato per registrare la riduzione di quantità nell'errore predittivo;

Odds ratio: questa misura rappresenta le probabilità nell'ottenere i differenti risultati di una variabile. Per esempio, facciamo riferimento alla Tabella 1.1. Se B è presente, allora le probabilità di trovare A nella stessa transazione sono: $\frac{a}{d}$. Se B è assente, d'altro canto, le possibilità di trovare A nella stessa transazione sono: $\frac{b}{d}$. Se non esiste alcun tipo di associazione tra A e B, allora le probabilità di trovare A in una transazione dovrebbero rimanere le stesse, indipendentemente dal fatto che anche B è presente in quella stessa transazione. Possiamo allora usare il rapporto di queste probabilità ($\frac{ad}{bc}$) per determinare il grado di associazione tra A e B;

Yule's Q e **Y-coefficients**: il valore della *odds ratio* va da 0 (correlazione perfettamente negativa) a ∞ (correlazione perfettamente positiva). I coefficienti Q ed Y di *Yule* sono varianti normalizzate della *odds ratio*, definiti in modo da spaziare da -1 a $+1$;

Kappa coefficient ($\kappa - coefficient$): questa misura cattura il grado di concordanza tra una coppia di variabili. Se le variabili concordano l'una con l'altra, allora i valori per a e d saranno grandi, valori che, a turno, sfoceranno in un ampio valore di κ;

Mutual information, **J-measure** e **Gini index**: l'entropia è in relazione con la varianza di una distribuzione della probabilità. L'entropia di una distribuzione uniforme è ampia, mentre l'entropia di una distribuzione asimmetrica è piccola. La *mutual information* è una misura basata sull'entropia per valutare le dipendenze tra le variabili. Essa rappresenta quanto si riduce il valore di entropia di una variabile quando è conosciuto il valore di una seconda variabile. Se le due variabili sono fortemente associate, allora il valore di *mutual information* sarà alto. Le altre misure definite secondo la distribuzione della probabilità delle variabili includono la *J-Measure* e il *Gini index*;

Gray and Orlowska(***Interestingness Weighting Dependency***): l'interesse [24] è usato per valutare la forza delle associazioni presenti tra insiemi di item nelle regole associative. Mentre il supporto e la confidenza sono utili per caratterizzare le regole associative, l'interesse contiene una componente discriminante la quale dà un'indicazione circa l'indipendenza dell'antecedente e del conseguente di una regola. L'interesse è quindi

dato dalla misura proposta da Gray ed Orlowska [24], in cui $\frac{a}{n_1 m_1}$ è la componente discriminante e k e m sono rispettivamente parametri atti a pesare la relativa importanza della componente discriminante e del supporto. Più alto è il valore di interesse risultato di questa misura, più le regole saranno interessanti;

Chi-Square Test: supponiamo di osservare il comportamento di due variabili A e B, proprio come quelle della Tabella 1.1. A può assumere solo R valori distinti a_1, a_2, ..., a_R e B può assumere solo C valori distinti b_1, b_2, ..., b_c. Le variabili A e B possono essere nominali o ordinali. Le variabili A e B sono indipendenti se e solo se, in riferimento alla Tabella 1.1, $a = n_1 m_1$. Diciamo anche che N_{ij} è il numero di osservazioni per la $i$esima riga e per la $j$esima colonna.
$N_{i.} = \sum_j N_{ij}$, $N_{.j} = \sum_i N_{ij}$, $N = \sum_i \sum_j N_{ij}$
e che p_{ij} sia la probabilità che un'istanza cada nella categoria a_j e nella categoria b_j; in particolare:
$p_{i.}$ è la probabilità che un'istanza cada nella categoria a_i.
$p_{.j}$ è la probabilità che un'istanza cada nella categoria b_j.
Le stime $\widehat{p_{i.}}$ e $\widehat{p_{.j}}$ di $p_{i.}$, $p_{.j}$ sono date da $\widehat{p_{i.}} = N_{i.}/N$, $\widehat{p_{.j}} = N_{.j}/N$.
A questo punto se A e B sono indipendenti, allora $p_{ij} = p_{i.}p_{.j}$ e il numero atteso di osservazioni nella categoria (a_i, b_j) è $Np_{i.}p_{.j}$. Così la stima del numero atteso di osservazioni nella cella (i, j) è dato da:

$$E_{ij} = N_{i.}N.j/N.$$

Per testare l'ipotesi di indipendenza di A e B, mettiamo semplicemente a confronto i numeri attesi con quelli osservati calcolando:

$$\chi^2 = \sum_i \sum_j \frac{(N_{ij} - E_{ij})^2}{E_{ij}}.$$

Sotto l'ipotesi di indipendenza statistica, la quantità di χ^2 ha una distribuzione chi-quadrata con $(R-1)(C-1)$ gradi di libertà. Chiaramente, se χ^2 è ampio, esso indica che i numeri osservati e attesi differiscono considerevolmente e l'ipotesi di indipendenza dovrebbe essere scartata. Più formalmente, il test chi-quadrato rifiuta l'ipotesi di indipendenza al livello $(1-\alpha)$ se:

$$\chi^2 > \chi^2_{(R-1)(C-1)}; 1-\alpha$$

dove $\chi^2 > \chi^2_{(R-1)(C-1)}; 1-\alpha$ è l'$1-\alpha$ percentile della distribuzione chi-quadrata con $(R-1)(C-1)$ gradi di libertà. Il test chi-quadrato di correlazione si applica all'intera tabella di contingenza, ma, comunque, per le regole associative è sufficiente un test di

dipendenza su una singola cella. Per una tabella di contingenza $2x2$, questi due test sono equivalenti semplicemente perché, se gli eventi A e B sono indipendenti, allora sono indipendenti anche A e $\neg B$, $\neg A$ e B e $\neg A$ e $\neg B$ e il risultato del test chi-quadrato supporterà l'ipotesi di indipendenza.

Il numero delle misure oggettive di interesse è molto ampio, e un elenco di queste ultime è presente nella Tabella 1.3 con riferimento alla tabella di contingenza presentata nella Tabella 1.1.

Durante il processo di *mining*, le misure di interesse possono essere usate in tre modi, modi che vengono chiamati i *ruoli* delle misure di interesse. Per prima cosa, le misure possono essere usate per scartare i pattern non interessanti durante il processo di estrazione delle regole, per aumentare così anche l'efficienza del processo stesso. In secondo luogo, le misure possono essere usate per ordinare i pattern in base ai valori di interesse prodotti dalle misure utilizzate. Da ultimo, le misure possono essere usate durante la fase di *postprocessing* per selezionare i pattern più interessanti. Per esempio, dopo il processo di mining, potremmo usare il *Chi-Square test* per selezionare tutte le regole aventi un livello di correlazione significativo.

Sulla base degli obiettivi dell'utente, gli esperti di *data mining* possono proporre l'uso di una misura di interesse appropriata, ma questo compito di selezione può non essere svolto solamente dagli esperti del dominio senza la collaborazione dell'utente. Questa scelta è difficile, dal momento che ogni misura è caratterizzata da qualità o difetti differenti rispetto alle altre, e non esiste una misura che possa essere definita *ottimale* in senso stretto. Un modo per risolvere questo problema è quello di cercare di trovare un buon compromesso.

Lo studio delle misure oggettive di interesse diventa quindi essenziale per cercare di trovare la misura migliore nell'interesse dell'utente che ha intrapreso l'analisi e del contesto della stessa.

1.6.5 Le proprietà delle misure oggettive di interesse

Le misure oggettive presentate nella Tabella 1.3 provengono da diversi campi quali la statistica, le scienze sociali, il *machine learning* e il *data mining* e sono utilizzate per valutare l'interesse delle regole associative. Comunque, in molte situazioni, queste misure possono fornire un'informazione contrastante circa il livello di interesse di una regola associativa e mentre una misura può risultare corretta in una determinata applicazione, un'altra può non esserlo. Ad aumentare il livello di difficoltà presente nello scegliere la misura più appropriata in uno specifico contesto, contribuisce il fatto che i risultati proposti dalle varie misure spesso non sono consistenti gli uni con gli altri. Così, per capire perché alcune delle misure sono inconsistenti, diventa necessario esaminare le proprietà di ciascuna misura.

Poniamo che $T(D) = \{t_1, t_2, ..., t_N\}$ sia l'insieme delle tabelle di contingenza $2x2$ derivabili da un data set D e che M sia l'insieme delle misure oggettive disponibili per la nostra analisi. Per ciascuna misura, $M_i \in M$ possiamo costruire un *vettore di interesse* $M_i(T) = \{m_{i1}, mi2, ..., m_{iN}\}$, dove ciascuna voce m_{ij} corrisponde al valore di M_i per la tabella t_j. Ciascun vettore di interesse può inoltre essere trasformato in un *vettore di classificazione* $O_i(T) = \{o_{i1}, o_{i2}, ..., oiN\}$, dove o_{ij} corrisponde all'ordine di m_{ij} e $\forall j, k : o_{ij} \leq o_{ik}$ se e solo se $m_{ik} \geq m_{ij}$.

Possiamo così definire la consistenza tra una coppia di misure nei termini della similitudine che intercorre tra i loro vettori di classificazione ed, in particolare, misureremo questa similitudine usando la correlazione di Paerson.

A questo punto la definizione di *consistenza* è quella che segue: due misure, M_1 e M_2 sono consistenti l'una rispetto all'altra in riferimento ad un data set D se la *correlazione* tra $O_1(T)$ e $O_2(T)$ è maggiore o uguale a una qualche soglia positiva t.

Piatetsky-Shapiro [25] ha proposto tre proprietà chiave che una buona misura M dovrebbe soddisfare:

- P1: $M = 0$ se A e B sono statisticamente indipendenti;
- P2: M cresce monotonicamente con a quando n_1 ed m_1 rimangono gli stessi;
- P3: M decresce monotonicamente con n_1 (o m_1) quando il resto dei parametri (a e n_1 o m_1) rimangono gli stessi.

Il primo principio (P1) stabilisce che una regola associativa che occorre per caso ha un valore di interesse pari a zero, cioè, non è interessante. Il secondo principio (P2) stabilisce che, più grande è il supporto per $A \rightarrow B$, più grande è l'interesse quando il supporto per A e per B è fisso, cioè, più positiva è la correlazione tra A e B, più interessante è la regola. Il terzo principio (P3) stabilisce che se il supporto per AB e B (o A) è fisso, più piccolo è il supporto per A (o B), più interessante è la regola.
In seguito, sono state proposte [19] altre cinque proprietà basate sulle operazioni riferibili ad una tabella di contingenza $2x2$.

- O1: M dovrebbe essere simmetrica permutando le variabili, cioè $M(A \rightarrow B) = M(B \rightarrow A)$;
- O2: M dovrebbe rimanere la stessa quando scaliamo qualsiasi riga o qualsiasi colonna di un fattore positivo;
- O3: M dovrebbe diventare $-M$ se sia le righe che le colonne sono invertite, cioè, i valori di interesse presenti nella tabella di contingenza dovrebbero cambiare segno scambiando le colonne con le righe o viceversa;
- O4: M dovrebbe rimanere la stessa se sono invertite sia le righe che le colonne;
- O5: M non dovrebbe avere alcun tipo di relazione con il conteggio dei record che non contengono né A né B.

Contrariamente ai principi di Piatetsky-Shapiro, queste proprietà non dovrebbero essere interpretate come delle affermazioni riguardo a ciò che è desiderabile. Al contrario, esse possono essere usate per classificare le misure in differenti gruppi. La proprietà (O1) stabilisce che le regole $A \rightarrow B$ e $B \rightarrow A$ dovrebbero avere lo stesso valore di interesse, cosa che, invece, non è affatto vera per molte applicazioni. Per esempio, la confidenza rappresenta la probabilità di un conseguente, dato un antecedente, ma non vice versa: ciò la rende una misura *asimmetrica*. Per fornire un maggior numero di misure simmetriche, gli autori di queste proprietà hanno trasformato ciascuna misura asimmetrica M in una misura simmetrica, prendendo semplicemente il valore massimo tra $M(A \rightarrow B)$ e $M(B \rightarrow A)$. Per esempio, hanno definito una misura di confidenza simmetrica, che è definita come: $\max(\frac{a}{n_1}, \frac{n_1}{a})$. Esempi di misure simmetriche sono: *cosine*, l'*interest factor*, la *odds ratio* e la *correlazione*. La proprietà (O2) richiede stabilità quando vengono scalate o le righe o le colonne di una tabella di contingenza. In particolare, una misura oggettiva M è stabile quando vengono scalate o le righe o le colonne se $M(T) = M(T')$ dove T è una tabella di contingenza caratterizzata da questi conteggi di frequenza $[f_{11}; f_{10}; f_{01}; f_{00}]$, T' è una tabella di contingenza con questi conteggi di frequenza opportunamente scalati $[k_1k_3f_{11}; k_2k_3f_{10}; k_1k_4f_{01}; k_2k_4f_{00}]$ e k_1, k_2, k_3, k_4 sono costanti positive. Esempi di misure che rispettano questa proprietà sono: la *odds ratio* e i coefficienti Q e Y di *Yule*. La proprietà (O3) stabilisce che $M(A \rightarrow B) = -M(A \rightarrow \neg B) = -M(\neg A \rightarrow B)$. Questa proprietà significa che la misura può identificare sia le correlazioni positive che le correlazioni negative. Esempi di misure che rispettano questa proprietà sono: *Piatetsky-Shapiro*, la *correlazione* e i coefficienti Q ed Y di *Yule*. La proprietà (O4) stabilisce che $M(A \rightarrow B) = M(\neg A \rightarrow \neg B)$. La proprietà (O3) è difatti un caso speciale della proprietà (O4) poiché, se la permutazione delle righe (o delle colonne) cambia il segno una volta e la permutazione delle colonne

(o delle righe) lo cambia ancora, allora il risultato generale nel permutare sia le righe che le colonne lascerà il segno inalterato. Esempi di misure che rispettano la proprietà (O4) sono: la *correlazione*, la *odds ratio*, *Piatetsky-Shapiro* e il *certainty factor*. La proprietà (O5) stabilisce che una misura dovrebbe tenere conto solo del numero di record che contengono A e B, oppure entrambi. Il supporto non soddisfa quest'ultima proprietà, mentre la confidenza sì. Altri esempi di misure che rispettano questa proprietà sono: *Laplace*, *Cosine* e *Jaccard*.

Ciò che abbiamo appena presentato suggerisce che non esiste nessuna misura che sia consistentemente migliore di un'altra nei domini di tutte le applicazioni. Questo accade perché differenti misure hanno differenti proprietà intrinseche, alcune delle quali possono essere auspicabili per certe applicazioni, ma non per altre. Così, per trovare la giusta misura, abbiamo bisogno di confrontare le proprietà di cui un'applicazione necessita con le proprietà che una determinata misura possiede. Questo può essere fatto calcolando la similitudine tra un vettore delle proprietà che rappresenta le proprietà necessarie ad un'applicazione con un vettore delle proprietà che rappresenta, invece, le proprietà intrinseche delle misure esistenti.

Gli autori che hanno sviluppato queste proprietà, hanno studiato 21 misure per cercare di capire se, tra di esse, ne esistevano alcune che fossero consistenti riguardo alle proprietà da loro introdotte e riguardo alle proprietà introdotte da Piatetsky-Shapiro. Il frutto della loro analisi può essere riassunto nella Tabella 1.5:

Tabella 1.5: Le proprietà delle misure oggettive di interesse secondo [19].

Misura	**P1**	**P2**	**P3**	**O1**	**O2**	**O3**	**O4**	**O5**
Correlation Analysis	Sì	Sì	Sì	Sì	No	Sì	Sì	No
Continua nella prossima pagina								

Tabella 1.5 – continua dalla pagina precedente

Misura	P1	P2	P3	O1	O2	O3	O4	O5
Goodman and Kruskal	Sì	No	No	Sì	No	No*	Sì	No
Odds ratio	Sì*	Sì	Sì	Sì	Sì	Sì*	Sì	No
Yule's Q	Sì	Sì	Sì	Sì	Sì	Sì	Sì	No
Yule's Y	Sì	Sì	Sì	Sì	Sì	Sì	Sì	No
Yule's Y	Sì	Sì	Sì	Sì	Sì	Sì	Sì	No
Cohen's	Sì	Sì	Sì	Sì	No	No	Sì	No
Mutual information	Sì	Sì	Sì	No**	No	No*	Sì	No
J-Measure	Sì	No	No	No**	No	No	No	No
Gini index	Sì	No	No	No**	No	No*	Sì	No
Support	No	Sì	No	Sì	No	No	No	No
Confidence	No	Sì	No	No**	No	No	No	Sì
Laplace	No	Sì	No	No**	No	No	No	No
Convinction	No	Sì	No	No**	No	No	Sì	No
Interest	Sì*	Sì	Sì	Sì	No	No	No	No
Cosine	No	Sì	Sì	Sì	No	No	No	Sì
Piatetsky-Shapiro	Sì	Sì	Sì	Sì	No	Sì	Sì	No
Certainty factor	Sì	Sì	Sì	No**	No	No	Sì	No
Added Value	Sì	Sì	Sì	No**	No	No	No	No
Collective strength	No	Sì	Sì	Sì	No	Sì*	Sì	No
Jaccard	No	Sì	Sì	Sì	No	No	No	Sì
Klosgen	Sì	Sì	Sì	No**	No	No	No	No

Dove:

Sì: sì se la misura è normalizzata.*

No: simmetrica con la permutazione delle righe o delle colonne.*

*No**: no, a meno che la misura non sia resa simmetrica prendendo il* $\max(M(A,B), M(B,A))$.

Guardando la Tabella 1.5, ciò che si può facilmente notare è che nessuna misura è consistente rispetto a tutte le altre riguardo al proprio valore di interesse nei confronti delle proprietà precedentemente esposte. Gli autori [19] hanno quindi cercato di capire quali fossero le situazioni che potevano influire in questa 'mancata consistenza'. Uno dei fattori che contribuisce ad accrescere questo problema è l'eliminazione dei pattern non frequenti sulla base della proprietà antimonotonica del supporto. Ciò che però è stato notato è che, quando la soglia minima di supporto usata per l'eliminazione degli itemset non frequenti viene usata insieme ad una soglia massima di supporto, i valori di correlazione tra le misure tendono a migliorare fino ad arrivare ad un punto in cui più del 71% delle coppie di misure analizzate ha un valore di correlazione superiore allo 0.85.

Un'altra soluzione proposta per rendere le misure consistenti è quella riguardante la standardizzazione delle tabelle di contingenza, cioè quando i valori non uniformi del supporto dei totali marginali di queste ultime vengono standardizzati e resi uguali (cioè: $n_1^* = m_1^* = n_2^* = m_2^* = N/2$)(vedi Tabella 1.6). Questo problema riscontrato nell'area delle regole associative è in qualche modo analogo a ciò che accade quando si usa l'*accuratezza* per valutare le performance di un modello di classificazione. Se, cioè, un data set contiene il 99% di esempi di una classe 0 e l'1% di esempi di una classe 1, allora un classificatore che produce un modello che classifica ciascun esempio di test come appartenente alla classe 0 è caratterizzato da un alto livello di accuratezza, nonostante si comporti in maniera pessima nel classificare gli esempi della classe 1. Così, l'accuratezza non è una misura affidabile poiché può essere facilmente oscurata dalle differenze nella distribuzione dei valori delle classi. Un modo per risolvere questo problema è quello di stratificare il data set, cosicché entrambe le classi abbiano una rappresentazione uguale durante la costruzione del modello. Una simile strategia di 'stratificazione' può essere usata per trattare le tabelle di contingenza caratterizzate da un

supporto non uniforme, cioè standardizzando i conteggi delle frequenza di una tabella di contingenza.

Nonostante i due approcci appena descritti presentino due scenari in cui

Tabella 1.6: La standardizzazione di una tabella di contingenza.

	$\boldsymbol{B}$	$\neg B$	
$\boldsymbol{A}$	a	b	n_1
$\neg A$	c	d	n_2
	m_1	m_2	N

$\rightarrow$

	$\boldsymbol{B}$	$\neg B$	
$\boldsymbol{A}$	a^*	b^*	n_1^*
$\neg A$	c^*	d^*	n_2^*
	m_1^*	m_2^*	N

$\rightarrow$

	$\boldsymbol{B}$	$\neg B$	
$\boldsymbol{A}$	x	$N/2 - x$	$N/2$
$\neg A$	$N/2 - x$	x	$N/2$
	$N/2$	$N/2$	N

molte delle misure diventano consistenti l'una con l'altra, tali scenari possono non essere validi per i domini di tutte le applicazioni. Per esempio, l'utilizzo di una soglia minima e massima di supporto, può non risultare utile per i domini che contengono variabili nominali, mentre, in altri casi, un potenziale utente può non sapere esattamente quale schema di standardizzazione sia più corretto seguire.

Viene quindi descritto [19] un approccio alternativo di tipo soggettivo per trovare la giusta misura, sulla base degli ordinamenti relativi forniti dagli esperti del dominio. Idealmente si vorrebbe che gli esperti ordinassero tutte le tabelle di contingenza derivabili da un determinato data set. Questo ordinamento generale potrebbe così aiutarci nell'identificare la misura che è più consistente rispetto alle aspettative degli esperti.

Sfortunatamente, però, chiedere agli esperti di ordinare manualmente tutte le tabelle è impraticabile. Un approccio più concreto sarebbe quello di fornire agli esperti un insieme più piccolo rispetto all'insieme totale contenente tutte le tabelle di contingenza. Per fare questo è però necessario identificare un piccolo sottoinsieme di tabelle di contingenza che ottimizzi i seguenti criteri:

1. il sottoinsieme deve essere piccolo abbastanza da permettere agli esperti di ordinarlo manualmente. D'altro canto, però, il sottoinsieme deve essere grande abbastanza da assicurare che scegliere la misura migliore da quel sottoinsieme sia quasi equivalente a sceglierla dall'insieme originale delle tabelle di contingenza;

2. il sottoinsieme deve essere abbastanza diversificato da cogliere quanti più possibili conflitti di ordinamento tra le differenti misure.

Il primo criterio è di solito determinato dagli esperti poiché sono loro quelli che decidono il numero di tabelle che riusciranno ad ordinare. Quindi, l'unico criterio che ci è concesso di ottimizzare algoritmicamente è quello riguardante la diversità del sottoinsieme da analizzare.
Per ottimizzare la diversità di questo sottoinsieme, gli autori delle cinque proprietà viste sopra [19] propongono due diversi algoritmi: *RANDOM* e *DISJOINT*.
Il primo seleziona in maniera random k delle N tabelle che dovranno essere presentate agli esperti. Ci si aspetterebbe che l'algoritmo *RANDOM* lavori malamente quando $k \leq N$. Tuttavia, i risultati ottenuti usando questo algoritmo sono piuttosto interessanti poiché essi possono servire come una linea di riferimento.
Il secondo algoritmo, per prima cosa, calcola la media e la deviazione standard degli ordinamenti di tutte le tabelle. In secondo luogo aggiunge all'insieme finale da presentare agli esperti la tabella di contingenza che è caratterizzata dalla più grande quantità di conflitti in termini di ordinamento. Andando avanti, *DISJOINT* calcola la 'distanza' tra ciascuna coppia di tabelle e cerca di trovare le k tabelle che sono più 'distanti' secondo i loro ordinamenti medi e che producono la più grande quantità di conflitti di ordinamento riguardo alla deviazione standard dei loro vettori di ordinamento.
È stato dimostrato [19] che i risultati che permette di ottenere *DISJOINT* su un piccolo campione costituito da 20 tabelle sono abbastanza consistenti con

l'ordinamento delle misure ottenibile qualora l'intero insieme delle tabelle fosse ordinato dagli esperti del dominio. Una controindicazione riguardo a questo algoritmo è quella riguardante però una sua possibile implementazione, la quale potrebbe risultare piuttosto costosa poiché sarebbe necessario calcolare la distanza tra tutte le $(Nx(N-1))/2$ coppie di tabelle.
Sempre con lo scopo di cercare di capire qual è la giusta misura da utilizzare in una determinata applicazione e in uno specifico contesto, sono state proposte [26] altre cinque proprietà atte a valutare le misure oggettive di interesse:

- (Q1) M è costante se non esiste un controesempio alla regola;
- (Q2) M descresce con b in modo lineare, concavo, o convesso intorno a 0+;
- (Q3) M aumenta quando aumenta il numero totale di record;
- (Q4) M la soglia minima è facile da fissare;
- (Q5) le semantiche della misura sono facili da esprimere.

La proprietà (Q1) stabilisce che le regole con una confidenza pari a 1 dovrebbero avere lo stesso valore di interesse, qualunque sia il loro supporto, proprietà che invece contraddice quanto già detto [19] riguardo al fatto che una misura dovrebbe combinare i due aspetti di supporto e confidenza. La proprietà (Q2) describe il modo in cui il valore di interesse decresce quando sono aggiunti un po' di controesempi. Se l'utente può tollerare qualche controesempio, è auspicabile un decrescimento concavo. Se invece il sistema richiede fortemente una confidenza pari a 1, allora è auspicabile un decrescimento convesso. La proprietà (Q3) descrive i cambiamenti nei valori di interesse che occorrono quando è aumentato il numero di record presente in un data set, assumendo che n_1, m_1 e a siano mantenuti costanti. La proprietà (Q4) stabilisce che, quando viene usata una soglia per una misura di interesse per

separare le regole interessanti e non, questa stessa soglia dovrebbe essere semplice da scegliere e la sua semantica dovrebbe essere espressa facilmente. La proprietà (Q5) stabilisce che la semantica di una misura di interesse dovrebbe essere facilmente comprensibile per l'utente.

1.6.6 Selezionare la misura di interesse oggettivo più appropriata

A causa dell'elevato numero di misure di interesse proposto nella Tabella 1.3, il processo di selezione della misura più appropriata diventa un problema di non poco conto. Ciò che abbiamo esaminato fino ad ora sono le proprietà di cui queste misure godono e abbiamo cercato di verificare la consistenza di ogni misura con le altre sulla base del comportamento che le caratterizza se testate in merito a queste proprietà. Riguardo alle medesime proprietà proposte, esistono due metodi per selezionare la misura più corretta, cioè l'*ordinamento* e il *raggruppamento*. Entrambi i metodi possono essere utilizzati anche conducendo delle valutazioni empiriche direttamente sul data set.

Un metodo per selezionare la misura più appropriata è basato su un approccio che prende in considerazione più criteri per effettuare la propria decisione finale [26]. In questo stesso approccio, segni e pesi vengono assegnati a ciascuna proprietà che l'utente ritiene sia importante. Per esempio, se l'utente necessita di una proprietà simmetrica, ad una misura in esame viene assegnato un 1 se è simmetrica, e uno 0 se invece è asimmetrica. Con ciascuna riga che rappresenta una misura e con ciascuna colonna che rappresenta una proprietà, viene così costruita una matrice di decisione. Una voce nella matrice rappresenta il segno dato alla misura secondo la proprietà richiesta. Applicando questo approccio multicriterio, possiamo così ottenere un ordinamento dei risultati. Grazie a questo metodo, all'utente non è richiesto di ordinare i pattern estratti. Piuttosto, egli dovrà identificare le proprietà di

cui necessita nell'analisi che sta conducendo e dovrà specificare il significato di queste stesse proprietà in una particolare applicazione.
Un ulteriore metodo per analizzare le misure è quello di dividere le stesse in gruppi [27]. Così come per l'approccio visto sopra, questo metodo basato sul raggruppamento può essere utilizzato sia in riferimento alle proprietà delle misure sia all'intero insieme di regole generato a partire dal data set iniziale. Nel primo caso, le misure vengono raggruppate sulla base della similitudine delle loro proprietà, e si lavora su una matrice di decisione in cui ogni riga rappresenta una misura e ciascuna colonna rappresenta una proprietà. Nel secondo caso, invece, si lavora su una matrice in cui ciascuna riga rappresenta una misura e in cui ciascuna colonna prende il significato di una misura applicata ad un insieme di regole. Ciascuna voce rappresenta un valore di similitudine tra le due misure sul già specificato insieme di regole. La similitudine è calcolata a partire dagli ordinamenti delle due misure sull'insieme di regole.

1.6.7 Le misure soggettive di interesse

Nelle applicazioni dove l'utente possiede una conoscenza del contesto d'analisi, i pattern ordinati in base alle misure oggettive di interesse e giudicati come interessanti possono in realtà non esserlo. Una misura soggettiva di interesse prende in considerazione sia i dati sia la conoscenza dell'utente. Tale misura è appropriata quando: (1) la conoscenza del contesto d'analisi da parte degli utenti varia, (2) gli interessi degli utenti variano, e (3) la conoscenza del contesto d'analisi da parte degli utenti evolve. A differenza delle misure oggettive considerate in precedenza, le misure soggettive possono non essere rappresentabili da semplici formule matematiche poiché la conoscenza dell'utente può non essere rappresentata in varie forme. Al contrario, esse sono solitamente incorporate nel processo di *mining*. Le misure soggettive di interesse sono basate sui criteri di *novità* e di *inaspettatezza*.

In particolare un pattern è *inaspettato* se contraddice le aspettative o la conoscenza di una persona. Può essere considerato *inaspettato* anche un pattern che è un'eccezione rispetto ad un pattern più generale che è già stato scoperto. I pattern *inaspettati* sono interessanti perché identificano delle falle nella conoscenza precedente e possono mettere in evidenza un aspetto particolare dei dati che necessita di studi aggiuntivi. La differenza tra l'inaspettatezza e la novità è che un pattern *nuovo* è appunto tale e non è contraddetto da alcun pattern già noto all'utente, mentre un pattern *inaspettato* contraddice la precedente conoscenza o le aspettative dell'utente. Un pattern è *nuovo* per una persona se lui o lei non lo conosceva precedentemente e non è possibile dedurlo da altri pattern già noti. Nessun sistema di data mining conosciuto è in grado di rappresentare tutto ciò che un utente conosce, e così, la novità può non essere misurata esplicitamente in riferimento alla conoscenza di un utente. Similmente, nessun sistema di data mining è in grado di rappresentare ciò che l'utente non conosce, e quindi, la novità può non essere misura esplicitamente in riferimento all'ignoranza dell'utente. Al contrario, la novità è rilevata quando l'utente identifica direttamente ed esplicitamente un pattern come *nuovo* oppure quando si accorge che un pattern può non essere dedotto da alcun pattern già noto e non contraddice i pattern già precedentemente scoperti. Nell'ultimo caso, i pattern scoperti vengono usati come un'approssimazione della conoscenza dell'utente.

Per trovare pattern inaspettati o nuovi all'interno dei dati, possono essere distinti tre approcci in base ai ruoli delle misure di inaspettatezza nel processo di *mining*: (1) l'utente fornisce una specificazione formale della propria conoscenza, e dopo aver ottenuto i risultati di *mining*, il sistema sceglie quali pattern inaspettati presentare all'utente; (2) secondo un *feedback* interattivo da parte dell'utente, il sistema rimuove i pattern non interessanti; e (3) il sistema applica le specifiche dell'utente come se fossero dei vincoli durante

il processo di *mining*, in modo da restringere lo spazio di ricerca e fornire un minor numero di risultati.

Usare le misure di interesse soggettivo per filtrare i pattern interessanti lavorando sui risultati del processi di *mining*

[28] hanno messo in relazione l'inaspettatezza con un sistema di convinzione. Per definire le convinzioni, hanno usato formule arbitrarie di predicati della logica del primo ordine, piuttosto che regole del tipo se-allora. Essi hanno anche classificato le convinzioni come *forti* o *deboli*. Una convinzione *forte* è un vincolo che non può essere cambiato da una nuova evidenza. Se l'evidenza (le regole estratte a partire dai dati) contraddice le convinzioni forti, deve essere stato fatto un errore nell'acquisire l'evidenza. Una convinzione *debole* è una convinzione che l'utente spera di cambiare man mano che vengono scoperti nuovi pattern. Gli autori hanno adottato un approccio *bayesiano* e hanno assunto che il grado di convinzione è misurato con una probabilità condizionale. Data un'evidenza E (pattern), il grado di convinzione in α è aggiornato con la regola di Bayes come segue:

$$P(\alpha|E,\, \xi) = \frac{P(E|\alpha,\, \xi)P(\alpha|\xi)}{P(\alpha|E,\, \xi)+P(\neg\alpha|E,\, \xi)P(\neg\alpha|\xi)},$$

dove ξ è il contesto che rappresenta la precedente evidenza che supporta α. Allora, la misura di interesse per un pattern p, relativa a un sistema di convinzione debole B, è definita come la differenza relativa tra probabilità precedenti e posteriori:

$$I(p, B) = \sum_{\alpha\in B} \frac{|P(\alpha|p,\xi)-P(\alpha|\xi)|}{P(\alpha|\xi)}.$$

[28] hanno presentato una struttura generale per definire una misura di interesse per i pattern. Consideriamo ora come questa struttura può essere applicata ai pattern nella forma di regole associative. In riferimento alla Tabella 1.7, definiamo una convinzione α come 'le persone che comprano latte, uova e pane'. Qui, ξ denota il data set D. Inizialmente, supponiamo

Tabella 1.7: Esempio di un data set transazionale

Latte	Pane	Uova
1	0	1
1	1	0
1	1	1
1	1	1
0	0	1

che l'utente specifica il grado di convinzione in α come $P(\alpha|\xi) = 2.5 = 0.4$, sulla base del data set, dal momento che due delle cinque transazioni in esso presenti supportano la convinzione α. Similmente, $P(\neg\alpha|\xi)$ = supporto = 0.4 e confidenza $= 2/3 \approx 0.67$. Il nuovo grado di convinzione in α, sulla base della nuova evidenza p nel contesto della vecchia evidenza ξ, è denotato da $P(\alpha|p, \xi)$. Esso può essere calcolato con la regola di Bayes data precedente, sempre se noi conosciamo i valori dei termini di $P(\alpha|\xi)$, $P(\neg\alpha|\xi)$, $P(p|\neg\alpha, \xi)$ e $P(p|\neg\alpha, \xi)$. I valori dei primi due termini sono già stati calcolati.

Gli altri due termini possono essere calcolati come segue. Il termine $P(p|\alpha, \xi)$ rappresenta la confidenza della regola p, data una convinzione α, cioè, la confidenza della regola *latte* $\rightarrow$ *uova* valutata sulle transazioni 3 e 4, dove il latte, le uova e il pane appaiono insieme. Dallta Tabella 1.7, noi otteniamo $P(p|\alpha, \xi) = 1$. Similimente, il termine $P(p|\neg\alpha, \xi)$ rappresenta la confidenza della regola p, data una convinzione $\neg\alpha$, cioè, la confidenza della regola *latte* $\rightarrow$ *uova* valutata sulle transazioni 1, 2 e 5 dove il latte, le uova e il pane non appaiono insieme. Dalla Tabella 1.7, noi otteniamo $P(p|\neg\alpha, \xi) = 0.5$. Usando la regola di Bayes, calcoliamo $P(\alpha|E, \xi) = \frac{1x0.4}{1x0.4+0.5x0.6} \approx 0.57$, e secondo il valore della misura di interesse I per la regola p è calcolato come $I(p, B) = \frac{|0.57-0.4|}{0.4} \approx 0.43$.

Per ordinare le regole di classificazione secondo la conoscenza esistente dell'utente, [29] hanno proposto due tipi di specificazioni (T1 e T2) per definire

la conoscenza vaga dell'utente, specificazioni chiamate *impressioni generali*. Una impressione generale del tipo T1 può esprimere una relazione positiva o negativa tra una variabile di stato e una classe, una relazione tra una gamma (o un sottoinsieme) di valori di variabili di stato e una classe, o la vaga impressione che esista una relazione tra una variabile di stato e una classe. T2 estende T1 dividendo la conoscenza dell'utente nel *cuore* e nell'*intorno*. Il cuore fa riferimento alla conoscenza dell'utente che può essere chiaramente rappresentata e l'intorno fa riferimento alla conoscenza dell'utente che può essere solo vagamente rappresentata. Il cuore e l'intorno sono entrambi convinzioni deboli perché essi possono non essere veri, e così necessitano di essere sia verificati che contraddetti. Sulla base di queste due specificazioni, sono stati proposti degli algoritmi combacianti con queste ultime per ottenere delle regole ad esse *conformi*, e, inoltre, per ottenere delle regole con conseguenti inaspettati e delle regole con condizioni inaspettate. Queste regole vengono poi ordinate in base al grado in cui combaciano usando delle misure di interesse. Nel processo di *matching* per una regola R, le impressioni generali sono separate in due insiemi: G_S e G_D. L'insieme G_S consiste di tutte le impressioni generali con lo stesso conseguente di R, e G_D consiste di tutte le impressioni generali con differenti conseguenti rispetto a R. Nel caso di conferma della regola, R è combinato con G_S. La misura di interesse calcola la somiglianza tra le condizioni R e G_D. Nel caso della regola con conseguente inaspettato, R è combinato con G_D. La misura di interesse determina la somiglianza tra le condizioni di R e G_D. Nel caso delle regole con condizioni inaspettate, R è ancora una volta combinato con G_S, e la misura di interesse calcola la differenza tra le condizioni di R e G_S. Così, gli ordinamenti delle regole con condizioni inaspettate sono l'opposto rispetto a quelle delle regole confermanti.

Facciamo ora uso di un esempio per illustrare il calcolo dei valori di interesse nel caso delle regole confermanti usando le specificazioni del tipo T1.

Assumiamo che abbiamo scoperto una regola classificativa r, e che vogliamo usarla per confermare le impressioni generali dell'utente:

$$r\text{:senza\_lavoro} = \text{no, risparmi} > 10{,}000 \rightarrow \text{approvato,}$$

la quale stabilisce che se una persona non è senza lavoro e i suoi risparmi sono superiori ai 10,000 euro, il suo prestito verrà approvato.
Assumiamo che l'utente fornisca le seguenti cinque impressioni generali:

(G1) risparmi$>\rightarrow$ approvato
(G2) età $| \rightarrow$ {approvato, non_approvato}
(G3) senza_lavoro{no} $\rightarrow$ approvato
(G4) senza_lavoro{sì} $\rightarrow$ non_approvato
(G5) risparmi$>$,senza_lavoro{sì} $\rightarrow$ approvato

L'impressione generale (G1) stabilisce che se i risparmi di chi richiede un prestito sono considerevoli, il prestito verrà approvato. L'impressione (G2) stabilisce che l'età di chi richiede un prestito in maniera non specificata è in relazione al risultato positivo o negativo della sua pratica di prestito, e (G3) stabilisce che se chi richiede un prestito ha un lavoro, il prestito verrà approvato. L'impressione (G4) stabilisce che, se chi richiede un prestito è senza lavoro, il prestito non verrà approvato, mentre (G5) stabilisce che se i risparmi di chi richiede un prestito sono considerevoli e questa stessa persona non ha lavoro, il prestito verrà approvato comunque. Così, G_S è {(G4)}, e G_D è {(G1), (G2), (G3), (G5)}.
Dal momento che vogliamo usare una regola r per confermare queste impressioni generali, vogliamo considerare solo (G1), (G2), (G3), e (G5) poiché (G4) è caratterizzata da un conseguente differente rispetto alla regola r. L'impressione (G2) non combacia l'antecedente della regola r, e così viene scartata. L'impressione (G5) combacia parzialmente con la regola r e il grado di *matching* è rappresentato come un valore compreso tra 0 ed 1. Assumendo che 10,000 euro sia considerato come un valore piuttosto grande,

(G1) e (G3) insieme combaciano completamente con la regola r, così il grado di *matching* è pari ad 1. Finalmente, prendiamo il massimo dei valori di confronto, cioè 1, come valore di interesse per la regola r. Così, la regola r conferma fortemente le impressioni generali. Se noi volessimo trovare regole con condizioni inaspettate rispetto a regole con condizioni che confermano le impressioni generali, la regola r avrebbe dovuto avere un basso punteggio poiché esso è consistente con le impressioni generali.

[30] hanno anche proposto un'altra tecnica per ordinare le regole di classificazione secondo la conoscenza di *background* dell'utente, la quale è rappresentata dalle regole *fuzzy*. Sulla base della conoscenza esistente dell'utente, possono essere estratti tre diversi tipi di regole interessanti: i pattern *inaspettati*, i pattern *di conferma* e i pattern *azionabili*. Un pattern inaspettato è pattern che è inaspettato o precedentemente sconosciuto per l'utente, il quale corrisponde ai nostri termini di *inaspettatezza* e *novità*. Una regola può essere un pattern inaspettato se essa contiene una condizione inaspettata, un conseguente inaspettato, oppure entrambi. Un pattern di conferma è una regola che combacia parzialmente o completamente con la conoscenza già esistente dell'utente, mentre un pattern azionabile è un pattern che può aiutare l'utente a far qualcosa ricavandone un vantaggio. Per permettere ai pattern azionabili di essere identificati, l'utente dovrebbe descrivere le situazioni in cui lui in quanto utente può agire. Per tutte queste tre categorie, l'utente deve fornire alcuni pattern, rappresentati nella forma di regole *fuzzy*, che riflettono la sua conoscenza. Il sistema confronta ciascun pattern scoperto con queste regole *fuzzy*. I pattern scoperti sono quindi ordinati in base al grado in cui essi combaciano con le regole *fuzzy*. [30] hanno proposto diverse misure di interesse per le tre categorie. Tutte queste misure sono basate sulle funzioni dei valori *fuzzy* che rappresentano il grado di *match* tra la conoscenza dell'utente e i pattern scoperti.

I vantaggi dei metodi [29] e [30] sono rappresentati dal fatto che essi ordi-

nano i pattern estratti in base alla conoscenza esistente dell'utente, così come in base al data set. Lo svantaggio è rappresentato dal fatto che all'utente è richiesto di rappresentare la propria conoscenza in termini di specificazioni, compito che potrebbe non essere affatto semplice.
Le specificazioni e gli algoritmi di *matching* di [29] e [30] sono creati per le regole di classificazione, e quindi non possono essere applicati alle regole associative. Comunque, il principio generale può essere usato per le regole associative se le nuove specificazioni e gli algoritmi di *matching* sono stati proposti per le regole associative con più item presenti come conseguenti.

Eliminare i pattern non interessanti

Per ridurre la quantità di calcolo e di interazione con l'utente nel filtrare le regole associative interessanti, [31] ha proposto un metodo che rimuove le regole non interessanti, piuttosto che selezionare quelle interessanti. In questo metodo non sono definite misure di interesse; piuttosto, l'interesse di un pattern è determinato dall'utente tramite un processo interattivo. Questo metodo consiste di tre passi: (1) La migliore *regola candidata* è selezionata come la regola con esattamente un attributo nell'antecedente ed esattamente un attributo nel conseguente avente la più ampia *cover list*. La *cover list* di una regola R è l'insieme di tutte le regole estratte contenenti l'antecedente ed il conseguente di R. (2) La migliore regola candidata è presentata all'utente per la classificazione nella forma di una di quattro possibili categorie: non-vero-non-interessante, non-vero-interessante, vero-non-interessante, e vero-e-interessante. [31] ha descritto una regola come non-interessante se essa rappresenta una 'conoscenza duffusa', cioè, se è essa non è *nuova* per usare la nostra terminologia. Se la miglior regola candidata R è non-vera-non-interessante o vera-non-interessante, il sistema rimuove lei e la sua cover list. Se la regola è non-vera-interessante, il sistema rimuove

questa regola così come tutte le regole nella sua cover list che hanno il suo stesso antecedente, e mantiene tutte le regole presenti nella cover list caratterizzate dalla presenza di antecedenti più specifici. Infine, se la regola è vera-interessante, il sistema la mantiene. Questo processo si ripete fino a che l'insieme delle regole è vuoto oppure l'utente ferma il processo. I pattern rimanenti sono veri e interessanti per l'utente.
Il vantaggio di questo metodo è che agli utenti non viene richiesto di fornire delle specificazioni; piuttosto, essi lavorano interattivamente con il sistema. Gli utenti hanno solo bisogno di classificare le regole semplici come vere o false e interessanti o non interessanti, e poi il sistema può eliminare un numero significativo di regole non interessanti. L'inconveniente di questo metodo è che, sebbene esso renda l'insieme delle regole più piccolo, non ordina le regole rimanenti in base al grado di interesse. Questo metodo può essere applicato anche alle regole classificative.

Porre dei vincoli allo spazio di ricerca

Piuttosto che selezionare le regole non interessanti dopo il processo di *mining*, [32] ha proposto un metodo per ridurre lo spazio di ricerca sulla base delle aspettative dell'utente. In questo metodo, non è definita alcuna misura di interesse. Qui, le convinzioni dell'utente sono rappresentate nello stesso formato delle regole estratte. Solo le regole inaspettate, cioè, le regole che contraddicono le convinzioni esistenti, sono estratte. L'algoritmo per trovare le regole inaspettate consiste di due parti: **ZoominUR** e **ZoomoutUR**. Per una data convinzione $X \rightarrow Y$, ZoominUR trova tutte le regole della forma X, $A \rightarrow \neg Y$ aventi valori di supporto e confidenza sufficienti e che sono le regole più specifiche con un conseguente in contraddizione rispetto alla convinzione data. A questo punto, ZoomoutUR generalizza le regole trovate da ZoominUR. Per la regola X, $A \rightarrow \neg Y$, ZoomoutUR trova

tutte le regole della forma X', $A \rightarrow \neg Y$, dove X' è un sottoinsieme di X.
Questo metodo è simile ai metodi [29] e [30], in cui l'utente deve fornire una specificazione della propria conoscenza. Comunque, questo metodo non ha bisogno di trovare tutte le regole con valori sufficienti di supporto e confidenza; piuttosto, deve trovare soltanto quelle regole che sono in conflitto con la conoscenza dell'utente, cosa che rende il processo di *mining* più efficiente. Lo svantaggio è che questo metodo non ordina le regole. Sebbene [32] abbiano proposto il loro metodo per le regole associative aventi un solo item presente nei loro conseguenti, questo stesso metodo può comunque essere facilmente applicato alle regole classificative.
Sulla base delle precedenti analisi, possiamo notare che se l'utente conosce quali tipi di pattern vuole confermare o contraddire, i metodi [29], [30] e [32] sono idonei. Se l'utente non vuole rappresentare esplicitamente la conoscenza riguardo al dominio, d'altro canto, il metodo interattivo [31] è molto appropriato.

1.6.8 Le misure semantiche

Una misura semantica prende in considerazione la semantica e il significato dei pattern. Dal momento che le misure semantiche coinvolgono la conoscenza del dominio da parte dell'utente, da alcuni sono considerate un tipo particolare di misure soggettive. L'utilità e l'azionabilità dipendono dalla semantica dei dati, e in tal senso possono essere considerate semantiche. Le misure basate sull'utilità, dove la semantica rilevante è l'utilità dei pattern nel dominio, sono il tipo più comune di misure semantiche (un pattern è utile se il suo utilizzo da parte di una persona contribuisce a far raggiungere un obiettivo). Per utilizzare una approccio basato sull'utilità, l'utente deve specificare della conoscenza aggiuntiva riguardante il dominio. Al contrario delle misure soggettive, dove la conoscenza del dominio riguarda i dati stessi ed è solitamente rappresentata in un formato simile a quello

dei pattern estratti, la conoscenza del dominio richiesta per le misure semantiche non è in relazione con la conoscenza o le aspettative dell'utente riguardo i dati. Piuttosto, essa rappresenta una funzione di utilità che riflette gli obiettivi dell'utente. Questa funzione dovrebbe essere ottiminzzata nei risultati estratti. Per esempio, il manager di un negozio preferirebbe le regole associative che sono in relazione con gli item ad alto profitto piuttosto che quelle che sono statisticamente significative.

Le misure basate sull'utilità

Una misura basata sull'utilità prende in considerazione non soltanto gli aspetti statistici dei dati grezzi, ma anche l'utilità dei pattern estratti. Motivati dalla teoria della decisione, [33] hanno stabilito che 'l'interesse di un pattern = probabilità + utilità'. Sulla base degli obiettivi specifici dell'utente e sull'utilità dei pattern estratti, gli approcci di *mining* basati sull'utilità possono essere utili nelle applicazioni reali, specialmente nei problemi dove bisogna saper prendere delle decisioni.
Il metodo più semplice per includere l'utilità è chiamato *estrazione pesata di regole associative*, il quale assegna un peso a ciascun item sulla base dell'importanza di quest'ultimo [34]. Questi pesi assegnati agli item sono anche chiamati pesi *orizzantali*. Essi possono rappresentare il prezzo o il profitto di un prodotto. In questo scenario, sono state proposte due misure per rimpiazzare il supporto. La prima è chiamata supporto pesato, $(\sum_{i_j \in AB} w_j) Support(A \rightarrow B)$, dove i_j denota un item che appare nella regola $A \rightarrow B$ e w_j denota il suo peso corrispondente. Il primo fattore della misura ha una preferenza nei confronti delle regole con più item. Quando il numero degli item è ampio, anche se tutti i pesi sono piccoli, il peso totale può essere grande. La seconda misura, il supporto pesato *normalizzato*, è stato proposto per ridurre questa preferenza ed è definito come $\frac{1}{k}(\sum_{i_j \in AB} w_j) Support(A \rightarrow B)$, dove k è il numero degli item presenti nella

regola. La misura tradizionale del supporto è un caso speciale di supporto pesato normalizzato perché quando tutti i pesi per gli item sono uguali a 1, il supporto pesato normalizzato è identico al supporto.

[35] hanno proposto un altro *data model* il quale assegna un peso a ciascuna transazione. Il peso rappresenta la significatività di una transazione presente nel data set. I pesi assegnati alle transazioni sono anche chiamati pesi *verticali*. Per esempio, il peso può riflettere il tempo di una transazione, cioè, alle transazioni relativamente più recenti possono essere assegnati pesi più grandi. Sulla base di questo modello, il supporto verticale pesato è definito come:

$$Support_v(A \to B) = \frac{\sum_{AB \subseteq r} w\_v_r}{\sum_{r \in D} w\_v_r},$$

dove w_v_r denota il peso verticale per la transazione r.

Il modello misto-pesato utilizza sia i pesi orizzontali che quelli verticali. In questo modello, a ciasun item è assegnato un peso orizzontale e a ciascuna transazione è assegnato un peso verticale. Il supporto misto-pesato è definito come:

$$Support_m(A \to B) = \frac{1}{k}(\sum_{i_j \in AB} w_j) Support_v(A \to B).$$

Sia il $support_v$ che il $support_m$ sono estensioni della misura tradizionale di supporto. Se tutti i pesi verticali e orizzontali sono posti ad 1, sia il $support_v$ che il $support_m$ sono identici al supporto.

L'estrazione di associazioni orientate ad un obiettivo e basate sull'utilità (OOA) permette all'utente di impostare degli obiettivi per il processo di *mining* [33]. In questo metodo, gli attributi sono divisi in due gruppi: gli attributi *target* e gli attributi *non-target*. Ad un attributo non-target (chiamato anche *attributo non-obiettivo*) è concesso di apparire soltanto tra gli antecedenti delle regole associative. Ad un attributo target (chiamato anche *attributo ogbiettivo*) è permesso di apparire soltanto tra i conseguenti delle regole. Alle coppie target attributo-valore sono assegnati dei valori di

utilità. Il problema di *mining* concerne il trovare itemset frequenti di attributi non-target tali che i valori di utilità delle loro corrispondenti coppie target attributo-valore siano al di sopra di una data soglia. Per esempio,

Tabella 1.8: Data set di esempio

Trattamento	Efficacia	Effetti collaterali
1	2	4
2	4	2
2	4	2
2	2	3
2	1	3
3	4	2
3	1	4
4	5	2
4	4	2
4	4	2
4	3	1
5	4	1
5	4	1
5	4	1
5	3	1

nella Tabella 1.8, *Trattamento* non è un attributo target, mentre *Efficacia* e *Effetti collaterali* sono attributi target. L'obiettivo del problema di *mining* è di trovare dei trattamenti ad alta efficacia e con pochi effetti collaterali.
La misura di utilità è definita come:

$$u = \frac{1}{support(A)} \sum_{A \subseteq r \wedge r \in DB} u_r(A),$$

dove A è un itemset non-target che deve essere trovato (le coppie attributo-valore *Trattamento* presenti nell'esempio), $support(A)$ denota il supporto di A nel data set D, r denota un record che soddisfa A, e $u_r(A)$ denota l'utilità di A in termini di record r. Il termine $u_r(A)$ è definito come:

$$u_r(A) = \sum_{A_i = v \in C_r} u_{A_i} = v.$$

dove C_r denota l'insieme degli item target in un record r, $A_i = v$ è una coppia attributo-valore di un attributo target, e $u_{A_i} = v$ denota l'ultima utilità associata. Se esiste solo un attributo target e il suo peso è uguale ad 1, allora $\sum_{A \subseteq r \wedge r \in DB} u_r(A)$ è identico al $support(A)$ e quindi u è uguale ad 1.

Proseguendo con l'esempio, assegniamo i valori di utilità alle coppie target

Tabella 1.9: Valori di utilità per *Efficacia* ed *Effetti collaterali*

Efficacia			Effetti collaterali		
Valore	Significato	Utilità	Valore	Significato	Utilità
5	Molto meglio	1	4	Molto serio	-0.8
4	Meglio	0.8	3	Serio	-0.4
3	Nessun effetto	0	2	Un po'	0
2	Peggio	-0.8	1	Normale	0.6
1	Molto peggio	-1			

Tabella 1.10: Data set di esempio

Itemset	Utilità
Trattamento = 1	-1.6
Trattamento = 2	-0.25
Trattamento = 3	-0.066
Trattamento = 4	-0.8
Trattamento = 5	1.2

attributo-valore mostrate nella Tabella 1.9, e secondo ciò otteniamo i valori di utilità per ciascun trattamento mostrato nella Tabella 1.10. Per esempio, il Trattamento 5 ha il più grande valore di utilità (1.2), e quindi, incontra al meglio gli obiettivi specificati dall'utente.

Questo *data model* è stato generalizzato in [36]. Gli attributi sono ancora una volta classificati in attributi target e non-target, chiamati rispettivamente *attributi segmento* e *attributi statistici*. Per un itemset X composto da attributi non-target, la misura di interesse, chiamata *statistica*, è definita

come *statistica* $= f(D_x)$, dove D_x denota l'insieme di record che soddisfano X. La funzione f calcola statistica a partire dai valori degli attributi target in D_x. Sulla base di questa struttura astratta, è stato proposto [36] un altro modello dettagliato chiamato *marketshare*. In questo modello, gli attributi target sono MSV e P. L'attributo MSV è un attributo categorico per il quale i valori di *marketshare* devono essere calcolati, per esempio *CompanyName*. P è un attributo continuo, come *GrossSales*, cioè la base per il calcolo di *marketshare* per MSV. La misura di interesse chiamata *marketshare* è definita come:

$$msh = \sum_{r \in D_x \wedge MSV_r = v} P_r / \sum_{r \in D_x} P_r,$$

dove P_r denota il valore P per il record r, e MSV_r denota il valore MSV per il record r. Una tipica semantica per questa misura è la percentuale di vendite P, per una specifica compagnia MSV, in certe condizioni X. Se P_r è impostato ad 1 per tutti i record r, allora msh è uguale alla $confidenza(X \rightarrow (MSV = v))$.

Tutte le misure di utilità discusse in questa sezione sono estensioni delle misure di supporto e confidenza, e la maggior parte di queste estendono l'algoritmo standard Apriori identificando delle proprietà di tipo *upper-bound* per il passo di *pruning*. Nessuna singola misura di utilità è adatta per ogni possibile applicazione, poiché diverse applicazioni hanno differenti obiettivi e *data model*.

L'azionabilità

Come già detto, un pattern azionabile può aiutare l'utente a fare qualcosa traendone un vantaggio. [29] hanno suggerito che, se vogliamo che i pattern azionabili siano identificabili, allora l'utente dovrebbe descrivere le situazioni in cui può agire. Usando questo approccio, l'utente fornisce alcuni pattern, nella forma di regole *fuzzy*, i quali rappresentano le possibili

azioni e le possibili situazioni con cui l'utente stesso si troverà ad avere a che fare. Così come con i pattern di conferma, il sistema confronta ciascun pattern estratto con le regole *fuzzy* e poi li ordina in base al grado in cui essi combaciano. Le azioni con i più alti gradi di abbinamento sono selezionate per essere eseguite.

[37] hanno proposto una misura per trovare le azioni ottimali per una gestione profittevole della relazione con il cliente. In questo metodo, viene estratto un albero di decisione a partire dai dati. I nodi che sono foglie corrispondono alle condizioni del cliente, mentre i nodi foglia sono in relazione con il profitto che può essere ottenuto sempre dal cliente. Viene assegnato un costo per cambiare la condizione di un cliente. Sulla base dei valori di *gain information* del costo e del profitto, il sistema trova l'*azione ottimale*, cioè, l'azione che massimizza il $profitto_{gain} - \sum_{costo}$. Dal momento che questo metodo lavora su un albero di decisione, è prontamente applicabile alle regole di classificazione, ma non alle regole associative.

[38] hanno suggerito un metodo integrato per estrarre le regole associative e per consigliare la regola migliore rispetto al profitto che l'utente ne può trarre. In aggiunta al supporto e alla confidenza, il sistema integra due altre misure: il *profitto della regola* e la *segnalazione del profitto*. Il profitto della regola è definito come il profitto totale ottenuto nelle transazioni per una regola che combacia con la regola. La segnalazione del profitto per una regola è la media del profitto per ciascuna transazione che combacia con la regola. Il sistema di segnalazione sceglie le regole secondo la segnalazione del profitto, il profitto della regola e la concisione. Questo metodo può essere applicato direttamente alle regole di classificazione se l'informazione riguardante il profitto è integrata in tutti gli attributi rilevanti.

1.6.9 La selezione delle regole più interessanti in termini di *neighborhood-based unexpectedness* [1]

L'*unexpectedness*, cioè l'interesse di una regola determinato dalla sua propria *inaspettatezza* riguardo a ciò che è già noto, è stato già discusso nelle sezioni precedenti. In questo caso, però, quello di cui vogliamo parlare è dell'inaspettatezza di una regola valutata rispetto ai *neighborhood* di quella regola, cioè rispetto ai suoi vicini. Ciò che si pensa è infatti che i vicini di una regola dovrebbero essere presi in considerazione quando si parla di inaspettatezza. Se facciamo un'analogia con le montagne, normalmente una persona non direbbe che tutte le cime dell'Himalaya che superano i 4000 metri di altezza sono più interessanti delle montagne più alte presenti nel Nord America e in Giappone, sebbene queste cime siano più alte delle più alte montagne esistenti nel Nord America e in Giappone. Piuttosto, l'interesse di una montagna dovrebbe dipendere dalla sua altezza così come dalla sua posizione nel suo 'intorno'; così, il giapponese Monte Fuji è famoso perché non esistono cime a lui comparabili nel suo intorno. Parlando con la terminologia propria delle regole associative, possiamo dire che l'interesse di una regola dovrebbe dipendere dalla sua confidenza così come dal grado di fluttuazione della confidenza nel suo intorno e sulla densità delle regole estratte in quel determinato intorno.

Per definire l'intorno di una regola è necessario introdurre delle funzioni che definiscano la distanza tra le stesse regole. La distanza può essere definita semanticamente o sintatticamente.

La distanza semantica tra due regole $X_1 \rightarrow Y_1$ e $X_2 \rightarrow Y_2$ è definita come: $Dist_{mset} = |m(X_1Y_1)| + |m(X_2Y_2)| - 2 * |m(X_1Y_1X_2Y_2)|$. Usando questa definizione, possono esserci due diverse regole R_1 e R_2, più precisamente $A \rightarrow B$ e $B \rightarrow A$, tali che $Dist_{mset}(R_1, R_2) = 0$. Così $Dist_{mset}$ non è una distanza metrica se usata su tutto l'insieme di potenziali regole. Questo può poi portare a intorni 'più affollati', e ciò può avere alcuni effetti sulla

inaspettatezza *neighborhood-based*.

La distanza definita sintatticamente è invece una distanza di tipo metrico e, per questo, sarà quella che useremo da questo punto in avanti. La 'nostra' funzione di distanza è definita in maniera tale che qualcuno potrebbe assegnare differenti scale di importanza alle differenze presenti in diverse parti delle regole. Le differenze tra gli itemset sono divise in tre parti: (1) la differenza simmetrica tra tutti gli item presenti nelle due regole, (2) la differenza simmetrica tra gli antecedenti delle due regole, (3) la differenza simmetrica tra i conseguenti delle due regole.

Quindi, dati 3 numeri reali non negativi $\delta_1, \delta_2, \delta_3$, definiamo la *itemset distance* tra due regole: $R_1 : X_1 \rightarrow Y_1$ e $R_2 : X_2 \rightarrow Y_2$ come:

$$Dist_{iset}(R_1, R_2) = \delta_1 * |(X_1Y_1) \ominus (X_2Y_2)| + \delta_2 * |X_1 \ominus X_2| + \delta_3 * |Y_1 \ominus Y_2|.$$

Per illustrare meglio questa definizione, consideriamo le seguenti regole:

$$R_1 : D \rightarrow BC$$
$$R_2 : AD \rightarrow BC$$
$$R_3 : BC \rightarrow D$$
$$R_4 : BC \rightarrow AD$$

A questo punto, la $Dist_{iset}(R_1, R_2) = \delta_1 + \delta_2$ e la $Dist_{iset}(R_3, R_4) = \delta_1 + \delta_3$. Sia $\delta_1 + \delta_2$ e $\delta_1 + \delta_3$ sono determinati da A, dal momento che gli item B, C e D non portano alcun contributo al calcolo delle due distanze.

Per meglio illustrare il fatto che diverse differenze di posizione portano differenti contributi al calcolo della distanza, consideriamo le regole seguenti:

$$R_5 : AB \rightarrow CD$$
$$R_6 : ADF \rightarrow CE$$

A questo punto, la $Dist_{iset}(R_5, R_6) = 3 * \delta_1 + 3 * \delta_2 + 2 * \delta_3$. Il contributo dei differenti δ è quello che segue:

- l'item A occorre in entrambe le regole e occorre sullo stesso lato delle due regole. Così A non porta alcun contributo al calcolo della distanza. Lo stesso accade per C;

- l'item D occorre in entrambe le regole, ma su differenti lati. Così D porta un contributo al calcolo della distanza $\delta_2 + \delta_3$;

- l'item B occorre in una regola e, in particolare, nell'antecedente di quest'ultima; esso non occorre nell'altra regola. Così B non porta alcun contributo al calcolo della distanza $\delta_1 + \delta_2$. Lo stesso accade con F;

- l'item E occorre in una regola, e nel suo conseguente; esso non occorre nell'altra regola. Così E non porta alcun contributo al calcolo della distanza $\delta_1 + \delta_3$.

Differente scelte di valori per δ_1, δ_2 e δ_3 possono essere usate per riflettere le preferenze dell'utente. Per il resto di questa presentazione noi imposteremo $\delta = 1$, $\delta_2 = \frac{n-1}{n^2}$ e $\delta_3 = \frac{1}{n^2}$, dove $n = |I|$, cioè al numero totale di item. La scelta di avere $\delta_1 > \delta_2 > \delta_3$ riflette la convinzione che tre tipi di differenze tra itemset dovrebbero contribuire in maniera differente al calcolo della distanza: la differenza totale $(X_1Y_1) \ominus (X_2Y_2)$ è qui più importante della differenza tra gli antecedenti $X_1 \ominus X_2$, la quale è a sua volta più importante della differenza tra i conseguenti $Y_1 \ominus Y_2$. In particolare, δ_1 e δ_2 sono rispettivamente impostati a $\frac{n-1}{n^2}$ e $\frac{1}{n^2}$ per fare in modo che le regole aventi insiemi identici di item siano più vicine una all'altra rispetto alle regole aventi insiemi di item differenti. Naturalmente, la distanza può essere variata impostando δ_1, δ_2 e δ_3 in maniera differente; tutti i valori qui esposti sono quelli originali risalenti all'esposizione di [1].
L'intorno di una regola è quindi così composto da quelle regole la cui distanza nei confronti di una certa regola è minore o uguale ad una soglia

prefissata definita dall'utente, cioè: un $r - intorno$ di una regola $R_0(r > 0)$, formalmente definito come $N(R_0, r)$, è il seguente insieme:

$$\{R : Dist_{iset}(R, R_0) \leq r, \text{con R una regola potenziale.}$$

Una volta che abbiamo definito questo intorno l'interesse delle regole trovate viene calcolato o in termini di *confidenza inaspettata* o in termini di *densità inaspettata*.

Per trovare la 'confidenza inaspettata', c'è bisogno di introdurre due misure riguardanti la fluttuazione dei valori della confidenza delle regole estratte in un intorno: la *confidenza media* e la *deviazione standard* della confidenza.

Supponiamo che M sia un insieme di regole estratte con un certa soglia minima di supporto e confidenza - rispettivamente, *min_support* e *min_confidence* - che R_0 sia una regola estratta in M e che r sia > 0. A questo punto:

- la *confidenza media* dell'r-intorno di R_0 è definita come la media dei valori della confidenza delle regole nell'insieme $M \cap N(R_0, r) - \{R_0\}$; per denotare questo valore usiamo la dicitura $avg\_conf(R_0, r)$;

- la *deviazione standard* dell'r-intorno di R_0 è definita come la deviazione standard dei valori della confidenza nell'insieme $M \cap N(R_0, r) - \{R_0\}$; per denotare questo valore useremo la dicitura $std\_conf(R_0, r)$.

Quando l'insieme $M \cap N(R_0, r) - \{R_0\}$ è vuoto, [1] decidono di impostare questi due valori a zero, sebbene siano possibili altre scelte ugualmente valide. Per identificare la confidenza inattesa di una regola R_0 in un r-intorno si ha la condizione che $|conf(R_0) - avg\_conf(R_0, r)|$ sia molto più ampia di $std\_conf(R_0, r)$.

Quindi, una regola R_0 è *interessante*, in termini di *confidenza inaspettata*, nel suo r-intorno se il valore di $||conf(R_0) - avg\_conf(R_0, r)| - std\_conf(R_0, r)|$ è grande rispetto ad una soglia pre-specificata dall'utente; in altre parole, se la confidenza di R_0 devia dalla $avg\_conf(R_0, r)$ molto di più della deviazione

media.

Per esempio, la Figura 1.8 dimostra che questa definizione cattura ef-

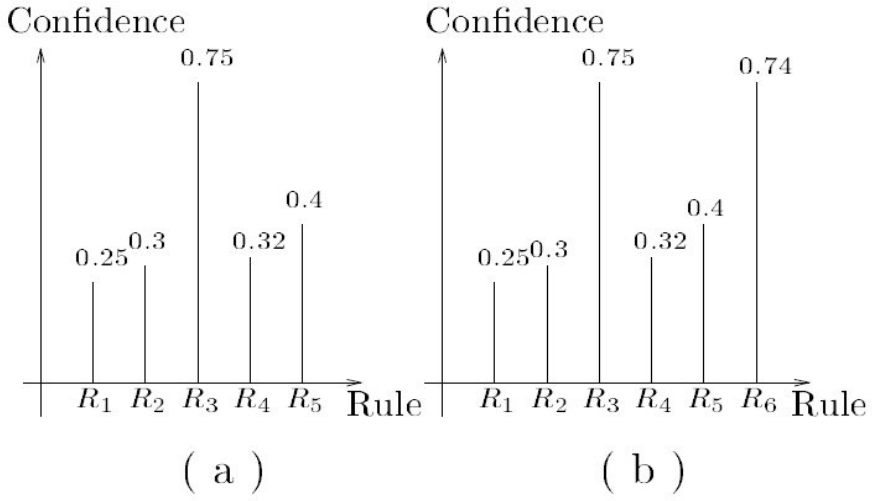


Figura 1.8: La confidenza inaspettata

fettivamente le regole con una confidenza inaspettata in un intorno. Supponiamo che esistano soltanto cinque regole estratte $R_1, ..., R_5$ in un certo r-intorno, i cui valori di confidenza sono rispettivamente 0.25, 0.3, 0.75, 0.32 e 0.4. Allora la $avg\_conf(R_3, r) = 0.3175$ e la $std\_conf(R_3, r) = (((0.25-0.3175)^2+(0.3-0.3175)^2+(0.32-0.3175)^2+(0.4-0.3175)^2)/4)^{\frac{1}{2}}$. Così, $conf(R_3) - avg\_conf(R_3, r) = 0.4325$, una differenza circa sedici volte più grande rispetto al valore di $std\_conf(R_3, r)$. Così, la confidenza di R_3 è inaspettata nel suo r-intorno.

Un secondo tipo di regole sono considerate interessanti perché esistono molte regole potenziali nei loro intorni, ma in questi stessi intorni ci sono poche regole estratte.

Una regola R_0 è *interessante*, del tipo *isolato*, se ha un r-intorno inaspettatamente sparso: se il numero di regole potenziali in $N(R_0, r)$ è grande ma il numero di regole estratte, cioè $|M \cap N(R_0, r)|$, è relativamente piccolo.

Se chiamiamo $\frac{M \cap N(R_0,r)}{N(R_0,r)}$ la *densità* dell'r-intorno di R_0, allora la condizione nella definizione precedente può essere riscritta come 'se il numero di regole potenziali in $N(R_0, r)$ è grande, ma la densità dell'r-intorno di R_0 è relativamente piccola.' In particolare, il significato di 'il numero di regole estratte è relativamente piccolo' può essere specificato usando una qualche soglia, la quale può essere sia specificata dall'utente sia calcolata a partire dall'applicazione in uso (per esempio $\frac{|M|}{N}$, dove N è il numero totale di regole potenziali).
Che una regola possa essere interessante o meno è inoltre dipendente dalle soglie di *min_support* e *min_confidence*. Per esempio, quando queste soglie sono aumentate le regole che prima non erano isolate possono diventarlo.

1.6.10 Una rappresentazione condensata delle regole associative

Abbiamo già fatto cenno - nel corso del paragrafo 1.6.3 - alla possibilità di estrarre delle *basi* per le regole associative. Quello che gli autori di questo tipo di approccio [39] cercano di mettere in luce è che, come abbiamo già visto, i problemi principali nella generazione delle regole associative a partire da un data set sono: la scoperta degli itemset frequenti e, appunto, la generazione delle regole associative a partire da questi ultimi.
Secondo [39] esistono tre tipi di approcci per risolvere il primo dei due problemi elencati. Nel primo gli itemset frequenti vengono trovati considerando simultaneamente tutti gli itemset frequenti di una data dimensione. Nel secondo l'estrazione degli itemset frequenti si basa sull'estrazione degli itemset frequenti massimali da cui viene derivato l'intero insieme degli itemset frequenti ed il supporto di questo insieme viene conteggiato all'interno del data set. Nel terzo, l'estrazione degli itemset frequenti si basa sulla estrazione degli itemset frequenti chiusi, estrazione definita secondo l'*operatura di*

chiusura di Galois; in questo caso vengono infatti prima estratti gli itemset frequenti chiusi e, in un secondo momento, da questi vengono derivati sia gli itemset frequenti che il loro relativo supporto, senza più accedere al data set. Quest'ultimo approccio sembrerebbe essere il migliore quando si lavora su dati densi o molto correlati.

Come si è già detto nel corso dei precedenti paragrafi, il secondo dei problemi esposti all'inizio è cruciale, dal momento che il numero di regole che possono essere estratte da un database può essere elevatissimo e molte di queste regole spesso risultano inutili ed ingannevoli perché ridondanti. Come può essere allora ridotto il numero delle regole arrivando ad un sottoinsieme di esse contenente soltanto le regole non ridondanti?

Secondo [39] esistono nuovamente tre tipi di approcci atti a risolvere questo problema. Come già visto, i primi forniscono dei meccanismi per filtrare le regole una volta che queste sono già state estratte, ma si tratta di metodi che, appunto, riducono il numero delle regole solo in una fase successiva di visualizzazione e le regole ridondanti vengono comunque estratte; gli altri due approcci estendono invece la definizione di regola associativa per fare in modo di non estrarre regole 'simili'. Il primo della seconda tipologia di approcci può usare una tassonomia per rappresentare gli item e da questi estrarre così regole associative più generalizzate, oppure può usare misure statiche, come il coefficiente di correlazione di Paerson o il test chi-quadrato, oppure, infine, può estrarre solo le regole con l'antecedente più grande tra quelle che hanno i medesimi valori di supporto o lo stesso conseguente.

Il secondo della tipologia di approcci sopra presentati fa invece uso della *chiusura della connessione di Galois* per estrarre delle *basi* per le regole associative. In particolare, una *base* è un insieme non ridondante che è minimale secondo alcune proprietà matematiche e dal quale tutte le regole associative sono derivabili, con relativo supporto e confidenza, senza più accedere al data set.

Lo scopo è quindi quello di migliorare la rilevanza e l'utilità delle regole associative estraendone il minor numero possibile senza perdita di informazione. Per fare ciò viene proposta una rappresentazione condensata per massimizzare l'informazione portata da ciascuna regola. In questo senso le regole associative più rilevanti sono quelle più generali e, come già detto, non ridondanti (una regola $r : a \rightarrow c$ è più generale di una regola $r' : a' \rightarrow c'$ se esse hanno identici valori di supporto e confidenza, se l'antecedente a di r è un sottoinsieme di a' e se il conseguente c di r è un sovrainsieme di c') e a cui viene dato il nome di *regole associative min-max*. Da queste ultime possono poi essere derivate e generate tutte le altre regole associative, con i relativi valori di supporto e confidenza.

Un data set è rappresentato da una tripla con $D = (O, I, R)$, dove O e I sono insiemi finiti, rispettivamente, di oggetti e di item e R è una relazione binaria. Ciascuna coppia (o, i) contenuta in R denota il fatto che l'oggetto o facente parte dell'insieme O è in relazione con l'item i facente parte dell'insieme I. Un itemset l è un insieme di item tale che l è un sottoinsieme di I e l non può essere vuoto.

La connessione di Galois di una relazione binaria finita è una coppia di applicazioni (Φ, Ψ). Φ associa ad un insieme di oggetti $O \subseteq O$ gli item in relazione con tutti gli oggetti $o \in O$ e Ψ associa ad un itemset $l \subseteq I$ gli oggetti in relazione con tutti gli item $i \in l$. Quando un oggetto o è in relazione con tutti gli item $i \in l$ diciamo che o *contiene* I.

L'operatore di chiusura $\gamma = \Phi \circ \Psi$ associa a un itemset l il più grande insieme di item in comune a tutti gli oggetti che contengono l: la chiusura di un itemset è uguale all'intersezione di tutti gli oggetti che lo contengono. Con l'aiuto di questo operatore possono essere definiti gli itemset frequenti chiusi.

Un itemset frequente $l \subseteq I$ è un itemset frequente chiuso se e solo se $\gamma(l) = l$. L'itemset chiuso più piccolo che contenga un itemset l è la sua chiusura $\gamma(l)$.

L'insieme degli itemset frequenti chiusi e il loro supporto è un insieme minimo e non ridondante per generare tutti gli itemset frequenti ed il loro supporto, così come per generare tutte le regole associative, e i loro relativi valori di supporto e confidenza.

Un itemset $g \subseteq I$ è un *generatore* di un itemset chiuso l se e solo se $\gamma(g) = l$ e non esiste un $g' \subseteq I$ con $g' \subset g$ tale che $\gamma(g') = l$. Un generatore di cardinalità k è un k-generatore.

A questo proposito, l'algoritmo *CLOSE* è un algoritmo iterativo per estrarre generatori ed itemset frequenti chiusi. Durante un'iterazione k, una lista di k-generatori candidati viene considerata; vengono quindi contati i loro valori di chiusura ed il loro supporto all'interno del data set e venogno scartati i generatori non frequenti. I generatori frequenti vengono quindi usati per costruire i $(k+1)$-generatori candidati. Le chiusure dei generatori frequenti sono gli itemset frequenti chiusi ed il supporto di un generatore è anche il supporto della sua chiusura.

L'algoritmo $CLOSE^+$ identifica invece gli itemset frequenti chiusi ed i generatori a partire dagli itemset frequenti senza accedere al data set. Gli itemset frequenti chiusi ed i generatori sono identificati tra gli itemset frequenti usando queste due proposizioni: *il supporto di un generatore è più piccolo del supporto di tutti i suoi sottoinsiemi* e *il supporto di un itemset chiuso è più grande del supporto di tutti i suoi sovrainsiemi.*

Usando la prima affermazione determiniamo se un k-itemset frequente I è un generatore di un itemset chiuso confrontando il suo supporto e il supporto dei suoi $(k-1)$-itesemt frequenti inclusi in l. La seconda proposizione permette invece di determinare se un k-itemset frequente l è chiuso confrontando il suo supporto e il supporto dei suoi $(k+1)$-itemset frequenti in cui l è contenuto.

Diciamo allora che $\mathcal{AR}$ è l'insieme di tutte le regole associative estratte. Una regola associativa $r : l_1 \rightarrow l_2 \in \mathcal{AR}$ è una regola associativa min-

max se e solo se $\nexists r' : l'_1 \rightarrow l'_2 \in \mathcal{AR}$ avente $supporto(r') = supporto(r)$, $confidenza(r') = confidenza(r)$, $l'_1 \subseteq l_1$ e $l_2 \subseteq l'_2$.

Sulla base di questa definizione possono essere descritte con esattezza le regole associative min-max esatte ed approssimate - rispettivamente quelle che hanno un valore di confidenza uguale al 100% e quelle che hanno un valore di confidenza inferiore al 100% - le quali rispettivamente costituiscono la *base min-max esatta* e la *base min-max approssimata*.

Per prima cosa, possiamo osservare che le regole associative esatte, nella forma $r : l_1 \rightarrow (l_2/l_1)$, sono regole tra due itemset frequenti $l_1 \subset l_2$ i quali hanno la stessa chiusura, infatti: $\gamma(l_1) = \gamma(l_2)$. Dal momento che la $confidenza(r) = 1$ abbiamo che il $supporto(l_1) = supporto(l_2)$, e, dal momento che $l_1 \subset l_2$, osserviamo che $\gamma(l_1) = \gamma(l_2)$. Le regole associative min-max sono quindi definite a partire dalle regole associative esatte appena descritte.

A questo punto diciamo che g è il generatore di $\gamma(l_1) = \gamma(l_2)$ tale che $g \subseteq l_1$. Dal momento che g è minimo, abbiamo che $g \subseteq l_1 \subset l_2 \subseteq \gamma(l_2)$. Inoltre, tutti gli itemset nell'intervallo $[g, \gamma(l_2)]$, definito per inclusione (l'intervallo $[l_1, l_2]$ contiene infatti tutti i sovrainsiemi di l_1 che nel contempo sono sottoinsiemi di l_2), hanno la stessa chiusura $\gamma(l_2)$ e quindi lo stesso supporto. Tra tutte le regole aventi la forma $r : l_1 \rightarrow (l_2 l_1)$ con $l_1, l_2 \in [g, \gamma(l_2)]$, la regola associativa min-max è la regola $g \rightarrow (\gamma(l_2)\ g)$. Questa regola ha un antecedente più piccolo, g, e un conseguente più grande, $\gamma(l_2)$, tra tutte quelle regole aventi lo stesso valore di supporto.

Diciamo quindi che *Closed* è l'insieme degli itemset frequenti chiusi estratti dal data set e, per ciascun itemset frequente chiuso f, diciamo che *Genf* è l'insieme dei generatori di f. La base min-max esatta è quindi:

$$MinMaxEsatto = \{r : g \rightarrow (f\ g) | f \in Closed \wedge g \in Genf \wedge g \neq f\}$$

Tutte le regole associative esatte, ed il loro supporto, possono essere dedotte a partire dalla base esatta min-max. Tutte le regole nella base esatta min-

max sono regole associative min-max.

L'algoritmo per derivare tutte le regole associative esatte riceve l'insieme *MinMaxEsatto* in input e ritorna in output l'insieme *TuttoEsatto* contenente tutte le regole associative esatte.

Le regole associative approssimate, nella forma $r : l_1 \rightarrow (l_2\ l_1)$, sono regole tra due itemset frequenti $l_1 \subset l_2$ tali che $\gamma(l_1) \subset \gamma(l_2)$. Dal momento che la $confidenza(r) < 1$ abbiamo che il $supporto(l_1) > supporto(l_2)$ e possiamo dedurre che $\gamma(l_1) \subset \gamma(l_2)$.

A questo punto possiamo dire che g_1 è il generatore dell'itemset frequente chiuso f_1 e che g_2 è il generatore dell'itemset frequente chiuso f_2 tali che $f_1 \subset g_2 \subseteq l_2 \subseteq f_2$. Tutte le regole nella forma $r : l_1 \rightarrow (l_2\ l_1)$ dove $l_1 \in [g_1, f_1]$ e $l_2 \in [g_2, f_2]$ hanno gli stessi valori di supporto e confidenza dal momento che g_1, l_1 e f_1 hanno lo stesso supporto di g_2, l_2 e f_2. Tra tutte queste regole, possiamo allora dedurre che la regola associativa min-max è $g_1 \rightarrow (f_2\ g_1)$. A tal proposito, g_1 è l'itemset minimo nell'intervallo $[g_1, f_1]$ e f_2 è l'itemset massimo nell'intervallo $[g_2, f_2]$.

La generalizzazione di questa proprietà a tutte le coppie di itemset frequenti l_1 e l_2 tali che $l_1 \subset l_2$ e $supporto(l_1) \neq supporto(l_2)$ definisce la base approssimata min-max contenente tutte le regole associative approssimate min-max sulla base di questa definizione:

$$MinMaxApprox = \{r : g \rightarrow (f\ g) | f \in Closed \wedge g \in Gen \wedge \gamma(g) \subset f\}$$

con *Gen* l'insieme dei generatore degli itemset frequenti chiusi in *Closed*.

Tutte le regole associative approssimate possono essere ugualmente dedotte, con i loro relativi valori di supporto e confidenza, dalla base approssimata min-max. Tutte le regole nella base approssimata min-max sono regole associative min-max.

L'input dell'algoritmo per generare tutte le regole associative approssimate è composto dagli insiemi *MinMaxApprox* e *MinMaxEsatto*, costituiti rispettivamente dalle regole min-max approssimate e dalle regole min-max esatte.

Il suo output è l'insieme *TuttoApprox* contenente tutte le regole approssimate.

Il numero delle regole associative approssimate può essere ridotto ulteriormente senza perdere la capacità di dedurre tutte le regole associative approssimate, sempre con i relativi valori di supporto e confidenza, rimuovendo le *regole associative min-max transisitive*.

Una regola associativa min-max $g \rightarrow (f\ g)$ con $\gamma(g) \subset f$ è transitiva se esiste un itemset frequente chiuso f' tale che $\gamma(g) \subset f' \subset f$. Diciamo che g' è il generatore di f' tale che $\gamma(g) \subset g' \subseteq f' \subset f$. Allora, abbiamo le seguenti due regole associative min-max approssimate: $g \rightarrow (f'\ g)$ e $g' \rightarrow (f\ g')$. La regola $g \rightarrow (f\ g)$ è la composizione transitiva delle due regole precedenti; il suo supporto è uguale a quello della seconda regola e la sua confidenza è pari al prodotto della confidenza delle due regole.

Questa caratterizzazione può essere generalizzata a tutte le triple consistenti di un generatore g, la sua chiusura f ed un sovrainsieme chiuso f' di f per definire la base approssimata non transistiva min-max, la quale è la riduzione transitiva della base approssimata min-max. Diciamo che $l_1 \lessdot l_2$ quando un itemset l_1 è l'immediato predecessore di un itemset l_2. Le regole approssimate non-transitive min-max sono nella forma $g \rightarrow (f\ g)$ dove f è un itemset frequente chiuso e g è un generatore frequente tale che $\gamma(g)$ è un predecessore immediato di f. Da questa definizione può essere derivata la base approssimata min-max come segue:

$$MinMaxReduc = \{r : g \rightarrow (f\ g) | f \in Closed \wedge g \in Gen \wedge \gamma(g) \text{ un immediato predecessore di } f.$$

L'algoritmo presentato [39] costruisce l'insieme *MinMaxReduc* fatto di regole non-transitive approssimate min-max usando gli itemset frequenti chiusi ed i loro generatori. Per ciascun generatore g determina tutti gli itemset frequenti chiusi f che sono immediati successori della chiusura di g e genera tutte le regole tra g e f che hanno sufficiente confidenza.

L'algoritmo per generare tutte le regole transitive genera l'insieme *MinMaxApprox* delle regole approssimate min-max usando l'insieme *MinMaxReduc* di regole non-transitive approssimate e la soglia minima di confidenza come input.

L'approccio è incrementale: vengono aggiunte nuove regole transitive min-max sino a che non viene creata nessuna nuova regola.

Risultati sperimentali condotti sia su data set sintetici che su data set operazionali dimostrano che l'estrazione di queste basi riduce considerabilmente il numero di regole estratte, particolarmente in caso di data densi o correlati. Il risultato è più facile da navigare e, dal momento che spesso sono eliminate le regole ridondanti e fuorvianti, anche notevolmente più utile.

Capitolo 2

Definizione del problema di business

2.1 Determinazione degli obiettivi di business

2.1.1 Il CAAF-CISL

Il CAAF-CISL opera sul territorio per fornire agli iscritti, lavoratori e pensionati, assistenza e consulenza completa nel campo fiscale e delle agevolazioni fiscali. Esso mira cioè ad essere un ponte nel rapporto tra il cittadino e la Pubblica Amministrazione, favorendo il miglioramento e la semplificazione di questo stesso rapporto.
Nato nel 1993, il CAAF-CISL ha aumentato costantemente ogni anno il numero di pratiche trattate (fino ad arrivare ai 5 milioni di pratiche gestite nell'ultimo anno), nonché la gamma dei servizi offerti e, ad oggi, assiste più di 3 milioni di utenti con oltre 2000 sedi diffuse in tutto il territorio nazionale.
I punti di forza dell'organizzazione sono costituiti da:

- la presenza capillare su tutto il territorio italiano;

- un sistema di elaborazione dati all'avanguardia e continuamente aggiornato per far fronte alle novità della normativa;
- un piano di formazione del personale continuo e consolidato negli anni;
- una copertura assicurativa completa ottenuta proprio in virtù dei livelli di qualità raggiunti.

Dal Settembre 2002 il CAAF-CISL è inoltre certificato secondo la norma UNI EN ISO 9001:2000 (Vision 2000), sinonimo dell'impegno costante volto ad offrire servizi di qualità e dell'attenzione riposta nel soddisfare ogni richiesta degli utenti in un'ottica di miglioramento continuo.
Caratteristiche comuni a tutti gli operatori del CAAF-CISL sono:

- la partecipazione alla missione sindacale della CISL;
- la responsabilità nel tutelare l'interesse del cittadino nel rispetto della normativa;
- la professionalità nel saper rispondere con competenza alle esigenze dell'utente;
- l'accoglienza di ogni persona con i suoi bisogni.

Gli oltre 2000 uffici del CAAF-CISL offrono una ampia varietà di servizi, tra i quali possiamo ricordare:

- MODELLO 730: dichiarazione dei redditi riguardante lavoratori dipendenti, collaboratori coordinati e continuativi e pensionati;
- MODELLO UNICO: dichiarazione dei redditi riguardante tutti coloro che non possono (o non vogliono) presentare il modello 730, purché non abbiano redditi da impresa;
- ICI: imposta del Comune su immobili, terreni agricoli, aree fabbricabili;

- RED: certificazione della situazione reddituale di cittadini da determinate prestazioni erogate dall'INPS;
- ISE: indicatore della situazione economica del cittadino, attraverso il quale accedere a prestazioni sociale e servizi di pubblica utilità;
- SUCCESSIONI: dichiarazione degli eredi del deceduto, da presentare all'Ufficio del Registro competente;
- CONTENZIOSO: assistenza al contribuente che riceve cartelle di pagamento le quali contestano il mancato pagamento di imposte o tasse.

2.1.2 Il Modello 730

Tra i servizi offerti dal CAAF-CISL, e presentati nel paragrafo precedente, quello che a noi interessa è quello relativo all'assistenza alla compilazione del modello 730, il quale rappresenta il soggetto della nostra analisi di *data mining*. Nelle righe che seguono ne illustriamo utilità e struttura in modo da poter contestualizzare le azioni che verranno prese e quindi descritte nelle fasi successive del progetto.
La dichiarazione dei redditi è l'atto formale attraverso il quale il contribuente espone i propri redditi, indica le spese e gli oneri per i quali vuole far valere i benefici fiscali (deduzioni e detrazioni) e calcola l'imposta a debito o a credito. Essa deve essere compilata e presentata nell'anno successivo a quello in cui i redditi sono stati percepiti o maturati.
Per le persone fisiche, il modello da utilizzare può essere quello *UNICO Persone Fisiche*, oppure, se il dichiarante è un lavoratore dipendente o un pensionato, il *modello 730*.
Il modello 730 è stato introdotto in Italia nel 1993 relativamente alla presentazione della dichiarazione dei redditi per l'anno 1992. Esso fu ideato e progettato nel 1992 dall'allora Segretario Generale del Ministero delle Finanze Giorgio Benvenuto e lo scopo principale era quello di provvedere im-

mediatamente al rimborso delle imposte a credito a favore dei dipendenti e pensionati per il tramite del sostituto d'imposta, anziché aspettare i lunghi tempi dei vecchi uffici delle imposte.
In particolare, il modello 730 può essere presentato solo dai contribuenti residenti in Italia che appartengono ad una delle seguenti categorie:

- lavoratori dipendenti;
- pensionati;
- percepenti indennità sostitutive dei redditi di lavoro dipendente (ad esempio indennità di mobilità);
- sacerdoti della Chiesa Cattolica;
- soci di cooperative di produzione e lavoro, di servizi agricoli e di prima trasformazione dei prodotti agricoli e di piccola pesca;
- giudici costituzionali, parlamentari nazionali e altri titolari di cariche pubbliche elettive (consiglieri regionali, provinciali, comunali, ecc.);
- soggetti impegnati in lavori socialmente utili e precari della scuola.

Il modello 730 può essere utilizzato per dichiarare le seguenti tipologie di reddito:

- redditi da lavoro dipendente;
- redditi assimilati a quelli da lavoro dipendente;
- redditi da capitale;
- redditi da lavoro autonomo per i quali non è richiesta la partita IVA;
- alcuni dei redditi diversi;
- alcuni dei redditi assoggettabili a tassazione separata.

La presentazione del modello 730 presenta alcuni vantaggi per il contribuente che decide di avvalersene:

- è relativamente semplice da compilare;
- non richiede calcoli perché sono effettuati dal soggetto che presta l'assistenza fiscale (il sostituto d'imposta o il CAF);
- eventuali rimborsi sono restituiti direttamente dal sostituto d'imposta con la retribuzione di luglio o, in caso di pensione, a partire dal mese di agosto;
- in caso di pagamento di imposte a saldo o in acconto non c'è necessità di recarsi in banca o in posta perché le somme dovute sono trattenute direttamente dal sostituto d'imposta in busta paga o sulla pensione;
- è possibile compensare eventuali crediti IRPEF con l'ICI da pagare;
- ci sono minori controlli sulla dichiarazione da parte dell'Amministrazione se il modello è presentato ad un CAF o ad un professionista abilitato;
- sono previste dalla legge garanzie assicurative che salvaguardano il contribuente in caso di errori imputabili al CAF o al professionista abilitato.

Il modello 730 è inoltre l'unico modello che permette di presentare la dichiarazione congiunta di marito e moglie. La dichiarazione congiunta diventa un vantaggio soprattutto quando uno dei due coniugi non dispone di un sostituto d'imposta che possa effettuare il conguaglio. In questo caso il coniuge indicato come dichiarante sarà quello che dispone di un sostituto di imposta in grado di effettuare il conguaglio.

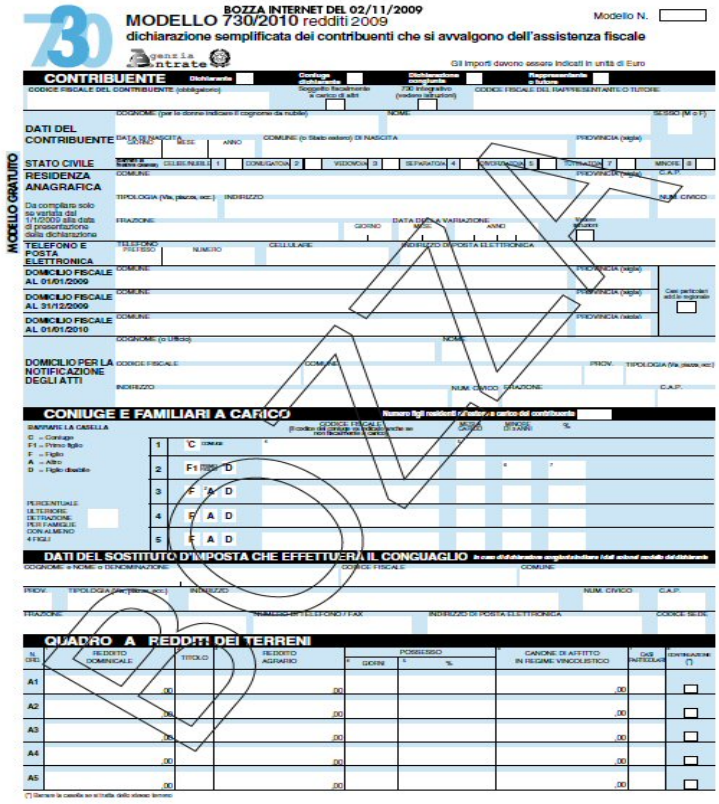


Figura 2.1: Un'immagine del modello 730 dell'anno 2010 per i redditi del 2009

2.1.3 Gli obiettivi di business

Soprattutto negli ultimi anni, il CAAF-CISL ha registrato un notevole aumento di clienti che scelgono di avvalersi del servizio di assistenza fiscale da quest'ultimo proposto per la compilazione del modello 730. Per meglio cercare di capire i motivi per i quali quest'ultima affermazione si realizza concretamente e per la curiosità e la volontà di analizzare questo andamento sono state scelte quattro sedi del CAAF-CISL corrispondenti alle province di Arezzo, Latina, Ragusa e Trento.

L'appello alle tecniche di data mining da parte del CAAF-CISL ha quindi

come obiettivo quello di rispondere alla seguente domanda:

esistono delle caratteristiche, fiscali o demografiche, che identificano i contribuenti che usufruiscono per la prima volta del servizio di assistenza fiscale alla compilazione del modello 730?

Sapere infatti se esiste qualche elemento, sia esso di tipo economico o fiscale oppure di tipo sociale o demografico, che accomuna una larga fetta dei nuovi clienti e li distingue dagli altri può rivestire un ruolo molto importante nella politica di business dell'ente: diventerebbe infatti possibile focalizzare le campagne pubblicitarie soprattutto sulle persone che sono maggiormente propense ad usufruire del servizio, utilizzando i media che meglio si adattano a raggiungere questa fetta della popolazione.
Queste informazioni devono essere dedotte tramite l'analisi dei modelli 730 compilati negli anni precedenti grazie all'assistenza degli uffici CAAF-CISL. In particolare, potendoci avvalere dei dati di queste dichiarazioni relative ai 5 anni che vanno dal 2005 al 2009 dovrebbe essere possibile definire un profilo o una serie di profili tipici relativi ai nuovi clienti.
Naturalmente, nel caso in cui non esistano tratti distintivi dei nuovi clienti, è possibile che l'analisi non porti a nessun risultato significativo in ottica di business.

2.1.4 Criteri di successo

Una fase importante nella pianificazione di un processo di data mining è la definizione dei criteri che definiscono se un progetto ha raggiunto o meno gli obiettivi sperati. Nel caso di studio in esame è però difficile identificare con precisione dei parametri di questo tipo: alla base dell'analisi di data mining non ci sono infatti degli obiettivi di business esatti, ma vi è l'obiettivo di identificare eventuali tratti caratteristici che si ripetono tra i clienti che compilano il 730 al CAAF-CISL per la prima volta.

Possiamo quindi etichettare come un successo la scoperta di caratteristiche che identificano con un discreto livello di precisione i contribuenti più predisposti di altri ad effettuare la dichiarazione dei redditi presso una sede del CAAF-CISL.
Può tuttavia rappresentare un'informazione importante anche il sapere e lo scoprire che tali caratteristiche in realtà non esistono e che quindi l'incremento di nuovi clienti degli ultimi anni è stato casuale o, comunque, non è derivato da cause identificabili dai dati messi a disposizione.

2.2 Analisi della situazione

2.2.1 Le risorse a disposizione

I dati

Dato l'obiettivo del progetto (la profilazione dei nuovi clienti nelle quattro province sopra elencate), il soggetto dell'analisi (il modello 730) e la durata temporale su cui basare il nostro studio (gli ultimi 5 anni), i dati messi a disposizione sono stati i tracciati dei cinque database contenenti i dati dei modelli 730 compilati dal 2005 al 2009.
Per quanto riguarda la tipologia di informazioni presenti nelle tabelle dei singoli database, queste si dividono in tre categorie:

- i valori con cui sono stati riempiti i campi del modello 730;
- i parametri interni per il calcolo di deduzioni e detrazioni d'imposta;
- i dati relativi ai documenti allegati al 730 (possono riferirsi a fabbricati, terreni, autocertificazioni, ecc.).

La struttura di ogni database è lo specchio della struttura di un modello 730 e, quindi, la composizione delle tabelle e dei loro attributi varia da database a database in base alle modifiche che ha subito il modello 730 negli ultimi

anni.

Tuttavia, la maggior parte delle modifiche nei modelli consistono nell'aggiunta di nuovi campi da compilare: la variazione del contenuto informativo dei database si può quindi vedere come una crescita incrementale negli anni. Dal punto di vista strutturale ciò si rispecchia nell'aggiunta di nuove tabelle e nel riposizionamento dei dati in tabelle diverse rispetto a quelle in cui questi ultimi erano registrati in passato.

Su questi stessi dati le operazioni di pre-processing erano già state effettuate a seguito del lavoro di [2] (lavoro che, però, era riferito agli anni che andavano dal 2004 al 2008, ma che è stato riapplicato ai nuovi dati del 2009) e si è quindi effettivamente lavorato su cinque viste contenenti soltanto i dati relativi alle tabelle di nostro interesse e selezionate a partire da una tabella generata a seguito del lavoro di definizione della struttura dei dati di [2] e che contiene i dati relativi ai dichiaranti per gli anni in esame.

Il lavoro precedente [2] già effettuato sui dati è effettivamente stato di grande importanza e di non poca difficoltà poiché sono stati evitati tutta una serie di problemi quali:

- il grande numero di tabelle e di attributi nelle fonti dei dati che poteva causare dispersione dell'informazione;
- la documentazione poco chiara nella descrizione degli attributi che avrebbe potuto portare a fraintendimenti nel significato degli attributi;
- le tabelle e gli attributi aventi nomi poco comprensibili che ugualmente potevano portare a fraintendimenti e a difficoltà nella loro comprensione;
- la presenza di attributi con nomi molto simili in tabelle diverse;
- lo stesso tipo di informazione ripetuta in tabelle diverse;

- lo stesso tipo di informazione rappresentata da tabelle diverse in database di anni diversi;
- errori nell'operazione di occultamento dei dati di contribuenti che potevano portare a problemi quali la divulgazione di dati sensibili riguardanti i contribuenti.

Per superare tutte queste difficoltà è stato necessario uno studio molto approfondito sia del dominio sia della struttura di tutte le basi di dati. Tutto questo è stato accompagnato da una collaborazione continua con l'esperto del dominio che ha fornito le indicazioni necessarie per interpretare al meglio, e senza errori, le sorgenti informative.
Ad alcune informazioni non è stato lecito accedere per motivi di privacy riguardante coloro che hanno compilato il modello 730 presso il CAAF-CISL: in particolare non abbiamo potuto avere a disposizione le informazioni relative a nome, cognome e codice fiscale dei contribuenti. Ogni persona è stata quindi rappresentata da un codice univoco e non reversibile che ne impediva l'identificazione.

Risorse software

Gli strumenti software utilizzati durante lo sviluppo del progetto sono i seguenti:

- *SQL Server Management Studio*: per il caricamento e la consultazione dei database;
- *Oracle SQL Developer*: per la consultazione, lo studio e l'applicazione degli algoritmi di mining sulle viste su cui si è scelto di lavorare;
- *SPSS Clementine 11.1*: per il controllo della qualità dei dati e la realizzazione delle analisi di data mining;

- *Microsoft Office Excel 2007*: per la realizzazione delle analisi di data mining;
- *Weka 3.6.1*: per il controllo della qualità dei dati, per la realizzazione delle analisi di data mining e per l'applicazione degli algoritmi di mining.

Glossario di termini relativi al dominio d'analisi

Acconto: è l'importo che il contribuente è tenuto a versare, solitamente in due rate (la prima nel mese di maggio e la seconda nel mese di novembre), come anticipo dell'imposta sui redditi dovuta per l'anno in corso.
Credito d'imposta: indica l'eventuale differenza, a favore del contribuente, tra l'imposta dovuta per l'anno a cui si riferisce la dichiarazione e quanto è stato già pagato sotto forma di ritenute, crediti ed acconti; tale eccedenza può essere utilizzata per compensare futuri debiti d'imposta, oppure, ne può essere chiesto il rimborso.
Deduzione: è la spesa che può essere sottratta dal reddito complessivo, con un beneficio rapportato all'aliquota marginale raggiunta dal contribuente. Le deduzioni operano pertanto in modo diverso dalle detrazioni le quali, invece, abbattono l'imposta da pagare.
Detrazione: sono le agevolazioni consistenti nella possibilità di sottrarre determinate somme dall'imposta lorda. Esse spettano in particolare ai contribuenti che hanno familiari a carico o che possiedono redditi di lavoro dipendente o di pensione, di lavoro autonomo o professionale o di impresa minore. A differenza della deduzione fiscale, la quale viene sottratta dalla base imponibile, la detrazione viene sottratta dall'imposta lorda con lo scopo di determinare l'imposta netta effettivamente dovuta.
Dichiarazione congiunta: può essere presentata dai coniugi per i redditi posseduti da ciascuno di essi. La dichiarazione congiunta comporta la determinazione di un'unica imposta IRPEF da versare per entrambi i coniugi

e, quindi, essa consente anche la compensazione del debito d'imposta di un coniuge con l'eventuale credito dell'altro.

Liquidazione: è una procedura dell'Amministrazione finanziaria diretta a rendere certo ed esigibile il credito erariale, credito sorto a seguito di un accertamento fiscale.

Ritenuta: è un metodo di riscossione delle imposte dirette il quale consiste nell'attribuire ad un soggetto (sostituto d'imposta) l'obbligo di trattenere e versare all'erario una parte delle imposte dovute da un altro soggetto.

Sostituti d'imposta: sono tutti quei soggetti che corrispondono retribuzioni, compensi di lavoro autonomo o redditi di capitale. Il sostituto è anche obbligato dalla legge al pagamento delle imposte in luogo dei soggetti a cui corrisponde le retribuzioni.

NO TAX AREA (deduzioni da lavoro e pensione): è un meccanismo che esclude una parte del reddito dalla tassazione. La parte di reddito non soggetta a tassazione è inversamente proporzionale al reddito percepito fino ad annullarsi dopo un certo limite. Con la Finanziaria del 2007 tali deduzioni sono state eliminate e sostituite da detrazioni.

NO TAX FAMILY: è stata introdotta nel 2005 e trasforma le precedenti detrazioni per carichi di famiglia in deduzioni. Vengono pertanto fissate delle misure base di 3.200 euro per il coniuge e 2.900 euro per i figli e gli altri familiari a carico. Questa ultima deduzione, inoltre, può essere aumentata sino a 3.450 euro per i figli di età inferiore a 3 anni, a 3.200 euro per il primo figlio in mancanza del coniuge e a 3.700 euro in caso di soggetto portatore di handicap.

2.3 Definizione degli obiettivi di data mining

In questa fase riprendiamo gli obiettivi di analisi definiti in termini di business e li riformuliamo secondo un'ottica di data mining. Lo scopo di questo progetto può essere definito come:

lo studio delle caratteristiche demografiche ed economiche delle persone che hanno compilato il modello 730 negli ultimi 5 anni: esiste un legame tra alcune di queste caratteristiche e il fatto che un cliente scelga di avvalersi del servizio per la prima volta? Se questo legame esiste, qual è?

Il genere di quesito appena formulato rientra nella classe dei problemi risolvibili tramite una tecnica di classificazione (vedi, inoltre, il paragrafo 1.2.1).

2.3.1 Le viste

Negli algoritmi di classificazione i dati in input, detti anche *training set*, consistono in record caratterizzati da attributi o caratteristiche multipli. Ogni record, o oggetto, è inoltre definito da una etichetta della classe a cui appartiene.
Nelle viste utilizzate per il nostro progetto di mining, inoltre, all'inizio, ogni record conterrà i dati relativi alla dichiarazione dei redditi di una specifica persona. In particolare:

- se un soggetto ha compilato il modello 730 tutti gli anni dal 2005 al 2009, allora la vista conterrà 5 record relativi a quella persona, ognuno avente i dati di uno specifico anno;

- in relazione ad ogni dichiarazione congiunta, la vista conterrà 2 record: uno con i dati dichiarati dalla moglie e uno con quelli del marito;

- la variabile target indicherà se il record contiene i dati della dichiarazione di un nuovo cliente (se cioè non esistono record di dichiarazioni relative allo stesso cliente presentate in anni precedenti a quello del record in esame).

2.3.2 Variabili target

Per il CAAF-CISL un cliente si classifica come nuovo in un anno x se valgono entrambe queste condizioni:

1. quel particolare cliente ha compilato il modello 730 nell'anno x;
2. quel particolare cliente non ha mai compilato il modello 730 negli anni precedenti all'anno x ($x-1, x-2, ..., x-n$ con $n \to \infty$).

In particolare, un cliente è nuovo se i suoi dati appaiono per la prima volta in una dichiarazione 730 del CAAF-CISL, sia questa individuale o congiunta. Pertanto, la variabile target di tipo *flag* della nostra analisi è una: *NUOVODICHIARANTE*, la quale, in corrispondenza dei dati di una specifica persona, assume i seguenti valori:

- 1 se tale persona è nuova agli archivi delle persone che hanno compilato la dichiarazione al CAAF-CISL;
- 0 altrimenti.

In relazione alla dichiarazione di una persona, oltre alla variabile NUOVODICHIARANTE, ogni record presenta anche un'altra variabile di tipo flag che descrive la posizione dell'eventuale coniuge corrispondente a quella specifica persona all'interno dei dati relativi a tutte le compilazioni delle dichiarazioni dei redditi.
In particolare, questa variabile (chiamata *NUOVOCONIUGE*) assume i seguenti valori:

- 1 se il cliente è coniugato ed il coniuge non ha fatto dichiarazioni 730 negli anni precedenti, ma l'ha fatta nell'anno in esame;
- 0 se il cliente non è coniugato oppure se il coniuge ha fatto una dichiarazione presso il CAAF-CISL negli anni precedenti, oppure se il coniuge non ha mai fatto la dichiarazione 730 negli anni precedenti ed in quello attuale.

La presenza di NUOVODICHIARANTE e di NUOVOCONIUGE, in combinazione alle variabili flag *DICHIARAZIONECONGIUNTA*, *COMPILA_DICH* e *CHURN*[1], ci permette di studiare i contribuenti in diverse situazioni. In particolare vediamo prima come sono definite le variabili flag DICHIARAZIONECONGIUNTA, COMPILA_DICH e CHURN:

- DICHIARAZIONECONGIUNTA:
  - 1 se i dati presenti nel record sono estrapolati da una dichiarazione congiunta;
  - 0 se i dati presenti nel record sono estrapolati da una dichiarazione singola.
- COMPILA_DICH:
  - 1 se il record è relativo ad un dichiarante in una dichiarazione congiunta oppure ad una dichiarazione singola;
  - 0 se il record è relativo ad un coniuge in una dichiarazione congiunta.
- CHURN:
  - 1 se il cliente non compilerà il 730 l'anno successivo rispetto a quello di cui si sta esaminando la dichiarazione;
  - 0 se il cliente compilerà il 730 l'anno successivo rispetto a quello di cui si sta esaminando la dichiarazione.

[1]Prendono il nome di *CHURN* quelle persone che sono propense ad abbandonare l'azienda di cui sono clienti. In molti mercati, come quello delle telecomunicazioni e quello assicurativo, è molto importante eseguire delle analisi di tipo churn che, basandosi su tecniche di data mining di tipo predittivo, forniscano una lista dei clienti che sono a rischio di abbandono. Tutto ciò può essere infatti molto utile per suggerire le strategie da attuare in modo da mantenere alta la profittabilità dell'azienda.

In relazione alla combinazione dei valori di NUOVODICHIARANTE, NUOVOCONIUGE, DICHIARAZIONECONGIUNTA, COMPILA_DICH e CHURN possiamo distinguere una persona come appartenente ad una delle seguenti situazioni:

Tabella 2.1: I possibili valori della variabile target.

CHURN	ND	NC	DC	CD	DESCRIZIONE
0	1	1	1	1	dichiarante in una dichiarazione congiunta in cui entrambi i coniugi sono nuovi clienti.
0	1	1	1	0	coniuge in una dichiarazione congiunta in cui entrambi i coniugi sono nuovi clienti.
0	1	1	0	1	dichiarante nuovo in una dichiarazione singola in cui il coniuge è nuovo cliente in un'altra dichiarazione.
0	1	1	0	0	caso non possibile.
0	1	0	1	1	dichiarante in una dichiarazione congiunta in cui il medesimo è nuovo cliente ma il coniuge no.
0	1	0	1	0	coniuge in una dichiarazione congiunta in cui il medesimo è nuovo cliente, ma il coniuge no.
0	1	0	0	1	dichiarante in una dichiarazione individuale in cui il medesimo è nuovo cliente e non coniugato oppure il coniuge è nuovo cliente.
0	1	0	0	0	caso non possibile.

Continua nella prossima pagina

Tabella 2.1 – continua dalla pagina precedente

CHURN	ND	NC	DC	CD	DESCRIZIONE
0	0	1	1	1	dichiarante in una dichiarazione congiunta in cui il medesimo non è nuovo cliente ma il suo coniuge sì.
0	0	1	1	0	dichiarante in una dichiarazione congiunta in cui il medesimo non è nuovo cliente ma il suo coniuge sì.
0	0	1	0	1	dichiarante in una dichiarazione individuale in cui il medesimo non è nuovo cliente ma il suo coniuge sì (sebbene abbia compilato un'altra dichiarazione).
0	0	1	0	0	caso non possibile.

La tabella si riferisce ai clienti che sono rimasti anche nell'anno successivo alla prima dichiarazione (con CHURN = 1 si avrebbe invece il caso contrario).

In essa abbiamo:

ND = NUOVODICHIARANTE.

NC = NUOVOCONIUGE.

DC = DICHIARAZIONECONGIUNTA.

CD = COMPILA_DICH.

Alla luce degli obiettivi d'analisi, dei dati e della struttura di questi ultimi, la tecnica che abbiamo deciso di usare è quella rappresentata dalle *regole di classificazione*: questa tecnica è infatti sicuramente più semplice e rapida nella implementazione e nella valutazione dei risultati rispetto a quella delle reti neurali.

Allo stesso modo essa è preferibile agli alberi di classificazione dal momento che questi ultimi sono più adatti a definire un tipo di modelli da utilizzare in ottica previsionale, cioè per classificare soggetti di cui non conosciamo il valore della variabile target.

In questo particolare progetto il nostro obiettivo è invece quello di mettere

in risalto le relazioni tra singoli attributi o tra gruppi di attributi con la variabile target: è questo che ci porta a determinare un insieme di caratteristiche comuni alle persone che si rivolgono al CAAF-CISL per la prima volta. La tecnica migliore è quindi quella delle regole di classificazione.

Capitolo 3

Definizione della struttura dei dati per le analisi

3.1 Descrizione dell'insieme dati per le analisi

All'inizio del nostro progetto le viste di cui abbiamo parlato nel precedente capitolo sono composte da una variabile target e da 154 attributi descrittivi che, per ogni dichiarazione, identificano il cliente e i dati da lui dichiarati nell'anno di presentazione del modello 730.
In particolare, i 154 attributi descrittivi (di cui 39 di tipo binario, 82 di tipo continuo e 33 di insiemi discreti) coprono 4 tipi di informazioni così distribuite:

- dati anagrafici;
- dati relativi a lavoro, pensione e redditi vari;
- dati relativi alle spese;
- altri tipi di dati (rapporti spese/redditi, ...).

La struttura di ogni vista è lo specchio della struttura di un modello 730 e, quindi, la composizione degli attributi varia da vista a vista in base alle

modifiche che ha subito il modello 730 negli anni.
Il lavoro di comprensione di ogni attributo è stato piuttosto lungo e laborioso, poiché è stato possibile solo in seguito alla profonda comprensione di ciò che quell'attributo andava concretamente a significare all'interno di un modello 730 in particolare e della normativa fiscale in generale. La fatica relativa a questo lavoro di comprensione era già stata svolta da [2], per cui il nostro lavoro di comprensione è risultato effettivamente più semplice, dal momento che è stato portato a termine grazie al già effettuato lavoro di mediazione e di traduzione da un linguaggio propriamente tecnico.
Ecco un elenco degli attributi di cui si è discusso sopra, divisi per categoria:

Dati anagrafici

Tabella 3.1: Attributi: i dati anagrafici.

Nome	Descrizione	Tipo
DichiarazioneCongiunta	Indica se la dichiarazione è congiunta	Flag
IsTutore	Indica se il dichiarante è il tutore legale di un altro soggetto	Flag
AnnoDiNascita	Anno di nascita del contribuente	Numerico
Età	Età	Alfanumerico
Sesso	Sesso	Alfanumerico
StatoCivile	Stato civile	Alfanumerico
ComuneDiNascita	Comune di nascita	Alfanumerico
ProvinciaDiNascita	Provincia di nascita	Alfanumerico
ComuneDiResidenza	Comune di residenza	Alfanumerico
ProvinciaDiResidenza	Provincia di residenza	Alfanumerico
CambioResidenzaUltimoAnno	Indica se nell'ultimo anno il contribuente ha cambiato residenza	Flag

Continua nella prossima pagina

Tabella 3.1 – continua dalla pagina precedente

Nome	Descrizione	Tipo
ComuneDomicilioFiscale	Comune del domicilio fiscale	Alfanumerico
ProvinciaDomicilioFiscale	Provincia del domicilio fiscale	Alfanumerico
NumeroFamiliariACarico	Numero di familiari a carico	Numerico
ResidenzaDiversaDalConiuge	Indica se la residenza del dichiarante è diversa da quella del coniuge	Flag
ConiugeACarico	Indica se il contribuente ha un coniuge a carico	Flag
NumeroFigliACarico	Numero di figli a carico del contribuente	Numerico
NumeroFigliDisabili	Numero di figli disabili a carico del contribuente	Numerico
NumeroFigliMinoriTreAnni	Numero di figli a carico del contribuente con età minore di tre anni	Numerico
HaPiuDiQuattroFigli	Indica se il contribuente ha più di quattro figli	Flag
NumeroFigliAd- AffidamentoCongiunto	Numero di figli ad affidamento congiunto	Numerico
IsSoggettoACaricoDiAltri	Indica se il contribuente è a carico di un altro soggetto	Flag
IsExFiglioACaricoDiCliente	Indica se in passato il contribuente è stato a carico di un altro soggetto	Flag
IsExConiugeACaricoDiCliente	Indica se in passato il contribuente è stato coniuge a carico di un altro soggetto	Flag
ProvinciaSostitutoImposta	Provincia del sostituto d'imposta	Alfanumerico

Continua nella prossima pagina

Tabella 3.1 – continua dalla pagina precedente

Nome	Descrizione	Tipo
ComuneSostituto-DiversoDaCliente	Il comune del sostituto d'impostsa nel caso in cui esso sia diverso dal comune di residenza del cliente	Alfanumerico
ProvinciaSostituto-DiversoDaCliente	La provincia del sostituto d'impostsa nel caso in cui essa sia diversa dalla provincia di residenza del cliente	Alfanumerico
PROV_PRATICA		Alfanumerico
COD_ATTIVITA	Codice dell'attività del sostituto d'imposta della persona che sta effettuando la dichiarazione	Alfanumerico
CAPresidenza	Indica il cap di residenza del contribuente	Numerico

Dati relativi a lavoro, pensione e redditi vari

Tabella 3.2: Attributi: dati relativi a lavoro, pensione e redditi vari.

Nome	Descrizione	Tipo
NumeroDiTerreniDichiarati	Numero di terreni dichiarati dal contribuente	Numerico
NumeroDiTerreniConReddito-Dominicale	Numero di terreni con reddito dominicale dichiarati dal contribuente	Numerico
NumeroDiTerreniConRedditoAgrario	Numero di terreni con reddito agrario dichiarati dal contribuente	Numerico

Continua nella prossima pagina

Tabella 3.2 – continua dalla pagina precedente

Nome	Descrizione	Tipo
NumeroDiTerreniPosseduti-AlCentoPerCento	Numero di terreni posseduti al cento per cento dal contribuente	Numerico
GiorniMediDiPossessoTerreno	Media dei giorni in cui il contribuente ha avuto il terreno in possesso durante l'anno di riferimento	Numerico
TitoloPrevalenteDeiTerreni	Tipo di utilizzo prevalente dei terreni	Insieme
TuttiITerreniConLoStessoTitolo	Indica se tutti i terreni dichiarati hanno lo stesso tipo di utilizzo	Flag
NumeroDiTerreniInAffitto	Numero di terreni concessi in affitto	Numerico
RedditoDominicaleTotale	Totale del reddito dominicale da terreni	Numerico
RedditoDominicaleMedio	Reddito dominicale medio	Numerico
RedditoAgrarioTotale	Totale del reddito agrario da terreni	Numerico
RedditoAgrarioMedio	Media del reddito agrario da terreni	Numerico
RedditoTotaleDaTerreni	Totale del reddito da terreni	Numerico
NumeroDiFabbricatiDichiarati	Numero di fabbricati dichiarati	Numerico
RenditaDaFabbricatiTotale	Rendita da totale da fabbricati dichiarati	Numerico
NumeroDiFabbricatiPosseduti-AlCentoPerCento	Numero di fabbricati posseduti al cento per cento	Numerico
GiorniMediDiPossessoFabbricato	Media dei giorni in cui il contribuente ha avuto il fabbricato in possesso durante l'anno di riferimento	Numerico

Continua nella prossima pagina

Tabella 3.2 – continua dalla pagina precedente

Nome	Descrizione	Tipo
UtilizzoPrevalenteDeiFabbricati	Tipo di utilizzo prevalente dei fabbricati	Insieme
NumeroUtilizziDiversiFabbricati	Numero di utilizzi diversi dei fabbricati	Numerico
CasiParticolari	Indica se ci sono stati degli eventi particolari in relazione ai fabbricati	Flag
CasoParticolarePrevalente	Evento particolare verificatosi più frequentemente in relazione al fabbricato	Flag
NumeroDiFabbricatiDatiInAffitto	Numero di fabbrifcati dati in affitto	Numerico
CreditiDiImpostaPerAcquisto-PrimaCasa	Credito di imposta per l'acquisto della prima casa	Numerico
CreditiDiImpostaPerCanoni-NonPercepiti	Credito di imposta per canoni non percepiti	Flag
PresenzaDiRedditi	Indica se sono presenti dei redditi	Flag
NumeroLavoriDipendentiSezione1	Numero di lavori svolti da dipendente dichiarati nel quadro C sezione 1	Numerico
TotaleRedditoQuadroCSezione1	Totale dei redditi da lavori dichiarati nel quadro C sezione 1	Numerico
LavoratoreInPensione	Indica se il contribuente è in pensione e continua a lavorare	Flag
PassaggioAllaPensione	Indica se nell'anno di riferimento il lavoratore è andato in pensione	Flag

Continua nella prossima pagina

Tabella 3.2 – continua dalla pagina precedente

Nome	Descrizione	Tipo
PrevalenzaLavoriTempo-DeterminatoIndeterminato	Indica se i lavori a tempo determinato prevalgono su quelli a tempo indeterminato nell'anno di riferimento	Flag
LavoratoreStabileSezione1	Indica se il contribuente ha lavorato per tutto l'anno di riferimento	Flag
NumeroLavoriDichiaratiSezione2	Numero di lavori dichiarati nel quadro C sezione 2	Numerico
TotaleRedditoQuadroCSezione2	Totale dei redditi da lavori dichiarati nel quadro C sezione 2	Numerico
RicezioneAssegniDaConiuge	Indica se il contribuente riceve assegni dal coniuge	Flag
PrevalenzaRedditiSezione2suSezione1	Indica se i redditi dichiarati nel quadro C sezione 2 prevalgono sui redditi dichiarati nel quadro C sezione 1	Flag
TotaleRedditoDaLavoro	Totale dei redditi generati da lavoro dipendente	Numerico
RitenuteIrpefSuLavoro-DipEAssimilati	Ritenute IRPEF sui lavori dichiarati nel quadro C sezione 1	Numerico
RitenuteAddComunale-SuLavoroDIpEAssimilati	Indica se è presente una ritenuta addizionale comunale sui lavori dichiarati nel quadro C sezione 1	Flag
RitenuteAddRegionale-	Ritenute addizionali regionali sui lavori dichiarati nel quadro C sezione 1	Numerico

Continua nella prossima pagina

Tabella 3.2 – continua dalla pagina precedente

Nome	Descrizione	Tipo
SuLavoroDIpEAssimilati		
PresenzaRedditiInQuadroD	Indica se sono presenti dei redditi nel quadro D	Flag
PresenzaDiRedditiInQuadroDSez1	Indica se sono presenti dei redditi nel quadro D sezione 1	Flag
PresenzaDiRedditiIn- QuadroDSez2(TassazSeparata)	Indica se sono presenti dei redditi a tassazione separata	Flag
UtiliEAltriProventiEquiparati	Utili ed altri proventi equiparati	Numerico
AltriRedditiDiCapitale	Altri redditi da capitale	Numerico
CompensiDiLavoroAutonomo- NonDerivantiDaAttProf	Compensi da lavoro autonomo derivanti da attività professionale	Numerico
RedditiDiversi	Redditi diversi	Numerico
PresenzaRedditiDaErede(TassSep)	Indica se sono presenti redditi da erede a tassazione separata	Flag
TotaleAltriRedditi (tot_quadroD)	Totale dei redditi dichiarati nel quadro D	Numerico
ImposteEdOneriAnnoPrec- RimborsEdAltri(TassSep)	Tipo di redditi a tassazione separata	Numerico
IsPensionato	Indica se il contribuente è in pensione	Flag
PresenzaLavoriConCodice3- InQuadroC	Indica se sono presenti lavori con codice 3 nel quadro D sezione 1	Flag

Dati relativi a spese

Tabella 3.3: Attributi: dati relativi a spese.

Nome	Descrizione	Tipo
PresenzaSpeseInSezione1	Indica se sono presenti delle spese nella sezione 1 del quadro E	Flag
SpeseSanitarieProprie-PatologieEsistenti	Spese sanitarie per patologie esistenti	Numerico
SpeseSanitarieProprie	Spese sanitarie proprie	Numerico
SpeseSanitarieFamiliariNonACarico	Spese sanitarie per familiari non a carico	Numerico
SpeseSanitariePerPortatori-DiHandicap	Spese sanitarie per portatori di handicap	Numerico
SpesePerAcquistoCaniGuida	Spese per l'acquisto di cani guida	Insieme
SpeseSanitarieAnni-PrecedentiRateizzate	Spese sanitarie di anni precedenti rateizzate	Numerico
TotaleSpeseSanitarie	Totale delle spese sanitarie	Numerico
RapportoSpeseSanitarie-RedditoTotale	Spese sanitarie / Reddito totale	Numerico
SpesePerMutuiPerAcquisto-AbitazionePrincipale	Spese mutui per acquisto abitazione principale	Numerico
SpesePerMutuiPerCostruzione-AbitazionePrincipale	Spese mutui per costruzione abitazione principale	Numerico
SpesePerMutuiPerAcquisto-	Spese per mutui per acquisto di altri immobili	Numerico

Continua nella prossima pagina

Tabella 3.3 – continua dalla pagina precedente

Nome	Descrizione	Tipo
AltriImmobili		
SpesePerMutui	Spese per mutui totali	Numerico
RapportoSpesePerMutui-RedditoTotale	Spese per mutui / Reddito totale	Numerico
RapportoSpeseMutui/SpeseTotali	Spese per mutui / Spese totali	Numerico
SpesePerAssicurazioni	Spese per assicurazioni	Numerico
RapportoSpeseSanitarieSpeseTotali	Spese sanitarie / Spese totali	Numerico
SpeseFunebri	Spese funebri	Numerico
SpeseDiIstruzione	Spese di istruzione	Numerico
AltreSpeseSez1	Altre spese dichiarate nel quadro E sezione 1	Numerico
TotaleSpeseConDetrazioneImposta19	Totale spese con detrazione di imposta al 19%	Numerico
PresenzaSpeseInSezione2	Indica se sono presenti delle spese nella sezione 2 del quadro E	Flag
SpesePerContributiPrevidenziali EdAssistenzialiDeducibili	Spese per contributi previdenziali ed assistenziali deducibili	Numerico
SpesePerContributiPerServizi-DomesticiEFamiliari	Spese per contributi per servizi domestici e familiari	Numerico
SpesePerErogazioniLiberali-AFavoreDiIstituzioniReligiose	Spese per erogazioni liberali a favore di istituzioni religiose	Numerico
SpesePerAssegnoAlConiuge	Spese per assegni al coniuge	Numerico
SpesePerPrevidenzaComplementare	Spese per previdenza complementare	Numerico

Continua nella prossima pagina

Tabella 3.3 – continua dalla pagina precedente

Nome	Descrizione	Tipo
SpeseAssistenzaPortatoriHandicap	Spese per assistenza ai portatori di handicap	Numerico
AltriOneriDeducibili	Altri oneri deducibili	Numerico
TotaleSpeseDeducibili-DalRedditoComplessivo	Totale delle spese deducibili dal reddito complessivo	Numerico
PresenzaSpeseInSezione3	Indica se sono presenti delle spese nel quadro D sezione 3	Flag
NumeroInterventiPerIlRecupero-DelPatrimonioEdilizio	Numero di interventi per il recupero del patrimonio edilizio	Numerico
TotaleSpeseConDetrazione-Imposta36o41	Totale delle spese con detrazione di imposta al 36% o 41%	Numerico
PresenzaSpeseInAltreDetrazioni	Indica se sono presenti delle altre spese detraibili	Flag
PresenzaDetrazioneSeSiTrasferisce-LaResidenzaDiLavoro	Indica se sono presenti delle detrazioni nel caso in cui si trasferisca la residenza di lavoro	Flag
PresenzaDetrazioneCasa-PrincipaleInAffitto	Presenza di detrazioni relative alla casa principale in affitto	Flag
RapportoSpeseConDetrazione-Imposta19SpeseTotali	Spese con detrazione di imposta al 19% / Spese totali	Numerico
RapportoSpeseConDetrazione-	Spese con detrazione di imposta al 36% o 41% / Spese totali	Numerico

Continua nella prossima pagina

Tabella 3.3 – continua dalla pagina precedente

Nome	Descrizione	Tipo
Imposta36o41SpeseTotali		
RapportoSpeseDeducibili-DalRedditoSpeseTotali	Spese deducibili dal reddito / Spese totali	Numerico
RapportoSpeseDetraibiliSpeseTotali	Spese detraibili / Spese totali	Numerico
RapportoSpeseTotaliRedditoTotale	Spese totali / Reddito Totale	Numerico
RapportoSpeseTotali-RedditoDaLavoroOPensione	Rapporto spese totali / reddito da lavoro in pensione	Numerico
RapportoSpeseTotali-RedditoDaProprieta	Rapporto spese totali / reddito da proprietà	Numerico
SpeseTotali	Spese totali	Numerico
TotaleSpeseDetraibili	Spese detraibili totali	Numerico

Altri dati

Tabella 3.4: Attributi: altri dati.

Nome	Descrizione	Tipo
CreditoDiImpostaPerIncrementoNel-lOccupazione	Credito di imposta per incremento nell'occupazione	Numerico
RedditoEstero	Reddito estero	Numerico
RelazioneRedditoEstero-RedditoTotale	Relazione reddito estero con reddito totale	Numerico

Continua nella prossima pagina

Tabella 3.4 – continua dalla pagina precedente

Nome	Descrizione	Tipo
RedditoInStatoEsteroConRegime- FiscalePrivilegiato	Reddito in stato estero con un regime fiscale privilegiato	Numerico
AccontoIrpefAnnoPrecedente	Acconto IRPEF dell'anno precedente	Numerico
AltreRitenute	Altre ritenute	Numerico
EccedenzeIrpefDaPrecedenti- Dichiarazioni	Eccedenza IRPEF da precedenti dichiarazioni	Numerico
RitenuteEAccontiSospesiPer- EventiEccezionali	Ritenute acconti sospesi per eventi eccezionali	Numerico
NonEffettuarePagamento- AccontoIrpef	Indica se il contribuente deve effettuare o meno il pagamento dell'acconto IRPEF	Flag
PagamentoAccontoIrpef- InManieraMinore	IRPEF da pagare in presenza di agevolazioni	Numerico
RapportoRedditoImponibile- RedditoComplessivo	Reddito imponibile / reddito complessivo	Numerico
RapportoTotDetrCreditoDi- ImpostaRedditoImponibile	Totale detrazione credito di imposta / reddito imponibile	Numerico
RapportoTotDetrCreditoDi- ImpostaRedditoComplessivo	Totale detrazione credito di imposta / reddito complessivo	Numerico

Continua nella prossima pagina

Tabella 3.4 – continua dalla pagina precedente

Nome	Descrizione	Tipo
RapportoImpostaNetta-ImpostaLorda	Imposta netta / imposta lorda	Numerico
RapportoRitenuteImpostaLorda	Ritenute / imposta lorda	Numerico
RapportoTotaleDetrazioni-RedditoComplessivo	Totale detrazioni / reddito complessivo	Numerico
RapportoRedditoTerreni-RedditoComplessivo	Reddito terreni / reddito complessivo	Numerico
RapportoRedditoFabbricati-RedditoComplessivo	Reddito fabbricati / reddito complessivo	Numerico
RapportoRedditoPossedimenti-RedditoComplessivo	Reddito possedimenti / reddito complessivo	Numerico
RapportoRedditoLavoroSu-RedditoComplessivo	Reddito lavoro / reddito complessivo	Numerico
RapportoRedditoAltriRedditi-SuRedditoComplessivo	Reddito da altri redditi / reddito complessivo	Numerico
RapDetrazioniFamiliariACarico-TotaleDetrazioni	Detrazioni familiari a carico / altre detrazioni	Numerico
RapDetrazioniLavoroDipendente-TotaleDetrazioni	Detrazioni lavoro dipendente / totale detrazioni	Numerico

Continua nella prossima pagina

Tabella 3.4 – continua dalla pagina precedente

Nome	Descrizione	Tipo
RapTotDetrazioniCrediti-DiImpostaImpostaNetta	Totale detrazioni crediti di imposta / imposta netta	Numerico
RapCreditiImpostaImpostePagate-EstTotDetrCredImposta	Crediti di imposta per imposte pagate all'estero / totale crediti di imposta	Numerico

Capitolo 4

Preparazione dei dati

4.1 Esplorazione, selezione e trasformazione dei dati

A questo punto del progetto abbiamo quindi a disposizione le cinque viste contenenti tutte le informazioni che, in un primo momento, erano state ritenute interessanti per la realizzazione del modello di *data mining*.
Tuttavia non tutte le informazioni, e quindi gli attributi, qui presenti possono essere utilizzate direttamente per la modellazione poiché possono ancora essere presenti attributi ridondanti o poco utili per descrivere le caratteristiche dei nuovi clienti.
In questo capitolo arriveremo quindi alla definizione dell'insieme finale dei dati che verrà utilizzato per la realizzazione di un modello di qualità; la selezione degli attributi più utili per i nostri scopi sarà effettuata con la complicità finale degli esperti del dominio, i quali, avendo appunto una conoscenza completa e totale del dominio di analisi, saranno sicuramente in grado di aiutarci nell'arrivare a decisioni corrette riguardanti la discretizzazione di alcuni attributi e la selezione e l'eliminazione di altri.

4.1.1 Selezione degli attributi interessanti

Lo studio delle distribuzioni statistiche degli attributi e della correlazione tra questi ultimi ha portato alla eliminazione o alla ulteriore trasformazione di alcuni di questi attributi. L'eliminazione di alcuni di questi si è resa necessaria poiché alcuni attributi si sono rivelati essere o poco utili per descrivere i nuovi clienti (come il comune di nascita), o caratterizzati da uno scarso contenuto informativo (come l'attributo riguardante le spese per l'eventuale presenza di un cane guida) oppure, ancora, ridondanti l'uno con l'altro perché altamente correlati (come gli attributi età ed anno di nascita). Altri attributi sono stati invece trasformati in modo tale da sfruttare al meglio il loro potenziale informativo: così, per esempio, l'attributo relativo al comune di residenza riguardante la persona che effettua la dichiarazione è stato trasformato in un attributo di tipo binario, pendolare, recante l'informazione necessaria per capire se la persona che effettua la dichiarazione è, appunto, pendolare o meno.
L'insieme degli attributi sopravvissuti a questa prima selezione consta così di 69 elementi; essi si possono considerare realmente interessanti dal momento che contengono un tipo di informazione facile da reperire e favoriscono scelte azionabili in una futura ottica di marketing.
Ecco l'elenco degli attributi selezionati:

Tabella 4.1: Gli attributi selezionati.

TIPODICHIARAZIONE	REDDITO_TOT_TERRENI
SPESE_ISTRUZIONE	ALTRE_SPESE_SEZ1
DICHIARAZIONECONGIUNTA	NUM_FABBRICATI
COMPILA_DICH	REDDITO_FABBRICATI
SPESE_DETRAZIONE_19	PRESENZA_SPESE_SEZ2
PROVINCIA_PRAT	USO_PREV_FABBRICATI
Continua nella prossima pagina	

Tabella 4.1 – continua dalla pagina precedente

ETA	USO_DIVERSO_FABBRIC
SPESE_CON_PREVIDENZ	SPESE_DEDUCIBILI
SESSO	CASIPARTICOLARI
STATOCIVILE	AFFITTA_FABBRICATI
PRESENZA_SPESE_SEZ3	INTERVENTI_EDILIZI
CHURN	REDDITO_QUC_SEZ1
PENDOLARE	PASSAGGIO_A_PENSIONE
PRESENZA_SPESE_36O41	RAP_SPESE19_SPESETOT
COMUNEDIRESIDENZA	LAVORATORE_STABILE
CAMBIO_RESIDENZA	PRESENZA_LAVORI_SEZ2
RAP_SPESEDETR_SPESETOT	RAP_SPESETOT_REDDITOTOT
NUM_FAM_CARICO	REDDITO_DA_LAVORO
CONIUGE_A_CARICO	RITENUTE_IRPEF_LAV
RAP_SPESETOT_REDDITOLAV	SPESE_DETRAIBILI_TOT
NUM_FIGLI_CARICO	PRESENZA_REDDITI_QUD
FIGLI_MINORI_3ANNI	IS_PENSIONATO
ANNO	NUOVOCONIUGE
A_CARICO_DI_ALTRI	PRESENZA_SPESE_SEZ1
EX_FIGLIO_A_CARICO	SPESE_SANIT_PROPRIE
ECCEDENZE_IRPEF_PREC	NON_PAGARE_ACCONTO_IRPEF
EX_CONIUGE_A_CARICO	SPESE_SANITARIE
POSSIEDE_TERRENI	SPESE_MUTUI_CASA
RAP_REDDFABBR_REDDTOT	RAP_REDDLAV_REDDTOT
TIPO_PREVAL_TERRENI	SPESE_MUTUI
REDDITO_AGRARIO	RAP_SPESEMUTUI_SPESETOT
RAP_REDDPROPRIETA_REDDTOT	CAPresidenza
REDDITO_DOMINICALE	SPESE_ASSICURAZIONI

Continua nella prossima pagina

Tabella 4.1 – continua dalla pagina precedente

SPESE_FUNEBRI	RAP_SPESESANIT_SPESETOT
TIPO_SOSTITUTO_IMPOSTA	

4.1.2 Validazione degli esperti del dominio

Gli esperti del dominio sono stati consultati per una conferma riguardante gli attributi scelti e per un'ulteriore sotto-selezione degli attributi maggiormente interessanti. Inoltre, l'intervento degli esperti del dominio si è reso nuovamente necessario per una corretta discretizzazione degli attributi di tipo continuo quali l'età o il reddito. La discretizzazione è obbligatoria dal momento che il modello di mining che useremo, ovvero le regole di classificazione, ha come restrizione quella di elaborare soltanto valori discreti. Oltre a ciò, la discretizzazione permette inoltre di avere a che fare con risultati più chiari e facilmente interpretabili.
Agli esperti del dominio è stata così sottoposta la lista degli attributi presentata nel precedente paragrafo e, dati una descrizione riguardante il significato di ognuno di essi e un grafico delle distribuzioni dei valori per alcuni, agli esperti è stato chiesto di scegliere quali attributi mantenere, quali eliminare, quali aggiungere e quali, eventualmente trasformare.
A seguito di questa fase sono stati eliminati alcuni attributi e ne sono stati aggiunti altri. La lista finale degli attributi risulta così essere questa (gli attributi aggiunti sono evidenziati usando il grassetto):

Tabella 4.2: Gli attributi selezionati.

TIPODICHIARAZIONE	ANNO
NUOVOCONIUGE	SPESE_DETRAZIONE_19_FLAG
DICHIARAZIONECONGIUNTA	NUM_FABBRICATI_DISC
Continua nella prossima pagina	

Tabella 4.2 – continua dalla pagina precedente

USO_PREV_FABBRICATI_DISC	REDDITO_DA_LAVORO_DISC
ETA_DISCR	SPESE_CON_PREVIDENZ_FLAG
SESSO	PRESENZA_SPESE_SEZ3
STATOCIVILE	AFFITTA_FABBRICATI_FLAG
TIPO_SOSTITUTO_IMPOSTA	INTERVENTI_EDILIZI_FLAG
PENDOLARE	PASSAGGIO_A_PENSIONE
COMUNEDIRESIDENZA	LAVORATORE_STABILE
CAMBIO_RESIDENZA	PRESENZA_LAVORI_SEZ2
CONIUGE_A_CARICO_FLAG	**SPESE_INTERM_IMM**
NUM_FIGLI_CARICO_DISC	PRESENZA_REDDITI_QUD
FIGLI_MINORI_3ANNI_FLAG	IS_PENSIONATO
A_CARICO_DI_ALTRI_FLAG	PRESENZA_SPESE_SEZ1
EX_FIGLIO_A_CARICO	SPESE_SANIT_PROPRIE_FLAG
EX_CONIUGE_A_CARICO	SPESE_SANITARIE_FLAG
POSSIEDE_TERRENI_FLAG	SPESE_MUTUI_CASA_FLAG
SPESE_MUTUI_FLAG	**DETR_INQ_ALLOGGIO_PRINC**
RAP_SPESEMUTUI_SPESETOT	**SPESE_SPORT_RAGAZZI_FLAG**
SPESE_ASSICURAZIONI_FLAG	CAPresidenza
PRESENZA_SPESE_36O41	RAP_SPESETOT_REDDITOTOT
PRESENZA_SPESE_SEZ2	**INIZIO_INTERVENTI_EDILIZI_FLAG**
SPESE_PER_ONLUS_FLAG	**NON_DOVEVA_FARE_730_FLAG**

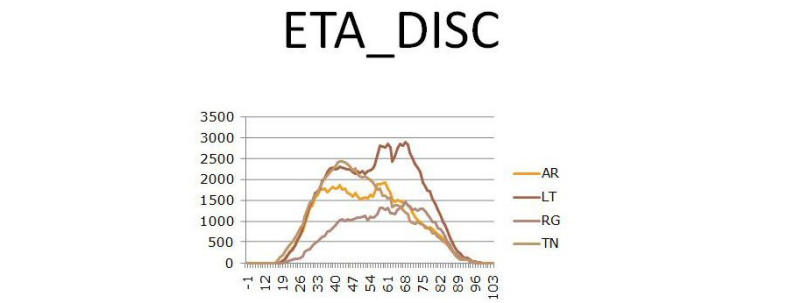
Gli attributi aggiunti dagli esperti del dominio sono 6:

- SPESE_INTERM_IMM;
- DETR_INQ_ALLOGGIO_PRINC;
- SPESE_SPORT_RAGAZZI_FLAG;
- INIZIO_INTERVENTI_EDILIZI_FLAG;

- SPESE_PER_ONLUS_FLAG;
- NON_DOVEVA_FARE_730_FLAG.

Il primo di essi dà un'informazione riguardante l'eventualità che, nell'anno della dichiarazione, ci siano state delle spese per una intermediazione immobiliare, il secondo ci dice se è presente una detrazione per l'abitazione principale, il terzo se ci sono state delle spese per le attività sportive dei figli, il quarto se, nell'anno della dichiarazione, ci sono stati iniziati degli interventi edilizi, il quinto se ci sono state delle spese verso delle ONLUS e il sesto ci dice se, nell'anno della dichiarazione, la persona soggetto della dichiarazione era esentato o meno dal presentare il 730.
Gli attributi che sono stati invece eliminati sono 27 e, tra l'eliminazione di questi ultimi e l'aggiunta di quelli appena descritti, si arriva a un totale di 48 attributi, 49 quando andremo a considerare anche l'attributo target binario NUOVOCLIENTE.

Come già detto in precedenza, gli esperti del dominio sono stati inoltre consultati per un problema riguardante la corretta discretizzazione di alcuni attributi numerici. Sono stati così analizzati i grafici riguardanti la distribuzione dei valori di questi attributi e, a partire da questi e da una conoscenza soggettiva degli esperti, sono stati scelti gli intervalli di valori più giusti in cui dividere i suddetti attributi.
Così, per esempio, l'attributo età è stato suddiviso in sei intervalli: $<= 25$, $26-35$, $36-45$, $46-55$, $56-65$ e > 65. La stessa operazione è stata ripetuta per altri attributi quali il reddito da lavoro, il numero dei fabbricati, l'uso prevalente dei fabbricati, il rapporto tra le spese totali e il reddito totale e il numero dei figli a carico.
Alla fine di questa fase si arriva così all'insieme dei dati finali pronti ad affrontare la fase di modellazione.



	AR	LT	RG	TN	Total
<=15	0,01%	0,03%	0,00%	0,00%	0,01%
16-25	2,25%	1,59%	0,75%	3,53%	2,11%
26-35	14,81%	10,76%	6,46%	14,69%	12,02%
36-45	20,34%	17,45%	14,94%	23,72%	19,34%
46-55	18,35%	17,13%	17,78%	21,79%	18,73%
56-65	19,82%	20,76%	20,35%	16,43%	19,35%
66-75	14,22%	19,73%	21,86%	11,21%	16,58%
76-85	8,24%	10,44%	14,85%	6,94%	9,74%
>=85	1,97%	2,11%	3,02%	1,69%	2,12%
Total	100,00%	100,00%	100,00%	100,00%	100,00%

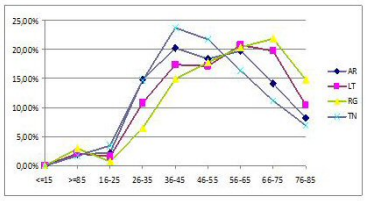


Figura 4.1: Un esempio: la distribuzione dell'attributo età al momento della prima discretizzazione presentata agli esperti del dominio.

Capitolo 5

La modellazione

Al termine della fase precedente abbiamo un insieme di 50 attributi con un contenuto informativo tale da poter essere considerati utili per la costruzione di modelli di *data mining*. In questo capitolo descriveremo tutte le fasi che abbiamo attraversato per arrivare alla estrazione delle informazioni desiderate da questi 50 attributi: dalla scelta della tecnica di *data mining* alla definizione dei parametri con cui applicare la tecnica e alla valutazione dei modelli realizzati.

5.1 Selezione della tecnica di data mining

Come già accennato nella sezione 2.3.2, la tecnica più adatta per il nostro progetto risulta essere quella rappresentata dalle regole di classificazione o regole classificative; infatti, il nostro scopo non è quello di creare un classificatore che ci dica se un contribuente compila il 730 per la prima volta, ma è piuttosto quello di definire delle caratteristiche distintive di nicchie di nuovi clienti. Queste caratteristiche possono poi essere rappresentate attraverso delle regole aventi come antecedente i caratteri distintivi della nicchia e come conseguente il fatto di essere un nuovo cliente.

5.2 Modellazione mediante estrazione di regole di classificazione

5.2.1 Le regole di classificazione

Le regole di classificazione sono estratte in modo distinto per ciascuna delle quattro province in esame - Arezzo, Latina, Ragusa e Trento - utilizzando quattro viste facenti riferimento agli anni 2008 e 2009.
Le regole che andremo ad estrarre risultano essere così composte:

$$A, B \rightarrow C$$

dove A è la parte discriminante della regola, ovvero una coppia o una serie di coppie attributo-valore che rappresentano una o più caratteristiche dei contribuenti (ad esempio, *età=26-35*); B è il contesto, ovvero un insieme di item che rappresenta un gruppo di contribuenti in cui studiare la densità dei nuovi clienti (ad esempio, *città=Arezzo*) e, infine, C è la classe, ovvero, nel nostro particolare caso, *nuovocliente=true*: essa ci dice infatti se le caratteristiche presenti nella parte sinistra della regola sono effettivamente riferibili alla classe di nostro interesse, cioè quella popolata solo e soltanto da nuovi clienti.
Un esempio di regola di classifcazione che andremo ad esaminare può quindi essere questo:

$$eta = 26 - 35, citta = arezzo \rightarrow nuovocliente = true$$

La struttura di queste regole deriva da studi già effettuati ([40] e [41]) riguardanti il problema di determinare, dato un dataset contenente un certo numero di record di decisioni storiche, una misura precisa del grado di discriminazione di cui è stato oggetto un particolare gruppo (per esempio: una minoranza etnica) in un dato contesto (per esempio: un'area geografica) rispetto ad una decisione (per esempio: il rifiuto di concedere un prestito).

Così, il problema principale nell'analisi della discriminazione è precisamente quello di quantificare il grado di discriminazione patito da un certo gruppo in un determinato contesto rispetto ad una certa decisione. Nell'approccio qui presentato, approccio a partire da cui è stata ricavata la struttura delle regole che verranno estratte nel proseguo di questo progetto, questo problema viene riformulato facendo riferimento a una terminologia più consona ad un contesto di *data mining* e, più in particolare, di regole. Se, quindi, **A** è la condizione (cioè, l'itemset) che caratterizza il gruppo sospettato di essere oggetto di discriminazione, se **B** è l'itemset che caratterizza il contesto e se **C** è l'item che riguarda la decisione (cioè la classe), allora l'analisi della discriminazione ha come scopo quello di studiare le regoli aventi questa forma: $A, B \rightarrow C$. Le regole aventi tale forma sono dette *PD rules*, cioè *potentially discriminatory rules*, ovvero regole potenzialmente discriminanti. In queste regole **A** è un itemset non vuoto di tipo **PD**, e **B** è un itemset di tipo **PND** (cioè, *potentially non discriminatory*: potenzialmente non discriminante). Nel caso delle regole appena citate però - ovvero le *potentially discriminatory rules* - la parte discriminante - cioè il PD itemset, cioè A - è già fissata poiché è un gruppo già stabilito che si deve verificare se è oggetto di azioni discriminanati o meno, mentre, nel nostro caso, l'itemset A non è già fissato, ma viene continuamente estratto proprio come avviene per l'itemset di contesto B.

Per cui, le regole con cui ci troviamo ad avere a che fare, condividono la forma ed il significato degli itemset delle regole PD, ma non la sostanza poiché le nostre regole non cercano di identificare una eventuale discriminazione di cui può essere oggetto un gruppo in esame, ma cercano delle particolarità, delle caratterizzazioni - sempre diverse - che i nuovi clienti possono avere se studiati in differenti contesti.

5.2.2 L'estrazione delle regole

L'estrazione delle regole è stata effettuata tramite l'utilizzo del tool DCUBE [42], integrato nell'ambiente Oracle SQL Developer.
Al suo avvio, esso richiede che siano individuate la sorgente dati a partire da cui verranno estratte le regole (essa può essere un file nel formato *ARFF* o *CSV*, oppure può essere *JDBC*; in quest'ultimo caso dev'essere inoltre specificata la query SQL di selezione dei dati) e la destinazione Oracle in cui popolare le regole estratte. Una svolta specificate la sorgente dei dati e la

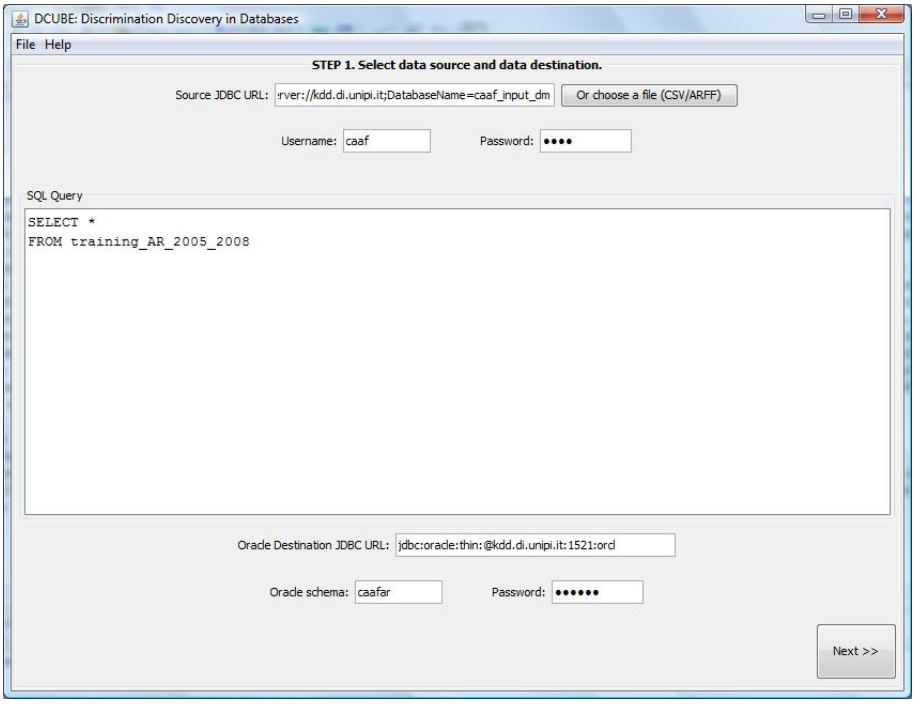


Figura 5.1: DCUBE: prima fase, la selezione della sorgente dati e della destinazione Oracle.

destinazione Oracle, dobbiamo selezionare la classe di interesse costituente il conseguente della regola. Nel nostro caso questa operazione consiste nello spostare l'item *'NUOVOCLIENTE=true'* nella selezione 'Class items'. A questo punto, dal momento che le regole di classificazione sono estratte a

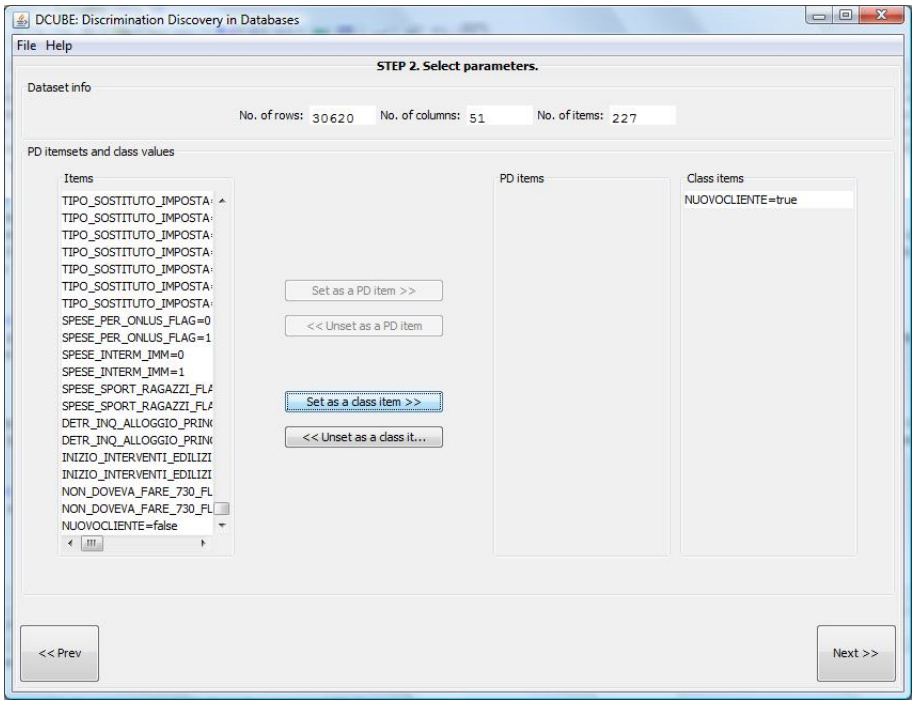


Figura 5.2: DCUBE: seconda fase, la selezione della classe.

partire dall'insieme di itemset frequenti e tale insieme può crescere in modo incontrollato (esponenziale nel numero di item), per limitare il numero di itemset frequenti sia da un punto di vista computazionale che di analisi successive, vengono proposti i seguenti parametri di estrazione:

- **Supporto minimo** degli itemset frequenti; può essere specificato sia in senso percentuale che assoluto. Il valore impostato nel nostro particolare progetto deriva da questa considerazione: mediamente i nuovi clienti sono il 15% del totale. Supponendo che esistano 20 regole capaci di caratterizzare i nuovi clienti, ognuna di queste, mediamente, dovrebbe coprire almeno lo $0,75\%$ dei clienti. Lo $0,75\%$ di supporto dovrebbe comprendere un numero sufficientemente alto di dichiarazioni, in modo da giustificare un'azione di marketing basata su una specifica regola. Assumendo che 200 persone sia un numero

adeguato, il valore di default del supporto minimo viene fissato in: $\max 0,75\%, 200 * num\_anni\_training$ (*num_anni_training* nel nostro caso è uguale a 2);

- *supporto massimo* degli itemset frequenti; in letteratura - [6] - è noto che item con supporto alto non sono significativi nelle regole estratte, ma, piuttosto, comportano un eccessivo carico computazionale nell'estrazione degli itemset frequenti. Per dimostrare ciò [6] ci propone un esempio basato sullo studio di un data set in cui gli item hanno una distribuzione asimmetrica, cioè in cui la maggior parte degli item ha un supporto moderatamente basso, ma esiste anche un piccolo insieme di item avente un supporto molto elevato. In particolare, in tale dataset più dell'80% degli item ha un supporto inferiore all'1%, mentre un piccolo gruppo di questi ha un supporto maggiore del 90%. Per meglio illustrare gli effetti della distribuzione asimmetrica sull'estrazione degli itemset frequenti, l'insieme degli item viene diviso in tre gruppi - G_1, G_2 e G_3 - secondo i loro corrispondenti valori di supporto. Il primo gruppo contiene i 1735 item che hanno un supporto inferiore all'1%, il secondo gruppo contiene i 358 item che hanno un supporto compreso tra l'1% e il 90% e il terzo gruppo comprende i 20 item che hanno un supporto superiore all'80%.

  La scelta della corretta soglia di supporto per il mining su questo data set può essere piuttosto difficile. Se viene impostata una soglia troppo elevata (per esempio, il 20%) possiamo perdere molti pattern interessanti che coinvolgono gli item a basso supporto contenuti in G_1. Nella *market basket analysis* tali item a basso supporto possono corrispondere a prodotti piuttosto costosi (come, per esempio, quelli di gioielleria) che sono comprati piuttosto raramente dagli acquirenti, ma il cui studio può essere comunque interessante per i rivenditori. Al contrario, se viene impostata una soglia di supporto troppo bassa, diventa

difficile trovare i pattern associativi per una delle seguenti ragioni. Per prima cosa, i requisiti computazionali e di memoria degli algoritmi di analisi associativa esistenti crescono considerevolmente con soglie di supporto troppo basse. Seconda cosa, il numero dei pattern estratti può crescere anch'esso sostanzialmente impostando soglie di supporto troppo basse. Terzo, possiamo correre il rischio di estrarre molti pattern spuri che mettono in relazione item ad alta frequenza come il latte con item a bassa frequenza come il caviale. Tali pattern, che chiamiamo *cross-support* pattern, cioè pattern a supporto incrociato, sono spuri nella maggior parte dei casi perché i loro livelli di correlazione tendono ad essere molto deboli. Per esempio, impostando una soglia di supporto pari allo 0.05%, ci sono $18,847$ pattern che coinvolgono item contenuti in G_1 e G_3. Di questi, il 93% sono pattern a supporto incrociato, cioè pattern che coinvolgono item contenuti sia in G_1 che in G_3. Il livello massimo di correlazione ottenuto a partire dai pattern a supporto incrociato è pari a 0.029, valore che è di molto inferiore al valore massimo di correlazione che si ottiene confrontando pattern estratti dallo stesso gruppo.

Con l'opzione di DCUBE appena presentata è quindi possibile eliminare dal training set gli item con supporto alto; come valore di default noi fissiamo al 90% la soglia massima di supporto;

- **differenza massima** tra due item. Due item A e B sono sinonimi al 10% se ogni 100 record in cui compare uno dei due (A o B) ci sono al più 10 casi in cui compare uno solo dei due (solo A o solo B). Formalmente, con riferimento alla tabella di contingenza 1.1, si ha $(b+c)/(a+b+c) <= 0.1$. Il sistema rimuove uno tra A e B, nel caso in cui questi siano sinonimi. L'item rimosso è detto sinonimo, mentre l'item mantenuto è detto rappresentante. Il sistema gestice il caso generale in cui A è sinonimo di più di un item in modo da rimuovere il

maggior numero possibile di item sinonimi (o in modo da mantenere il minor numero possibile di item rappresentanti); la lunghezza massima è impostata di default al 10%;

- **lunghezza massima** degli itemset frequenti. Dal momento che ragionare su regole con molti item non risulta conveniente e dal momento che estrarre itemset lunghi comporta tempi di calcolo più ampi, con questa opzione vogliamo limitare l'estrazione di itemset frequenti aventi una lunghezza massima prefissata. Impostiamo quindi il valore di default riferito alla lunghezza massima pari a 6, valore in cui è compreso anche l'item della classe.

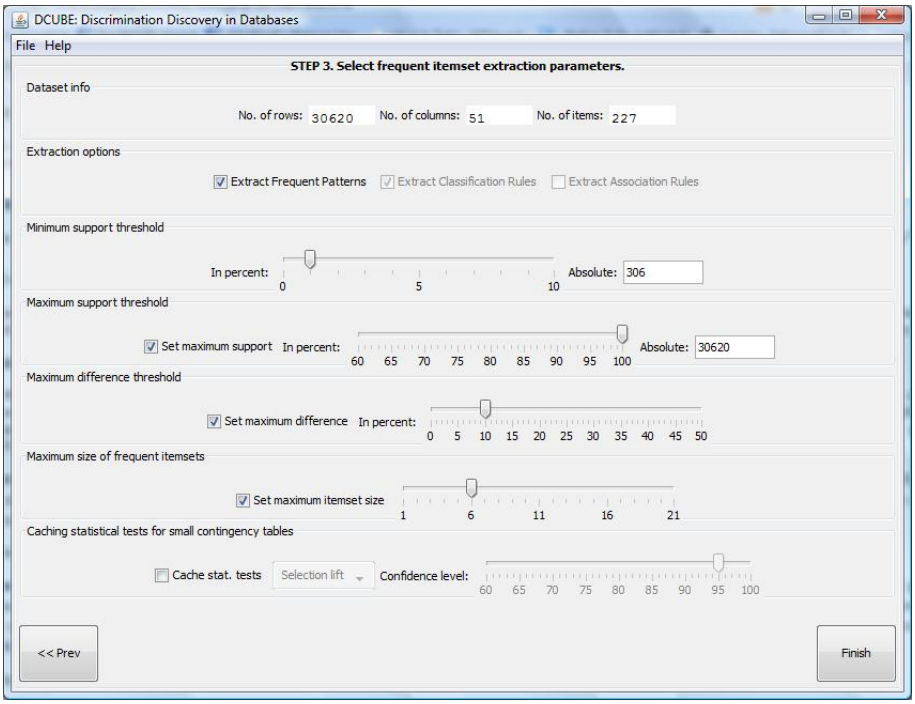


Figura 5.3: DCUBE: seconda fase, la selezione della classe.

Quindi, una volta impostati i valori di default per i parametri che abbiamo appena presentato, abbiamo atteso che, per ognuna delle quattro province, DCUBE popolasse le regole nella destinazione Oracle da noi indicata.

5.2.3 Quali regole possono considerarsi realmente interessanti?

La scelta delle misure di interesse

Una volta che DCUBE ha terminato il proprio lavoro, ci siamo trovati di fronte ad un numero veramente troppo grande di regole, numero che, come nella maggior parte dei processi di *mining* accade, ci rendeva impossibile esaminare singolarmente ogni regola per decidere se essa potesse essere interessante o meno. Per fare ciò, come abbiamo ampiamente parlato nel primo capitolo di questo lavoro, avevamo bisogno di trovare delle misure di interesse che, data una certa soglia minima da noi impostata, ci potessero aiutare a fare emergere solo le regole più interessanti secondo tale misura. Dopo una serie di studi e di prove, per la valutazione delle regole, la scelta è ricaduta su due misure in particolare: l'*elift* (*extended lift*) e il *4th quantifier of founded double implication*. Il loro utilizzo è facilmente attuaibile in DCUBE, dal momento che ad ogni regola è associata la propria corrispondente tabella di contingenza, a partire dalla quale possono poi essere calcolati i valori di qualsiasi misura scegliamo di implementare mediante metodi della classe di oggetti definiti da DCUBE e denominata CONTTABLE. In DCBUBE, oltre all'elift, sono di base implementate anche le seguenti misure: supporto, confidenza, coverage, selection lift e odd lift.
L'elift è il rapporto tra la confidenza della regola $A, B \rightarrow C$ e la confidenza media del solo contesto, ovvero di $B \rightarrow C$. In particolare se $A, B \rightarrow C$ è una regola di classificazione con $conf(B \rightarrow C) > 0$, allora l'elift della regola è:

$$elift(A, B \rightarrow C) = \frac{conf(A,B \rightarrow C)}{conf(B \rightarrow C)}.$$

Se abbiamo quindi a che fare con una regola del tipo $eta = 25 - 35, citta = arezzo \rightarrow nuovocliente = true$ con un valore di elift pari a 3, allora ciò significa che, tra gli aretini, restringendosi ai clienti nella fascia 25-35 si

triplica la frequenza di nuovi clienti rispetto alla media generale.
In riferimento alla tabella di contingenza 1.1, il 4th quantifier of founded equivalence è invece così definito:

$$4thQuantifierofFoundedDoubleImplication = \frac{a}{(n_1+c)}$$

Esso è cioè il rapporto tra il numero di casi che soddisfano sia A che B che C (il supporto della regola) fratto il numero di casi che soddisfano B ed almeno uno tra A e C. Se abbiamo quindi a che fare con una regola come la precedente con un valore di 4th quantifier of founded double implication pari all'80%, ciò significa che, tra gli aretini, esiste almeno l'80% di persone che ha un'età compresa tra i 25 e i 35 anni e fa parte del gruppo dei nuovi clienti.

La selezione delle regole basata sulle misure

Una volta scelte le due misure da utilizzare per la selezione delle regole abbiamo iniziato a lavorare concretamente nell'ambiente Oracle SQL Developer per capire se, in base alle misure scelte, emergevano delle caratteristiche interessanti capaci di caratterizzare diversi segmenti della popolazione cliente del CAAF-CISL.
Così, per ogni provincia, e per ognuna delle due misure scelte, abbiamo scritto una query che fosse in grado di fare emergere le prime 20 o 30 regole che rispettassero la soglia imposta per la misura in uso.

Nel caso dell'elift l'unico vincolo imposto all'inizio, oltre a quello relativo alla soglia minima, è stato quello di limitare la lunghezza del contesto ad un itemset composto da non più di due item per migliorare la leggibilità e l'interpretabilità delle regole.
Dopo una prima esecuzione della query ciò che emerge è una caratterizzazione che si può esprimere a parole come segue:

se entrambi i coniugi sono clienti CAAF-CISL allora entrambi sono nuovi clienti oppure nessuno dei due è nuovo.

```
SELECT d3pddecode(aset) AS aset, -- A
d3pnddecode(bset)     AS bset, -- B
'NUOVOCLIENTE=True'          AS c,     -- C
TRUNC(r.ct.elift(),2) AS M,
r.ct.support() AS Support
FROM pdrule r
JOIN pnditemsets pnd
ON r.bset = pnd.id
JOIN pditemsets pd
ON r.aset = pd.id
WHERE r.c =

  (SELECT D3ENCODE('NUOVOCLIENTE=True') FROM dual

  ) -- D3ENCODE calculated only once

AND r.ct.elift() > 5
AND pnd.len <= 2
ORDER BY r.ct.elift() DESC;
```

Figura 5.4: La selezione delle regole: esempio di una query.

Sulla globalità dei dati del 2008 e del 2009 esistono 29.346 coppie di coniugi che, congiuntamente o disgiuntamente, hanno fatto la dichiarazione presso CAAF-CISL. Di questi 2.512 coppie (pari al 8,6%) sono di entrambi nuovi clienti, 25.003 coppie (pari al 85,2%) sono di entrambi vecchi clienti, e 1.831 coppie (pari al 6,2%) sono con un nuovo cliente ed uno vecchio. Di queste ultime, però:

- in 768 casi (pari al 2,6% del totale delle coppie) il nuovo cliente della coppia era già coniuge a carico in una passata dichiarazione del coniuge vecchio cliente;
- in 747 casi (pari al 2,5% del totale delle coppie) il nuovo cliente della coppia non è un lavoratore stabile (cioè il numero giorni in cui ha

lavorato è inferiore ai 365).

Nei rimanenti 316 casi (pari al 1,1% del totale delle coppie), che quindi rappresentano i *veri* casi in cui uno dei due coniugi è nuovo e l'altro non lo è, il nuovo cliente della coppia risulta essere un pensionato.
Per questa ragione è stato deciso di eliminare, sia dalla parte discriminante della regola sia dal contesto, gli itemset che contenessero l'item 'NUOVOCONIUGE'.
Successivamente, il procedimento adottato è stato quello di diminuire progressivamente la soglia di elift fino ad arrivare ad un valore al di sotto del quale non sarebbe stato opportuno andare, il tutto congiuntamente alla eliminazione progressiva di attributi che caratterizzassero buona parte delle regole emerse ad ogni nuova esecuzione della query modificata.
Così, dopo aver scartato l'item 'NUOVOCONIUGE', ciò che è emerso è che buona parte dei nuovi clienti erano caratterizzati dal fatto di aver cambiato residenza nell'ultimo anno, così è stato eliminato, sia dalla parte discriminante delle regole sia dal loro contesto, l'item 'CAMBIO_RESIDENZA'.
Come già detto, questo modo di procedere è stato portato avanti sino a che non sono continuate ad emergere caratteristiche significative che accomunassero buona parte dei nuovi clienti e sino a quando la soglia minima continuava ad essere su livelli significativi tali da poter rendere una regola effettivamente interessante.
Lo stesso procedimento appena descritto è stato adottato anche per l'altra misura scelta, vale a dire il 4th quantifier of founded double implication. Anche qui all'inizio, anche se in un modo meno significativo rispetto all'elift, è emersa la stessa caratterizzazione riguardante l'attributo 'NUOVOCONIUGE', il quale è stato nuovamente e prontamente scartato.
Per quanto riguarda l'elift, a seguito di un'osservazione approfondita delle regole fin qui selezionate, abbiamo notato che, avendo a che fare con una regola $A, B \rightarrow C$, se, in essa, $B \rightarrow A$, allora la regola $A, B \rightarrow C$ è equivalente

alla regola $B \rightarrow C$. Per questa ragione emergono regole del tipo:

$$AFFITTA\_FABBRICATI\_FLAG = 0, NUM\_FABBRICATI\_DISC = 0 \rightarrow NUOVOCLIENTE = true$$

in cui è piuttosto ovvio che il non possedere fabbricati comporterà il fatto di non avere fabbricati concessi in affitto. L'affermazione appena fatta è confermata dalle tabelle di contingenza che andiamo ad esaminare, rispettivamente per le regole $A, B \rightarrow C$ e $B \rightarrow C$, cioè per $AFFITTA\_FABBRICATI\_FLAG = 0, NUM\_FABBRICATI\_DISC = 0 \rightarrow NUOVOCLIENTE = true$, ovvero la tabella 5.1, e per $NUM\_FABBRICATI\_DISC = 0 \rightarrow NUOVOCLIENTE = true$, ovvero la tabella 5.2.

In particolare, dobbiamo prendere in considerazione, per ogni tabella, la

Tabella 5.1: Tabella di contingenza per $AFFITTA\_FABBRICATI\_FLAG = 0, NUM\_FABBRICATI\_DISC = 0 \rightarrow NUOVOCLIENTE = true$

B	***C***	$\neg C$	
A	2035	5956	7991
$\neg A$	0	0	0
	0	0	7991

Tabella 5.2: Tabella di contingenza per $NUM\_FABBRICATI\_DISC = 0 \rightarrow NUOVOCLIENTE = true$

	C	$\neg C$	
B	2035	5956	7991
$\neg B$	1939	20690	22899
	3974	26646	30890

prima riga partendo dall'alto, poiché, per quanto riguarda la prima regola, in essa non si fa riferimento alle persone che posseggono fabbricati, come invece lo si fa nella seconda e nella terza riga della seconda.

Così vediamo che, il numero di persone che non ha fabbricati dati in affit-

to, che non possiede fabbricati e che è nuovo cliente è identico al numero di persone che non possiede fabbricati e è nuovo cliente, senza l'ulteriore caratterizzazione rappresentata dal fatto di concedere fabbricati in affitto. Allo stesso modo, il numero di persone che non ha fabbricati dati in affitto, che non possiede fabbricati e che non è nuovo cliente è identico al numero di persone che non possiede fabbricati e che non è nuovo cliente, senza, anche questa volta, l'ulteriore discriminante rappresentata dal fatto di concedere fabbricati in affitto.

Per quanto appena esposto è stata testata un'altra metodologia di selezione delle regole. Anche qui, ancora una volta, tra le regole emerse ad ogni nuovo valore della soglia minima, abbiamo cercato una particolare caratterizzazione che accomunasse buona parte delle regole nella parte discriminante. Una volta trovata, come, per esempio, nel caso dell'item 'CAMBIO_RESIDENZA=true', è stato imposto il vincolo che tale item fosse obbligatoriamente presente nella sua accezione negativa, vale a dire 'CAMBIO_RESIDENZA=false', nel contesto delle successive regole che saremmo andati a fare emergere. In questo caso, ottenevamo delle strutture ad albero in cui ad ogni nodo corrispondevano gli attributi scelti per lo splitting e i cui rami potevano portare o a un nodo foglia in cui si trovavano le regole trovate o ad un altro nodo in cui vi era un altro attributo deputato a un ulteriore splitting; il processo di costruzione di questo 'albero' si ferma quando non vi sono più attributi su cui effettuare lo splitting, ma nei nodi foglia si trovano solo regole.

Per esempio, se prendiamo come provincia Arezzo e come misura di filtraggio delle regole l'elift possiamo imporre che la soglia minima di elift sia pari a 3, che la parte A e B delle regole non possa contenere più di un item e che né in A né in B possa essere contenuto l'attributo 'NUOVOCONIUGE' (per le ragioni di questa scelta si veda la sezione 6.1). Se andiamo così ad analizzare le regole emerse dopo questa prima selezione notiamo che l'at-

tributo che più caratterizza queste ultime, in diversi contesti, è l'aver cambiato residenza nell'ultimo anno. A questo punto procediamo abbassando la soglia minima di elift a 1, eliminando l'attributo 'CAMBIO_RESIDENZA' dalla parte discriminante A ed imponendo il vincolo che nella parte B del contesto sia presente l'item 'CAMBIO_RESIDENZA=False'. Così facendo non otteniamo nessuna regola il cui contesto sia costituito dalle persone che nell'ultimo anno non hanno cambiato residenza. Con questo risultato, la costruzione di questo primo albero basato sullo splitting dell'attributo 'CAMBIO_RESIDENZA' termina qui poiché ogni diramazione dell'albero porta a un nodo foglia in cui si trovano o non si trovano soltanto regole.
Torniamo così da dove siamo partiti, cioè ad esaminare le regole aventi una soglia minima di elift pari a 3 e con un solo item presente nella parte discriminante e nel contesto. Notiamo quindi che un'altra caratterizzazione importante presente in questo regole è il fatto che esse coinvolgono i giovani aventi un'età inferiore o uguale ai 25 anni. Così, come avvenuto prima, procediamo abbassando la soglia minima di elift a 1, eliminando l'attributo 'ETA_DISCR=<= 25' dalla parte discriminante A ed imponendo il vincolo che nella parte B del contesto siano presenti tutte le altre fasce d'età diverse da quella appena menzionata. Così facendo otteniamo ben 70 regole la cui caratterizzazione principale è rappresentata dal fatto di non avere spese per assicurazioni. La costruzione dell'albero quindi procede eliminando l'attributo 'SPESE_ASSICURAZIONI_FLAG=0' dalla parte discriminante A ed imponendo il vincolo che nella parte del contesto B sia presente il suo opposto. In questo modo non otteniamo alcuna regola, per cui la costruzione dell'albero si può fermare qui.
Ancora una volta torniamo così al punto di partenza, cioè ad analizzare le regole aventi una soglia minima di elift pari a 3 ed un solo item nella parte discriminante A e nel contesto B. Notiamo così un'ultima caratterizzazione presente in queste regole: l'avere cioè chiesto la detrazione per gli inquili-

ni di alloggi adibiti ad abitazione principale. Così, come accaduto per la costruzione dei due precedenti alberi, procediamo abbassando la soglia minima di elift a 1, eliminando l'attributo 'DETR_INQ_ALLOGGIO_PRINC=1' dalla parte discriminante A ed imponendo il vincolo che nella parte B del contesto sia presente il suo opposto, vale a dire lo stesso attributo, però il cui valore sarà impostato a zero. Così facendo non ricaviamo alcuna regola il cui contesto sia composto dalle persone che non hanno richiesto la detrazione per gli inquilini di alloggi adibiti ad abitazione principale. Con questo risultato, la costruzione di quest'ultimo albero basato sullo splitting dell'attributo 'DETR_INQ_ALLOGGIO_PRINC' termina qui poiché ogni ramo dell'albero porta a un nodo foglia in cui si trovano o non si trovano soltanto regole.

Una volta discusse le metodologie utilizzate c'è un'ulteriore notazione da

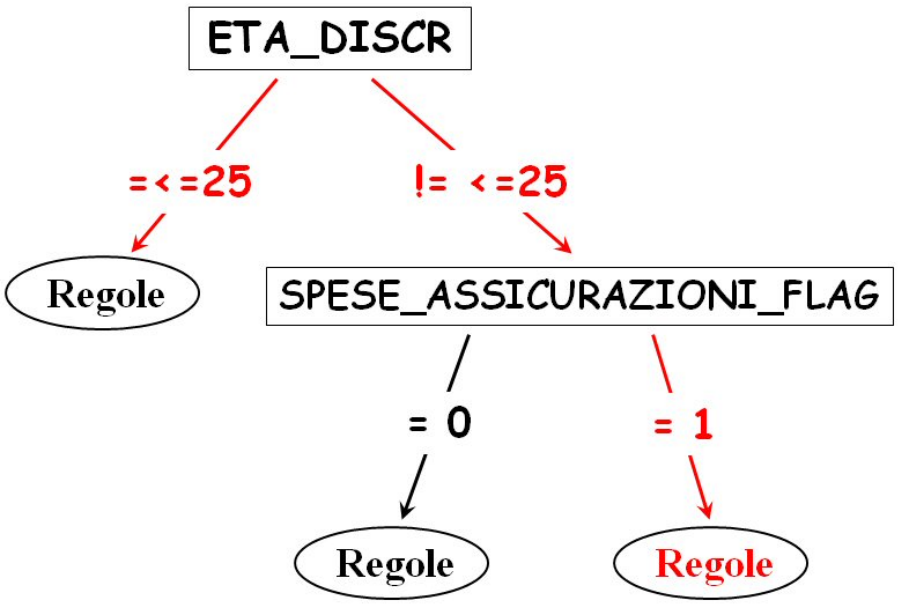


Figura 5.5: Un esempio di albero rappresentante lo splitting degli attributi

fare: durante il progressivo procedimento di selezione delle regole, la soglia di supporto massimo impostata inizialmente in DCUBE è passata dall'80%,

al 90% e, infine, al 100%. Questo è avvenuto per verificare che non perdessimo item e, conseguentemente, combinazioni di item, che potessero essere significativi.

La selezione delle regole basata sul contesto

Una volta terminata la fase precedente in cui la selezione delle regole si basava soltanto sull'utilizzo delle due misure descritte sopra, eravamo in grado di capire quali erano alcune delle caratteristiche più significative della popolazione costituita dai nuovi clienti del CAAF-CISL.
A questo punto, avendo conoscenza di tali caratteristiche e avendo due manopole sulle quali potevamo avere un'incidenza totale - vale a dire la parte discriminante delle regole e il loro contesto -, abbiamo deciso di selezionare le regole decidendo noi quale doveva essere il contesto e, successivamente, quale doveva essere la parte discriminante della regola.
La selezione basata sull'itemset di contesto avviene mettendo in *join* la tabella delle PDRULE con quella dei PNDITEMSETS e restringendosi a contesti con particolari caratteristiche, ad esempio:

- richiedendo che il contesto B contenga l'item 'STATOCIVILE=1' si ottengono delle regole supportate dai soli clienti non coniugati;
- fissando la lunghezza del contesto a 0 si ottengono regole riferite all'intera popolazione (provincia);
- fissando la lunghezza del contesto a nostra discrezione, in base al numero di item che volevamo fossero contenuti in esso;
- fissando il contesto al singoletto 'SESSO=F' si ottengono regole riferite soltanto alla popolazione femminile. La medesima operazione è stata condotta in maniera analoga per altri item, item che o rappresentavano una caratteristica significativa tra quelle emerse in precedenza

o che potevano poi essere considerati *azionabili* per future azioni di marketing.

La selezione delle regole basata sulla parte discriminante

La selezione basata sull'itemset costituente la parte discriminante della regola avviene mettendo in join la tabella delle PDRULE con quella dei PDITEMSETS e restringendosi a parti discriminanti composte o interamente costituite da particolari caratteristiche; ad esempio:

- fissando il contesto al singoletto '$ETA\_DISCR =<= 25$' si ottengono regole riferite soltanto ai giovani con un'età inferiore o uguale ai 25 anni. La stessa operazione è stata condotta in maniera analoga per altri item, item che o rappresentavano una caratteristica significativa tra quelle emerse in precedenza o che potevano poi essere considerati *azionabili* per future azioni di marketing;
- fissando la lunghezza della parte discriminante a nostra discrezione, in base al numero di item che volevamo fossero contenuti in essa;
- richiedendo che la parte discriminante A contenga soltanto l'item 'STATOCIVILE=2' si ottiene una regola supportata da clienti che sono coniugati.

Capitolo 6

La valutazione dei risultati

Una volta terminata la fase di modellazione e di selezione delle regole, i risultati sono stati valutati soggettivamente sia in funzione dell'incontro con gli esperti del dominio sia, successivamente, proprio insieme a questi ultimi. Prima dell'incontro con essi si è cercato infatti di capire quali fossero le caratteristiche maggiormente distintive per ciascuna provincia e per ciascuna misura utilizzata, in modo tale da essere già in grado di navigare con esperienza tra le regole emerse e di presentare una sintesi in qualche modo significativa di esse e in modo tale da essere in grado di rispondere alle domande posteci dagli esperti del dominio. Nel caso in cui poi gli esperti avessero identificato, a loro volta, delle caratteristiche - cioè degli attributi - interessanti, poiché azionabili, lo strumento Oracle SQL Developer era pronto per effettuare delle query in cui andare a regolare, eventualmente, le manopole del contesto e della parte discriminante delle regole.
Per ogni regola abbiamo presentato i valori della misura su cui era stata valutata e del supporto, valore che, come gli esperti del dominio ci avevano fatto capire, era comunque molto importante nell'ottica delle future azioni di marketing.
Nel presentare i risultati agli esperti del dominio si sono verificati anche dei problemi di comunicazione, problemi che non mancheremo di spiegare

più nel dettaglio; ciò che infatti a noi poteva sembrare comprensibile letto nella forma di una regola di classificazione, per loro non lo era e, solo con l'esperienza acquisita nel corso della riunione, abbiamo trovato una forma di presentazione tale da rendere chiaro sia il significato delle regole, sia le modalità di ricerca e di selezione di queste ultime.

6.1 Risultati interessanti

In questa sezione presentiamo i risultati più interessanti, divisi per provincia e per misura.

6.1.1 Arezzo

Elift

Nella prima esecuzione della query di cui abbiamo presentato un esempio in precedenza non abbiamo scartato nessun attributo e abbiamo filtrato solo gli output che presentavano un elift superiore ad una soglia minima di 5.0. Il risultato di questa selezione è stato quello di cui si è parlato nel precedente capitolo, cioè con un output totalmente modellato da questo tipo di caratterizzazione:

se entrambi i coniugi sono clienti CAAF-CISL allora entrambi sono nuovi clienti oppure nessuno dei due è nuovo.

Per questa ragione, successivamente, ci è sembrato necessario ripetere l'analisi senza considerare l'attributo 'NUOVOCONIUGE': in questo modo sarebbe stato più semplice individuare delle informazioni importanti che nello studio precedente erano state relegate in secondo piano dai dati di questo attributo. La soglia minima è stata abbassata a *3.5* ed è stata limitata la lunghezza degli itemset del contesto e della parte discriminante ad un solo item (o a zero) così da migliorare l'essenzialità e l'interpretabilità delle regole. Anche in questo secondo caso risalta una caratteristica che discrimina i nuovi

clienti in alcuni contesti molto più delle altre: l'aver cioè cambiato residenza durante l'ultimo anno.
Visualizziamo ora gli output più interessanti di questo secondo studio:

Tabella 6.1: Le regole: secondo studio.

A	B	Elift	Supporto
CAMBIO_RESIDENZA=true	STATOCIVILE=2	4.28	344
CAMBIO_RESIDENZA=true		3.77	738

Presentare le regole in questa forma non sembrava però alimentare la comunicazione tra noi e gli esperti del dominio: la veste suggeritaci dall'ambiente Oracle SQL Developer rendeva chiaro il significato delle regole a noi, ma non a loro anche, e soprattutto, per la scarsa esperienza con le 'legende' degli attributi e per il significato dei valori proposti in termini di elift e di supporto.
Ciò che sembrava essenziale, oltre a capire quello che la regola voleva esprimere, pareva essere il supporto: il capire cioè quante erano le persone per cui quella particolare caratteristica risultava valida e quanto grande era la fetta rappresentata da queste persone nell'intero composto dai nuovi clienti. Abbiamo così cercato di esporre le regole in una forma maggiormente discorsiva ricordando però la loro forma originale e rendendo ben chiari i numeri che le corredavano. Ecco, cioè, come le due regole sopra presentate possono essere riformulate guardando alle tabelle 6.2 e 6.3.

Da questo punto della presentazione in poi continueremo ad utilizzare questa forma di esposizione delle regole poiché è quella finale con cui abbiamo trovato un punto d'incontro con gli esperti del dominio.
Per completezza di analisi, abbiamo deciso di ripetere l'analisi una terza volta, questa volta togliendo anche l'attributo che rappresentava il cambio

Tabella 6.2: Arezzo: regola 1.

Regola 1			
A	**B**	**Elift**	**Supporto**
CAMBIO_RESIDENZA=true	STATOCIVILE=2	4.28	344
Descrizione			
Se consideriamo i coniugati, concentrarsi sulle persone che hanno cambiato residenza nell'ultimo anno aumenta del 328% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che si sono sposati ed hanno cambiato residenza nell'ultimo anno sono 765: di questi 344 sono nuovi clienti, con una media del 45% (cioè, 4.28 volte la media generale del 13% di nuovi clienti totali).			

Tabella 6.3: Arezzo: regola 2.

Regola 2			
A	**B**	**Elift**	**Supporto**
CAMBIO_RESIDENZA=true		3.77	738
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno cambiato residenza nell'ultimo anno aumenta del 277% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno sono 1506: di questi 738 sono nuovi clienti, con una media del 49% (cioè, 3.77 volte la media generale del 13% di nuovi clienti totali).			

di residenza nell'ultimo anno. La soglia minima di elift è stata abbassata a 2.5.

Anche in questo terzo caso risalta una caratteristica che discrimina i nuovi clienti in alcuni contesti molto più delle altre: l'avere cioè un'età inferiore o uguale ai 25 anni. Gli output più interessanti di questo terzo studio sono ancora una volta visibili nella tabella 6.4.

Arrivati a questo punto, abbiamo nuovamente deciso di ripetere l'analisi una quarta volta, questa volta eliminando anche l'attributo riguardante i giovani. La soglia minima di elift è stata abbassata a 1.5.

In questo quarto caso le caratteristiche che discriminano i nuovi clienti in diversi contesti sono due e sono rappresentate da: l'avere un'età compresa

Tabella 6.4: Arezzo: regola 3.

Regola 3			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=<= 25		3.81	363
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sui giovani con un'età inferiore o uguale ai 25 aumenta del 281% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, aventi un'età inferiore o uguale ai 25 anni sono 734: di questi 363 sono nuovi clienti, con una media del 49% (cioè, 3.81 volte la media generale del 13% di nuovi clienti totali).			

tra i 26 ed i 35 anni e l'essere single. Gli output più interessanti sono visualizzati nelle tabelle 6.5 e 6.6.

In particolare, per meglio caratterizare alcune di queste caratteristiche

Tabella 6.5: Arezzo: regola 4.

Regola 4			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35		1.85	1040
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno un'età compresa tra i 26 ed i 35 anni aumenta dell'85% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa in questa fascia sono 4328: di questi 1040 sono nuovi clienti, con una media del 24% (cioè, 1.85 volte la media generale del 13% di nuovi clienti totali).			

appena emerse, se andiamo ad analizzare le caratteristiche di coloro che hanno meno di 25 anni notiamo che:

- **sono single**: i nuovi clienti dichiaranti che hanno un'età inferiore o uguale ai 25 anni sono 363, di questi 314 sono single;
- **non sono lavoratori stabili**: i nuovi clienti dichiaranti che hanno un'età inferiore o uguale ai 25 anni sono 363, di questi 357 non sono lavoratori stabili.

Tabella 6.6: Arezzo: regola 5.

Regola 5			
A	**B**	**Elift**	**Supporto**
STATOCIVILE=1		1.72	1160
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sui single aumenta del 72% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa in questa fascia sono 5176: di questi 1160 sono nuovi clienti, con una media del 22.4% (cioè, 1.72 volte la media generale del 13% di nuovi clienti totali).			

Una volta visionati questi risultati, gli esperti del dominio hanno però espresso la richiesta di vedere come erano suddivisi i clienti in base alle varie fasce d'età. Abbiamo così fissato la parte discriminante in modo tale da avere solo le regole che contessero in essa l'età. Rispetto a tale discriminante le caratteristiche che fanno balzare di almeno 1.5 volte la probabilità di essere nuovo cliente (rispetto alla probabilità media della fascia) sono le seguenti:

1. **inferiore o uguale ai 25 anni**: esistono 363 nuovi clienti, pari al 9% dei clienti della fascia.

2. **26-35 anni**: esistono 1.040 nuovi clienti, pari al 24% dei clienti della fascia. Si veda però la tabella 6.7;

Tabella 6.7: Arezzo: regola 6.

Regola 6			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35	STATOCIVILE=2	2.25	472
Descrizione			
Se consideriamo i coniugati, concentrarsi sulle persone rientranti nella fascia di età che va dai 26 ai 35 anni aumenta del 125% la probabilità di trovare nuovi clienti. In particolare, i nuovi clienti che hanno un'età compresa in questa fascia sono 1040: di questi 472 sono coniugati, con una media del 45.4% (cioè, quasi si raddoppia la media generale del 24% di nuovi clienti aventi un'età compresa tra i 26 ed i 35 anni).			

3. **36-45 anni**: esistono 826 nuovi clienti, pari al 13.1% dei clienti della fascia: se però ci restringiamo ai clienti che non hanno fabbricati (sono 467 degli 826) allora la percentuale balza al 21.6%;

4. **46-55 anni**: esistono 483 nuovi clienti, pari al $12,1\%$ dei clienti della fascia;

5. **56-65 anni**: esistono 440 nuovi clienti, pari al 11.1% dei clienti della fascia;

6. **maggiore o uguale ai 65 anni**: esistono 822 nuovi clienti, pari al 10.7% dei clienti della fascia: se però ci restringiamo ai clienti che non hanno spese nella sez.1 (sono 467 degli 822) allora la percentuale di nuovi clienti balza al 18.1%.

4th quantifier of founded double implication

In riferimento alla seconda misura di interesse utilizzata, ovvero il 4th quantifier of founded double implication, i risultati sono stati, per così dire, più deludenti poiché meno indicativi delle caratteristiche e dei contesti specifici in cui cercare i nuovi clienti. Data una soglia minima del 50% e data la iniziale eliminazione dell'attributo 'NUOVOCONIUGE' di cui abbiamo abbondantemente discusso sopra, abbiamo trovato che un contesto ricorrente in cui potevano essere ricercati i nuovi clienti era quello comprendente le persone che avevano cambiato residenza nell'ultimo anno.
Visualizziamo di seguito nelle tabelle 6.8 e 6.9 i risultati più significativi che abbiamo ottenuto.

Per completezza di analisi, abbiamo deciso di ripetere l'analisi una seconda volta, questa volta togliendo l'attributo che rappresentava il cambio di residenza nell'ultimo anno. La soglia minima di 4th quantifier of founded double implication è stata abbassata al 50%.

Tabella 6.8: Arezzo: regola 7.

Regola 7			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
PRESENZA_SPESE_SEZ3==False	CAMBIO_RESIDENZA==True NUM_FABBRICATI-DISC=0	63%	437
Descrizione			
Se consideriamo le persone che hanno cambiato residenza nell'ultimo anno e che non possiedono fabbricati sappiamo che almeno il 63% di queste persone non ha spese nella sezione 3 ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno, che non possiedono fabbricati e che non hanno spese nella sezione 3 sono 703: di questi 437 sono nuovi clienti, appunto con una media del 63%.			

Anche in questo secondo caso abbiamo individuato tre contesti ricorrenti in cui potevano essere ricercati i nuovi clienti, questi erano rappresentati da: il non possedere fabbricati, dal non avere spese e dall'avere un'età inferiore o uguale ai 25 anni. Visualizziamo ancora una volta gli output più interessanti di questo secondo studio nelle tabelle 6.10 e 6.11.

6.1.2 Latina

Elift

Ancora una volta, anche nel caso di Latina, il nostro studio è cominciato scartando l'attributo 'NUOVOCONIUGE' e filtrando solo gli output che presentavano un elift superiore ad una soglia minima del 1.5. Molte delle combinazioni superavano tale soglia e quindi ci fornivano delle caratteristi-

Tabella 6.9: Arezzo: regola 8.

Regola 8			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
CONIUGE_A_CARICO-_FLAG=False	CAMBIO_RESIDENZA==True REDDITO_DA_LAVORO-_DISC=5000-15000	58%	318
Descrizione			
Se consideriamo le persone che hanno cambiato residenza nell'ultimo anno e che hanno un reddito compreso tra i 5000 ed i 15000 euro sappiamo che almeno il 58% di queste persone non ha il coniuge a carico ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno, che hanno un reddito compreso tra i 5000 ed i 15000 euro e che non hanno il coniuge a carico sono 509: di questi 318 sono nuovi clienti, appunto con una media del 58%.			

che che, verificate in alcuni contesti, miglioravano la percentuale di nuovi clienti di almeno il 200%.

Elementi ricorrenti tra tutti questi risultati appaiono essere le caratteristiche $ETA\_DISCR =<= 25$, ETA_DISCR=26-35, COMUNEDIRESIDENZA=FONDI e PENDOLARE=1: questo indica che sapere se un contribuente ha un'età inferiore od uguale ai 25 anni o è compreso nella fascia 26-35 o è pendolare, oppure ancora, risiede a Fondi favorisce notevolmente la ricerca di nuovi clienti in particolari contesti.

Presentiamo di seguito, nelle tabelle 6.12, 6.13, 6.14, 6.15, 6.16 e 6.17, i risultati più significativi ottenuti.

Una volta visionati questi risultati, gli esperti del dominio hanno però espresso la richiesta di vedere come erano suddivisi i clienti in base alle varie fasce d'età. Abbiamo così fissato la parte discriminante in modo tale da

Tabella 6.10: Arezzo: regola 9.

Regola 9			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
POSSIEDE_TERRENI-_FLAG=0	NUM_FABBRICATI-_DISC=0 RAP_SPESETOT-_REDDITOTOT= =NESSUNA_SPESA	55%	758
Descrizione			
Se consideriamo le persone che non possiedono fabbricati e che non hanno spese sappiamo che almeno il 55% di queste persone non possiede terreni. In particolare, i clienti, sia vecchi che nuovi, che non possiedono fabbricati, che non hanno spese e che non hanno terreni sono 1353: di questi 758 sono nuovi clienti, appunto con una media del 55%.			

avere solo le regole che contessero in essa l'età. Rispetto a tale discriminante le caratteristiche che fanno balzare di almeno 1.5 volte la probabilità di essere nuovo cliente (rispetto alla probabilità media della fascia) sono le seguenti:

1. **inferiore o uguale ai 25 anni**: esistono 472 nuovi clienti, pari al 5% dei clienti della fascia;
2. **26-35 anni**: esistono 1.570 nuovi clienti, pari al 18% dei clienti della fascia;
3. **36-45 anni**: esistono 1.601 nuovi clienti, pari al 19% dei clienti della fascia;
4. **46-55 anni**: esistono 1.193 nuovi clienti, pari al 14% dei clienti della fascia;

Tabella 6.11: Arezzo: regola 10.

Regola 10			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE-_STABILE=0	ETA_DISCR=<= 25 NUM_FABBRICATI-_DISC=0	52%	314
Descrizione			
Se consideriamo le persone che hanno un'età inferiore o uguale ai 25 anni e che non possiedono fabbricati sappiamo che almeno il 52% di queste persone non ha un lavoro stabile ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età inferiore o uguale ai 25 anni, che non possiedono fabbricati e che non sono lavoratori stabili sono 598: di questi 314 sono nuovi clienti, appunto con una media del 52%.			

5. **56-65 anni**: esistono 1.242 nuovi clienti, pari al 15% dei clienti della fascia;

6. **maggiore o uguale ai 65 anni**: esistono 2.493 nuovi clienti, pari al 29% dei clienti della fascia.

4th quantifier of founded double implication

In riferimento alla seconda misura di interesse utilizzata, ovvero il 4th quantifier of founded double implication, i risultati sono stati, ancora una volta, più deludenti poiché meno indicanti delle caratteristiche e dei contesti specifici in cui cercare i nuovi clienti. Data una soglia minima del 50% e data la iniziale eliminazione dell'attributo 'NUOVOCONIUGE', abbiamo trovato che un contesto ricorrente in cui potevano essere ricercati i nuovi clienti era quello comprendente le persone aventi un'età inferiore o uguale ai 25 anni. Visualizziamo nella tabella 6.18 i risultati più significativi che abbiamo ot-

Tabella 6.12: Latina: regola 1.

Regola 1			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=<= 25		3.15	472
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno un'età inferiore o uguale ai 25 anni aumenta del 215% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età inferiore o uguale ai 25 anni sono 782: di questi 472 sono nuovi clienti, con una media del 60% (cioè, 3.15 volte la media generale del 19% di nuovi clienti totali).			

Tabella 6.13: Latina: regola 2.

Regola 2			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35	STATOCIVILE=2	2.18	845
Descrizione			
Se consideriamo i coniugati, concentrarsi sulle persone che hanno un'età compresa tra i 26 ed i 35 anni aumenta del 118% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni e che sono coniugati sono 2309: di questi 845 sono nuovi clienti, con una media del 37% (cioè, 2.18 volte la media generale del 19% di nuovi clienti totali).			

tenuto.

Per completezza di analisi, abbiamo deciso di ripetere l'analisi una seconda volta, questa volta togliendo l'attributo che includeva i giovani.
Anche in questo secondo caso abbiamo individuato cinque contesti ricorrenti in cui potevano essere ricercati i nuovi clienti, questi erano rappresentati da: il non possedere fabbricati, dall'avere un reddito basso o assente, dall'avere un'età compresa tra i 26 ed i 35 anni, dal non avere spese e dall'essere single. Visualizziamo ancora una volta nelle tabelle 6.19, 6.20 e 6.21 gli output più interessanti di questo secondo studio.

Tabella 6.14: Latina: regola 3.

Regola 3			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35		1.86	1570
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno un'età compresa tra i 26 ed i 35 anni aumenta dell'86% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni sono 4391: di questi 1570 sono nuovi clienti, con una media del 36% (cioè, 1.86 volte la media generale del 19% di nuovi clienti totali).			

Tabella 6.15: Latina: regola 4.

Regola 4			
A	**B**	**Elift**	**Supporto**
COMUNEDIRESIDENZA=FONDI		1.8	601
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che risiedono a Fondi aumenta dell'80% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che risiedono a Fondi sono 1741: di questi 601 sono nuovi clienti, con una media del 34.5% (cioè, 1.8 volte la media generale del 19% di nuovi clienti totali).			

6.1.3 Ragusa

Elift

Ancora una volta, anche nel caso di Ragusa, il nostro studio è cominciato scartando l'attributo 'NUOVOCONIUGE' e filtrando solo gli output che presentavano un elift superiore ad una soglia minima di 2.0. Molte delle combinazioni superavano tale soglia e quindi ci fornivano delle caratteristiche che, verificate in alcuni contesti, miglioravano la percentuale di nuovi clienti di almeno il 400%.

Elementi ricorrenti tra tutti questi risultati appare essere le caratteristiche CAMBIO_RESIDENZA=True: questo indica che sapere se un contribuente

Tabella 6.16: Latina: regola 5.

Regola 5			
A	**B**	**Elift**	**Supporto**
STATOCIVILE=1		1.71	1607
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sui single aumenta del 71% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che sono single sono 4900: di questi 1607 sono nuovi clienti, con una media del 32.8% (cioè, 1.71 volte la media generale del 19% di nuovi clienti totali).			

Tabella 6.17: Latina: regola 6.

Regola 6			
A	**B**	**Elift**	**Supporto**
PENDOLARE=1		1.65	484
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sui pendolari aumenta del 65% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che sono pendolari sono 1538: di questi 484 sono nuovi clienti, con una media del 31.8% (cioè, 1.65 volte la media generale del 19% di nuovi clienti totali).			

ha cambiato residenza nell'ultimo favorisce notevolmente la ricerca di nuovi clienti in particolari contesti.
Nelle tabelle 6.22 e 6.23 presentiamo i risultati più significativi ottenuti.

Per completezza di analisi, abbiamo deciso di ripetere l'analisi una seconda volta, questa volta togliendo l'attributo indicante il cambio di residenza. Anche in questo secondo caso abbiamo individuato tre contesti ricorrenti in cui potevano essere ricercati i nuovi clienti, questi erano rappresentati da: l'avere figli con un'età inferiore ai 3 anni, l'avere 3 figli e dall'avere un'età compresa tra i 26 ed i 35 anni. La soglia minima di elift è stata abbassata a 1.5. Visualizziamo ancora una volta nelle tabelle 6.24, 6.25 e 6.26 gli output più interessanti di questo secondo studio.

Tabella 6.18: Latina: regola 7.

Regola 7			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE_STABILE=0	ETA_DISCR=<= 25	59%	452
Descrizione			
Se consideriamo le persone che hanno un'età inferiore o uguale ai 25 anni sappiamo che almeno il 59% di queste persone non ha un lavoro stabile ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età inferiore o uguale ai 25 anni e che non hanno un lavoro stabile sono 746: di questi 452 sono nuovi clienti, appunto con una media del 59%.			

In particolare, per meglio caratterizzare questo segmento, se andiamo ad analizzare le caratteristiche di coloro che hanno figli piccoli o hanno tre figli a carico notiamo che:

- **sono perlopiù uomini**: i nuovi clienti dichiaranti che hanno figli piccoli sono 473, di questi 399 sono uomini. I nuovi clienti dichiaranti che hanno tre figli a carico sono 350, di questi 303 sono uomini;

- **sono coniugati**: i nuovi clienti dichiaranti che hanno figli piccoli sono 473, di questi 448 sono coniugati. I nuovi clienti dichiaranti che hanno tre figli a carico sono 350, di questi 339 sono coniugati;

- **hanno il coniuge a carico**: i nuovi clienti dichiaranti che hanno figli piccoli sono 473, di questi 325 hanno il coniuge a carico. I nuovi clienti dichiaranti che hanno tre figli a carico sono 350, di questi 281 hanno il coniuge a carico.

Tabella 6.19: Latina: regola 8.

Regola 8			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE_STABILE=0	NUM_FABBRICATI-_DISC=0 REDDITO_DA-_LAVORO_DISC= =BASSO/ASSENTE	57%	530
Descrizione			
Se consideriamo le persone che non possiedono fabbricati e che hanno un reddito da lavoro molto basso o totalmente assente sappiamo che almeno il 57% di queste persone non ha un lavoro stabile ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che non possiedono fabbricati, che hanno un reddito da lavoro molto basso o totalmente assente e che non hanno un lavoro stabile sono 887: di questi 530 sono nuovi clienti, appunto con una media del 57%.			

Ripetendo la medesima operazione per le parti discriminanti riguardanti il cambio di residenza e l'avere un'età compresa tra i 26 ed i 35 anni abbiamo rispettivamente che.
Coloro che hanno cambiato residenza:

- **sono perlopiù uomini**: i nuovi clienti dichiaranti che hanno cambiato residenza nell'ultimo anno sono 951, di questi 641 sono uomini;
- **sono perlopiù coniugati**: i nuovi clienti dichiaranti che hanno cambiato residenza nell'ultimo anno sono 951, di questi 699 sono coniugati;
- **non sono lavoratori stabili**: i nuovi clienti dichiaranti che hanno cambiato residenza nell'ultimo anno sono 951, di questi 786 non sono lavoratori stabili;

Tabella 6.20: Latina: regola 9.

Regola 9			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE-_STABILE=0	ETA_DISCR=26-35 RAP_SPESETOT-_REDDITOTOT= =NESSUNA_SPESA	54%	520
Descrizione			
Se consideriamo le persone che hanno un'età compresa tra i 26 ed i 35 e che non hanno spese sappiamo che almeno il 54% di queste persone non ha un lavoro stabile ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni, che non hanno spese e che non hanno un lavoro stabile sono 962: di questi 520 sono nuovi clienti, appunto con una media del 54%.			

- **sono divisi perlopiù uniformemente per fasce d'età**: i nuovi clienti dichiaranti che hanno cambiato residenza nell'ultimo anno sono 951, di questi 59 hanno un'età inferiore o uguale ai 25 anni, 230 hanno tra i 26 ed i 35 anni, 277 hanno tra i 36 ed i 45 anni, 161 hanno tra i 46 ed i 55 anni, 76 hanno tra i 56 ed i 65 anni e 148 hanno più di 65 anni.

Coloro che hanno tra i 26 ed i 35 anni:

- **sono perlopiù uomini**: i nuovi clienti dichiaranti che hanno tra i 26 ed i 35 anni sono 632, di questi 415 sono uomini;
- **sono perlopiù coniugati**: i nuovi clienti dichiaranti che hanno tra i 26 ed i 35 anni sono 632, di questi 425 sono coniugati;
- **non sono lavoratori stabili**: i nuovi clienti dichiaranti che hanno tra i 26 ed i 35 anni sono 632, di questi 630 non sono lavoratori stabili.

Tabella 6.21: Latina: regola 10.

Regola 10			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
NUM_FABBRICATI-_DISC=0	STATOCIVILE=1 PRESENZA_SPESE-_SEZ1=False	51%	614
Descrizione			
Se consideriamo i single e le persone che non hanno spese nella sezione 1 sappiamo che almeno il 51% di queste persone non possiede fabbricati ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che sono single, che non hanno spese nella sezione 1 e che non hanno fabbricati sono 1093: di questi 614 sono nuovi clienti, appunto con una media del 51%.			

Una volta visionati questi risultati, gli esperti del dominio hanno però espresso la richiesta di vedere come erano suddivisi i clienti in base alle varie fasce d'età. Abbiamo così fissato la parte discriminante in modo tale da avere solo le regole che contessero in essa l'età. Rispetto a tale discriminante i numeri e le caratteristiche che fanno balzare di almeno 1.5 volte la probabilità di essere nuovo cliente (rispetto alla probabilità media della fascia) sono i seguenti:

1. **inferiore o uguale ai 25 anni**: esistono 123 nuovi clienti, pari al 4% dei clienti della fascia;

2. **26-35 anni**: esistono 632 nuovi clienti, pari al 18% dei clienti della fascia;

3. **36-45 anni**: esistono 745 nuovi clienti, pari al 22% dei clienti della fascia; se però ci restringiamo ai nuovi clienti che non hanno spese

Tabella 6.22: Ragusa: regola 1.

Regola 1			
A	**B**	**Elift**	**Supporto**
CAMBIO_RESIDENZA=True		4.99	951
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno cambiato residenza nell'ultimo anno aumenta del 399% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno sono 1200: di questi 951 sono nuovi clienti, con una media del 79% (cioè, 4.99 volte la media generale del 15.9% di nuovi clienti totali).			

Tabella 6.23: Ragusa: regola 2.

Regola 2			
A	**B**	**Elift**	**Supporto**
CAMBIO_RESIDENZA= =True	CONIUGE_A- _CARICO_FLAG=True	4.2	466
Descrizione			
Se consideriamo le persone che hanno il coniuge a carico, concentrarsi sulle persone che hanno cambiato residenza aumenta del 320% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza ed hanno il coniuge a carico sono 533: di questi 466 sono nuovi clienti, con una media dell'82% (cioè, 4.2 volte la media generale del 15.9% di nuovi clienti totali).			

nella sez.1 (sono 538 dei 745) allora la percentuale balza al 41.6% e ai nuovi clienti che hanno reddito nella fascia che vai dai 5000 ai 15000 euro (sono 429 dei 745) allora la percentuale balza al 36.9%;

4. **46-55 anni**: esistono 547 nuovi clienti, pari al 16% dei clienti della fascia;

5. **56-65 anni**: esistono 506 nuovi clienti, pari al 15% dei clienti della fascia;

6. **maggiore o uguale ai 65 anni**: esistono 906 nuovi clienti, pari al 26% dei clienti della fascia.

Tabella 6.24: Ragusa: regola 3.

Regola 3			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35		2.68	632
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno un'età compresa tra i 26 ed i 35 anni aumenta del 168% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni sono 1484: di questi 632 sono nuovi clienti, con una media del 43% (cioè, 2.68 volte la media generale del 15.9% di nuovi clienti totali).			

Tabella 6.25: Ragusa: regola 4.

Regola 4			
A	**B**	**Elift**	**Supporto**
FIGLI_MINORI_3ANNI_FLAG=1		2.37	473
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno figli piccoli (con un'età inferiore o uguale ai 3 anni) aumenta del 137% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno figli piccoli sono 1258: di questi 473 sono nuovi clienti, con una media del 38% (cioè, 2.37 volte la media generale del 15.9% di nuovi clienti totali).			

4th quantifier of founded double implication

Data una soglia minima del 50% e data la iniziale eliminazione dell'attributo 'NUOVOCONIUGE', abbiamo individuato tre contesti ricorrenti in cui potevano essere ricercati i nuovi clienti: questi sono quello comprendente le persone che hanno cambiato residenza nell'ultimo anno, quello comprendente le persone che hanno tre figli a carico e, infine, quello coinvolgente le persone che risiedono nel comune di Vittoria.
Visualizziamo nelle tabelle 6.27, 6.28, 6.29 e 6.30 i risultati più significativi che abbiamo ottenuto.

Tabella 6.26: Ragusa: regola 5.

Regola 5			
A	**B**	**Elift**	**Supporto**
NUM_FIGLI_CARICO_DISC=3		2.32	350
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno tre figli a carico aumenta del 132% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno tre figli a carico sono 950: di questi 350 sono nuovi clienti, con una media del 37% (cioè, 2.32 volte la media generale del 15.9% di nuovi clienti totali).			

6.1.4 Trento

Elift

Ancora una volta, come per le tre precedenti province, anche nel caso di Trento, la nostra analisi è cominciata scartando l'attributo 'NUOVOCONIUGE' e filtrando solo gli output che presentavano un elift superiore ad una soglia minima di 2.0. Molte delle combinazioni superavano tale soglia e quindi ci fornivano delle caratteristiche che, verificate in alcuni contesti, miglioravano la percentuale di nuovi clienti di almeno il 230%.
Elementi ricorrenti tra tutti questi risultati appaiono essere le caratteristiche ETA_DISCR=<= 25, DETR_INQ_ALLOGGIO_PRINC=1 e RAP_SPESE TOT_REDDITOTOT=NESSUNA_SPESA: questo indica che sapere se un contribuente ha un'età inferiore o uguale ai 25 anni, se un cliente ha chiesto la detrazione '*inquilino di alloggio principale*' o, ancora, se il cliente in questione non ha alcuna spesa, ciò favorisce notevolmente la ricerca di nuovi clienti in particolari contesti.
Presentiamo nelle tabelle 6.31, 6.32 e 6.33 i risultati più significativi ottenuti.

Successivamente, ci è sembrato interessante ripetere l'analisi senza con-

Tabella 6.27: Ragusa: regola 6.

Regola 6			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE-_STABILE=0	CAMBIO-_RESIDENZA=True	70%	786
Descrizione			
Se consideriamo le persone che hanno cambiato residenza nell'ultimo anno sappiamo che almeno il 70% di queste persone non ha un lavoro stabile ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno e che non sono lavoratori stabili sono 969: di questi 786 sono nuovi clienti, appunto con una media del 70%.			

siderare gli attributi relativi all'età inferiore o uguale ai 25 anni, alla detrazione inquilino di alloggio principale e alle spese; in questo modo sarebbe stato più semplice individuare delle informazioni importanti che, però, nello studio precedente, sono state relegate in secondo piano dai dati di questi attributi. La nuova soglia minima di elift scelta è stata 1.5.
Tuttavia, anche in questo secondo caso risaltano due caratteristiche che discriminano i nuovi clienti in alcuni contesti molto più delle altre: l'avere cioè un'età compresa tra i 26 ed i 35 anni e l'essere single.
Visualizziamo nelle tabelle 6.34 e 6.35 gli output più interessanti di questo secondo studio.

Una volta visionati questi risultati, gli esperti del dominio hanno però espresso la richiesta di vedere come erano suddivisi i clienti in base alle varie fasce d'età. Abbiamo così fissato la parte discriminante in modo tale da avere solo le regole che contessero in essa, appunto, l'età. Rispetto a tale discriminante i numeri e le caratteristiche che fanno balzare di almeno 1.5

Tabella 6.28: Ragusa: regola 7.

A	B	4th Quantifier of founded double implication	Supporto
STATOCIVILE=2	CAMBIO_RESIDENZA=True	62%	699
Descrizione			
Se consideriamo le persone che hanno cambiato residenza nell'ultimo anno sappiamo che almeno il 62% di queste persone è coniugato ed è nuovo cliente. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno e che sono coniugati sono 870: di questi 699 sono nuovi clienti, appunto con una media del 62%.			

volte la probabilità di essere nuovo cliente (rispetto alla probabilità media della fascia) sono i seguenti:

1. **26-35 anni**: esistono 1338 nuovi clienti, pari al 26.9% dei clienti della fascia; se però ci restringiamo ai nuovi clienti che non hanno spese nella sez.1 (sono 417 dei 1338) allora la percentuale balza al 48.2%;

2. **36-45 anni**: esistono 1325 nuovi clienti, pari al 15.9%.

4th quantifier of founded double implication

Data una soglia minima del 50% e data la iniziale eliminazione dell'attributo 'NUOVOCONIUGE', abbiamo individuato quattro contesti ricorrenti in cui potevano essere ricercati i nuovi clienti: questi sono quello comprendente le persone che hanno un'età inferiore o uguale ai 25 anni, quello comprendente le persone che hanno un'età compresa tra i 26 ed i 35 anni e, infine, quello coinvolgente i single.
Visualizziamo di seguito i risultati più significativi che abbiamo ottenuto.

Tabella 6.29: Ragusa: regola 8.

Regola 8			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
CAMBIO-_RESIDENZA=True	COMUNEDIRESIDENZA==VITTORIA	60%	384
Descrizione			
Se consideriamo le persone che risiedono a Vittoria sappiamo che almeno il 60% di queste persone ha cambiato residenza nell'ultimo anno. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno e che risiedono a Vittoria sono 418: di questi 384 sono nuovi clienti, appunto con una media del 60%.			

6.2 Interpretando i risultati

Prima dell'incontro con gli esperti del dominio dei quattro CAAF-CISL locali, uno sguardo onnicomprensivo rispetto alle quattro province esaminate e alle regole selezionate ci fa notare come la grande maggioranza dei nuovi clienti si trovi nella parte più giovane della popolazione. In particolare, come segmenti più importanti si distinguono i ragazzi aventi un'età inferiore o uguale ai 25 anni e quelli aventi un'età compresa tra i 26 ed i 35 anni: sono i ragazzi alla loro prima esperienza lavorativa, che magari si sono appena allontanati da casa e cercano una certa indipendenza dalla famiglia, che possono essersi sposati di recente ed hanno appena iniziato ad avere una loro famiglia.

Quindi, l'informazione forse più interessante che deriviamo da queste regole è che i nuovi clienti sono soprattutto persone che nell'anno precedente hanno cambiato notevolmente il loro stile di vita: parliamo sia di giovani al loro primo impiego lavorativo ma anche di single (celibi, divorziati o separati) o di persone sposate da poco che cambiano casa e iniziano una nuova vita.

Tabella 6.30: Ragusa: regola 9.

Regola 9			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
TIPO_SOSTITUTO- _IMPOSTA= = AMMINISTRAZIONE PUBBLICA E DIFESA - ASSICURAZIONE SOCIALE OBBLIGATORIA	NUM_FIGLI_- CARICO_DISC=3 52%		274
Descrizione			
Se consideriamo le persone che hanno tre figli a carico sappiamo che almeno il 52% di queste persone ha indicato come sostituto di imposta 'AMMINISTRAZIONE PUBBLICA E DIFESA - ASSICURAZIONE SOCIALE OBBLIGATORIA'. In particolare, i clienti, sia vecchi che nuovi, che hanno indicato questo come sostituto di imposta e che hanno tre figli a carico sono 452: di questi 274 sono nuovi clienti, appunto con una media del 52%.			

Tra queste persone è molto diffuso il cambio di residenza recente, un lavoro non stabile ed un reddito da lavoro basso (fino a 15000 euro): questo fa pensare che siano alla prima esperienza di lavoro in assoluto o comunque in inattività da alcuni anni.

Il cambio di residenza appare quindi in generale molto discriminante per la compilazione del modello 730. Questa tendenza è forte sia per i giovani (alla loro prima esperienza lontano da casa) che per le persone leggermente più mature e sposate: incentrare delle politiche di marketing verso i cittadini che sono in procinto di cambiare casa potrebbe quindi dare risultati molto positivi.

A seguito dell'incontro con gli esperti del dominio delle quattro province,

Tabella 6.31: Trento: regola 1.

Regola 1			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=<= 25		2.98	586
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno un'età inferiore o uguale ai 25 anni aumenta del 198% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età inferiore o uguale ai 25 anni sono 1265: di questi 586 sono nuovi clienti, con una media del 46% (cioè, 2.98 volte la media generale del 15.5% di nuovi clienti totali).			

Tabella 6.32: Trento: regola 2.

Regola 2			
A	**B**	**Elift**	**Supporto**
DETR_INQ_ALLOGGIO_PRINC=1		2.59	594
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che hanno chiesto la detrazione '*inquilino di alloggio principale*' aumenta del 159% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno chiesto la detrazione '*inquilino di alloggio principale*' sono 1476: di questi 594 sono nuovi clienti, con una media del 40% (cioè, 2.59 volte la media generale del 15.5% di nuovi clienti totali).			

scopriamo invece che per due di esse, Latina e Ragusa, esistono risposte differenti rispetto a quelle sopra esposte riguardanti le caratterizzazioni che le regole selezionate lasciano emergere.

Per quanto riguarda la provincia di Latina, infatti, il fatto che siano presenti regole che riguardano il comune di Fondi è spiegabile avendo conoscenza del fatto che il comune di Latina è molto 'forte' in quei comuni che si trovano al confine del territorio della provincia, proprio come il comune in questione. L'affermazione appena fatta è validata da una seconda regola: quella riguardante i molti pendolari che, guardando alla popolazione totale della provincia di Latina, sono nuovi clienti. Questo accade perché molti clienti

Tabella 6.33: Trento: regola 3.

Regola 3			
A	**B**	**Elift**	**Supporto**
RAP_SPESETOT_REDDITOTOT= =NESSUNA_SPESA		2.11	1294
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che non hanno spese aumenta del 111% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che non hanno spese sono 3950: di questi 1294 sono nuovi clienti, con una media del 33% (cioè, 2.11 volte la media generale del 15.5% di nuovi clienti totali).			

Tabella 6.34: Trento: regola 4.

Regola 4			
A	**B**	**Elift**	**Supporto**
ETA_DISCR=26-35		1.73	1338
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone hanno un'età compresa tra i 26 ed i 35 anni aumenta del 73% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni sono 4969: di questi 1338 sono nuovi clienti, con una media del 27% (cioè, 1.73 volte la media generale del 15.5% di nuovi clienti totali).			

pendolari, non residenti nella provincia di Latina, per questioni, forse, di comodità, preferiscono usufruire dei servizi di un CAAF-CISL situato invece proprio sul territorio della provincia di Latina.

Per quanto riguarda invece la provincia di Ragusa, il gran numero di persone che sembrerebbero avere come tratto distintivo il fatto di aver cambiato residenza non sono in realtà persone residenti a Ragusa che hanno cambiato residenza sempre all'interno dello stesso Comune: sono infatti immigrati che l'anno scorso non erano proprio residenti nel territorio italiano. Questi nuovi clienti sono stati 'catturati' da promozioni specifiche a loro rivolte e i risultati fatti emergere dalle regole dimostrano il grande successo avuto

Tabella 6.35: Trento: regola 5.

Regola 5			
A	**B**	**Elift**	**Supporto**
STATOCIVILE=1	REDDITO_DA_LAVORO_DISC=		
STATOCIVILE=1	=5000-15000	1.65	578
Descrizione			
Se consideriamo gli aventi un reddito compreso tra i 5000 ed i 15000 euro, concentrarsi sui single aumenta del 65% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che hanno un reddito compreso tra i 5000 ed i 15000 euro e sono single sono 1681: di questi 578 sono nuovi clienti, con una media del 34% (cioè, 1.65 volte la media generale del 15.5% di nuovi clienti totali).			

dalle scorse campagne di marketing.

L'incontro con gli esperti del dominio specifici delle quattro province ci ha quindi permesso di meglio comprendere i risultati che le regole lasciavano emergere; quello che noi potevamo fare, infatti, era solo cercare di interpretare questi risultati con l'ausilio della nostra intelligenza e con un po' di senso pratico. Non potevamo, invece, essere a conoscenza delle campagne promozionali messe in atto nel corso degli anni per attirare nuovi clienti, dei bonus di cui questi stessi clienti potevano usufruire e delle convenzioni che potevano essere state messe in atto, per esempio, con determinate aziende.

Tabella 6.36: Trento: regola 7.

Regola 7			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
LAVORATORE-_STABILE=0	ETA_DISCR=<= 25 PRESENZA_SPESE_SEZ2= =False	51%	463
Descrizione			
Se consideriamo le persone che hanno un'età inferiore o uguale ai 25 anni e che non hanno spese nella sez.2 sappiamo che almeno il 51% di queste persone non ha un lavoro stabile. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età inferiore o uguale ai 25 anni, che non hanno spese nella sez.2 e che non sono lavoratori stabili sono 898: di questi 463 sono nuovi clienti, appunto con una media del 51%.			

Tabella 6.37: Trento: regola 8.

Regola 8			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
NUM_FABBRICATI-_DISC=0	ETA_DISCR=26-35 PRESENZA_SPESE_SEZ1= =False	50%	372
Descrizione			
Se consideriamo le persone che hanno un'età compresa tra i 26 ed i 35 e che non hanno spese nella sez.1 sappiamo che almeno il 50% di queste persone non possiede fabbricati. In particolare, i clienti, sia vecchi che nuovi, che hanno un'età compresa tra i 26 ed i 35 anni, che non hanno spese nella sez.1 e che non posseggono fabbricati sono 696: di questi 372 sono nuovi clienti, appunto con una media del 50%.			

Tabella 6.38: Trento: regola 9.

Regola 9			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
NUM_FABBRICATI-_DISC=0	STATOCIVILE=1 RAP_SPESETOT-_REDDITOTOT= =NESSUNA_SPESA	52%	354
Descrizione			
Se consideriamo le persone non coniugate e che non hanno spese sappiamo che almeno il 52% di queste persone non possiede fabbricati. In particolare, i clienti, sia vecchi che nuovi, non coniugati, che non hanno spese e che non posseggono fabbricati sono 612: di questi 354 sono nuovi clienti, appunto con una media del 52%.			

Capitolo 7

L'utilizzo dei risultati: i media per l'advertising.

Nonostante il nostro studio si sia arrestato alla fase relativa alla valutazione dei risultati, questi ultimi saranno comunque destinati ad essere impiegati per realizzare delle campagne pubblicitarie coerenti con quanto emerso.

Comunque, il fatto di non sapere come le regole selezionate saranno poi concretamente utilizzate dagli esperti di marketing non ci impedisce di fare una carrelata relativa ai media pubblicitari esistenti e di fornire degli esempi tratti dalla nostra analisi relativi al possibile impiego di questi ultimi. Nella presentazione futura è comunque bene tenere conto del fatto che gli esempi di applicazione proposti saranno limitati ai media presentati nella sezione del media mix, con l'esclusione di Internet, poiché gli esperti del dominio ci hanno già riferito che le decisioni su cui si baseranno le future campagne pubblicitarie saranno prese guardando soltanto ai media lì descritti.

Ora vediamo quindi qual è lo stato dell'arte dello specifico settore del marketing relativo ai media per l'advertising.

7.1 I media per l'advertising: stato dell'arte

Della parola *marketing* non è mai esistita una traduzione accreditata in lingua italiana. Secondo Kotler il concetto di *marketing* fa riferimento alla *superiore capacità dell'impresa rispetto ai concorrenti di raggiungere i suoi obiettivi attraverso lo sviluppo, il trasferimento e la comunicazione di valore al suo mercato obiettivo* [43]. Ponendo soprattutto enfasi sull'idea di valore per il cliente e sulla creazione di una strategia di mercato idonea a perseguirlo, il *marketing concept* (orientamento al cliente) viene a contrapporsi ad altri orientamenti gestionali quali quello del *product concept* (orientamento al prodotto) o del *selling concept* (orientamento alle vendite). A differenza degli altri approcci gestionali, la teoria di *marketing* pone alla base della creazione e del sostenimento del vantaggio competitivo il cliente e propone una strategia finalizzata alla comprensione dei suoi bisogni e alla realizzazione di un sistema di offerte coerente con gli stessi.

7.2 Dal *marketing concept* al *marketing management*

Se quindi il *marketing concept* porta la comprensione dei bisogni del cliente al centro della strategia aziendale, la realizzazione del concetto di *marketing* è propria del *marketing management*. Quest'ultimo consta di tre processi aziendali:

1. l'analisi del mercato;
2. la formulazione delle strategie;
3. l'implementazione delle decisioni operative.

Per creare valore per i clienti è innanzitutto necessario operare un'attenta analisi dei loro desideri, bisogni e preferenze. A questo scopo, per prima cosa,

è necessario stabilire le finalità dell'analisi che si intende intraprendere; se cioè quest'ultima è:

- un'analisi di tipo *esplorativo*: laddove la conoscenza del problema è particolarmente ridotta e l'obiettivo è quello di una sua più precisa identificazione;
- un'analisi di tipo *descrittivo*: laddove esiste già una conoscenza iniziale, sia pur generale del problema e dei fattori che possono contribuire alla sua soluzione ma si ritiene necessario approfondirne le caratteristiche;
- un'analisi di tipo *causale*: quando la conoscenza del problema è già approfondita e l'obiettivo è quello di misurare l'esatta relazione tra il problema e le sue cause e, quindi, comprendere a fondo come intervenire su specifiche variabili per ottenere i risultati desiderati.

Solo a questo punto può essere individuata con chiarezza la fonte più appropriata per il reperimento delle informazioni cercate. Nella fattispecie, i dati possono essere raccolti direttamente dall'operatore che ha maturato una specifica esigenza analitica, attraverso indaginis volte ad hoc (*dati primari*), oppure possono essere già disponibili (*dati secondari*), in quanto raccolti nell'ambito di analisi precedentemente sviluppate, accessibili in alcuni casi a titolo gratuito, in altri esclusivamente a pagamento. Il limite dei dati secondari è quello di essere stati generati per scopi spesso almeno in parte diversi da quelli che muovono il ricercatore ad approfondire uno specifico problema di *marketing* e, quindi, l'opportunità di un loro integrale utilizzo deve essere di volta in volta valutata con estrema cautela.
Successivamente alla comprensione analitica della situazione complessiva di mercato, è fondamentale accorpare le informazioni in un piano strategico dettagliato, che risulti interpretativo della situazione e delle azioni da affrontare. Il *marketing strategico* è cadenzato in tre fasi:

1. la segmentazione;

2. il *targeting*;

3. il *positioning*.

Il punto di partenza è rappresentato dalla segmentazione, intesa come l'accorpamento dei clienti in gruppi relativamente omogenei sotto il profilo dei benefici ed attributi da essi ricercati nei processi di acquisto e consumo.
In seguito alla segmentazione del mercato, la strategia del *marketing* richiede di operare il *targeting*, definito come l'identificazione del segmento di clienti-obiettivo che si vuole soddisfare mediante la strategia di mercato. Il *targeting* segue abitualmente i concetti di omogeneità, stabilità, economicità e raggiungibilità.
L'ultimo passo riguarda il *positioning*, cioè il posizionamento, ovvero la formulazione di un sistema di offerta coerente con il segmento-obiettivo e differenze da quanto proposto dai concorrenti che competono per lo stesso segmento.
Una volta pianificato il sistema di offerta, l'impresa può passare alla fase di implementazione, selezionando le variabili operative che consentono la realizzazione della strategia di *marketing*. Come evidenziato da McCarthy [44], le variabili in questione possono essere denominate le *quattro P*: *product* (prodotto), *price* (prezzo), *place* (distribuzione) e *promotion* (promozione). Tali variabili rientrano nel *marketing mix*, poiché devono comporre insieme una miscela esplosiva per realizzare il concetto di *marketing* e avere di conseguenza una maggiore probabilità di conquistare il mercato attraverso un livello crescente di soddisfazione del cliente.

7.3 Il mix di attività promozionali: il *promotion mix*

Il concetto di *promotion mix* si riferisce alla *combinazione integrata, sinergica e coerente di diversi tipi di comunicazione personale e non personale messa in essere dall'azienda durante un determinato periodo di tempo* [45]. Il *promotion mix* consta di cinque elementi. Quattro - pubblicità, promozione delle vendite, pubbliche relazioni, *direct marketing* - sono forme di comunicazione non personale; una - la vendita personale - è una forma di comunicazione personale. Vediamole sinteticamente una per una.

1. La *pubblicità* (*advertising*) è una forma di comunicazione non personale soggetta ad un pagamento. Verte sull'azienda, i suoi prodotti, o le sue attività e viene trasmessa mediante un mezzo di massa ad un pubblico di riferimento. Essa permette di raggiungere molti compratori simultaneamente, è efficace per creare l'immagine della merca e vi è una grande varietà di media tra i quali scegliere. Allo stesso tempo però raggiunge molte persone che non sono potenziali acquirenti, viene spesso evitata dalla gente ed il costo complessivo può essere elevato;

2. la *promozione delle vendite* (*sales promotion*) è un'attività o un mezzo materiale che offre alla clientela, agli addetti alla vendita o ai rivenditori un incentivo diretto all'acquisto di un prodotto. Tale stimolo può assumere la forma di buoni sconto, estrazioni a premi, rimborsi, espositori;

3. con *pubbliche relazioni* (*public relations*) si intende quella tipologia di comunicazione non personale che mira ad influenzare gli atteggiamenti e le opinioni di clienti e non-clienti nei confronti dell'azienda. Un genere molto in uso è la *publicity*, che è una forma non-pagata di comu-

nicazione non personale sull'organizzazione e i suoi prodotti, veicolata attraverso un mezzo di massa sotto forma di testi informativi;

4. il *marketing diretto* (*direct marketing*) si avvale di comunicazioni dirette con i clienti. Può assumere la forma di invio di posta diretto (*direct mail*), di *marketing telefonico* o *via web* (*e-mail marketing*), di catalogo, di *telemarketing* e di pubblicità a risposta diretta. Similmente alla vendita personale, si configura come dialogo interattivo fra l'operatore di *marketing* ed il cliente. Tra i vantaggi di questa forma di comunicazione registriamo il fatto che il messaggio può essere concepito e personalizzato rapidamente ed il fatto che può facilitare la relazione con il cliente; allo stesso tempo, però, il mantenimento e la gestione di database aggiornati possono essere dispendiosi e spesso il tasso di risposta del cliente può essere limitato;

5. la *vendita personale* (*personal selling*) è una comunicazione faccia a faccia con i potenziali clienti, con lo scopo di informarli sul prodotto di un'azienda a persuaderli all'acquisto.

Naturalmente gli operatori di *marketing* si adoperano per definire la corretta combinazione degli elementi della promozione, il giusto mix, al fine di assicurarsi che il loro prodotto trovi buona accoglienza. Il mix di attività promozionali ovviamente subirà aggiustamenti e modifiche continui per adattarsi ai cambiamenti nel mercato e all'adozione di nuove strategie. Approntando il proprio mix di attività promozionali, l'azienda dovrebbe prendere in considerazione tre fattori cruciali: (1) il ruolo della promozione nel *marketing mix* complessivo; (2) la natura del prodotto e (3) la natura del mercato.

7.4 La pubblicità

La pubblicità punta a promuovere il prodotto dell'azienda mediante messaggi - su carta ed elettronici - diramati dai mezzi di comunicazione. La considerazione alla base della pubblicità è che i messaggi hanno la capacità di raggiungere un ampio numero di persone e di renderle consapevoli, persuaderle, e informarle dell'offerta dell'azienda.

Gli obiettivi propri della pubblicità vertono quindi sul creare la *consapevolezza*, facilitare la *comprensione*, determinare la *convinzione*, e sollecitare l'*ordine*.

Il processo pubblicitario parte dalla creazione di messaggi rivolti a un target sotto forma di parole, idee, suoni e altre forme di stimoli audiovisivi, concepiti per influenzare il comportamento del consumatore cui ci si rivolge. Ciò significa che buona parte dell'attività pubblicitaria prevede la gestione di un processo di comunicazione. Per essere efficace, il messaggio dovrebbe quindi rispettare due principi generali: (1) tenere in considerazione i principi base della comunicazione; (2) basarsi su una valida teoria del comportamento del consumatore.

Il processo base di comunicazione è composto da tre elementi: (1) l'emittente o fonte della comunicazione; (2) la comunicazione o messaggio; (3) il ricevente o *audience*. L'attività di conversione del messaggio di *marketing* in una comunicazione efficace è chiamata *codifica*.

I messaggi pubblicitari devono essere diffusi tramite canali appositi, noti come mezzi o veicoli pubblicitari (*advertising media*). Tali mezzi differiscono tra loro, in generale, per efficacia, selettività e costo. Alcuni sono preferiti ad altri grazie alla minor presenza di 'rumori', così il messaggio viene recepito e compreso più facilmente. Per esempio, una specifica pubblicità veicolata su un quotidiano dovrà competere con altri messaggi pubblicitari, illustrazioni o testi presenti nella stessa pagina. Nel caso della radio e della televisione possono essere presenti altri elementi di distrazione, come nel caso in cui si

stia guidando e al contempo si ascolta la radio.
Il rapporto esistente tra la pubblicità ed il comportamento del consumatore è abbastanza intuitivo. Per molti prodotti o servizi la pubblicità può influenzare la scelta finale del consumatore ed indurlo ad acquistare un prodotto. È però chiaro che i consumatori sono soggetti a vari tipi di influenza, e dunque si crea il problema di definire quanto sia importante - o quanto possa essere importante - la pubblicità. È a questo punto che l'esperto in attività pubblicitarie dovrebbe apportare il proprio contributo, in quanto conoscitore dei principi teorici del comportamento d'acquisto del consumatore.
La corretta pianificazione di una campagna pubblicitaria, e quindi la creazione di messaggi che risultino alla fine persuasivi, richiede nozioni di *marketing* e conoscenze teoriche di creatività. Per quanto riguarda le nozioni di *marketing*, prima di lanciare una campagna pubblicitaria è necessario disporre di alcune importanti informazioni:

1. *chi* sono i clienti attuali e potenziali dell'azienda;
2. *quali* sono le loro caratteristiche demografiche, economiche e psicologiche;
3. *quanti* sono i clienti attuali e potenziali;
4. *quali* individui - parte i clienti attuali e potenziali - *influenzano* la decisione d'acquisto.

7.5 Il mix di mezzi pubblicitari (*media mix*)

La selezione dei media da utilizzare non è un compito facile. Vi sono numerosi tipi, e soprattutto numerose combinazioni, di mezzi pubblicitari fra i quali scegliere. I principali sono elencati qui di seguito.
Nel settore della pubblicità il metodo più comunemente usato per la valutazione dell'efficienza o della produttività è il cosiddetto costo per mille o

CPM: questo parametro si riferisce al costo da sostenere per raggiungere mille clienti futuri. Il pregio principale del CPM è costituito dalla sua semplicità e dal fatto che permette di determinare una base comune di confronto per differenti tipi di media. La semplicità del CPM è però anche il suo principale svantaggio, poiché non contempla al suo interno altri fattori per determinare il successo di una pubblicità in un determinato medium quale il livello di coinvolgimento del consumatore.
Per la valutazione dell'efficacia della pubblicità si fa riferimento anche ad elementi quale la circolazione del messaggio, l'ampiezza dell'audience e il numero di televisori in uso per minuto di messaggio pubblicitario. Si possono ovviamente definire diverse graduatorie di media a seconda del metodo di misurazione preso in considerazione. Un altro problema legato a questo tema è costituito dal cosidetto 'effettivo raggiungimento' del cliente potenziale. La copertura (*reach*) è definita dal numero di spettatori target esposti al messaggio pubblicitario almeno una volta in un periodo di tempo determinato. Un fattore altrettanto importante è il numero di volte in cui, in media, tali spettatori sono stati esposti al messaggio in esame, nel periodo di tempo considerato. Questo tasso di esposizione è chiamato frequenza media (*frequency*). Poiché gli operatori di *marketing* devono rispettare i limiti di budget loro imposti, essi dovranno decidere se aumentare la copertura a spese della frequenza media o viceversa.

7.5.1 I quotidiani

Più della metà di tutta la pubblicità che appare sui quotidiani è a risposta diretta; i settori che fanno maggiormente ricorso a questo media sono quello dei servizi finanziari, delle auto, della vendita per corrispondenza e il *retail*. I vantaggi derivati dall'utilizzo di questo medium sono rappresentati da:

1. la copertura intensiva di mercati locali e nazionali;

2. l'alta credibilità che il medium stampa assicura al messaggio pubblicitario;

3. l'accettazione e l'utilizzo molto diffusi da gran parte della popolazione e la flessibilità e la tempestività propria del mezzo in sé;

4. un basso tasso di CPM;

5. ha un duplice effetto: di rafforzamento dell'immagine di marca, ma anche di risposta al messaggio pubblicitario da parte del consumatore.

Al contempo, la scelta di questo medium comporta anche alcuni svantaggi quali:

1. la breve durata dell'esposizione del messaggio pubblicitario;

2. il basso tasso di risposta da parte del consumatore, spesso dovuto alla lettura affrettata tipica del lettore medio che molto frequentemente ignora il messaggio pubblicitario;

3. le possibilità creative molto limitate;

4. il medium in questione ha una scarsa audience occasionale.

Sulla base delle caratteristiche emerse, ed esemplificate attraverso le regole di classificazione presentate nel precedente capitolo, non crediamo che il medium quotidiano possa esssere uno strumento idoneo per le future campagne pubblicitarie: quanto, infatti, può emergere dalle regole è, per sua natura, un modello locale e non globale, facente riferimento a un gruppo ristretto di persone caratterizzate da certe particolarità. L'audience a cui i quotidiani si riferiscono è invece un'audience generica, non particolarizzata e, proprio questa caratteristica, potrebbe causare una dispersione del messaggio pubblicitario.

7.5.2 Le riviste

I vantaggi derivati dall'utilizzo di questo medium sono rappresentati da:

1. l'alta selettività geografica e demografica, grazie alla quale è possibile segmentare l'audience in base ai suoi interessi;

2. la possibilità di enfatizzare maggiormente il messaggio pubblicitario rispetto al medium quotidiani dal momento che è possibile utilizzare colori per la marca o il messaggio;

3. l'attenzione del lettore è molto più viva.

Al contempo, la scelta di questo medium comporta anche alcuni svantaggi quali:

1. bisogna sostenere un alto costo di CPM;

2. possiede sempre e comunque i limiti creativi propri della stampa;

3. i tempi di attesa per vedere pubblicato il proprio messaggio pubblicitario sono piuttosto lunghi (dalle sei alle otto settimane);

4. non è garantita una posizione di eccellenza per il proprio messaggio (a meno che non venga pagato un sovrapprezzo).

In questo caso, l'audience a cui le riviste mirano è molto selezionata sia geograficamente che demograficamente, facilmente segmentabile in base a certe caratteristiche. Dalle regole non emergono degli interessi specifici caratterizzanti determinate nicchie della popolazione e che potrebbero far ricadere la scelta su una rivista piuttosto che su un'altra: ciò che si evince, piuttosto, è che, guardando alla popolazione totale, tendono ad essere nuovi clienti in tutte le province i giovani (fino ai 35 anni) e, in quasi tutte le province, le persone che hanno cambiato residenza nell'ultimo anno.

Tali caratteristiche potrebbero quindi essere sfruttate ponendo la nuova pubblicità in riviste lette in larga scala dai giovani oppure in riviste contenenti gli annunci immobiliari provinciali.

7.5.3 Gli inserti

Gli inserti possono essere di vario tipo, 'sciolti' o rilegati, e vengono allegati a molti quotidiani, supplementi e riviste. In essi è compreso un tagliando o un numero di telefono per la risposta. Sono in particolare quella montagna di carta che sempre più spesso riempie i quotidiani durante la settimana. Un errore che fanno frequentemente le persone che si accostano per la prima volta al *marketing* è quello di ritenere che gli inserti non servano a niente, perché vengono gettati direttamente nella spazzatura. Le cose non stanno proprio così.
Gli inserti distribuiti con i giornali, assieme agli opuscoli distribuiti porta a porta o da prelevare a cura dei consumatori, sono uno di quei media a cui gli operatori di *marketing* generico prestano scarsa attenzione, ma che possono invece rivelarsi essere uno strumento molto utile.
Gli inserti funzionano perché attirano l'attenzione su sé stessi prima che il lettore possa essere 'distratto' dagli articoli del giornale o della rivista in cui sono contenuti. Come minimo il lettore dà loro un'occhiata, prima di buttarli, che è molto di più di quanto spesso accada per le pubblicità diffuse a mezzo stampa. I punti di forza di questo medium sono costituiti da:

1. hanno un maggiore impatto rispetto alla stampa dal momento che attirano in maniera superiore l'attenzione del lettore;

2. permettono una maggiore libertà creativa rispetto alla stampa poiché si possono utilizzare diversi colori, forme e dimensioni, e in più è possibile inserirvi una grande quantità di informazioni;

3. generano più risposte rispetto alle pubblicità diffuse sui quotidiani.

I punti di debolezza sono invece rappresentati da:

1. vengono spesso considerati come 'spazzatura';
2. i tempi di produzione sono lunghi;
3. hanno un costo per mille contatti (CPM) maggiore rispetto a quello dei quotidiani.

Gli inserti in larga parte riprendono quanto detto sopra per le riviste. Infatti, essi consentono una buona segmentazione dell'audience, proprio come avviene per queste ultime, ma con la possibilità di poter anche sfruttare l'audience generale fornita per esempio dai quotidiani. Così, possono essere allegati degli inserti sia in quotidiani locali generici, sia in quotidiani locali che hanno come argomento principale la compravendita di immobili.

7.5.4 La televisione

Mettendo a confronto i vari media, bisogna ammettere che la televisione è unica. Nient'altro permette infatti all'impresa di dimostrare, di provare e, soprattutto, di drammatizzare come fa la televisione. I vantaggi rappresentati dall'utilizzo di questo medium sono costituiti da:

1. l'utilizzo di una combinazione di suoni, immagini e movimento in grado di catturare e 'inchiodare' l'attenzione dello spettatore/acquirente;
2. l'ampia copertura dell'audience di massa che la televisione permette di raggiungere;
3. è in grado di generare il *passaparola*;
4. gode di *autorità* e di *credibilità* presso lo spettatore, il quale molto spesso ritiene che se una comunicazione viene dalla televisione, deve per forza essere un qualcosa di 'importante' o di 'vero';
5. ha un basso costo per mille contatti (CPM);

6. il *targeting* è migliorato grazie alla proliferazione del digitale;
7. non ha eguali quando si tratta di costruire l'immagine di marca e ottenere risposte simultaneamente.

La scelta di questo medium comporta però anche alcuni svantaggi, in particolare rappresentati da:

1. un basso tasso di risposta, dovuto soprattutto allo stato di passività in cui lo spettatore si trova quando sta guardando la televisione;
2. i messaggi devono essere semplici ed esposti con chiarezza;
3. è un medium sempre costoso ed è difficile raggiungere le audience di maggior valore;
4. a livello analogico non vi è ancora nessuna selettività dell'audience;
5. conosce tempi piuttosto lunghi, sia per la produzione della pubblicità, sia per la trasmissione della medesima;

Per la televisione vale lo stesso discorso fatto per i quotidiani: essa fa infatti ancora riferimento ad un'audience troppo generica, soprattutto a livello analogico locale non consente alcuna segmentazione della popolazione e, non è quindi in grado di dare giustamente un corrispettivo di marketing a quanto emerso dalle regole di classificazione presentate nel precedente capitolo.

7.5.5 La radio

La radio ha alcune delle caratteristiche delle trasmissioni televisive. Può generare il passaparola e un certo entusiasmo, ma spesso non agli stessi livelli della televisione. Ciononostante, questo media non andrebbe sottovalutato: le pubblicità radiofoniche possono generare notorietà e credibilità, elementi che andrebbero considerati come parte delle caratteristiche necessarie per stimolare la risposta diretta. I punti di forza derivati dall'utilizzo di questo medium sono rappresentati da:

1. è economica: con la radio si può raggiungere un ampio audience a basso costo, in particolare, tra tutti, è il medium più conveniente;
2. funziona molto bene in abbinamento con altri media;
3. si può operare un'ottima selettività dell'audience tramite la scelta dell'emittente;
4. ha un basso costo per unità di tempo;
5. è flessibile geograficamente;
6. è un media apprezzato dal pubblico.

Gli svantaggi conseguenti alla scelta di questo medium sono invece costituiti da:

1. un basso tasso di risposta al messaggio pubblicitario soprattutto perché il medium radio viene spesso utilizzato come sottofondo: le persone la ascoltano mentre stanno facendo altro;
2. si possono utilizzare solo messaggi semplici, brevi, stimolanti e facili da memorizzare poiché è molto difficile recepire informazioni diffuse solo oralmente rispetto a quelle che sono anche caratterizzate da una componente visiva;
3. è necessario usare messaggi piuttosto creativi e ripetuti di frequente per non annoiare l'ascoltatore con la ripetizione della stessa pubblicità.

La radio può essere considerata un medium particolarmente adatto per rappresentare quanto emerso dalle regole di classificazione; infatti, essa esiste ed è saldamente presente in qualsiasi contesto locale, anche con la possibilità di segmentare l'audience scegliendo la stazione più adatta, per esempio, a raggiungere i giovani fino ai 35 anni. Lo stesso messaggio può essere costruito in tal senso, fatto, cioè, per attirare l'attenzione del pubblico più giovane,

da poco emancipato dalla famiglia e che ha appena cambiato stile di vita (per esempio cambiando residenza).

7.5.6 Consegna di materiale pubblicitario nelle abitazioni

Si tratta di una distribuzione mirata di opuscoli *senza indirizzo*, o pacchetti di materiale pubblicitario, nella cassetta delle lettere. Il materiale pubblicitario consegnato porta a porta viene distribuito da imprese specializzate o dal servizio postale pubblico.
Alcuni sostengono che la distribuzione porta a porta offra molti dei vantaggi del *direct mail*, ma a costi inferiori. Non dovendo sostenere spese postali e per l'acquisto di liste di destinatari, i costi di distribuzione sono dieci volte inferiori rispetto a quelli per il *direct mail*.
Ciononostante questo media ha importanti svantaggi rispetto al *direct mail*. Il *targeting*, ad esempio, è meno accurato di quello realizzabile con il *direct mail*, essendo basato su profili geodemografici (e/o sugli stili di vita), invece che su indirizzi individuali. Di conseguenza dovrebbe forse essere utilizzato solo da imprese che vendono prodotti di interesse universale.
I punti di forza tipici dell'utilizzo di questo medium sono:

1. il costo per mille contatti (CPM) è più basso rispetto al *direct mail*;
2. ha una certa versatilità creativa come il *direct mail*: in esso è possibile trasmettere una grande quantità di informazioni e in un'ampia varietà di formati, purché entri nella cassetta delle lettere;
3. ha un alto impatto, rispetto ai media di ampia scala;
4. permette un *targeting* migliore rispetto agli inserti;
5. è un medium notevole per le imprese per le quali è importante l'audience locale.

I punti di debolezza derivanti dalla scelta di questo medium sono invece:

1. un tasso di risposta inferiore rispetto al *direct mail*;

2. il messaggio pubblicitario non può essere personalizzato;

3. da molte persone viene spesso considerato come 'pubblicità spazzatura';

4. non suscita un'impressione di autorevolezza nel ricevente;

5. vi è una relativa difficoltà nel controllo dalla qualità dal momento che non è inconsueto che addetti al recapito poco seri gettino via gli opuscoli invece di consegnarli;

6. non può essere utilizzato con il duplice scopo di suscitare una risposta e creare l'immagine di marca.

La consegna di materiale pubblicitario nelle abitazioni potrebbe forse essere considerato come un medium di salvataggio nel caso in cui, dalle nostre regole, non fosse emersa alcuna caratterizzazione sensata atta a segmentare la popolazione di riferimento; dal momento che, però, delle particolarità discriminanti ci sono, questo medium potrebbe forse essere utilizzato con un senso più simile a quello del *direct mail*, ma senza alcun indirizzo né possibilità anagrafica da poter utilizzare. Guardando alle regole selezionate provincia per provincia, ci si potrebbe concentrare nel consegnare materiale pubblicitario nelle abitazioni appena costruite dove i nuovi inquilini si sono da poco trasferiti, oppure, guardando a regole come quelle delle tabelle 7.1 e 7.2.

ci si potrebbe concentrare su particolari comuni di una determinata provincia.

Tabella 7.1: Latina: regola 4.

Regola 4			
A	**B**	**Elift**	**Supporto**
COMUNEDIRESIDENZA=FONDI		1.8	601
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sulle persone che risiedono a Fondi aumenta dell'80% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che risiedono a Fondi sono 1741: di questi 601 sono nuovi clienti, con una media del 34.5% (cioè, 1.8 volte la media generale del 19% di nuovi clienti totali).			

7.5.7 La cartellonistica

Un cartellone è un foglio di carta stampato che si affigge in un luogo pubblico con lo scopo di comunicare qualcosa, o di fare pubblicità. I vantaggi rappresentati dall'utilizzo di questo medium sono:

1. il grande impatto visivo che è capace di creare sui destinatari;
2. la relativa assenza di pubblicità concorrenti;
3. la possibilità di essere visto da molte persone e per molte volte;
4. lo sforzo minimo da parte dei destinatari (che non devono, ad esempio, sintonizzarsi su una rete o acquistare un giornale come invece accade per altri media);
5. la possibilità di essere posizionati in maniera strategica (ad esempio in prossimità di un punto vendita o comunque in aree ben precise);
6. è relativamente poco oneroso.

Gli svantaggi sono invece costituiti da:

1. le limitazioni alla creatività;

Tabella 7.2: Ragusa: regola 8.

Regola 8			
A	**B**	**4th Quantifier of founded double implication**	**Supporto**
CAMBIO-_RESIDENZA=True	COMUNEDIRESIDENZA==VITTORIA	60%	384
Descrizione			
Se consideriamo le persone che risiedono a Vittoria sappiamo che almeno il 60% di queste persone ha cambiato residenza nell'ultimo anno. In particolare, i clienti, sia vecchi che nuovi, che hanno cambiato residenza nell'ultimo anno e che risiedono a Vittoria sono 418: di questi 384 sono nuovi clienti, appunto con una media del 60%.			

2. l'estrema sinteticità (di solito i poster sono messi in luoghi di passaggio, ad esempio lungo una strada, dove i tempi a disposizione per la lettura sono minimi - ma non necessariamente: basti pensare ai poster alle fermate della metropolitana);

3. nessuna selettività dell'audience: a differenza di altri media è infatti molto più difficile prevedere chi vedrà un cartellone;

4. l'eventuale immagine negativa del medium che può riflettersi sulla pubblicità (talvolta la cartellonistica ambientale è stata ritenuta deturpante per il paesaggio o l'estetica urbana e distraente per chi guida un veicolo).

Anche la cartellonistica sembrerebbe essere un medium molto adatto ad esprimere in forma pubblicitaria quanto emerso dalle nostre regole. Questo è infatti uno strumento che può essere considerato molto generico riguardo alla forma di espressione del messaggio, che, cioè, può non essere costruito per essere notato solo da certe categorie di persone, ma che può però essere posizionato in punti strategici per la ricezione di quest'ultimo.

Così, i nostri cartelloni potrebbero essere posizionati in nuovi quartieri residenziali, oppure, seguendo la proposta indicata da regole come quella della tabella 7.3, i cartelloni potrebbero essere localizzati nei treni o nelle stazioni frequentati dalla maggior parte dei pendolari di quella provincia.

Tabella 7.3: Latina: regola 6.

Regola 6			
A	**B**	**Elift**	**Supporto**
PENDOLARE=1		1.65	484
Descrizione			
Se consideriamo la popolazione totale, concentrarsi sui pendolari aumenta del 65% la probabilità di trovare nuovi clienti. In particolare, i clienti, sia vecchi che nuovi, che sono pendolari sono 1538: di questi 484 sono nuovi clienti, con una media del 31.8% (cioè, 1.65 volte la media generale del 19% di nuovi clienti totali).			

7.5.8 Internet

La comunicazione on line ha per obiettivo la diffusione di informazioni al fine di incidere opportunamente sulla relazione tra acquirente e rivenditore. Le principali differenze rispetto alla comunicazione tradizionale sono da ascriversi, dal lato dell'utente, alla possibilità di interattività con lo stesso messaggio e ad una progressiva frammentazione dell'attenzione dovuta al moltiplicarsi degli emittenti; dal lato degli operatori, alla segmentazione spinta fino alla personalizzazione in una logica *one-to-one* che Internet consente di perseguire e, al contempo, alle maggiori difficiltà nel sollecitare il coinvolgimento emotivo degli utenti.

Le forme più innovative di pubblicità in Rete inducono a sviluppare appropriate strategie per portare il messaggio direttamente al consumatore, nel momento in cui egli ne ha più bisogno. La comunicazione in ambiente digitale deve, in altre parole, tendere ad essere sempre più contestuale.

La prima forma di pubblicità nata in Internet era rappresentata dallo stesso

sito, ma al moltiplicarsi degli indirizzi Web, la semplice creazione di una presenza in Rete ha cominciato a non essere più sufficiente per raggiungere una vasta audience. Sono nate così le prime forme di pubblicità on line: *banner* e *button*. Il progressivo affollamento di tali formule pubblicitarie sul Web ha finito tuttavia col ridurne sensibilmente l'efficacia, stimolando la nascita di nuove modalità di comunicazione on line. In particolare, la Rete rende oggi possibile lo sviluppo di due fondamentali modalità di pubblicità e promozione. Una prima è di tipo *multimedia* ed è necessariamente confinata al web. Rientrano in questa categoria i già citati *banner* e *button*, ma anche *interstitial* e *daughter window*, *e-coupon* e *loyalty program*, mini-siti, *sponsorship* e *advertorial*. La seconda modalità è invece di tipo *text-based* e può essere diffusa attraverso qualunque strumento di connessione ad Internet, dal cellulare al PDA; tipicamente viene usata al fine di sollecitare una risposta diretta (*direct marketing*); la pubblicità tramite posta diretta o *e-mail marketing*, fino alla sua evoluzione in *viral marketing*, rappresenta un classico esempio di questa forma di comunicazione.

Banner e *button*

Il *banner* - insieme alla sua versione ridotta, il cosiddetto *button* - è il più classico e noto degli strumenti pubblicitari utilizzati in Rete. Si tratta di uno spazio grafico che può assumere varia forma e dimensione ed è in generale collocato in una posizione rilevante all'interno di una pagina Web. Rispetto alla cartellonistica, il *banner* ha il vantaggio di consentire di combinare immagini in movimento, ma con la possibilità di interazione diretta in tempo reale da parte dell'utente. La probabilità che il *banner* abbia successo dipende tanto dal traffico all'interno del sito in cui è collocato, quanto dal suo specifico contenuto. Il *banner* può comparire in un'ampia varietà di formati e dimensioni e, soprattutto, può presentarsi in tre diverse modalità di erogazione del messaggio:

- banner statici;
- banner dinamici;
- banner interattivi.

I primi rappresentano la forma più semplice di *banner* ed il loro utilizzo è giustificato soprattutto da obiettivi di *branding*, ovvero di costruzione dell'immagine, in prevalenza da parte di aziende già note. Il vantaggio sta nella sua rapidità di visualizzazione per l'utente e, per l'azienda, nella sua facilità ed economicità di realizzazione, nonché nella sua universale accettazione da parte di qualunque sito.
I *banner* animati prevedono che siano presenti immagini in movimento all'interno di uno spazio analogo a quello del *banner* statico. *Banner* di questo tipo sono quindi più lenti da caricare, ma consentono la veicolazione di messaggi più articolati e completi, permettendo di far cogliere tempestivamente all'utente novità contenute nel sito oggetto di promozione e all'azienda di aggiornare informazioni evidenziate nell'annuncio modificando uno solo dei *frame*, ma lasciando inalterato l'intero impianto di animazione.
I *banner* interattivi si differenziano dai precedenti in quanto finalizzati al coinvolgimento dell'utente in attività che possono spaziare dal semplice inserimento di un'informazione o espressione di una preferenza alla partecipazione a un gioco o al riempimento di un *form* di intervista, senza entrare nel sito pubblicizzato, ma limitando l'interazione nell'ambito dello stesso *banner*.

Link sponsorizzati

I *link* di testo sponsorizzati da specifici investitori sono tra le formule di pubblicità on line più semplici, meno intrusive e, al contempo, molto spesso più efficaci. Trovano in genere collocazione all'interno di motori di ricerca, dove vengono di volta in volta proposti in abbinamento a parole chiave

predefinite digitate dall'utente. Possono comparire all'interno di *button*, accanto ai risultati della ricerca, oppure come primo risultato di ricerca, sotto forma di puro *link* di testo.

Interstital e *daughter window*

Gli *interstitial* sono finestre pubblicitarie a tutta pagina che vengono visualizzate dall'utente nel corso della navigazione, senza che ne venga fatta da questi alcuna richiesta - ragione per cui sono spesso chiamati anche *pop-up*. Proprio per la loro intrusività, ne è stata da tempo introdotta una versione che occupa solo una porzione della pagina web, definita *daughter* o *sister window*. Tanto nell'uno quanto nell'altro caso, la durata della finestra sullo schermo dell'utente varia in base alla programmazione della *pop-up window*; l'utente può comunque eliminare una simile finestra in ogni momento, cliccando al di fuori della sua area o, comunque, sul bottone di chiusura in alto a destra. Perché possano avere successo devono presentare un elevato contenuto di intrattenimento o di informazione.

Mini-siti

Sono siti che nascono intenzionalmente con una durata limitata, funzionale alla promozione di uno specifico evento o prodotto.

Sponsorship e *advertorial*

La sponsorizzazione on line può assumere varie forme e sta acquisendo una crescente popolarità. Consiste nell'abbinamento di un marchio ad un evento, spesso sotto forma di un concorso, a un contenuto affine e, pertanto, idoneo a rafforzare l'immagine del primo e a diffonderne l'implicito messaggio commerciale, senza necessariamente mirare a generare un incremento di traffico nel suo sito. L'obiettivo per la sponsorizzazione è quindi di *branding*. Una differente tipologia di *sponsorship* è rappresentata dal-

l'*advertorial*, ovvero un messaggio che combina la presenza di uno sponsor con contenuti editoriali.

7.5.9 Il *social network advertising*

Social network advertising è un termine che viene usato per descrivere una forma di pubblicità on line che si focalizza sui siti di *social network*. Uno dei principali benefici del fare pubblicità su un sito di *social network* quale *Facebook*, *Myspace*, *Friendster*, o *Bebo* è che gli inserzionisti possono trarre vantaggio delle informazioni demografiche rilasciate dagli utenti e segmentare appropriatamente i loro messaggi.
eMarketer ha affermato che nel 2008 sono stati spesi in tutto il mondo circa 1.5 bilioni di dollari con la pubblicità sui siti di *social network* e che questo mercato continuerà a crescere raggiungendo i 2.6 bilioni di dollari spesi nel 2012, la maggior parte dei quali saranno spesi nei soli Stati Uniti.
Il *social network advertising* si articola in tre principali classificazioni:

- **la pubblicità diretta basata sulla rete di amici di una persona target**: questo può essere il formato più efficace ma, di sicuro, anche il più controverso. Un esempio di questa modalità di azione è il *Facebook beacon project*[1]. Esso si basa sulle attività che gli amici di questa persona target stanno compiendo, quindi, per esempio, si potrebbe vedere companire un messaggio come questo tra il feed delle news: 'Mario ha appena comprato un CD dei Radiohead da MusicWorld'. Questa può essere una modalità di *advertising* estremamente efficace, dal momento che spesso le persone tendono a prendere la decisione riguardo al comprare qualcosa o al fare qualcosa basandosi sulle decisioni già prese dai loro amici. Comunque, un'ampia dose di controversie circonda questa metodologia poiché può essere considerata intruisiva riguardo la relazione personale che una persona ha con i propri amici ed, inoltre,

[1] http://www.facebook.com/advertising/

solleva importanti questioni riguardo alla privacy degli utilizzatori del *social network* in questione;

- **la pubblicità diretta collocata su un sito di *social network***: questo è un approccio più tradizionale di pubblicità on line; è un concetto simile a quello di piazzare un *banner* su un altro sito che non sia un sito di *social network*. Questi *advertising* sono per esempio collocati nell'angolo in alto a destra di una qualsiasi pagina di *Myspace*, o inseriti con la forma di un *banner* alla sinistra di un qualunque profilo *Facebook*, e così via. Esistono comunque due differenze sostanziali rispetto al collocamento di un *banner* in una qualunque altra pagina web che non sia di *social network*. Per prima cosa, posizionando un messaggio pubblicitario su un sito di *social network* si può trarre vantaggio dalle informazioni demografiche presenti nel profilo degli utenti registrati ed il messaggio può quindi essere personalizzato per il singolo utente a cui è diretto. In secondo luogo, questi tipi di messaggi pubblicitari possono essere anche inseriti da singoli imprenditori nelle pagine delle applicazioni che loro hanno sviluppato attraverso reti pubblicitarie quali *AdParlor*. Questi messaggi hanno accesso ai dati degli utenti e possono generare profitti per gli sviluppatori delle applicazioni, dando loro così un'ulteriore motivazione per creare altre applicazioni e fornendo inoltre agli inserzionisti un metodo più attraente per raggiungere gli utenti dei *social network*;

- **la pubblicità indiretta realizzata attraverso la creazione di *gruppi* o pagine**: questa è una tecnica di *marketing* innovativa in cui una compagnia andrà a creare una *pagina* o un *gruppo* a cui gli utenti potranno liberamente decidere se unirsi o meno. Le compagnie interessate useranno questo metodo per costruirsi una base di 'abbonati' o di 'fan' ed useranno quest'ultima per lanciare un concorso,

un nuovo prodotto, o semplicemente per accrescere la consapevolezza riguardo al proprio *brand*. Questi gruppi possono crescere velocemente riguardo al numero di iscritti, i quali possono diventare in fretta uno strumento di *marketing* molto efficace.

7.6 Il marketing diretto (*direct marketing*)

Mediante il *marketing* diretto l'azienda comunica direttamente con la clientela, on line, per corrispondenza postale diretta, con cataloghi, tramite pubblicità a risposta diretta o vendita personale.
Le tecniche di *marketing* diretto tradizionali sono oggi maneggiabili con maggiore efficienza ed efficacia, grazie all'ausilio di computer e database. Sicuramente nell'ultimo decennio la tecnologia ha impresso un'accelerazione notevole alle attività di *marketing* diretto. Grazie ad essa, oggi gli operatori di *marketing* sono in grado di personalizzare i meccanismi di comunicazione, creando connessioni e dialoghi *one-to-one* con i clienti.
Un altro ovvio fattore di crescita del *marketing* diretto va rintracciato in Internet, sempre più utilizzato dai consumatori per acquistare diversi tipi di prodotti.
Ai consumatori afflitti da 'mancanza di tempo', il *marketing* diretto offre diversi vantaggi. Può far risparmiare tempo e spesso anche denaro, può far ottenere servizi migliori, maggiore tutela della propria privacy, e molti consumatori lo trovano anche divertente. Molto spesso le attività di *direct marketing* si dimostrano molto efficaci per innescare il processo d'acquisto, dal momento in cui il cliente chiede maggiori informazioni su un prodotto o un servizio.

7.6.1 La posta o *direct mail*

Il *direct mail* è la consegna di materiale pubblicitario nella cassetta della posta di specifici e pre-selezionati destinatari. I vantaggi che derivano dall'utilizzo di questo medium sono:

1. l'estrema selettività dell'audience;
2. il fatto che non vi è alcuna possibilità di competizione da parte di pubblicità concorrenti all'interno dello stesso medium;
3. la grande personalizzazione che può essere attivata: la posta diretta può raggiungere i clienti singolarmente e può quindi essere segmentata secondo i loro bisogni in base a precedenti transazioni avute con l'ezienda. L'evoluzione delle tecnologie di stampa permette inoltre la personalizzazione delle immagini così come del messaggio in sé e dei dati personali (come il nome o l'indirizzo). Per esempio, tutti gli uomini destinatari di un'offerta possono ricevere un pacchetto personalizzato con la fotografia di un uomo sulla copertina, mentre tutti i destinatari donne riceveranno un pacchetto con un'immagine femminile;
4. la flessibilità;
5. può essere molto profittevole, se ben utilizzata.

Gli svantaggi invece comprendono:

1. il costo relativamente alto che comporta la scelta di questo medium rispetto a qualsiasi altro medium pubblicitario di massa; i costi possono anche salire per l'acquisizione di database e per il mantenimento e l'aggiornamento di questi ultimi;
2. il fatto che spesso i consumatori non prestano molta attenzione ai messaggi pubblicitari e li gettano senza neanche concedere loro un'occhiata.

7.6.2 L'*e-mail marketing* e il *viral marketing*

La comunicazione di tipo *text-based*, prevalentemente centrata sullo strumento della posta elettronica, mira a gestire la relazione con il cliente e ad alimentarne la fedeltà una volta che egli si è già trasformato in acquirente. La comunicazione multimediale e quella *text-based* (*permission-based*) appaiono, quindi, in quest'ottica, come due approcci complementari.
L'uso dell'*e-mail marketing* può essere molto efficace qualora gli utenti abbiano dichiarato le proprie preferenze all'interno di un sito, nell'ambito del quale hanno talvolta già effettuato uno o più acquisti, e hanno comunque espressamente dato il proprio consenso al fine di continuare ad essere informati relativamente a specifiche aree di interesse o categorie di prodotto, ed eventualmente a particolari offerte speciali all'interno di ciascuna di esse.
Sotto il profilo della capacità di incidere sulle scelte e i comportamenti utenti, ancora superiore tende ad essere l'effetto di una peculiare evoluzione dell'*e-mail marketing*: il cosiddetto *viral marketing*. Il *marketing* virale altro non è che la formalizzazione del classico fenomeno del 'passaparola' adattato al Web, con l'obiettivo di moltiplicare i contatti e legittimare la qualità dell'offerta attraverso il consiglio diretto di chi l'ha già sperimentata.

Conclusioni

In questa tesi abbiamo affrontato il seguente problema:

quali sono le particolarità che caratterizzano i clienti che per la prima volta decidono di avvalersi del servizio di assistenza fiscale per la compilazione della dichiarazione dei redditi offerto dal CAAF-CISL?

Lo scopo del progetto di tesi è stato quindi quello di dare una risposta a questo quesito e di cercare di scoprire quelle eventuali caratteristiche (fiscali o demografiche) in grado di identificare e differenziare i nuovi contribuenti rispetto ai vecchi.
L'intero ciclo di lavoro è durato 6 mesi; gli obiettivi stabiliti ad inizio del lavoro possono essere riassunti in due punti principali:

1. **Studio delle tecniche di selezione delle regole associative**: una volta compreso qual era l'obiettivo di interesse per le analisi e, a seguito di uno studio condotto con la finalità di capire qual era la risposta al problema in termini di tecniche di *data mining*, siamo giunti alla conclusione che la soluzione risiedeva nella tecnica delle regole associative, più in particolare delle regole di classificazione. La metodologia con cui le regole sarebbero state estratte ci era chiara: si trattava di un tool - DCUBE - già utilizzato in precedenza con questi stessi scopi. Il problema rimanente era però quello di cercare di capire come fare a far 'galleggiare' soltanto le regole più interessanti, dal momento che il numero di regole emerse era talmente grande da rendere impossibile

esaminarle soggettivamente una ad una. A questo scopo è stato così condotto uno studio molto approfondito riguardante lo stato dell'arte delle misure di interesse esistenti per valutare le regole associative: misure di interesse oggettive, soggettive e semantiche. A seguito di questo approfondimento la scelta è ricaduta su due misure di interesse oggettive: l'*extended lift* (*elift*) e il *4th quantifier of founded implication*; esse sono state così implementate all'interno di DCUBE ed utilizzate per la selezione finale delle regole;

2. **caratterizzazione dei nuovi clienti**: dal momento in cui ci era risultata chiara la soluzione teorica al problema in esame, abbiamo dato inizio alla vera e propria fase di analisi dei dati. La parte di pre-processing su di essi era già stata svolta fino ai dati relativi al 2008 grazie al lavoro di [2] ed i processi di pre-processing di questo stesso lavoro si sono comunque rivelati essere flessibili e quindi riutilizzabili per quest'altra nuova analisi. Una volta avuti a disposizione i dati in formato corretto, grazie anche al lavoro di collaborazione con gli esperti del dominio che ci hanno aiutato nella parte relativa alla comprensione del significato degli attributi e nella fase di selezione degli stessi, abbiamo dato il via alla fase di estrazione delle regole tramite l'utilizzo del tool DCUBE, fase il cui scopo era quello di cercare delle eventuali caratteristiche dei contribuenti le quali possono favorire maggiormente la presenza di nuovi clienti all'interno di un particolare contesto. La selezione delle regole più interessanti è avvenuta grazie a due misure di interesse oggettivo in particolare tra tutte quelle elencate nella prima parte della trattazione: vale a dire l'elift e il 4th quantifier of founded implication.

Terminato il lavoro, i risultati sono stati presentati agli esperti del dominio del CAAF-CISL così da avere un riscontro sull'effettiva utilità delle informazioni estratte. Durante questo stesso incontro abbiamo dovuto ripen-

sare la forma con cui avevamo deciso di presentare le informazioni estratte, dal momento che questo nostro modo di esprimere e di rappresentare le regole di classificazione non consentiva loro di cogliere il vero significato di esse. Una volta superato questo problema enunciando l'informazione portata dalle regole in termini più discorsivi, è potuta iniziare una collaborazione più fruttuosa. Abbiamo capito che, oltre alla conoscenza che la regola era in grado di fare scoprire, essi erano interessati al numero di persone per cui quella stessa regola era valida: vale a dire al suo supporto, valore la cui importanza superava anche, ai loro occhi, quello della misura di interesse. Nello spiegare loro le regole è stato così sempre messo in primo piano il supporto, e, con queste intenzioni, ci siamo mossi con loro per cercare di trovare nuove regole che meglio approfondissero certe caratteristiche (attributi) a cui sembravano essere particolarmente interessati, poiché azionabili. Alla fine dell'incontro il loro giudizio è stato piuttosto favorevole, ma si sono comunque rivelati essere disorientati in quanto sono state scoperte delle caratteristiche e delle tendenze seguite dai nuovi clienti che essi non si aspettavano di vedere emergere dai dati. Essi pensavano infatti di poter condurre delle campagne pubblicitarie più orientate alla famiglia e, quindi alle persone mature, invece i nuovi clienti, in tutte le province, appaiono essere i giovani che hanno di recente cambiato residenza e si sono emancipati dalla famiglia.

Inoltre, grazie all'incontro con gli esperti del dominio del CAAF-CISL delle quattro province interessate, abbiamo avuto modo di capire il perché alcune caratteristiche emergevano dai dati provincia per provinca; questo era possibile dal momento che essi potevano dirci se erano state effettuate delle promozioni pubblicitarie particolari o se era stata stipulata una qualche convenzione con una determinata azienda.

L'interesse degli esperti del dominio si è anche manifestato nella volontà di proseguire nella realizzazione di un altro studio di data mining riguardante

questi stessi dati: nel cercare cioè di capire dove si trovano i nuovi clienti, di cercare quindi di localizzarli, provincia per provincia, in una mappa della città estraendo così dei modelli geo-referenziati. Grazie ai risultati di questo studio essi potrebbero per esempio comprendere quale sede del CAAF-CISL risulti essere la più frequentata e perché, e, magari, dare il via a delle campagne promozionali per cercare di valorizzare anche le altre sedi provinciali del CAAF-CISL.

Dal momento che, però, questo lavoro si conclude senza poter partecipare alla fase di realizzazione delle campagne pubblicitarie, la parte finale della tesi si occupa di dare un'idea di che cosa sia il marketing mix in generale e la parte relativa al promotion mix in particolare. Vengono così descritti tutti i media potenzialmente utilizzabili per sfruttare ciò che è emerso dalle regole presentate alla fine di questo progetto e, in conclusione ad ogni descrizione, viene mostrato un piccolo esempio il cui scopo è quello di illustrare se il medium in questione possa essere adatto o meno per raggiungere al meglio il segmento di persone identificato dalle regole selezionate.

Bibliografia

[1] Guozhu Dong and Jinyan Li. Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In *PAKDD '98: Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining*, pages 72–86, London, UK, 1998. Springer-Verlag.

[2] Carlo Licciardi e Claudio Petroni. Classificazione della clientela nei servizi fiscali. Master's thesis, Università di Pisa, 2009. Tesi di laurea specialistica in Informatica per l'economia e l'azienda.

[3] Antonio Albano. Basi di dati di supporto alle decisioni, 2008.

[4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[5] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. *CRISP-DM 1.0: Step-by-step data mining guide.* SPSS, 2000.

[6] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining.* Addison Wesley, 2006.

[7] Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT '99:*

Proceedings of the 7th International Conference on Database Theory, pages 398–416, London, UK, 1999. Springer-Verlag.

[8] J.Pei, J.Han, and R.Mao. Closet: an efficient algorithm for mining frequent closed itemsets. In *Proceeding of the 2000 ACM-SIGMOD international workshop data mining and knowledge discovery (DMKD'00)*, pages 11–20, 2000.

[9] MJ Zaki and CJ Hsiao. Charm: an efficient algorithm for closed itemset mining. In *Proceeding of the 2002 SIAM international conference on data mining (SDM'02)*, pages 457–473, 2002.

[10] J.Wang, J.Han, and J.Pei. Closet+: searching for the best strategies for mining frequent closed itemsets. In *Proceeding of the 2003 ACM SIGKDD international conference on knowledge discovery and data mining (KDD'03)*, pages 236–245, 2003.

[11] G.Grahne and J.Zhu. Efficiently using prefix-trees in mining frequent itemsets. In *Proceeding of the ICDM'03 international workshop on frequent itemset mining implementations (FIMI'03)*, pages 123–132, 2003.

[12] Guimei Liu, Hongjun Lu, Wenwu Lou, and Jeffrey Xu Yu. On computing, storing and querying frequent patterns. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 607–612, New York, NY, USA, 2003. ACM.

[13] Roberto J. Bayardo, Jr. Efficiently mining long patterns from databases. In *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 85–93, New York, NY, USA, 1998. ACM.

[14] Doug Burdick. Mafia: A maximal frequent itemset algorithm for transactional databases. In *ICDE '01: Proceedings of the 17th International Conference on Data Engineering*, page 443, Washington, DC, USA, 2001. IEEE Computer Society.

[15] Krzysztof J. Cios, Witold Pedrycz, Roman W. Swiniarski, and Lukasz A. Kurgan. *Data mining: a knowledge discovery approach.* Hardcover, 2007.

[16] Jef Wijsen and Robert Meersman. On the complexity of mining quantitative association rules. *Data Min. Knowl. Discov.*, 2(3):263–281, 1998.

[17] F.Bonchi, F.Giannotti, A.Mazzanti, and D.Pedreschi. Exante: anticipated data reduction in constrained pattern mining. In *Proceeding of the 7th European conference on principles and pratice of knowledge discovery in databases (PKDD'03).*, pages 59–70, 2003.

[18] J.Wang, J.Han, Y.Lu, and P.Tzvetkov. Tfp: An efficient algorithm for mining top-k frequent closed itemsets. *IEEE Trans Knowl Data Eng*, 17:652–664, 2005.

[19] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41, New York, NY, USA, 2002. ACM.

[20] S.Brin, R.Motwani, J.D.Ullman, and S.Tsur. Dynamic itemset counting and implication rules for market basket data. In *The Proceedings of SIGMOD*, pages 255–264, 1997.

[21] Khalil M. Ahmed, Nagwa M. El-Makky, and Yousry Taha. A note on 'beyond market baskets: Generalizing association rules to correlations'. *SIGKDD Explorations*, 1:46–48, 2000.

[22] W.Klosgen. Problems for knowledge discovery in databases and their treatment in the statistics interpreter explora. *Int. J. Intell. Systems*, 7:649–673, 1992.

[23] L.A. Goodman and W.H. Kruskal. Measures of association for cross-classifications. *Journal of the American Statistical Association*, 49:732–764, 1968.

[24] B.Gray and M.E.Orlowska. Clustering categorical attributes into interesting association rules. In *Proceeding of the 2003 second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '98)*, pages 132–143, 1998.

[25] G.Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229–248. MIT Press, 1991.

[26] P.Lenca, P.Meyer, B.Vaillant, and S.Lallich. *A multicriteria decision aid for interestingness measure selection.* LUSSI, 2004.

[27] B.Vaillant, P.Lenca, and S.Lallich. A clustering of interestingness measures. In *Proceedings of the 7th International Conference on Discovery Science (DS 2004).*

[28] Avi Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowl. and Data Eng.*, 8(6):970–974, 1996.

[29] B.Liu, W.Hsu, and S.Chen. Using general impressions to analyze discovered classification rules. In *Proceedings of the 3rd International Con-*

ference on Knowledge Discovery and Data Mining (KDD-97), pages 31–36, 1997.

[30] Bing Liu, Wynne Hsu, Lai-Fun Mun, and Hing-Yan Lee. Finding interesting patterns using user expectations. *IEEE Trans. on Knowl. and Data Eng.*, 11(6):817–832, 1999.

[31] Sigal Sahar. Interestingness via what is not interesting. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 332–336, New York, NY, USA, 1999. ACM.

[32] B.Padmanabhan and A.Tuzhilin. A belief-driven method for discovering unexpected patterns. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, pages 94–100, 1998.

[33] Yi-Dong Shen, Zhong Zhang, and Qiang Yang. Objective-oriented utility-based association mining. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, page 426, Washington, DC, USA, 2002. IEEE Computer Society.

[34] C. H. Cai, A. W. C. Fu, C. H. Cheng, and W. W. Kwong. Mining association rules with weighted items. In *IDEAS '98: Proceedings of the 1998 International Symposium on Database Engineering & Applications*, page 68, Washington, DC, USA, 1998. IEEE Computer Society.

[35] Songfeng Lu, Heping Hu, and Fan Li. Mining weighted association rules. *Intell. Data Anal.*, 5(3):211–225, 2001.

[36] Hong Zhang, Balaji Padmanabhan, and Alexander Tuzhilin. On the discovery of significant statistical quantitative rules. In *KDD '04: Pro-*

ceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 374–383, New York, NY, USA, 2004. ACM.

[37] Charles X. Ling, Tielin Chen, Qiang Yang, and Jie Cheng. Mining optimal actions for profitable crm. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining*, page 767, Washington, DC, USA, 2002. IEEE Computer Society.

[38] K.Wang, S.Zhou, and J.Han. Profit mining: From patterns to actions. In *Proceedings of the 8th Conference on Extending Database Technology (EDBT 2002)*, pages 70–87. IEEE Computer Society, 2002.

[39] Nicolas Pasquier, Rafik Taouil, Yves Bastide, Gerd Stumme, and Lotfi Lakhal. Generating a condensed representation for association rules. *J. Intell. Inf. Syst.*, 24(1):29–60, 2005.

[40] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Integrating induction and deduction for finding evidence of discrimination. In *ICAIL '09: Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 157–166, New York, NY, USA, 2009. ACM.

[41] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Integrating induction and deduction for finding evidence of discrimination. In *9th SIAM Conference on Data Mining (SDM 2009)*, pages 581–592, 2009.

[42] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Dcube: Discrimination discovery in databases. In *ACM International Conference on Management of Data (SIGMOD 2010)*, 2010.

[43] Philip Kotler. *Marketing Management. Millenium Edition.* Prentice-Hall., 2000.

[44] Philip Kotler. *Marketing Management.* Homewood, 1972.

[45] J.Paul Peter and James H.Donnelly Jr. *Marketing.*

[46] Michael J. A. Berry and Gordon S. Linoff. *Data Mining.* Apogeo Editore, 2002.

[47] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.

[48] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2005.

[49] Emanuela Prandelli e Gianmario Verona. *Marketing in rete. Analisi e decisioni nell'economia digitale.* McGraw-Hill, 2002.

[50] Damiano Milanato. Demand planning. processi, metodologie e modelli matematici per la gestione della domanda commerciale. In *Capitolo 8 - Regole Associative*. Springer Verlag, 2008.

[51] Naresh K. Malhotra. *Marketing Research: an applied orientation.* Pearson Prentice Hall, 2007.

[52] A. Parasuraman, Dhruv Grewal, and R. Krishnan. *Marketing Research.* Houghton Mifflin College, 2004.

[53] FADI THABTAH. A review of associative classification mining. *Knowl. Eng. Rev.*, 22(1):37–65, 2007.

[54] Zijian Zheng, Ron Kohavi, and Llew Mason. Real world performance of association rule algorithms. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 401–406, New York, NY, USA, 2001. ACM.

[55] Geoffrey I. Webb. Efficient search for association rules. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–107, New York, NY, USA, 2000. ACM.