

# Data Mining per la Business Intelligence

Casi di studio

M. Nanni, KDD Lab, ISTI-CNR, Pisa

Draft 18-04-2006



# Indice

<b>1</b>	<b>Customer Segmentation</b>	<b>5</b>
1.1	Obiettivi . . . . .	5
1.2	Processo di KDD e algoritmi di DM utilizzati . . . . .	6
1.3	Data Preprocessing . . . . .	6
1.3.1	Selezione dei dati . . . . .	8
1.3.2	Preparazione/trasformazione dei dati . . . . .	8
1.4	Data Mining . . . . .	11
1.5	Analisi dei risultati . . . . .	12
1.6	Deployment: profiling dei cluster . . . . .	15
<b>2</b>	<b>Fraud Detection</b>	<b>23</b>
2.1	Obiettivi . . . . .	23
2.2	Dati di origine . . . . .	23
2.3	Modello dei costi . . . . .	24
2.4	Preparazione dei dati . . . . .	25
2.5	Costruzione del modello predittivo . . . . .	25
2.6	Valutazione del modello . . . . .	27
2.7	Conclusioni . . . . .	33
<b>3</b>	<b>Competitive Intelligence</b>	<b>35</b>
3.1	Gli obiettivi . . . . .	35
3.2	I dati di input . . . . .	36
3.3	Analisi relazionale dei dati . . . . .	37
3.4	Un esempio applicativo . . . . .	39



# Capitolo 1

## Customer Segmentation: caso di studio AMRP

In questo capitolo riassumeremo metodologie seguite e risultati ottenuti in un caso di studio svolto dalla compagnia Loyalty Consulting – partner canadese di IBM – per migliorare il *CRM* (Customer Relationship Management) di una banca. Lo strumento di data mining utilizzato in questa esperienza è l'IBM Intelligent Miner versione 2.

### 1.1 Obiettivi

Ogni metodologia di CRM, finalizzata allo svolgimento di un marketing ottimale nei confronti della clientela, è caratterizzata da numerosi aspetti in qualche modo quantificabili, che includono almeno: profitto indotto dal cliente, valore del cliente a lungo termine, fedeltà del cliente. Un punto fermo in questo senso, accettato da chiunque gestisca le attività di CRM è il fatto che non tutti i clienti siano uguali dal punto di vista dell'azienda, e quest'ultima deve concentrare i propri sforzi nel trattenere quelli che già sono i clienti migliori, aumentando allo stesso tempo il profitto indotto da quelli che potenzialmente possono diventarlo e limitando le risorse sprecate nel trattare la clientela meno promettente. Di conseguenza, per il CRM diventa fondamentale possedere strumenti che aiutino a formulare questa differenziazione della clientela, e tra questi lo strumento strumento base risulta essere la segmentazione della clientela, naturalmente basata sulle caratteristiche del cliente centrali al lavoro di CRM.

Nel caso di studio qui descritto, l'obiettivo della banca è quello di creare una segmentazione avanzata della clientela che consenta una miglior comprensione del comportamento dei clienti. Inoltre, tale segmentazione dovrà poi essere confrontata con quella già posseduta dalla banca e creata attraverso una analisi RFM (Recency-Frequency-Monetary) standard.

La precedente attività di analisi della banca si è molto concentrata sull'attività di warehousing, risultando in un data warehouse pulito e già contenente tutte

le informazioni necessarie per le analisi richieste. In particolare, è stato possibile derivare quelle variabili che, ritenute di maggior interesse per gli analisti della banca stessa, misurano secondo diversi punti di vista il valore del cliente. Tali variabili sono essenzialmente ottenute combinando le grandezze tipicamente in uso negli approcci RFM e descritte in maggior dettaglio in Sezione 1.3.2.

## 1.2 Processo di KDD e algoritmi di DM utilizzati

I passi seguiti sono sinteticamente illustrati in Figura 1.1 e, come lo schema mostra chiaramente, sono stati adottati due diversi approcci al clustering per ricavare la segmentazione cercata, i cui risultati sono poi stati confrontati. Il primo approccio consiste nell'utilizzo di un algoritmo di clustering demografico che richiede di discretizzare le variabili quantitative in gioco, mentre il secondo approccio è basato su clustering a reti neurali direttamente applicabili a variabili continue.

Le fasi del processo di analisi sono le seguenti:

- Definizione di valore del cliente (per l'azienda ed i suoi azionisti).
- Selezione e preparazione dei dati. Nel caso dell'approccio demografico, questo passo include la discretizzazione delle variabili continue.
- Clustering (demografico o neurale).
- Analisi dei risultati del clustering.
- Classificazione dei cluster tramite alberi di decisione, onde ottenerne una caratterizzazione.
- Confronto dei due risultati ottenuti.
- Selezione di cluster e/o segmenti specifici per analisi più approfondita.

Nel caso specifico di questa esperienza applicativa, l'approccio demografico ha fornito risultati più significativi e quindi nel seguito verrà discusso solo questo.

## 1.3 Data Preprocessing

In questa fase rientra sia la selezione del sottinsieme di dati interessanti (ai fini dell'analisi) contenuti del data warehouse, che la preparazione dei dati, inclusiva di pulizia e calcolo di variabili derivate.

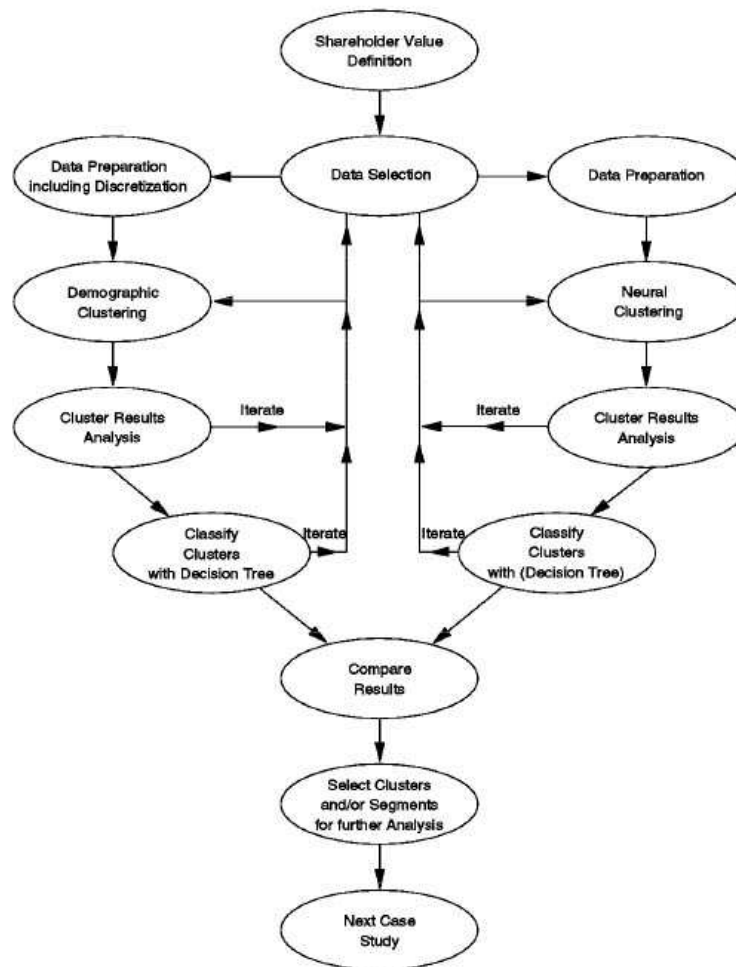


Figura 1.1: Processo seguito per la segmentazione della clientela

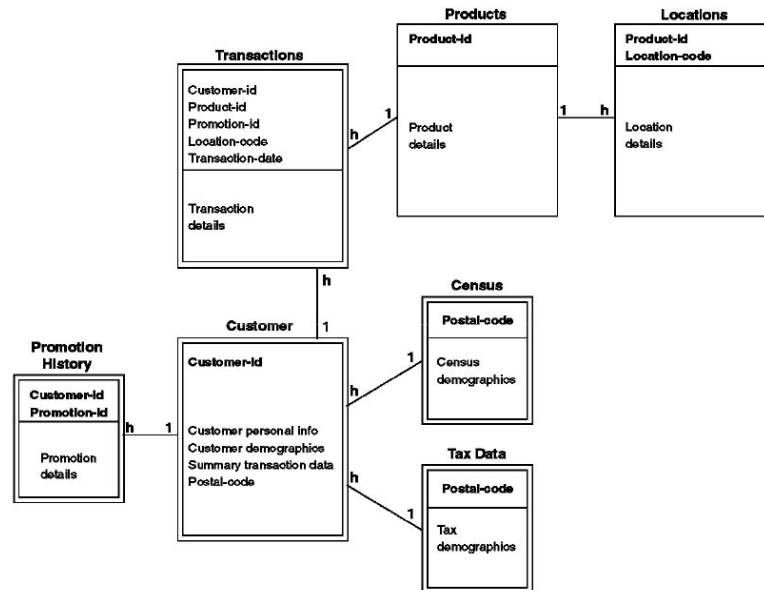


Figura 1.2: Modelli dei dati seguito

### 1.3.1 Selezione dei dati

Il modello dei dati seguito, illustrato in Figura 1.2, copre sostanzialmente due aree: da una parte le caratteristiche del cliente, soprattutto in termini di anagrafica, e dall'altra la sua attività con l'azienda, in termini di transazioni effettuate con relative informazioni sui prodotti acquistati/servizi usufruiti.

In particolare, il campione dei dati utilizzato nell'analisi di seguito descritta copre 50.000 clienti e le loro transazioni lungo un periodo di 12 mesi.

### 1.3.2 Preparazione/trasformazione dei dati

**Pulizia.** La fase di pulizia dei dati in questa esperienza è stata resa pressoché banale dalla estrema pulizia del data warehouse di partenza, richiedendo solo il riempimento con degli zeri dei campi nulli relativi a clienti che non hanno effettuato alcuna transazione su un certo prodotto.

**Preparazione.** Conclusa la pulizia dei dati, si è passati alla loro preparazione, ovvero al loro arricchimento con variabili derivate da quelle note. In particolare, sono state estratte due famiglie di attributi derivati:

- Variabili ad ampio spettro aventi un significato piuttosto generico e prive di obiettivi specifici. Tra queste si includono: totali di transazioni, totali su singoli quadrimestri, nonché differenze e rapporti tra queste.



- Variabili che descrivono il valore del cliente (*shareholder value*, ovvero valore per l'azionista), individuate e richieste esplicitamente dagli analisti della banca e sulle quali dovrà basarsi la segmentazione dei clienti. Più precisamente, sono state utilizzate le seguenti:
  - Numero di prodotti usati dal cliente nel corso della propria vita (sottinteso: vita di cliente)
  - Numero di prodotti usati dal cliente negli ultimi 12 mesi
  - *Revenue* (introito lordo) indotto dal cliente nel corso della propria vita
  - *Revenue* indotto dal cliente negli ultimi 12 mesi
  - Credito più recente del cliente
  - *Tenure* del cliente (lunghezza del rapporto con l'azienda, dalla data della prima operazione a quella della più recente), espressa in mesi
  - Rapporto N. Prodotti / *Tenure*
  - Rapporto Revenue / *Tenure*
  - *Recency* (inverso del tempo trascorso dall'ultima operazione ad oggi)

**Trasformazione.** Due tipi di trasformazione sono stati applicati alle variabili continue: una discretizzazione di tutte variabili continue per consentire l'uso del clustering demografico e una trasformazione logaritmica di alcune di esse per agevolare l'utilizzo degli algoritmi neurali.

La discretizzazione è stata effettuata scegliendo in modo arbitrario, guidato dall'esperienza, i seguenti quantili: 10, 25, 50, 75, 90 e ricavandone i 6 corrispondenti intervalli di valori. Tali fasce, poi, sono state aggiustate manualmente esaminando le distribuzioni che ne scaturivano e cercando di ottenere distribuzioni unimodali (quindi con un singolo picco) o quanto meno monotone. Ciò allo scopo di ottenere poi risultati più facilmente interpretabili. Infine, è stata effettuata una analisi di correlazione delle variabili, eliminando quelle che risultano essere ridondanti in quanto fortemente correlate ad altre. I risultati della discretizzazione si possono osservare e confrontare con le variabili originali in Figura 1.3. In particolare, si può notare come per la maggior parte degli attributi numerici si siano così risolti problemi di multi-modalità (esempio: NUMPROD12 diventata NPROD12) ed eccessiva sparsità (esempio: REVENUE12 diventata REVENUEL12).

Infine, si è operata una trasformazione logaritmica per standardizzare alcune variabili numeriche dalla distribuzione irregolare, onde consentirne un più efficace uso da parte degli algoritmi di clustering neurale. Come mostrato in Figura 1.4, i dati così trasformati sono molto meno sparsi e più facilmente visualizzabili degli originali (vedi Figura 1.3 precedente).

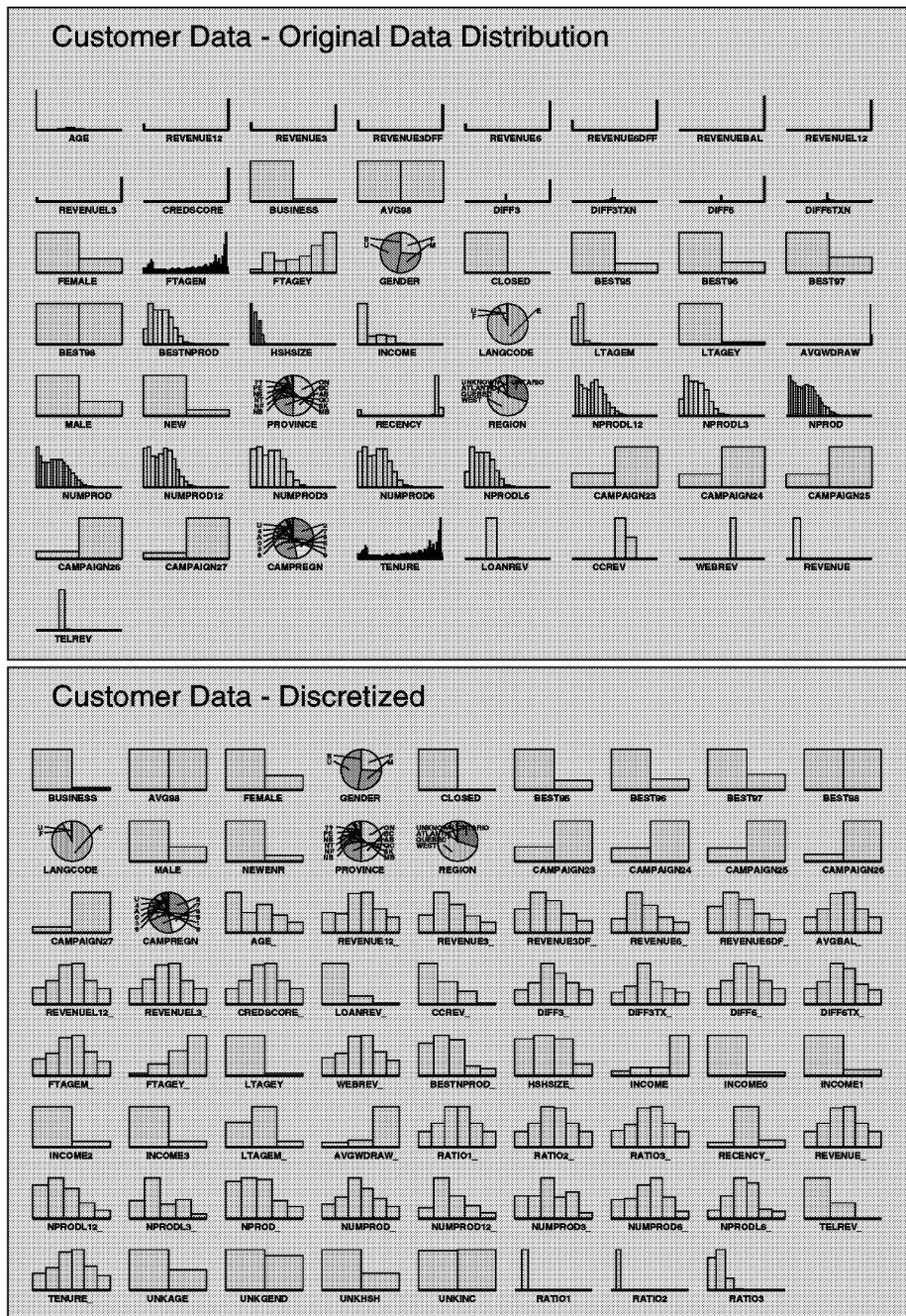


Figura 1.3: Dati prima e dopo la discretizzazione

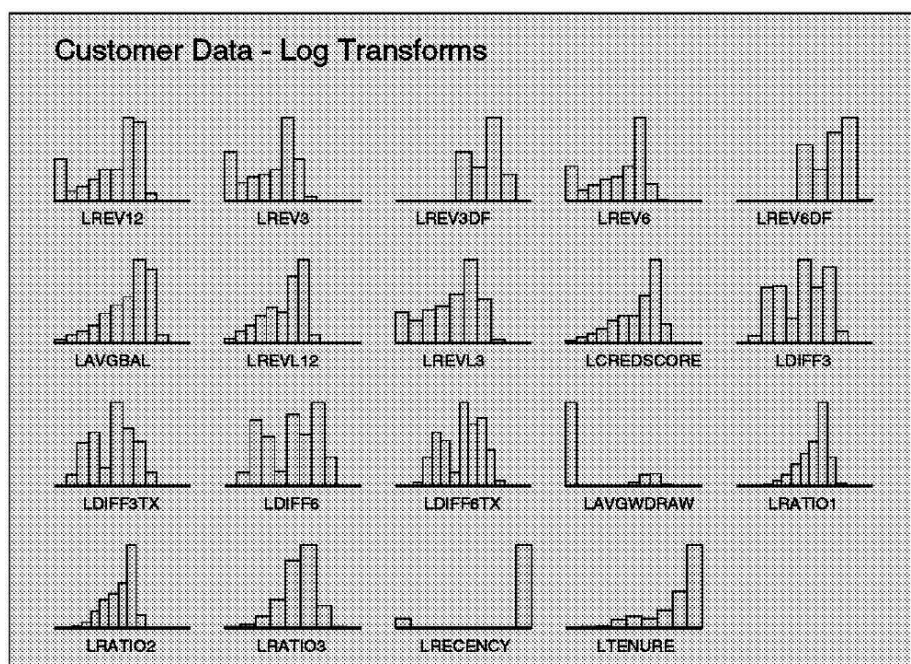


Figura 1.4: Attributi continui dopo la trasformazione logaritmica

## 1.4 Data Mining

L'applicazione degli algoritmi di clustering richiede di specificare diversi parametri, inclusa la lista degli attributi da utilizzare attivamente nella generazione dei cluster.

**Parametri.** L'algoritmo di clustering demografico, parte integrante di IBM Intelligent Miner, richiede i seguenti parametri di base: numero massimo di cluster da estrarre (non necessariamente raggiunto dall'algoritmo), numero massimo di iterazioni (ovvero scansioni del dataset), accuratezza (ovvero una misura della variazione della configurazione di cluster riscontrata tra due iterazioni consecutive: una variazione inferiore alla accuratezza specificata indica il raggiungimento di una configurazione essenzialmente stabile) e soglia di similarità (elementi aventi similarità superiore a tale soglia sono considerati uguali). Senza entrare nei dettagli dell'algoritmo, notiamo che un alto numero di scansioni e una bassa accuratezza portano a risultati più precisi ma anche a computazioni molto più lente. Inoltre, nel contesto bancario in cui ci troviamo, si è ritenuto ragionevole utilizzare un numero massimo di cluster – ovvero segmenti di clientela – pari a 9. Un numero più alto porterebbe ad una segmentazione difficilmente gestibile all'atto pratico. Infine, l'algoritmo neurale richiede due soli parametri di base, equivalenti al numero massimo di cluster ed al numero di passi da eseguire.

**Attributi di input.** Intelligent Miner consente di dividere gli attributi di input (già elencati nella Sezione 1.3.2) in due gruppi, quelli attivi e quelli inattivi o ausiliari. I primi vengono utilizzati dall'algoritmo per estrarre i cluster, mentre i secondi vengono usati soltanto per fornire una descrizione a posteriori del cluster. Secondo le specifiche fornite dai committenti (ovvero gli analisti della banca), le linee guida nella formazione di segmenti di clientela devono essere le misure del valore del cliente, per cui solo esse sono state utilizzate come attributi attivi, mentre gli altri attributi (aggregati, differenze e rapporti di vario tipo) sono servite come ausilio alla descrizione dei cluster ottenuti.

## 1.5 Analisi dei risultati

Esaminiamo qui in dettaglio i risultati ottenuti dopo diverse iterazioni dell'algoritmo demografico e di quello neurale. In Figura 1.5 viene visualizzato l'output riassuntivo offerto da Intelligent Miner per il clustering demografico.

Ogni fascia orizzontale del grafico rappresenta un cluster, il cui ID è indicato sull'estrema destra della fascia e la cui popolosità (in percentuale rispetto all'intera popolazione) è indicata a sinistra. Gli attributi sono poi ordinati da sinistra verso destra per importanza rispetto al cluster (quello più a sinistra è l'attributo che più discrimina un cluster dagli altri). Gli attributi inattivi, ovvero che non hanno contribuito alla creazione dei cluster, hanno il nome tra parentesi quadrate. Infine, per ogni variabile si hanno due grafici sovrapposti (istogrammi per le variabili continue, diagrammi a torta per gli altri): quelli in grigio scuro sullo sfondo indicano la distribuzione della variabile nell'intero dataset, gli altri, in colore più chiaro e in primo piano, indicano la distribuzione limitatamente al singolo cluster.

I risultati indicano che ci sono almeno 9 cluster nei dati (forse di più, ma 9 era limite massimo che abbiamo impostato), sono distribuiti in modo ragionevole (non c'è un solo grande cluster) e la distribuzione delle variabili all'interno dei cluster tende ad essere ben differenziata dalla distribuzione globale, rendendo i cluster stessi meglio differenziati dagli altri. Infine, le variabili Best98, Revenue e CreditScore risultano essere importanti per molti cluster.

Per confronto, Figura 1.6 riporta il riassunto dei cluster ottenuti con l'algoritmo neurale.

Le caratteristiche più evidenti di quest'ultimo risultato sono le seguenti: i cluster sono distribuiti in modo meno uniforme rispetto ai cluster demografici, le variabili più importanti per i cluster sono più o meno le stesse nei due risultati e le variabili discretizzate risultano essere le più importanti anche con l'approccio neurale. Vista la somiglianza dei due risultati e la qualità leggermente migliore del clustering demografico, nel resto si discuterà nel dettaglio solo quest'ultimo.

**Analisi dettagliata dei cluster.** Un primo risultato importante consiste nel fatto che la variabile booleana Best98 appaia come importante per molti dei cluster ottenuti, mostrando distribuzioni interne ai cluster molto sbilanciate verso il TRUE (= cluster con maggioranza di clienti Best98) o verso il FALSE (= cluster con quasi nessun cliente Best98). Infatti, Best98 è una variabile prece-





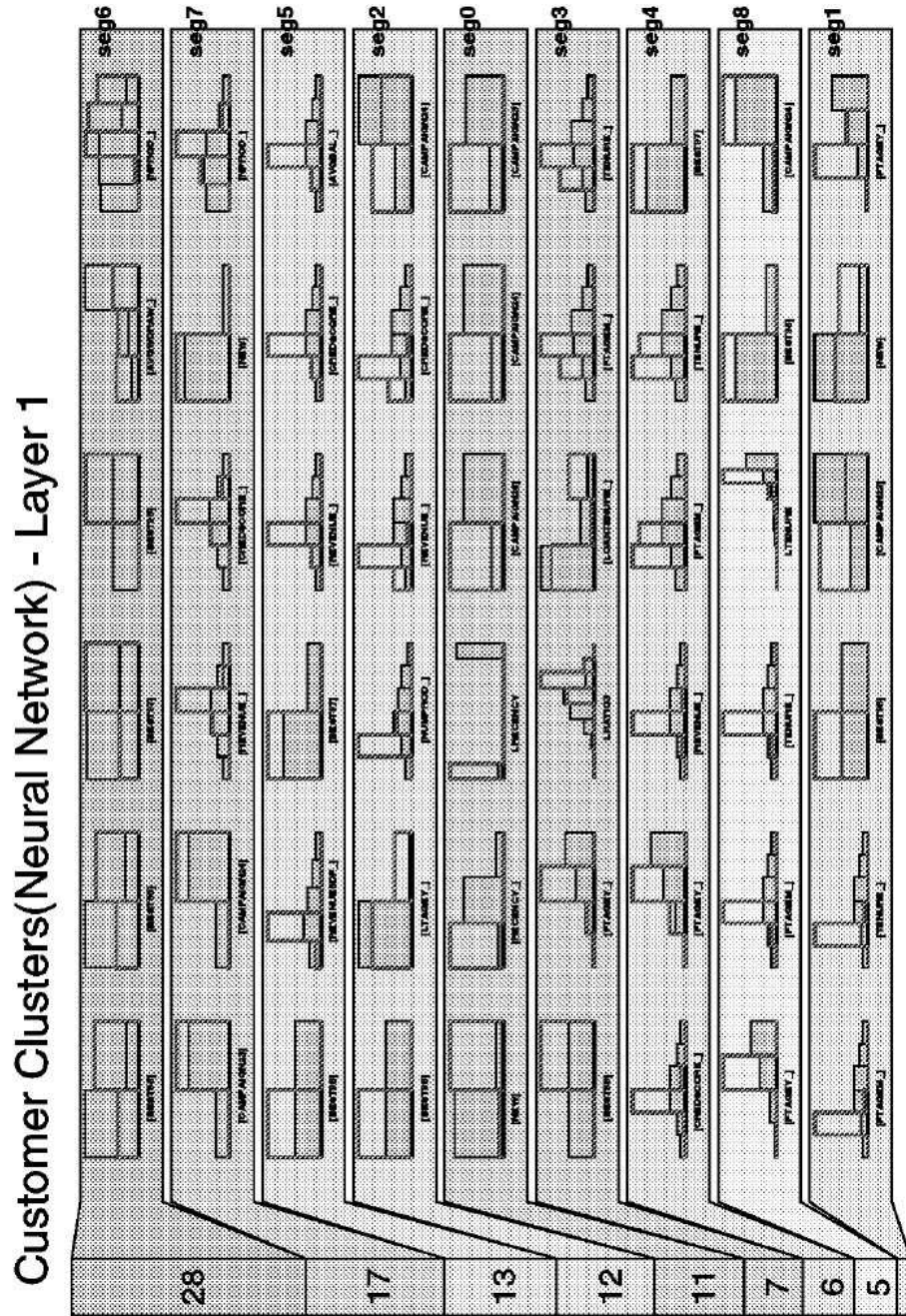


Figura 1.6: Output clustering neurale

dentemente calcolata dagli analisti della banca con mezzi e strumenti specifici, indicante i clienti migliori nel database. Considerando che tale attributo è stato utilizzato come non attivo, e quindi non ha preso parte alla costruzione dei cluster, abbiamo sostanzialmente ottenuto a costo zero una divisione in segmenti di clientela molto coerente con questa ripartizione di clienti preesistente e con la quale gli analisti della banca hanno già confidenza. Confermare la conoscenza già posseduta dagli esperti del dominio è un primo importante passo per guadagnare la fiducia dell'utente nelle nuove soluzioni proposte, come appunto quelle basate su data mining.

Basandoci sulle misure di valore del cliente (attributi non attivi del clustering), e in particolar modo su Best98, forniamo qui una caratterizzazione dei cluster più interessanti:

**Cluster 6 (Figura 1.7):** Questo può essere interpretato come il cluster popolato solo (o quasi) da clienti Best98, il cui credito, revenue negli ultimi 12 mesi, revenue mensile e numero di prodotti usati mensilmente sono nel 50-imo e 75-imo percentile, ovvero medio alti. Il cluster copre il 24% della popolazione circa.

**Cluster 3 (Figura 1.8):** Quasi tutti i clienti sono non-Best98, e i loro revenue, credito, revenue negli ultimi 12 mesi, revenue mensile e numero di prodotti usati mensilmente sono nel 25-imo e 50-imo percentile, ovvero medio-bassi. Il cluster rappresenta il 23% della popolazione.

**Cluster 5 (Figura 1.9):** I valori di revenue, credito e numero di prodotti mensile sono tutti almeno al 75-imo percentile e quasi tutti al 90-imo percentile. Inoltre, Best95, Best96 e Best97, analoghe di Best98 calcolate nei tre anni precedenti, mostrano una crescita progressiva di clienti Best nei tre anni. Di conseguenza, questo risulta essere un cluster di clienti ad alto profitto attuale ed in crescita. Esso copre il 9% della popolazione.

**Cluster 1 (Figura 1.10):** Questo rappresenta decisamente il cluster dei nuovi clienti, dato che la variabile NEW (indicante appunto se il cliente è nuovo o meno) ha grande importanza ed è fortemente spostata sul valore TRUE nel cluster. Inoltre mostra valori bassi di recency (ovvero nessuna transazione recente) e di tenure, forse perché non hanno ancora iniziato ad utilizzare il conto da poco aperto presso la banca. Può quindi essere interessante monitorare il progresso di questi clienti nel futuro.

## 1.6 Deployment: profiling dei cluster

Concludiamo la descrizione di questo caso di uso con una esemplificazione di quella che è l'ultima fase del processo di knowledge discovery: l'utilizzo delle informazioni estratte per pianificare azioni (in questo caso di marketing) mirate.

Un punto di riferimento per tali pianificazioni può essere il cruscotto in Figura 1.11, che indica per ognuno dei cluster trovati alcune informazioni facilmente

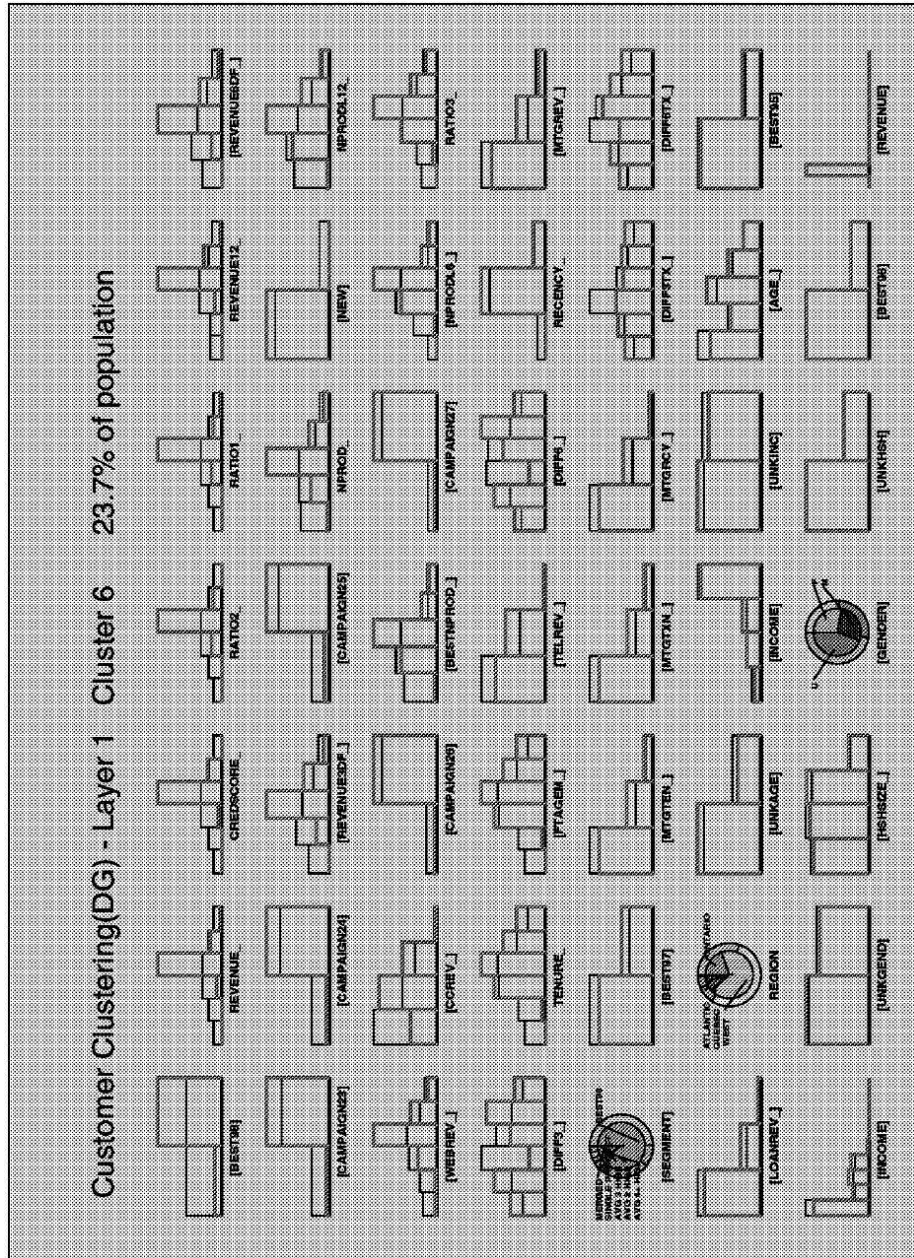


Figura 1.7: Dettaglio del cluster 6





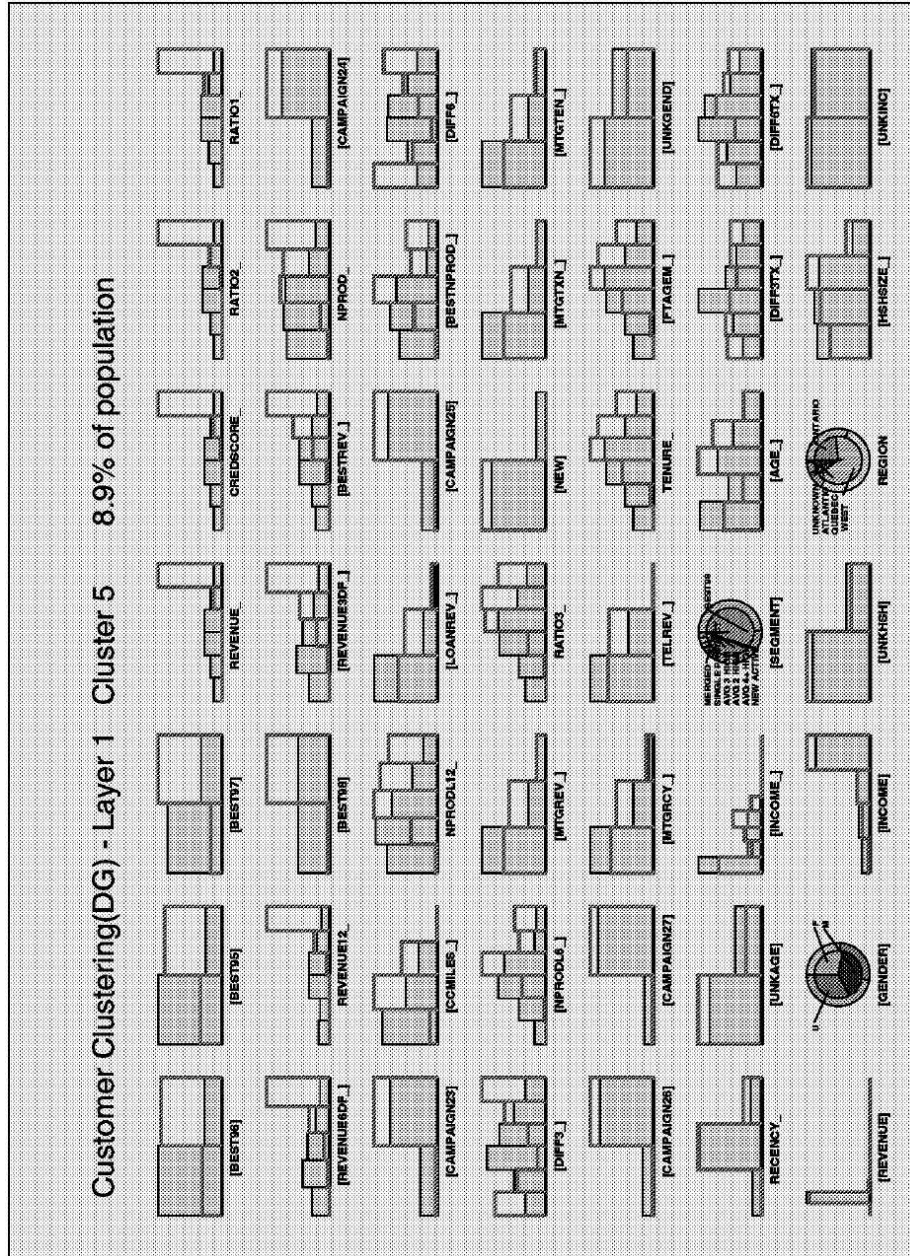


Figura 1.9: Dettaglio del cluster 5

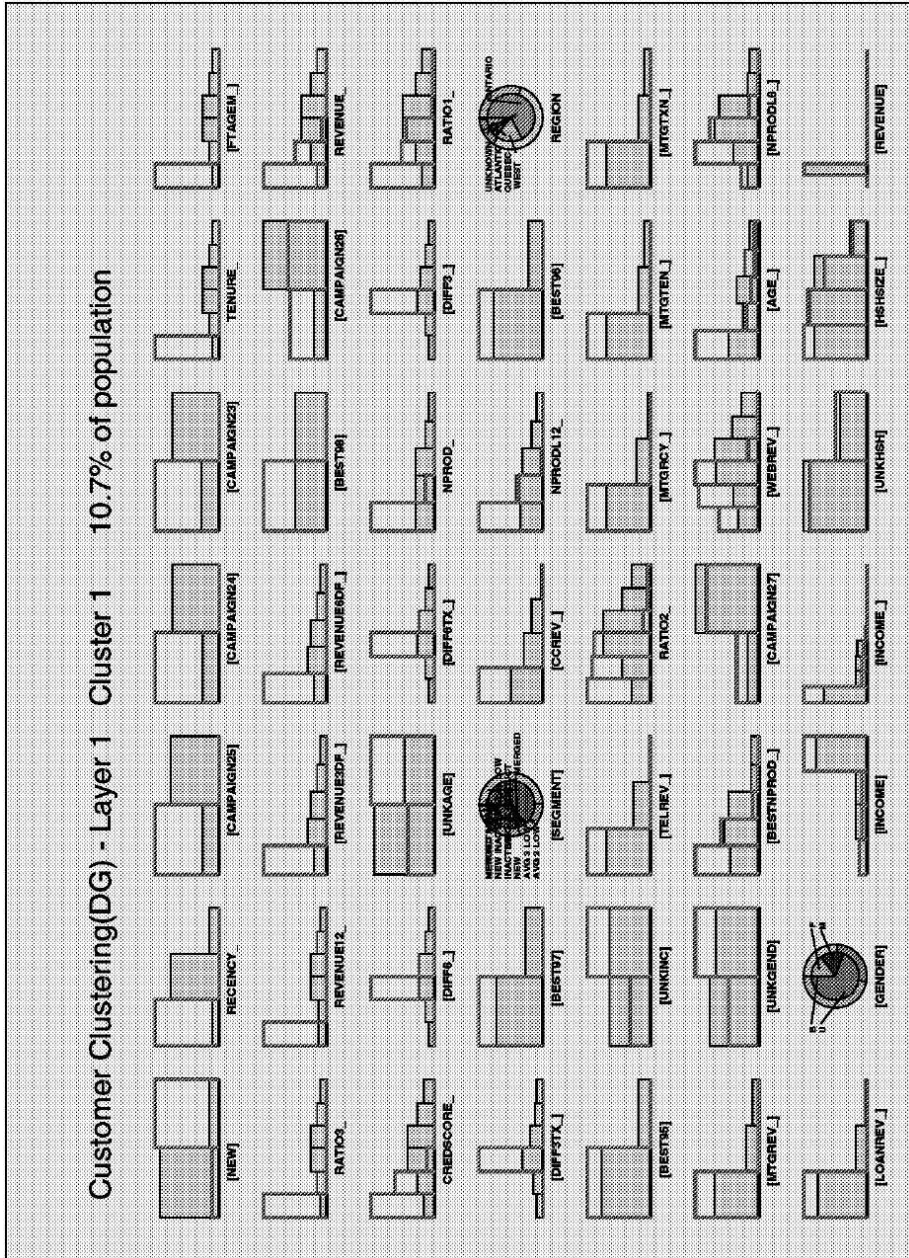


Figura 1.10: Dettaglio del cluster 1

*Table 1. Customer Revenue by Cluster*

Cluster ID	Revenue	Customer	Product index	Leverage	Tenure
5	34.74%	8.82%	1.77	3.94	60.92
6	26.13%	23.47%	1.41	1.11	57.87
7	21.25%	10.71%	1.64	1.98	63.52
3	6.62%	23.32%	0.73	0.28	47.23
0	4.78%	3.43%	1.45	1.40	31.34
2	4.40%	2.51%	1.46	1.75	61.38
4	1.41%	2.96%	0.99	0.48	20.10
8	0.45%	14.14%	0.36	0.03	30.01
1	0.22%	10.64%	0.00	0.02	4.66

Figura 1.11: Revenue, prodotti acquistati, tenure nei 9 cluster

estraibili dal data warehouse: revenue complessiva del cluster in percentuale rispetto al totale, popolarità del cluster (Customer), rapporto tra il numero medio di prodotti acquistati nel cluster e il numero medio globale (Product index), il rapporto Revenue/Customer (Leverage) e la tenure.

Si nota immediatamente che il cluster 5 è il più proficuo, in quanto porta il 34% dell'introito con solo il 9% della popolazione, risultante in un alto valore di leverage. Inoltre, si può anche notare che tale proficuità aumenta sia con l'aumentare del tenure che del numero di prodotti acquistati.

Dalla tabella appena descritta si possono derivare diverse strategie di business, partendo dalla semplice osservazione che i migliori clienti sono attualmente nei cluster 2, 5 e 7, come indicato dal valore di leverage. Alcune possibili strategie ad alto livello sono le seguenti:

- Attuare strategie di retention per i clienti dei cluster 2, 5 e 7, in quanto sono i clienti migliori.
- Effettuare cross-selling per i cluster 2, 6 e 0 confrontandoli con i cluster 5 e 7: i due gruppi mostrano un product index simile, quindi sembra verosimile poter convertire i clienti di 2, 6 e 0 in clienti ottimi, come sono quelli dei cluster 5 e 7. La strategia da seguire consiste nel confrontare i prodotti concretamente acquistati dai due gruppi di cluster, alla ricerca di prodotti mancanti nel primo gruppo, mancanze sulle quali si può forse agire tramite iniziative di cross-selling.
- Analogamente, si può effettuare cross-selling sui cluster 3 e 4 confrontandoli coi cluster 2, 6 e 0.

- Adottare una strategia di attesa per il cluster 1: essendo clienti nuovi non si posseggono informazioni sufficienti a pianificare azioni precise, ma in generale può essere utile fornire loro informazioni su prodotti e servizi al fine di renderli proficui più rapidamente.
- Infine, il cluster 8 sembra contenere solo persone che, nonostante la lunga permanenza presso la banca, portano pochi acquisti e pochi introiti, e quindi non è conveniente investire denaro in alcuna azione a loro diretta.

*Fonte: Intelligent Miner for Data. Applications Guide.*  
<http://www.redbooks.ibm.com>



## Capitolo 2

# Fraud Detection: Lotta all'evasione fiscale

In questo capitolo presentiamo brevemente i risultati ottenuti in un caso di studio su Fraud Detection, ovvero rilevazione/previsione di frodi. In particolare, questa attività è stata parte di un progetto, portato avanti con la collaborazione del Ministero delle Finanze, orientato alla verifica della adeguatezza e sostenibilità del knowledge discovery nella rilevazione dell'evasione fiscale. Lo strumento adottato è Angoss Knowledge Studio, ed in particolare il modello predittivo usato è un classificatore basato su alberi di decisione.

### 2.1 Obiettivi

L'obiettivo generale in questo caso di studio è il miglioramento delle strategie di scelta delle verifiche fiscali da porre in atto. Infatti tali strategie devono rispettare alcuni vincoli sulle risorse disponibili, umane e finanziarie, per cui devono affrontare due problemi contrastanti:

- Massimizzare gli introiti derivanti dalle verifiche, ovvero selezionare i soggetti da verificare in modo da massimizzare il recupero delle tasse evase.
- Minimizzare i costi dovuti alle verifiche stesse, ovvero selezionare i soggetti da verificare in modo da minimizzare le risorse necessarie ad eseguirle concretamente.

### 2.2 Dati di origine

I dati utilizzati nel caso di studio consistono di informazioni provenienti da dichiarazioni dei redditi integrate con altre sorgenti, quali bollette di telefono ed elettricità, contributi previdenziali per gli impiegati, ecc. Ogni riga nel dataset

corrisponde ad una compagnia (di seguito riferito semplicemente come il *soggetto*) di medie o grandi dimensioni che ha sottomesso una dichiarazione dei redditi in un certo periodo di tempo. Il dataset di partenza consiste di 80643 tuple di 175 attributi numerici, di cui un numero esiguo di tipo categorico (essenzialmente codici numerici). Di queste, 4103 tuple corrispondono a soggetti sottoposti a verifica: il risultato è registrato in un dataset separato contenente 4103 tuple e 7 attributi. In particolare uno di tali attributi è detto *recovery* e rappresenta l'ammontare delle tasse evase accertato dalla verifica, pari a zero se nessuna frode è stata riscontrata.

Si fa notare che i soggetti sottoposti a verifica sono stati selezionati dall'autorità fiscale seguendo una procedura ben formalizzata e basata sulla conoscenza degli esperti in materia. Analogamente, quindi, per poter risultare accettabili dall'autorità stessa, i modelli predittivi ottenuti nel caso di studio qui descritto dovevano essere il risultato di un processo di analisi chiaro e deterministico, nonché replicabile su altri dati di partenza di uguale tipo.

## 2.3 Modello dei costi

L'identificazione di un modello dei costi associato alle verifiche fiscali costituisce una parte fondamentale del caso di studio. Infatti, le verifiche fiscali sono operazioni molto costose e quindi è importante riuscire a orientare tali verifiche principalmente verso soggetti che porteranno verosimilmente ad un alto introito.

Il modello dei costi adottato può essere spiegato molto semplicemente per mezzo di un esempio nel contesto di una compagnia di vendite per corrispondenza. In questo esempio si assume di proporre una promozione del tipo chi risponde riceve un regalo, e si assegna un costo o un profitto alle possibili reazioni dei clienti:

- Se il cliente risponde facendo un ordine si ha un profitto pari a: valore dell'ordine - spese postali - costo del regalo.
- Se il cliente senza alcun ordine c'è solo un costo pari a: spese postali + costo del regalo.
- Se il cliente non risponde c'è sol un costo pari alle spese postali.

In accordo con questo approccio si è prima definito un nuovo attributo - *audit\_cost* derivato dagli altri attributi. Esso rappresenta una stima fornita dagli esperti del costo di una verifica fiscale in rapporto alla dimensione e alla complessità del soggetto da verificare. Di conseguenza si è definito un ulteriore attributo *actual\_recovery*:

$$actual\_recovery(i) = recovery(i) - audit\_cost(i)$$

rappresentante il recupero di tasse al netto dei costi. L'obiettivo è di utilizzare tale quantità non solo per valutare a posteriori la bontà di un modello predittivo, ma anche per discriminare i casi positivi e negativi all'atto di costruire i modelli



stessi. Più precisamente, la variabile target nell'analisi svolta viene derivata da `actual_recovery` nel seguente modo:

$$car(i) = \begin{cases} negative & \text{se } actual\_recovery(i) \leq 0 \\ positive & \text{se } actual\_recovery(i) > 0 \end{cases}$$

La variabile `car` (class of actual recovery) indica quali sono i soggetti da considerare casi positivi nella costruzione del modello predittivo. Si noti che non vengono qui presi in considerazione i vincoli sulle risorse che limitano il numero di verifiche realisticamente effettuabili, vincoli che in qualche misura verranno considerati nelle successive fasi del caso di studio.

## 2.4 Preparazione dei dati

La raccolta ed integrazione delle diverse sorgenti da cui provenivano i dati necessari all'analisi effettuata costituiscono un passo estremamente costoso in termini di tempo, principalmente a causa della eterogeneità delle sorgenti di dati, la loro mole e l'inconsistenza di alcune unità di misura e scale adottate. Tali operazioni vanno al di là degli obiettivi di questa descrizione, per cui non verranno qui discusse nel dettaglio.

Al fine di agevolare la costruzione dei modelli predittivi sono state eliminate quelle tuple che costituivano del rumore, ovvero quelle tuple contenenti valori di attributi eccessivamente devianti dalla norma e quelle contenente troppi valori nulli. Tale operazione ha portato alla selezione di 3880 tuple, di cui l'82% di casi positivi e il 18% di casi negativi. Inoltre sugli attributi è stata effettuata un'analisi di significatività e di correlazione, onde identificare gli attributi privi di valori significativi e quelli essenzialmente derivati da altri e quindi ridondanti. Ciò ha portato alla eliminazione di numerosi attributi non utili, passando dai 175 di partenza a 20 attributi selezionati.

Essendo l'obiettivo finale la costruzione e validazione di modelli predittivi è necessario stabilire una appropriata ripartizione delle tuple di input in un training set (da usare per la costruzione dei modelli) e un test set (da usare per valutare la bontà dei modelli estratti). In particolare va preso in considerazione il pericolo di overfitting dei modelli predittivi che, in alcuni contesti, emerge quando il training set è troppo grande. Nel caso specifico tale pericolo è stato valutato eseguendo una serie di esperimenti in cui si utilizzano dimensioni del training set crescenti, pari al 10%, 20%, 33%, 50%, 66% e 90% dell'intero dataset di input. Il risultato di tali esperimenti conferma che nel caso specifico non c'è rischio di overfitting e dimensioni più grandi del training set portano a modelli predittivi migliori. Di conseguenza è stato ripartito il dataset in un training set di 3514 tuple e un training set di 366 tuple.

## 2.5 Costruzione del modello predittivo

Il modello richiesto in questa applicazione è un classificatore binario avente la variabile `car` come attributo target, e quindi in grado di distinguere tra soggetti

positivi (per i quali è conveniente effettuare la verifica) e soggetti negativi (per i quali una verifica porterebbe ad una perdita economica). Nel nostro caso è interessante, oltre alla accuratezza del classificatore, anche il valore complessivo di `actual_recovery` ottenuto applicando il classificatore al test set e sottoponendo a verifica tutti e soli i soggetti suggeriti dal classificatore. Tale valore, quindi, potrà essere confrontato con il recupero di tasse evase realmente ottenuto sul campo, ricavabile dalle 366 tuple del test set:

$$\begin{aligned} \text{actual\_recovery}(\text{real}) &= 159.6 \\ \text{audit\_cost}(\text{real}) &= 24.9 \end{aligned}$$

dove costi e recupero sono espressi in milioni di euro.

Seguendo questa direzione i classificatori ottenuti nei vari esperimenti sono stati valutati sulla base di varie metriche:

**matrice di confusione:** questa ricava una ripartizione del test set in quattro gruppi di tuple, di cui si forniscono le dimensioni:

- quelle correttamente classificate come casi positivi (true positive, o TP);
- quelle correttamente classificate come casi negativi (true negative, o TN);
- quelle erroneamente classificate come casi positivi (false positive, o FP);
- quelle erroneamente classificate come casi negativi (false negative, o FN);

**indice di misclassificazione:** rappresenta la percentuale di tuple del test set classificate in modo non corretto, ovvero, equivalentemente, la dimensione percentuale di  $FP \cup FN$ .

**actual\_recovery:** rappresenta il recupero complessivo di tasse evase, espresso come la somma degli `actual_recovery` di tutte le tuple del training set.

**audit\_cost:** rappresenta la somma dei costi di tutte le tuple del training set.

**profitability:** rappresenta il valore medio di `actual_recovery` per singola verifica effettuata, quindi pari al rapporto tra `actual_recovery` del classificatore (definito sopra) e numero di casi classificati come positivi.

**relevance:** variabile artificiale che mette in relazione `profitability` e indice di misclassificazione, calcolata come:

$$\text{relevance} = 10 \times \frac{\text{profitability}}{\text{indice\_misclassificazione}}$$

Notiamo in particolare che le prime due misure sono indipendenti dal dominio applicativo, le successive quattro sono dipendenti dal dominio e l'ultima mette in relazione una misura di un tipo con una dell'altro.

Nella costruzione del modello predittivo sono stati seguiti due approcci complementari. Nel primo ci si pone l'obiettivo di minimizzare i falsi positivi (FP), allo scopo di minimizzare le spese in verifiche fiscali inutili (in quanto la frode non sussiste). Nel secondo, l'obiettivo è minimizzare i falsi negativi (FN), allo scopo di perdere meno evasori possibile e quindi massimizzare il recupero di tasse evase. In situazioni reali questi due obiettivi sono in conflitto tra loro, dato che il primo tende a minimizzare l'impiego di risorse (per fare verifiche) anche a scapito di qualche evasore mancato, mentre il secondo tende a far grande uso di risorse. La soluzione ottimale consiste nel trovare un giusto compromesso tra le due. I seguenti parametri di tuning della costruzione di modelli forniscono il mezzo per raggiungere il suddetto compromesso:

- Livello di pruning: come già detto non ci sono problemi di overfitting, quindi è stato mantenuto un basso livello di pruning, pari al 10%.
- Pesì di misclassificazione: questi sono pesi assegnati ai due tipi di misclassificazione, FP e FN, usati dagli algoritmi di costruzione del modello. Più alto è un peso, minore sarà il numero di misclassificati del corrispondente tipo, per cui questo costituisce lo strumento fondamentale per determinare in quale misura minimizzare FP o FN.
- Replicazione della classe minoritaria: quando le due classi da discriminare hanno popolosità molto sbilanciate nel training set, la classe più frequente tende a prevalere – nel nostro caso la classe negativa – a volte facendo scomparire l'altra dal classificatore. Un modo per affrontare il problema consiste nel duplicare le tuple appartenenti alla classe minoritaria fino a che non si raggiunge un bilanciamento approssimativo.
- Adaptive boosting: tecnica che costruisce una sequenza di classificatori, in cui il  $k$ -esimo viene creato a partire dagli errori del  $(k-1)$ -esimo e una nuova tupla viene classificata con il voto di maggioranza di tutti i classificatori, pesati secondo la propria precisione. Questo porta tipicamente ad una riduzione sensibile del numero di tuple misclassificate.

## 2.6 Valutazione del modello

Tra gli esperimenti effettuati nel corso del caso di studio, descriviamo i seguenti quattro particolarmente significativi, relativi ad altrettanti classificatori ottenuti, cercando di valutarne la bontà in riferimento agli obiettivi preposti. Nei primi due casi viene seguita la politica minimizza FP (e quindi minimizza risorse sprecate), negli ultimi due, invece, viene seguita la politica minimizza FN (più dispendiosa ma che cattura più evasori).

### Classificatore A

In questo classificatore si accentua la politica minimizza FP semplicemente utilizzando il training set originale, il quale, contenendo soprattutto casi negativi,

tende a classificare anche le nuove tuple come negative, riducendo così indirettamente il rischio di assegnare erroneamente etichette positive. Di conseguenza, non sono stati utilizzati pesi di misclassificazione per indurre una politica rispetto all'altra. Per ridurre la misclassificazione globale è stato applicato un adaptive boosting con 10 classificatori (numero scelto dopo diversi tentativi), ottenendo la seguente matrice di confusione:

Negative	Positive	← Classe assegnata
237	11	Classe reale = Negative
70	48	Classe reale = Positive

Il classificatore, quindi, suggerisce 59 verifiche di cui 11 inutili, ed ottiene i seguenti valori degli indicatori:

- indice di misclassificazione = 22% (81 errori)
- actual\_recovery = 141.7 MEuro
- audit\_cost = 4 MEuro
- profitabilty = 2.401 MEuro/verifica
- relevance = 1.09

Il modello risulta avere una alta profitability, e se confrontato con i valori del caso reale, si osserva un recupero dell'88% con solo il 16% delle verifiche, coerentemente con la politica al risparmio che si voleva indurre. Questo è sintetizzato anche dal grafico comparativo in Figura 2.1.

## Classificatore B

In questo classificatore la politica minimizza FP viene portata all'estremo, utilizzando sia il training set originale che pesi di misclassificazione che fanno pesare di più gli errori di tipo FP. Più esattamente agli errori FP viene dato peso doppio rispetto a quelli FN. Inoltre, anche qui è stato applicato un adaptive boosting, questa volta con 3 classificatori (numero sempre scelto dopo diversi tentativi), ottenendo la seguente matrice di confusione:

Negative	Positive	← Classe assegnata
246	2	Classe reale = Negative
108	10	Classe reale = Positive

Il classificatore suggerisce solo 12 verifiche di cui 2 inutili, ed ottiene i seguenti valori degli indicatori:

- indice di misclassificazione = 30% (110 errori)
- actual\_recovery = 15.5 MEuro

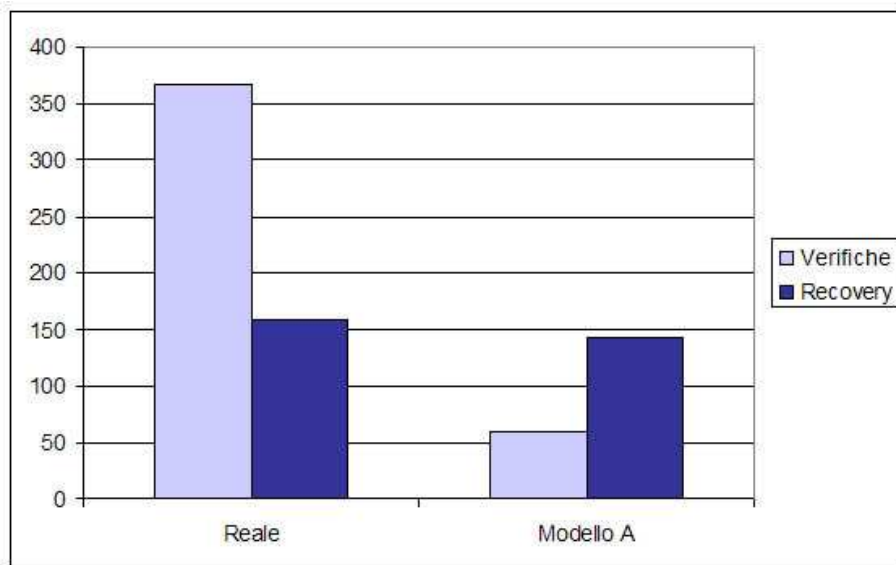


Figura 2.1: Performance modello A

- $\text{audit\_cost} = 1.1 \text{ MEuro}$
- $\text{profitabilty} = 1.291 \text{ MEuro/verifica}$
- $\text{relevance} = 0.43$

Il grafico comparativo in Figura 2.2 mostra i valori di recovery e numero di verifiche effettuate.

Come è evidente, le pochissime verifiche suggerite portano anche a mancare molti evasori, rendendo il modello utile in quei casi in cui le risorse a disposizione sono limitatissime.

### Classificatore C

In questo caso si adotta la politica minimizza FN, cercando di condizionare il classificatore in favore di classificazioni come casi positivi: le tuple positive sono state duplicate fino a raggiungere un bilanciamento tra le due classi nel training set, e sono stati adottati dei pesi di misclassificazione che assegnano agli errori FN un peso triplice rispetto a quelli FP. Anche qui è stato applicato un adaptive boosting con 3 classificatori ottenendo la seguente matrice di confusione:

Negative	Positive	← Classe assegnata
150	98	Classe reale = Negative
28	90	Classe reale = Positive

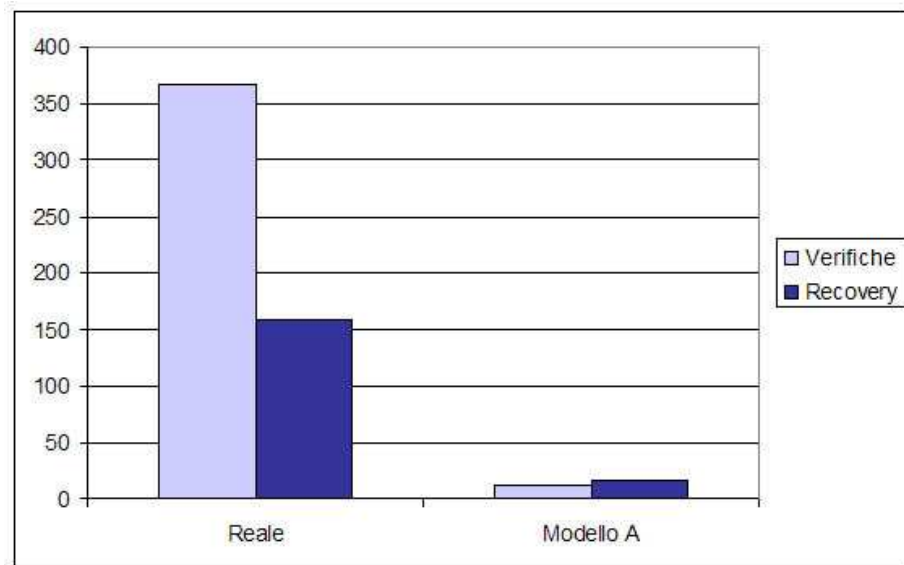


Figura 2.2: Performance modello B

Il classificatore suggerisce 188 verifiche di cui più della metà inutili, ed ottiene i seguenti valori degli indicatori:

- indice di misclassificazione = 34% (126 errori)
- actual\_recovery = 165.2 MEuro
- audit\_cost = 12.9 MEuro
- profitabilty = 0.878 MEuro/verifica
- relevance = 0.25

Il grafico comparativo in Figura 2.3 mostra i valori di recovery e numero di verifiche effettuate.

Sorprendentemente il recupero netto del modello C è superiore al caso reale con solo il 50% delle verifiche effettuate. Questo è dovuto naturalmente ai tanti casi TP (soggetti fraudolenti scoperti) ottenuti. Si noti come la profitability sia comunque molto più bassa dei modelli A e B, quanto essi si concentravano su pochi soggetti altamente proficui.

### Classificatore D

Come per il classificatore C, si adotta la politica minimizza FN bilanciando le due classi nel training set, questa volta adottando dei pesi di misclassificazione

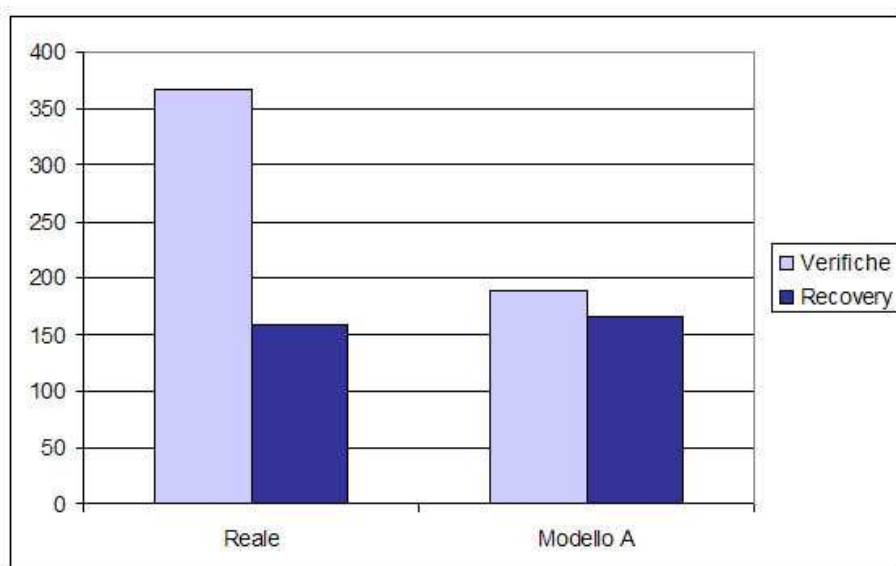


Figura 2.3: Performance modello C

che assegnano agli errori FN un peso quadruplo (anziché triplice come nel classificatore C) rispetto a quelli FP. In questi caso non è stato applicato alcun boosting. La matrice di confusione ottenuta è la seguente:

Negative	Positive	← Classe assegnata
135	113	Classe reale = Negative
21	97	Classe reale = Positive

Il classificatore suggerisce 210 verifiche di cui circa metà inutili, ed ottiene i seguenti valori degli indicatori:

- indice di misclassificazione = 36.6% (126 errori)
- actual\_recovery = 163.5 MEuro
- audit\_cost = 14.4 MEuro
- profitabilty = 0.778 MEuro/verifica
- relevance = 0.21

Rispetto al modello C gli FN diminuiscono, e quindi si catturano più evasori, ma il maggior numero di FP (verifiche effettuate a vuoto) compensano il guadagno conquistato, portando il recupero globale ad una piccola riduzione. Il grafico comparativo in Figura 2.4 mostra i valori di recovery e numero di verifiche effettuate.

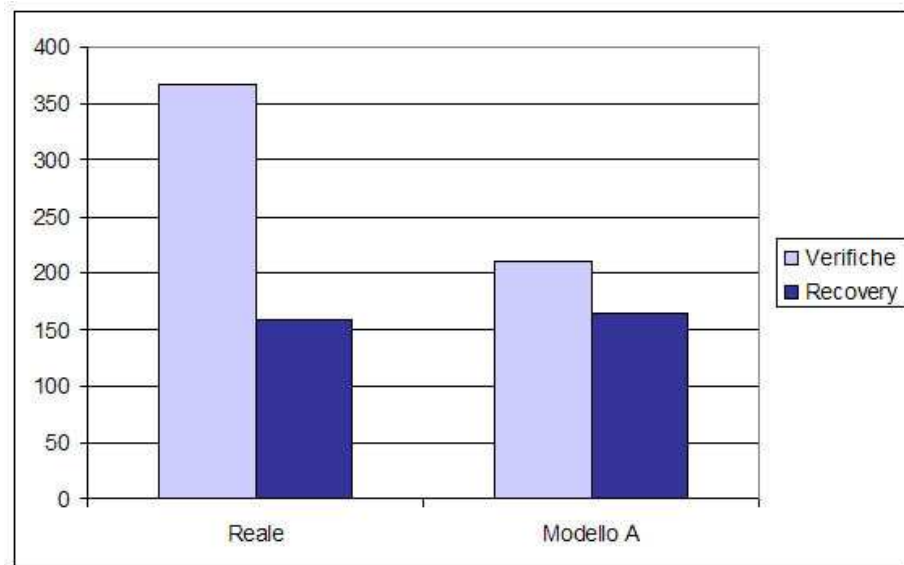


Figura 2.4: Performance modello D

### Combinazione di Classificatori

Il metodo adaptive boosting è stato fin qui sfruttato per migliorare la precisione dei classificatori, lasciando gestire la combinazione di classificatori agli algoritmi, in modo automatico. Un metodo analogo può essere seguito combinando manualmente diversi classificatori allo scopo di forzare una politica di classificazione rispetto alle altre, o di combinarle in modi più flessibili. I modi più semplici per combinare due classificatori A e B ottenendone un altro C sono essenzialmente due:

- Congiunzione di due classificatori:  $C = A \wedge B$ . Un soggetto è classificato come positivo solo quando sia A che B lo classificano come positivo.
- Disgiunzione di due classificatori:  $C = A \vee B$ . Un soggetto è classificato come positivo quando uno o entrambi tra A e B lo classificano come positivo.

A titolo di esempio mostriamo i risultati di  $Modello_A \wedge Modello_C$ , che suggerisce 58 verifiche contro le 59 di A da solo:

- actual\_recovery = 163.5 MEuro
- audit\_cost = 14.4 MEuro
- profitabilty = 0.778 MEuro/verifica



Altre combinazioni più complesse sono possibili, che cercano in qualche modo di conciliare i diversi condizionamenti che i diversi modelli impongono, come  $A \wedge (C \vee D)$ , ecc.

## 2.7 Conclusioni

Come mostrato dagli esperimenti, l'uso di classificatori nella selezione di soggetti da sottoporre a verifica porta ad ottenere generalmente buoni recuperi economici con un ridotto numero di verifiche, mentre ponendo l'accento sul numero di evasori rintracciati o, inversamente, sulle limitate risorse a disposizione si possono ottenere diversi risultati che forniscono un compromesso tra copertura degli evasori e risparmio nell'eseguire le accertazioni.

*Fonte:* A classification-based methodology for planning audit strategies in fraud detection. *KDD 1999*.



## Capitolo 3

# Competitive Intelligence: il caso di studio Derwent @ CINECA

### 3.1 Gli obiettivi

L'applicazione di tecniche di data mining per estrarre conoscenza da banche dati di tipo tecnico-scientifico consente di effettuare studi di technology watch (monitoraggio tecnologico) o competitive intelligence (monitoraggio dell'attività della concorrenza).

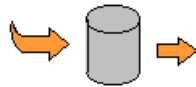
L'obiettivo, espresso in forma generica, è rispondere ai seguenti interrogativi:

- Quali sono gli orientamenti del mercato
- Quali sono le aree tecnologiche emergenti
- Quali aziende stanno investendo sulle nuove tecnologie
- Quali saranno i miei concorrenti nei prossimi anni
- In quale area un mio concorrente sta preparando nuovi prodotti da mettere sul mercato
- Quale area abbandonerà nei prossimi anni

Le fonti di informazione più affidabili sono le banche dati disponibili online che contengono documenti tecnico-scientifici. Tra queste, una delle più utilizzate è la banca dati Derwent che raccoglie tutti i brevetti che sono stati depositati in tutto il mondo negli ultimi 10 anni. Una ricerca in questa banca dati (per argomento, per azienda, o per anno) può portare ad estrarre centinaia, a volte anche migliaia, di documenti. Diventa così necessaria un'elaborazione automatica che raggruppi i documenti individuando le principali aree tematiche,

## Raccolta dei Documenti

### esempio di documento brevettuale



1/3881 - (C) Derwent Info 1994

AN: 94-364398 [45]

TI: Television with function for enlarging picture by variation of deflection frequency - has microprocessor for controlling system synchronous signal output, horizontal and vertical frequency drive circuit, sync. signal counter, signal detector.

DC: W03

PA: (GLDS) GOLDSTAR CO LTD

IN: O.KEITH

NP: 1

PR: 88KR-011143 880831

IC: H04N-005/262; C08J-005/18; G11B-005/704

PN: KR940043 B1 940120 DW9445

AB: ..... abstract .....

Figura 3.1: Esempio di informazioni presenti per ogni brevetto

metta in evidenza le sinergie e le relazioni tra le diverse aree e consenta di analizzare l'evoluzione temporale in ogni area e la strategia dei concorrenti. La grande mole di dati contenuti in ciascun documento e la loro tipologia testuale rende indispensabile l'uso di strumenti di data mining.

Il processo di estrazione di conoscenza richiede, in questo contesto, grande cura nelle prime fasi di individuazione delle fonti e di estrazione dei documenti, mentre la parte di pre-processing risulta molto meno impegnativa essendo le banche dati in input di ottima qualità. La tecnica appropriata in questo tipo di applicazione è una particolare tecnica di segmentazione (clustering) che si basa sull'analisi relazionale, descritta in seguito.

Prima di descrivere i risultati ottenibili attraverso un esempio concreto, è opportuno fare una premessa metodologica che spieghi brevemente come si ottengono i risultati.

### 3.2 I dati di input

I documenti raccolti sono dei testi strutturati in campi. Un brevetto, per esempio, è un testo strutturato in una serie di campi tra cui possiamo riconoscere: il titolo, l'azienda depositante (ed eventuale holding), il nome dell'inventore, la data di deposito, l'abstract e alcuni codici di classificazione (vedi Figura 3.1).

Ogni brevetto è infatti caratterizzato da un numero variabile di codici appartenenti a diversi sistemi di classificazione, che descrivono il contenuto e l'area

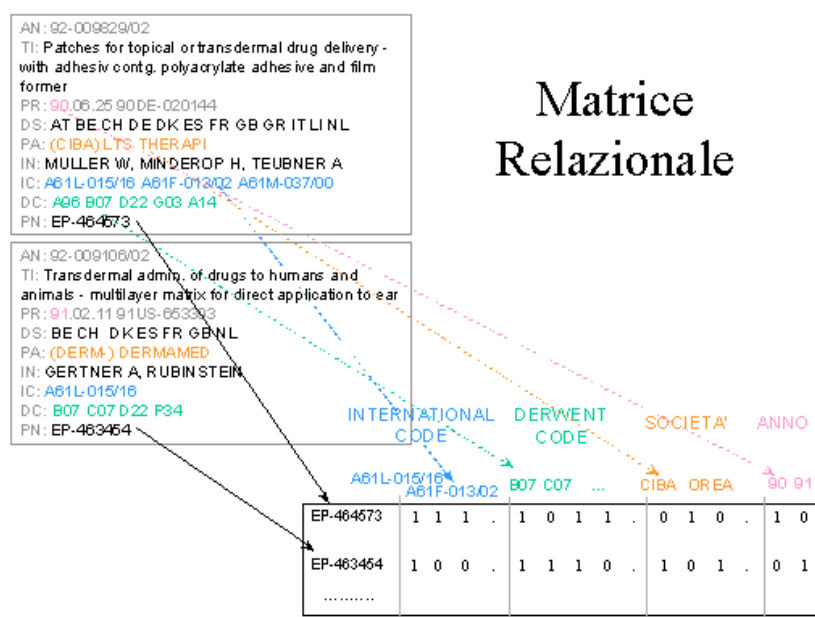


Figura 3.2: Trasformazione di un insieme di documenti in una matrice binaria

applicativa del brevetto. Questi codici sono parzialmente sovrapposti e ridondanti, così che non è facile, anche per un esperto, riconoscere l'importanza di un brevetto e le sue relazioni con gli altri e con altre aree applicative.

### 3.3 Analisi relazionale dei dati

Il Centro di Matematica Applicata di Parigi dell'IBM ha messo a punto una tecnica per analizzare questo tipo di dati che si basa sull'analisi relazionale.

Il contenuto di ciascun documento viene trasportato in una matrice binaria (vedi Figura 3.2) in cui ogni riga rappresenta un brevetto ed ogni colonna una variabile descrittiva (un codice, una parola contenuta nel titolo, l'anno di deposito, ecc.). In ogni casella della matrice, un 1 indica la presenza di quella particolare variabile come attributo descrittivo di quel particolare documento, uno 0 indica l'assenza di quel particolare attributo descrittivo nel documento in questione.

Questa matrice è il punto di partenza per poter mettere in relazione i documenti. Il confronto avviene, inizialmente, per coppie di documenti: per ciascuna coppia viene calcolato un indice di somiglianza. Tale indice aumenta all'aumentare degli 1 in comune (cioè aumenta quando i due documenti condividono lo stesso attributo descrittivo) e diminuisce all'aumentare degli attributi che li differenziano.

L'algoritmo di segmentazione usa gli indici di somiglianza per individuare la

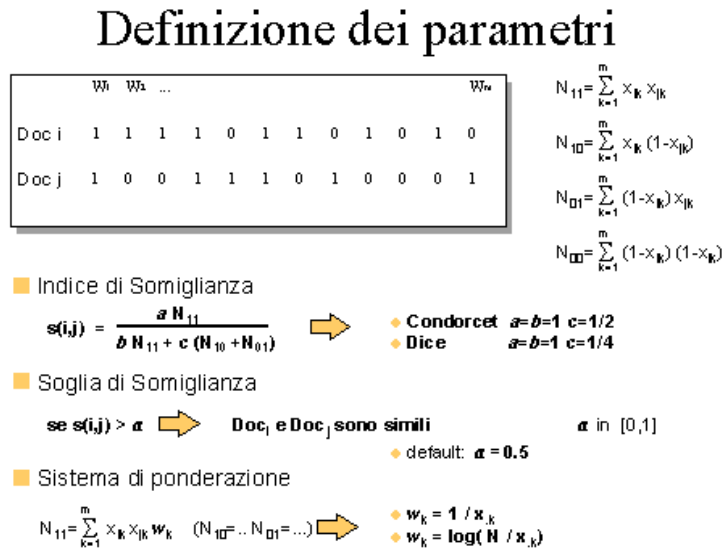


Figura 3.3: Parametri del criterio di somiglianza del clustering

partizione migliore. Quella cioè che dà luogo a raggruppamenti di documenti il più omogenei possibile, al loro interno, e il più separati possibile dagli altri raggruppamenti ottenuti. Il numero di raggruppamenti (o cluster) non è determinato a priori, come avviene nella cluster analysis classica (si veda il metodo classico K-means): è l'algoritmo che individua i raggruppamenti impliciti nei dati in maniera automatica. Questo consente da una parte di eliminare qualsiasi arbitrarietà e forzatura esterna, dall'altra di individuare ed identificare anche i raggruppamenti più piccoli che spesso sfuggono all'analisi e sono di estremo interesse in quanto possono indicare potenziali di mercato non sufficientemente sviluppati.

Questa tecnica è resa flessibile, ed adattabile a qualsiasi struttura di dati, dalla presenza di diversi parametri nella definizione di similarità tra documenti:

- Per quanto riguarda l'indice di somiglianza  $s(i, j)$  per due documenti  $i$  e  $j$ , la formula generale (mostrata in Figura 3.3) indica una famiglia di indici normalizzati (che variano tra 0 e 1). Il valore di tali indici è proporzionale alla presenza di 1 in comune ( $N_{11}$ ) e inversamente proporzionale alla presenza concomitante di 1 e 0 ( $N_{10}$  e  $N_{01}$ ). La concomitanza di 0 non ha nessun effetto sul valore dell'indice (l'assenza di un attributo descrittivo in entrambi i documenti non dà in effetti alcuna informazione rispetto alla loro somiglianza).

Il valore effettivamente assunto dall'indice dipende dal peso che si vuole assegnare agli attributi comuni e agli attributi che differenziano. All'inter-

no di questa famiglia di indici, uno dei più utilizzati è l'indice di Condorcet che attribuisce peso unitario alla presenza di attributi comuni e peso pari a  $1/2$  alla presenza di attributi discordanti. Un indice ancora meno rigido, in quanto attribuisce un peso inferiore alle differenze ( $1/4$ ), è noto col nome di Dice ed è particolarmente utile quando oggetto di confronto sono dei documenti testuali che presentano molte parole diverse. In questo caso l'indice di Condorcet porterebbe ad individuare ben pochi documenti simili.

- Inoltre è possibile alzare od abbassare la soglia che di solito è fissata a 0,5. Due documenti sono considerati simili se l'indice di somiglianza supera il valore soglia. Abbassare tale valore consente di definire un criterio meno rigido. Anche questo risulta utile quando la matrice binaria è sparsa, contiene cioè molti 0.
- Infine è possibile utilizzare un sistema di ponderazione che assegna pesi diversi agli attributi. L'importanza di un attributo è inversamente proporzionale alla sua frequenza in tutto l'insieme di documenti. In assenza di un sistema di ponderazione ogni attributo (sia raro che frequente) ha la stessa importanza nel definire la somiglianza (o la dissomiglianza) tra documenti. Nel nostro caso, può essere utile assegnare un peso maggiore agli attributi rari (parole nei titoli, per esempio, che compaiono raramente, dovrebbero avere maggiore importanza nel definire i gruppi rispetto alle parole frequenti).

Una volta definiti i parametri, ha inizio la fase di data mining vera e propria che dà come risultato i principali raggruppamenti tematici, ottenuti tramite individuazione delle ricorrenze di parole (contenute nel titolo dei documenti) e/o di codici classificatori.

La fase successiva, di analisi e valutazione dei risultati ha come punto di partenza la mappa dei cluster: una rappresentazione grafica dei gruppi individuati tramite la quale è possibile accedere alla descrizione completa di ciascun cluster.

### 3.4 Un esempio applicativo

Un importante centro di ricerca francese nel campo della cosmesi era interessato a conoscere gli sviluppi del mercato del cerotto medicale (patch technology). La ricerca di documenti relativi al cerotto medicale ha portato ad individuare 146 brevetti. Poiché questo studio è stato effettuato nel 1992, i documenti coprivano l'arco temporale 1979 - 1991. Erano stati depositati da 105 diverse aziende in 12 paesi e contenevano 94 diversi codici di classificazione internazionale (e 52 codici Derwent).

L'applicazione dell'algoritmo di data mining ha consentito di individuare 20 gruppi tematici, rappresentati in Figura 3.4.

## Patch technology- mappa dei clusters

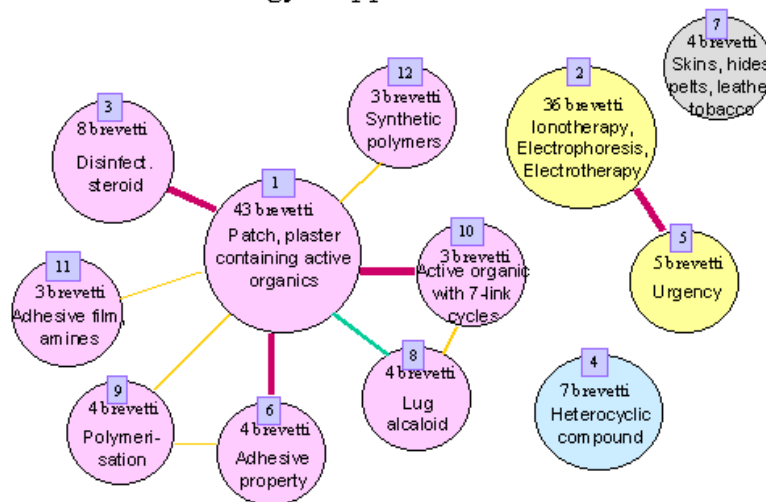


Figura 3.4: Gruppi di brevetti estratti

La mappa qui riprodotta ne presenta i primi 12. Ogni cerchio rappresenta un gruppo di documenti ed è caratterizzato da un numero identificativo (che ne indica l'importanza in termini di dimensione), dal numero di documenti che contiene e da alcune parole che ne caratterizzano l'argomento (sono le parole più frequenti all'interno del gruppo).

I legami tra gruppi sono rappresentati da linee il cui colore e spessore indica la forza del legame. Un insieme di gruppi tra loro collegati rappresenta una macro tecnologia (o macro area). La mappa fornisce una prima visione d'insieme degli argomenti individuati e delle loro relazioni. L'interfaccia del sistema sviluppato all'uopo, con un clic del mouse sull'argomento di interesse è possibile accedere alla descrizione completa del gruppo di documenti e, da qui, ai singoli documenti. A titolo d'esempio è riportata la descrizione (sintetica) del cluster n.2 (Figura 3.5).

La descrizione mette in evidenza i codici di classificazione (e la relativa descrizione) che compaiono in questo gruppo di documenti, i nomi delle aziende depositanti (il 42% dei brevetti contenuti in questo gruppo sono stati depositati dalla Drug Delivery System) e l'anno di deposito.

La rappresentazione grafica di Figura 3.6 consente di valutare l'attività di ciascuna azienda nel tempo e in ciascuna area tecnologica. L'evoluzione temporale indica che l'interesse su questo argomento (Elettroforesi) è andato aumentando nel tempo, la maggior parte dei brevetti è infatti stata depositata negli ultimi anni. Si tratta quindi di una tecnologia, almeno al momento dello studio,



### Patch technology- descrizione del cluster n.2


#### Classificazione Internazionale:

A61N-001/30 Electrotherapy; Appliances of electrical power by contact electrodes; Ionotherapy or electrophoresis devices  
A61M-037/00 Therapeutic patch

#### Classificazione Derwent:

S05 Electromedical  
P34 Health, Electrotherapy

#### Società proprietarie:

	DRUG DELIVERY SYST	42%
	BASF AG	36%
	KOREA RES INST CHEM	16%
	MEDTRONIC INC	6%

Anno	n. brevetti
1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

Figura 3.5: Descrizione di un cluster di esempio

in espansione.

Si può notare, sempre nel secondo cluster e quindi sempre relativamente alla elettroforesi, che, mentre per la BASF si tratta di un settore di ricerca consolidato nel tempo e su cui mantiene un'attività di ricerca più o meno costante nel tempo, per la Drug Delivery System si tratta di un settore nuovo, sul quale sta investendo pesantemente.

L'esplorazione dei risultati può procedere in varie direzioni: approfondendo il contenuto del secondo cluster tramite esame dei singoli documenti, passando ad argomenti collegati (in questo caso il quinto cluster tratta un argomento collegato all'elettroforesi), tornando alla mappa per selezionare un'altra area tematica o, infine, analizzando la presenza delle aziende nei diversi cluster e la caratterizzazione temporale di ciascuna area tematica. In Figura 3.7 è rappresentata la distribuzione delle prime 20 aziende nei cluster.

Le aziende in tutto sono 105, l'algoritmo seleziona automaticamente quelle più presenti. La prima barra (T) mostra la loro distribuzione percentuale nell'insieme dei documenti, le altre mostrano la loro distribuzione percentuale all'interno di ciascun gruppo tematico. L'estensione di ciascun colore indica la quota percentuale dell'azienda all'interno del cluster. Si può notare che la Drug Delivery System (colore arancione scuro) è presente, oltre che nel secondo cluster, anche nel quinto, che, come si è visto in precedenza, è un argomento collegato. E' assente invece da ogni altra area di ricerca. BASF (colore arancione chiaro) è impegnata anche nelle aree identificate dai cluster 11 e 19. Medtronic

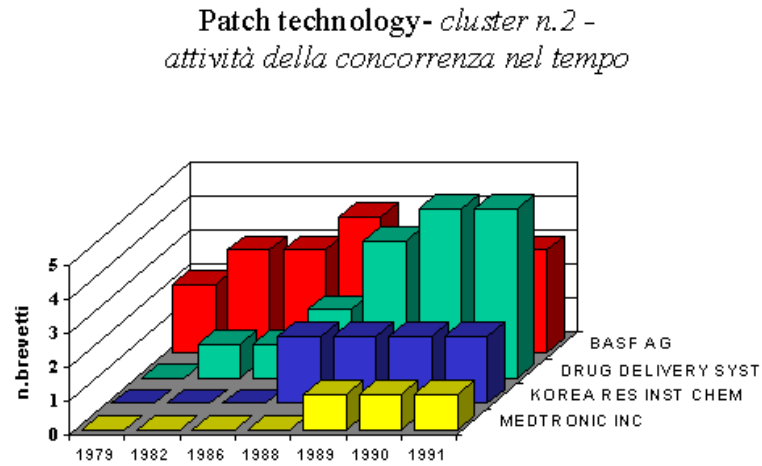


Figura 3.6: Andamento nel tempo dell'attività di alcune aziende

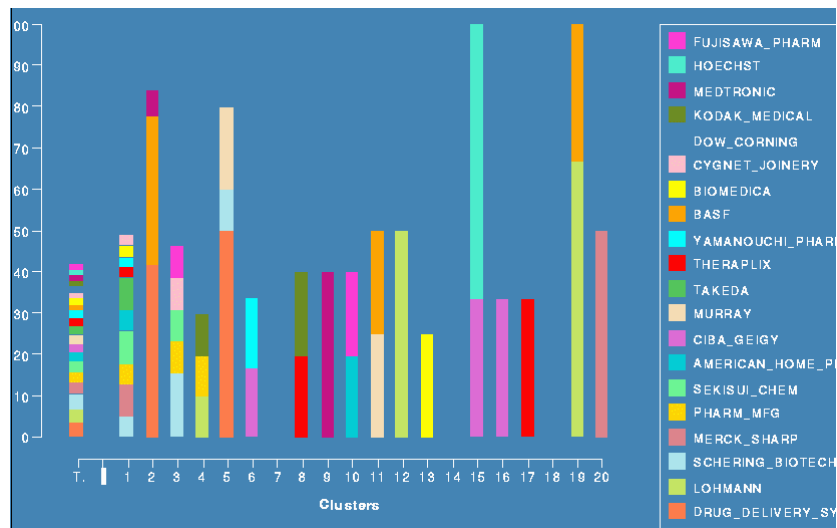


Figura 3.7: Distribuzione delle prime 20 aziende nei cluster

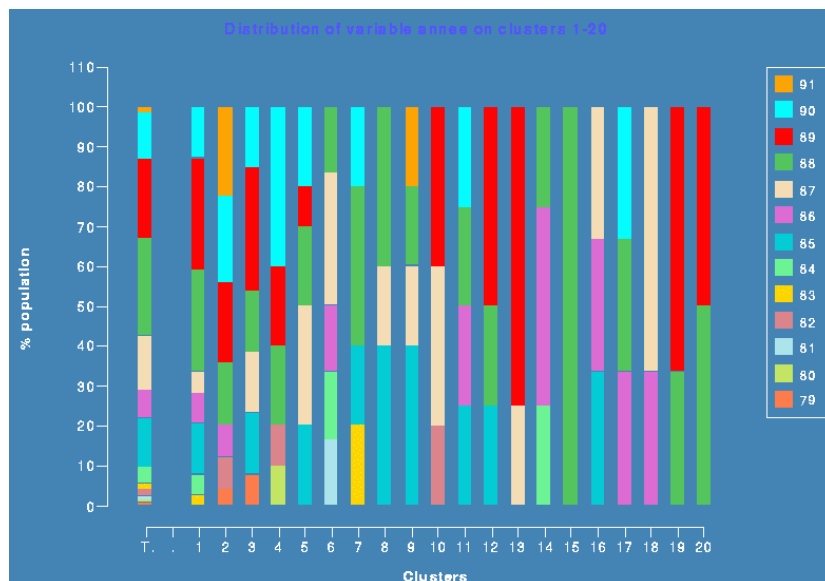


Figura 3.8: Distribuzione degli anni di deposito dei brevetti

(colore rosso scuro), che era poco presente nel campo dell'elettroforesi, è invece impegnata nell'area 9 (polimerizzazione) dove risulta la depositante del 40

La distribuzione dell'anno di deposito sul totale dei documenti (Figura 3.8, prima barra a sinistra), mostra come la maggior parte dell'attività di ricerca nel campo del cerotto medicale sia stata effettuata negli anni 88 (verde scuro) e 89 (rosso).

La maggiore presenza del colore azzurro e/o ocra in alcuni cluster indica le aree di ricerca più recenti. I cluster 16 e 18 rappresentano aree di ricerca che con tutta probabilità sono state abbandonate (l'attività è ferma al 1987).

Fonte: <http://open.cineca.it/datamining/dmCineca/>