Strumenti e tecniche di Business Intelligence per applicazioni CRM

di

Stefano Bordoni

Luglio 2004

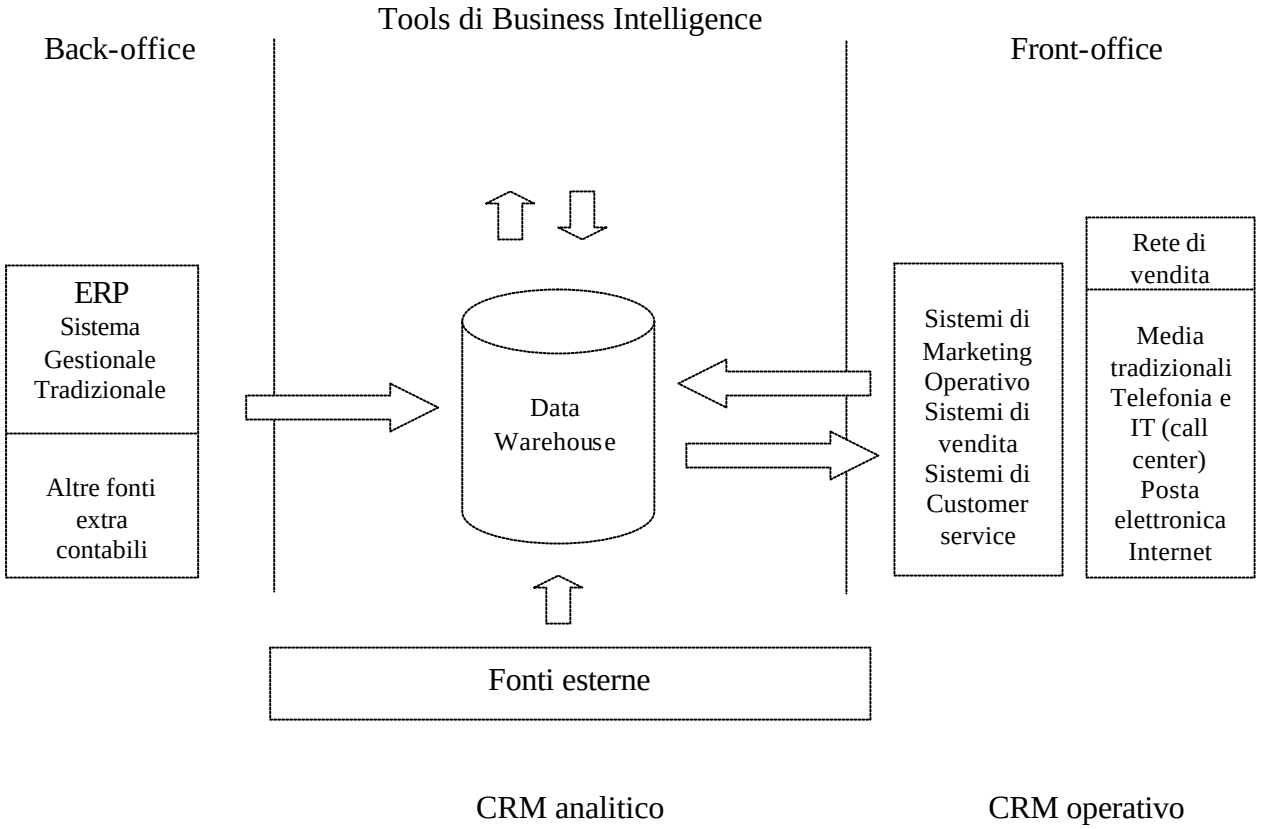
Università degli Studi di Modena e Reggio Emilia
Dipartimento di Economia Politica
Viale Berengario,51
41100 Modena (Italia)
e-mail: bordoni.stefano@unimore.it

Introduzione

Il problema che il CRM (Customer Relatioship Management) è chiamato a risolvere si può riassumere attraverso lo spot pubblicitario di un broker di servizi bancari statunitense [1]. Una coppia di giovani sposi è seduta ad un tavolo e dialoga con un impiegato di banca. "Sicuro di non poter abbassare i tassi"? "No", risponde l'impiegato, "l'offerta è buona". La coppia sorride e inaspettatamente, licenzia il bancario con la frase: " Spiacenti di averle rubato del tempo. Avanti un altro".
La situazione descritta dallo spot riassume le profonde modifiche nella relazione impresa consumatore che caratterizzano il passaggio in atto da un'economia di offerta verso un'economia di domanda, dove la consapevolezza e il ruolo del consumatore obbliga l'impresa a strategie di vendita orientate ai clienti piuttosto che ai prodotti.
Il consumatore di oggi ha caratteristiche molto diverse da quello di epoche precedenti. Il suo potere d'acquisto si è praticamente dimezzato rispetto agli anni 60, ma è molto aumentata la sua capacità cognitiva su prodotti e servizi. E' meno omologabile, meno manipolabile, più consapevole, essenziale, sobrio ed attento all'atteggiamento dell'impresa che offre ciò che gli serve rispetto a criteri di equità, trasparenza e reciprocità (soddisfazione etica).
Può usare il web ed efficaci strumentazioni tecnologiche per informarsi ed effettuare *comparison shopping*, recuperando velocemente e a costi prossimi allo zero informazioni che gli consentono di mettere a fuoco in modo critico l'offerta delle imprese rispetto alle proprie necessità ed aspettative di acquisto.
Dal punto di vista del mercato e dell'offerta, gli ultimi anni sono stati caratterizzati da un aumento della pressione concorrenziale, dovuta alla globalizzazione dei mercati, alla diffusione delle nuove tecnologie e alla flessibilità delle produttività.
Per la sua vicinanza al pubblico e al mercato, la funzione aziendale più coinvolta a recepire questi sviluppi sociali ed economici è senza dubbio il marketing.
Di fronte all'evoluzione dei processi di acquisto e all'aumentata pressione della concorrenza, la funzione aziendale del marketing si è conseguentemente trasformata fino a svolgere un ruolo strategico (orientato a soddisfare il consumatore) anziché strettamente operativo (cioè orientato alla transazione e al mercato) come in passato.
Questo diverso approccio all'evoluzione della domanda, ai rapporti impresa mercato e alle relazioni instaurate con i consumatori ha finito per interessare la strategia complessiva dell'azienda per il raggiungimento della leadership.
La strategia di successo di un'impresa che è stata focalizzata negli anni 80 sulla qualità dei prodotto (Total Quality Management) e negli anni 90 sulla riduzione dei costi (Business Process Reengineering) verte oggi sulla capacità di istituire relazioni stabili con i clienti.
Il CRM rappresenta la strategia aziendale chiamata ad interpretare e soddisfare le esigenze e i bisogni dei clienti attraverso un'offerta strutturata e personalizzata sulla base delle necessità specifiche del singolo individuo. Può essere quindi definito come un processo integrato e strutturato per la gestione delle relazioni con la clientela con lo scopo di costituire relazioni di lungo periodo con il cliente, in grado di aumentare la soddisfazione dei clienti e il valore del cliente per l'impresa.
Si propone di individuare e selezionare i propri clienti, valutandone ed aumentandone il significato economico attraverso i concetti di costo di acquisizione, grado di soddisfazione, valore, fedeltà, fidelizzazione, ciclo di vita, individuazione dei bisogni e personalizzazione dell'offerta in termini di prodotti e servizi (mass customization).
Per comprendere il valore aggiunto di una strategia di CRM si consideri il seguente esempio.
Non conoscendo la composizione dettagliata della propria clientela, l'impresa si indirizza normalmente verso un prodotto (una gamma di prodotti) destinati a soddisfare il cliente in base alle caratteristiche medie della popolazione complessiva. Viceversa, un'indagine più approfondita

[1] Lo spot è della Lendigtree.com ed è citato in CRM, P. Greenberg APOGEO

potrebbe evidenziare *p* sotto popolazioni con caratteristiche diverse verso cui orientare in modo più mirato le decisioni di business.
Anziché soddisfare le esigenze di un unico cliente medio, si tenterà di diversificare l'offerta e le politiche di marketing in relazione alle necessità e alle aspettative dei *p* clienti teorici, migliorando profitti e relazioni con i clienti.
Per quanto detto, appare chiaro come il successo di una strategia di CRM dipenda essenzialmente dalle caratteristiche e dalla consistenza del sistema informativo e informatico dell'impresa e dalla sua capacità di recuperare dati validi sui propri clienti e rilevanti da un punto di vista strategico.
Questo processo di acquisizione ed analisi dei dati della clientela è tutt'altro che scontato anche in imprese tecnologicamente avanzate, in quanto le caratteristiche dei software gestionali o ERP (Enterprise Resource Planning) e CRM non sono affatto simili.
La base dell'ERP e del sistema informativo aziendale è costituita da funzioni internamente stabili e da processi prevedibili. Tra le funzioni di back office svolte dagli ERP ci sono il controllo della produzione, del magazzino e della distribuzione, del settore contabile e finanziario, delle risorse umane, ma non la catena della domanda e, all'interno di questa, le informazioni di front office sulle vendite, sul marketing e sul canale di assistenza al cliente.
La vendita di un prodotto genera attraverso l'ERP una fattura e crea un conto debitorio con le conseguenze produttive, gestionali, contabili e finanziarie del caso.
Ma l'unico contatto con il front office avviene e si interrompe al momento della segnalazione iniziale che avvia la transazione.
L'obbiettivo di fondo dell'ERP è l'integrazione delle funzioni di back office sui dati *interni* per diminuire le strozzature e le incompatibilità tra software home made e/o frammentati per attività.
Il CRM si basa invece sui dati *esterni* dei clienti per rispondere in tempo reale alla domanda di consumo in costante movimento che non è minimamente controllata dall'interno.
La convivenza di un sistema informativo interno con uno basato su informazioni residenti altrove è tutt'altro che semplice e si fonda sulla buona integrazione delle infrastrutture tecnologiche aziendali, in particolare tra le applicazioni di back office a sostegno del CRM analitico e quelle di front office a sostegno del CRM gestionale.
Se questa integrazione avviene, le informazioni esterne recuperate off line (caratteristiche socio demografiche, preferenze e sequenze di acquisto, storia del rapporto con l'azienda, dati psicografici) e on line (log file, email, form web, comunità virtuali) vengono collegate con quelle interne provenienti dal sistema ERP e da altri database aziendali in unico archivio integrato e scalabile (Customer o Data Warehouse).
Per concretizzare le potenzialità delle informazioni raccolte, è necessario predisporre una serie di strumenti di Business Intelligence in grado di interpretare le informazioni e sviluppare la conoscenza utile all'attivazione di politiche e strategie decisionali.
La conoscenza sviluppata in questa fase viene convogliata all'interno dei sistemi informativi dedicati al CRM operativo per gestire la relazione con il cliente, l'attività operativa di marketing, l'attività di vendita e di Customer Service (call center, web site, sales force automation).
Le tecniche, i processi e gli strumenti principali di Business Intelligence che consentono l'elaborazione dei dati grezzi contenuti nel Data Warehouse costituiscono l'oggetto di questo documento.
Nella categoria di Business Intelligence rientrano tutti gli strumenti e le tecniche in grado di supportare la fase decisionale, indipendentemente dal loro grado di complessità.
Le tecniche tradizionali di analisi dei dati, come i fogli elettronici, le interrogazioni (query) SQL, l'analisi multidimensionale (OLAP, On line analytical program), i report e le tecniche di visualizzazione sintetica dei dati, appartengono a questa categoria con lo stesso diritto di tecniche più evolute come ad es. i Sistemi esperti e le tecniche di Data Mining.



Questi strumenti differiscono, viceversa, per la diversa capacità di estrarre conoscenza da un problema e di sintetizzare informazioni indirizzate ai business analyst e immediatamente spendibili in un processo decisionale, più che agli amministratori di sistemi e di database.
In un processo di automazione della forza di vendita (Sales Force Automation), l'obbiettivo è una attenta, completa, trasparente (in una parola, robusta) gestione dei clienti effettivi e potenziali (contatti) attraverso l'uso corretto degli strumenti tecnologici e l'elaborazione di interrogazioni e di grafici di sintesi sull'andamento delle opportunità e dei processi di vendita (sales pipeline)[2].
Un grafico del fatturato delle vendite per area, per prodotto o di previsioni di vendita di un agente elabora in modo opportuno i dati e fornisce informazioni utili e corrette, ma non estrae informazioni (conoscenza) ulteriori rispetto a quelle già possedute dall'impresa.
Per quanto efficace, un grafico o una tecnica di data retrieval non possono aggiungere informazioni sulla identità dei clienti, sulle loro necessità, sulle preferenze e abitudini di acquisto, sulla propensione a rimborsare un prestito o tentare una frode assicurativa.
Ciò nonostante, la conoscenza e la possibilità di applicare strumenti diversi e teoricamente alternativi permette di adattare il processo di analisi di un problema decisionale in modo creativo e flessibile, ricavando informazioni spesso complementari e più approfondite rispetto ai diversi contesti e alle diverse finalità.
Gli esempi contenuti in questo documento hanno esattamente l'obbiettivo di fornire il panorama più esaustivo possibile sui diversi strumenti di Business Intelligence, per consentire un approccio critico e flessibile alla soluzione dei problemi di analisi dei dati per il supporto alle decisioni.
Gli strumenti e le tecniche esaminate sono stati suddivisi in quattro categorie principali. Le tecniche tradizionali di visualizzazione e data retrieval non vengono trattate in questo documento, che verte principalmente sui Sistemi Esperti, sulle principali tecniche di Data Mining e su alcune semplici tecniche statistiche esplorative.

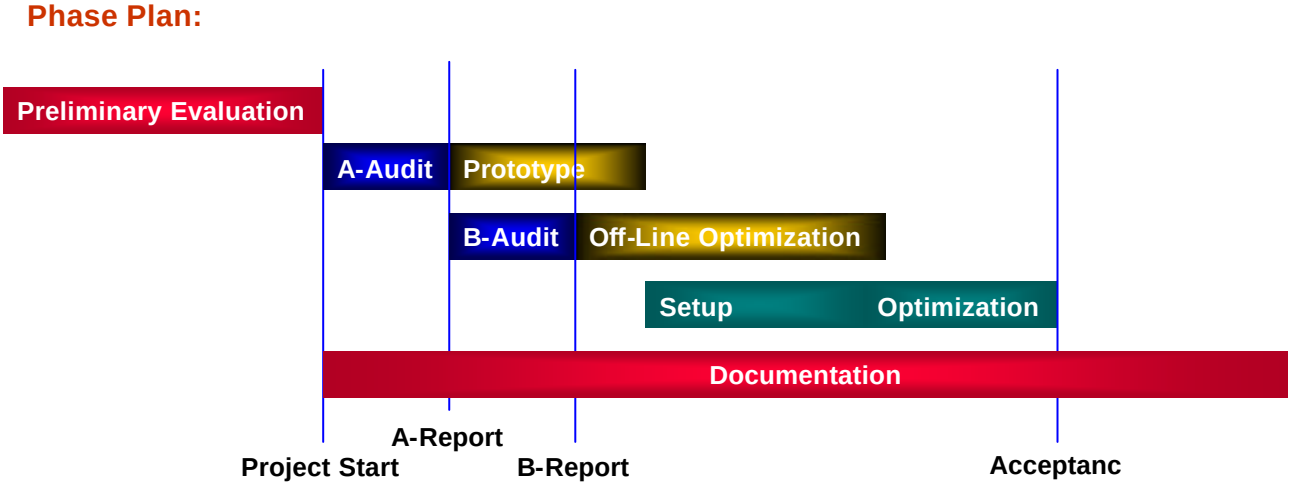
[2] Si veda, ad es. http://www.salesforce.com

Questo raggruppamento è del tutto arbitrario ed è in relazione con l'utilizzo e l'ambito di applicazione proposto delle tecniche che vengono presentate in queste pagine.
Va però detto, a chiarimento del rapporto tra tecniche statistiche e tecniche di Data Mining, che quasi ogni problema analizzato con tecniche di Data Mining possiede un'equivalente soluzione statistica, da cui il Data Mining consegue in modo diretto.
La scelta di privilegiare strumenti di Data Mining al posto di strumenti statistici deriva quasi esclusivamente dalla maggior semplicità d'uso e dalla più facile gestione di dati qualitativi e quantitativi insieme.
Gli strumenti e le tecniche di Data Mining non sono altro che la riorganizzazione funzionale di tecniche statistiche, in relazione alle specifiche finalità di analisi di ampi database aziendali (spesso contenenti dati secondari, oltre che primari) e di ricerca di informazioni in assenza di ipotesi formulate a priori.
Proprio la presenza o meno di specifiche ipotesi di ricerca distingue in letteratura due distinti approcci di analisi, denominate confermative (top-down, tipiche dei metodi statistici tradizionali) ed esplorative (bottom-up, tipico del Data Mining)[3].
Definito il *problem solving space* (ad es. la ricerca di regolarità nelle combinazioni o nelle sequenze di acquisto), il Data Mining si propone di estrarre dai dati aziendali informazioni (conoscenze) del tutto ignote e spesso contro intuitive (es. forte correlazione nelle vendite di pannolini e birra nel fine settimana) per fornire un supporto utile e applicabile ai processi decisionali strategici.
In relazione agli obbiettivi di CRM e più in generale di Market e Risk Management, verranno presentate in queste pagine alcune tecniche fondamentali per la profilazione (demographic clustering) e per la classificazione (decisonal tree) della clientela, nonchè per la ricerca di relazioni nelle abitudini d'acquisto (link analysis).
I sistemi esperti rule based (e l'addestramento neurale degli stessi) presentati in questo documento permettono di valutare differenze ed analogie rispetto alle finalità, all'ambiente di utilizzo e ai risultati delle tecniche di Data Mining.
Come descritto dagli esempi esposti, la principale e discriminante differenza risiede nel processo di acquisizione della conoscenza. I sistemi esperti traducono in modo deduttivo la conoscenza posseduta dagli esperti in algoritmi decisionali, mentre le tecniche di Data Mining ricavano le informazioni in modo induttivo attraverso la sola analisi dei dati.
La scelta di una o dell'altra tecnica viene normalmente condizionata dalle risorse a disposizione dell'impresa (dati affidabili e/o una consolidata esperienza umana) piuttosto che dalla natura del problema da risolvere.
Nonostante le differenze metodologiche, le tecniche possono essere usate in modo complementare per incrociare le conoscenze ricavate con metodi e presupposti diversi, ma aventi la stessa finalità.
Nelle pagine che seguono lo stesso problema di classificazione della clientela per rischio di frode viene affrontato con entrambe le tecniche per evidenziare analogie e differenze.
Nella parte finale del documento vengono esaminate le caratteristiche e i principi di progettazione di un database relazionale per facilitare la costruzione di un proprio archivio dove applicare ed esercitare le tecniche descritte in questo documento in modo critico e flessibile.
Lo scopo di queste pagine è dunque quello di esplorare la natura e l'organizzazione dei dati aziendali, la loro predisposizione al fine di consentirne l'analisi, lo studio e l'applicazione delle tecniche di business intelligence rispetto ai diversi contesti e la valutazione dei risultati da un punto di vista delle strategie di business.

[3] Giudici P. "Data Mining" McGraw-Hill

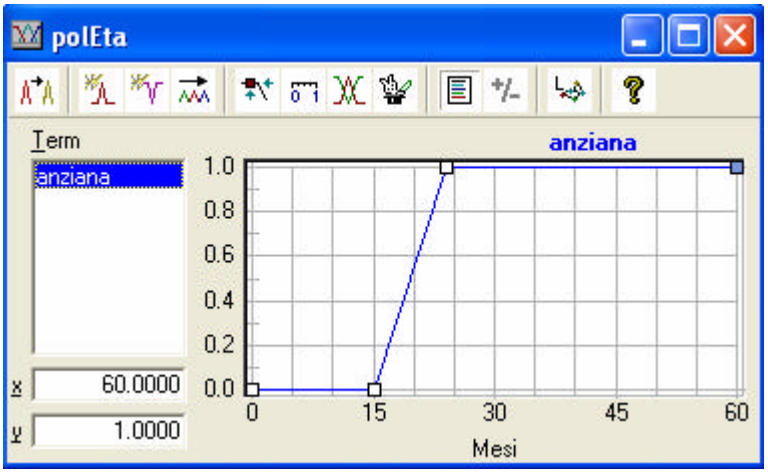
1. Sistemi Esperti rule-based e Sistemi Esperti Fuzzy

Lo scopo comune delle tecniche di KDD (Knowledge Discovery In Databases) utilizzate nei processi di Business Intelligence consiste nel valutare i dati di un problema per recuperare informazioni e conoscenze utili alla soluzione di problemi decisionali.
Tra le tecniche possibili, i Sistemi Esperti vengono utilizzati sia nel caso di assenza di dati validi che permettano di analizzare il problema con tecniche esplorative (ad esempio perché mancanti o legati a comportamenti passati e quindi obsoleti) che in quello in cui si disponga di consulenti particolarmente esperti in quel settore.
L'obbiettivo di un sistema esperto è recepire in modo essenzialmente deduttivo la base di conoscenza posseduta dall'esperto per emularne il comportamento (umano) nell'affrontare un dato problema decisionale. La base di conoscenza viene solitamente riprodotta in forma di regole di produzione all'interno di una shell (motore inferenziale e interfaccia utente) adatta allo scopo.
Per le sue caratteristiche di riproducibilità e distribuzione periferica del processo decisionale, l'ambito di applicazione di un sistema esperto è quello dell'analisi remota e autonoma su decisioni che devono essere prese velocemente e in relazione a grandi quantità di dati (mercato azionario, autorizzazione transazioni con carta di credito, concessione sul credito bancario...)
La natura prettamente deduttiva (top-down) dell'attività di interpretazione e acquisizione della conoscenza obbliga il programmatore (ingegnere di conoscenza) alla traduzione trasparente ed esplicita dei processi decisionali complessi, aggregando l'esperienza di più persone in un unico sistema. Questa fase istruttoria del processo conoscitivo, spesso svolto insieme a persone diverse della stessa azienda (es. direttori di filiali diverse della stessa banca) ha spesso un forte valore aggiunto per l'azienda stessa nel comprendere come effettivamente vengano interpretate e utilizzate procedure e linee guida nella diverse realtà in cui questa opera.
Ciò non toglie che il KE (knowledge engineer) non possa ricavare parte delle regole di produzione in modo induttivo (bottom-up), osservando gli esperti nella soluzione di un certo problema reale od ipotetico. La principale differenza tra sistemi deduttivi di analisi dei dati (o confermativi, a cui appartengono i Sistemi Esperti) e quelli induttivi (o esplorativi, come nel caso delle tecniche di data mining) risiede nel fatto che i primi testano sui dati premesse e conoscenze di comportamento del fenomeno note e fissate a priori, mentre i secondi cercano nei dati le regolarità e la conoscenza su un certo problema.
Va sottolineato che i due metodi non sono sempre alternativi, ma possono essere utilizzati in modo complementare per descrivere meglio e per ricavare informazioni più accurate sul problema.
La produzione di un sistema esperto inizia con una fase di analisi, nella quale viene attivato il processo di brainstorming che porta all'acquisizione e alla scrittura della logica decisionale da trasferire al sistema esperto.



In questo documento viene utilizzato una particolare tecnica di costruzione di Sistemi Esperti che utilizza come motore inferenziale un sistema fuzzy di produzione e controllo delle variabili, nonché di inferenza sui dati e di calcolo dei risultati.
Nonostante non cambi la logica generale di scrittura di un sistema esperto, la logica fuzzy risulta essere particolarmente efficace nella programmazione della logica decisionale (base di conoscenza), per la sua capacità di scrivere ed utilizzare leggi generali di comportamento ed interpretazione dei dati relativi alle variabili di un problema.
Questo documento non entra nel merito delle tecniche di costruzione di un sistema fuzzy, per i quali si rimanda a testi specifici[4].
In generale, la semplicità di scrittura di una KB (knowledge base) di un sistema esperto con motore inferenziale fuzzy deriva dall'uso degli insiemi fuzzy per rappresentare i termini coi quali viene solitamente valutata una variabile nel linguaggio naturale.

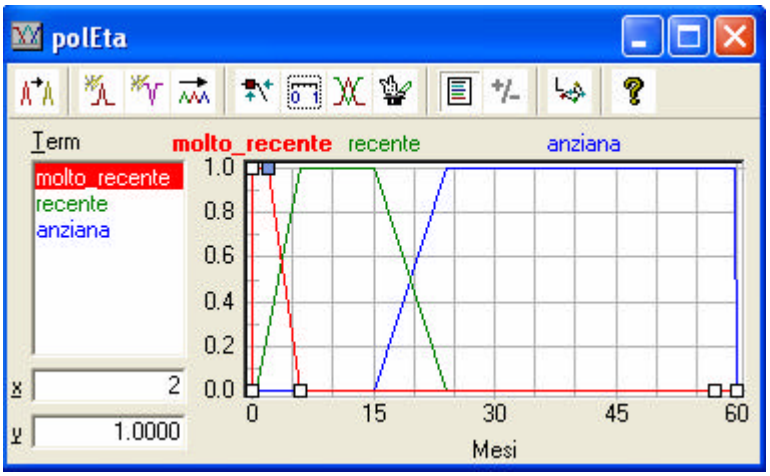
Si consideri ad esempio il caso della variabile "Età della polizza". Per essere valutata con la logica tradizionale booleana, l'anzianità della polizza dovrebbe essere confrontata con tanti intervalli (insiemi) quanti sono i valori discriminanti che questa può assumere. Più è alto il numero di intervalli, maggiore è il grado di precisione col quale il valore assunto dalla variabile viene processato. E' importante sottolineare che in un sistema esperto rule based, ad ogni confronto (cioè ad ogni intervallo), corrisponde una regola da inserire nella KB.
Nei sistemi fuzzy, viceversa, l'anzianità della polizza viene valutata attraverso l'utilizzo di un solo insieme (fuzzy) che corrisponde all'attributo (termine) linguistico "polizza anziana", risolvendo il problema del confronto con un numero di intervalli alto a piacere. Il confronto del valore assunto dalla variabile con questo insieme (MBF, funzione di appartenenza) identifica con che frequenza (percentuale) tale variabile appartiene all'insieme "polizza anziana". L'analisi sull'anzianità della polizza viene dunque effettuata nella KB da una regola unica, con grandi vantaggi in termini di trasparenza ed efficienza nella scrittura e nella manutenzione della KB.



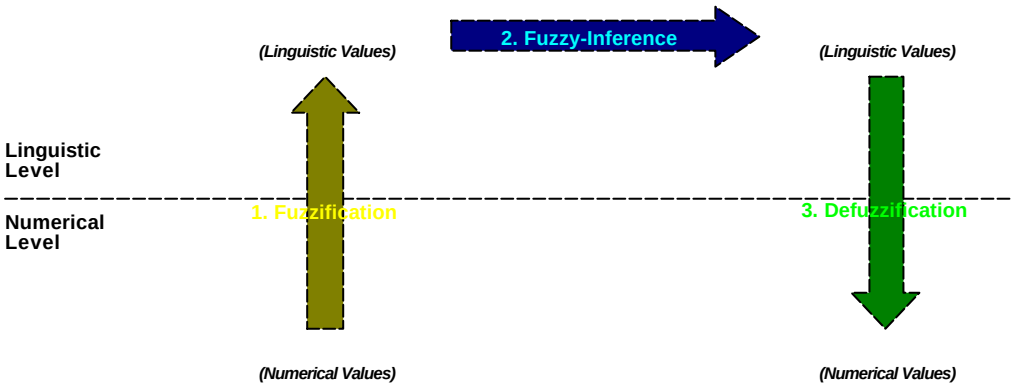
Nell'esempio considerato, il valore assunto in mesi dalla variabile "Età della polizza" viene confrontato con il termine "anziana". Per un valore inferiore ai 15 mesi, la polizza verrà considerata anziana con grado di appartenenza (frequenza) pari a zero. Tra 15 e 24 mesi, il grado di appartenenza cresce in modo lineare e rimane pari ad 1 per valori superiori a 24 mesi. La regola generale "polizza anziana" è in questo caso l'unico insieme necessario per confrontare (e valutare) qualsiasi valore possibile all'interno del dominio della variabile col suo valore linguistico.
Nel caso la variabile in questione possa assumere altri valori linguistici (es. molto recente, recente), questi verranno inseriti tra le proprietà della variabile di input in forma di altrettante MBF. Durante

[4] Von Altrock C. "Fuzzy logic and Neurofuzzy applications in business and finance" Prentice Hall PTR

il processo di elaborazione dei dati il valore assunto dalla variabile verrà quindi confrontata con ciascuno dei termini linguistici normalmente utilizzati dagli esperti nella fase di valutazione, producendo un valore sul grado di appartenenza a ciascuno di questi insiemi. Nel caso di una polizza di 18 mesi, questa verrà valutata principalmente recente (il grado di appartenenza all'insieme "recente" è del 67%) e contemporaneamente anziana, anche se con frequenza inferiore (il grado di appartenenza all'insieme "anziana" è del 33%).



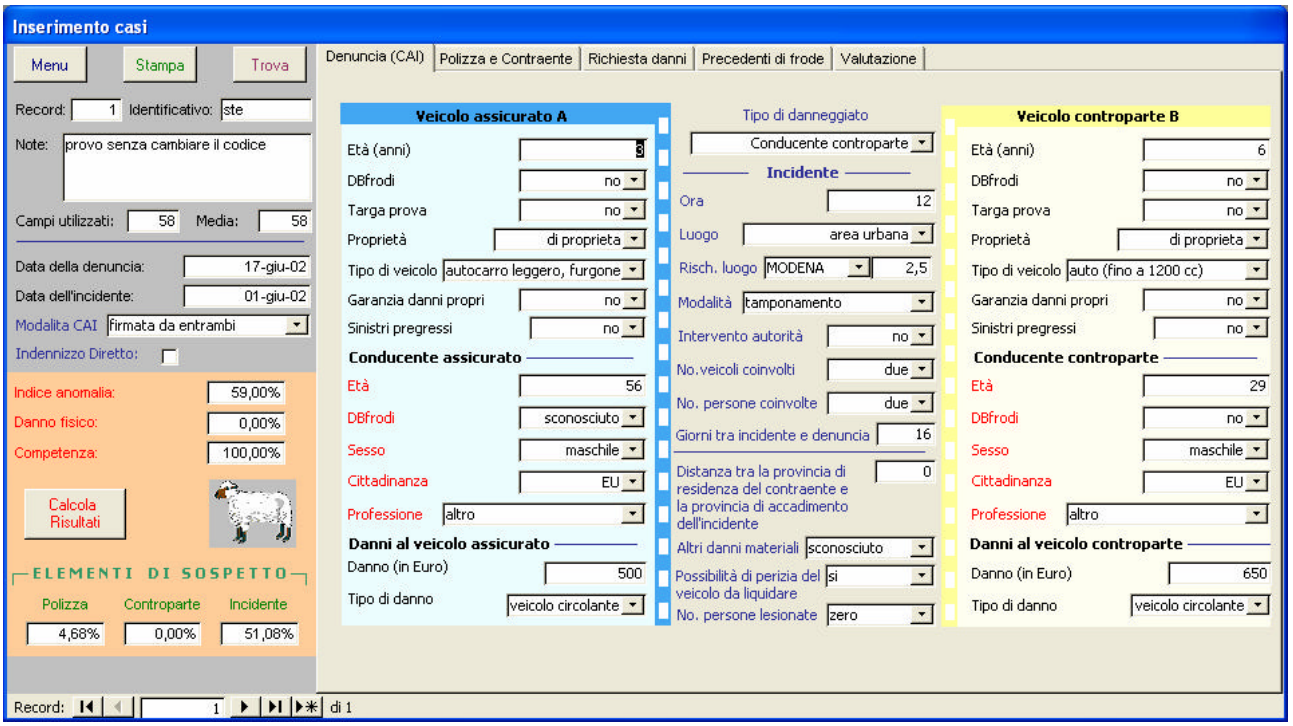
La valutazione linguistica, cioè il confronto dei dati di un record con i corrispondenti termini linguistici, corrisponde alla prima fase di elaborazione di un sistema fuzzy. Una volta fuzzificati, i valori di input vengono elaborato dalla base di conoscenza per ottenere un risultato linguistico e, in seguito, numerico (fase di defuzzificazione).



Caso: **Fraudulent claims**
Tecnica: Sistema esperto Fuzzy
Software: Fuzzytech
Obiettivo: classificare le richieste di rimborso di sinistri RCA in relazione al rischio di frode
Campi: età della polizza, numero dei sinistri, frazionamento della polizza, bonus-malus, età veicolo assicurato, danno veicolo controparte, rischiosità della provincial, modalità incidente
Dati di esempio: Fraud management.xls

Il modello presentato è una riduzione semplificata a 8 variabili di input di un sistema complesso a 72 variabili di input[5].

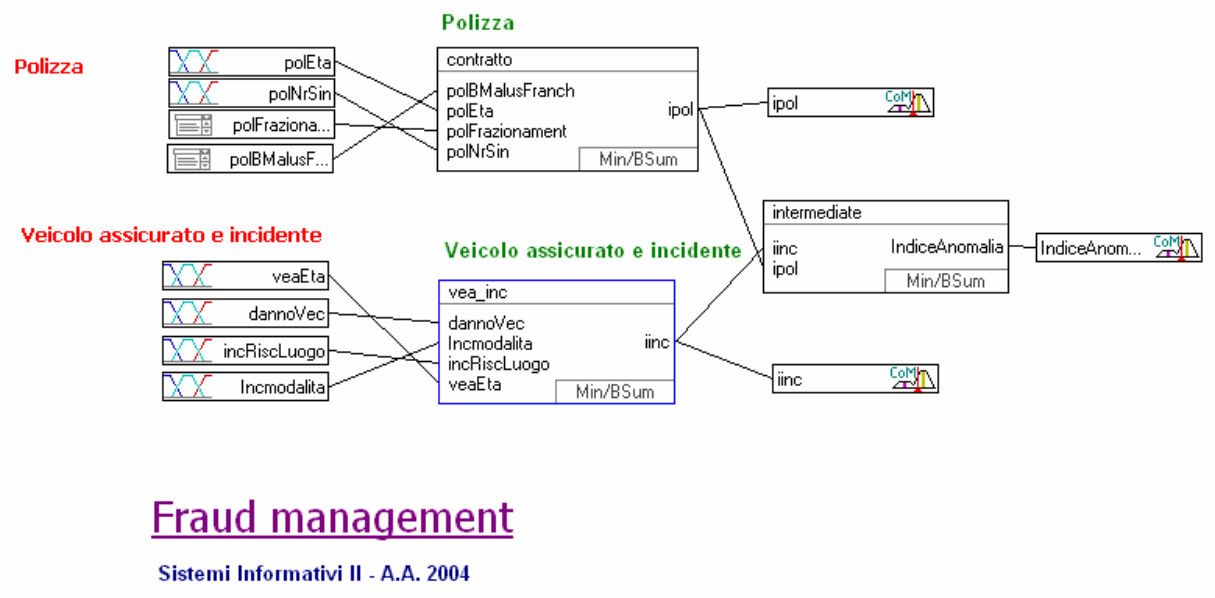
[5] Bordoni s. - Facchinetti G. (2001): " Insurance fraud evaluation. A fuzzy expert system"



Nell'esempio riportato il modello viene realizzato, per definizione, sulla base delle descrizione del problema fornita dall'esperto anti frode della Compagnia Assicurativa. In questo modo vengono identificate le variabili di input, di output e le regole di analisi relative alla valutazione della genuinità del caso in esame.
In questo modo viene costruita una prima versione della logica decisionale e un tester che funzioni come prototipo per l'analisi della corretta e completa traduzione delle procedure di valutazione dei dati comunicate degli esperti.
Il progetto realizzato contiene 8 variabili di input, due risultati parziali sulla valutazione della sospettabilità del contratto e dell'incidente e un indice complessivo di anomalia del caso esaminato. La base di conoscenza viene tradotta in 131 regole di produzione, come evidenziato dal prospetto sulle caratteristiche del progetto:

Input Variables	8
Output Variables	3
Intermediate Variables	0
Rule Blocks	3
Rules	131
Membership Functions	28

La struttura della logica decisionale può essere rappresentata in forma grafica:



Un primo test sulla completezza dalla logica inserita nel modello e sulla sua capacità discriminante di valutare in modo coerente i dati, può essere svolta attraverso un'analisi di sensitività del modello, collegando un foglio elettronico alla versione sorgente del sistema esperto:

Fraud claims

Analisi di sensitività

Polizza (contratto)					Veicolo ass. e incidente					Index
Età della polizza	Numero di sinistri	Frazionamento	Bonus-malus	Polizza	Età vea	Danno vec	Rischiosità provincia	Modalità incidente	Vea-Inc	Anomalia
0	0	1	1	67%	3	3500	4	14%	28%	46,1%
10	0	1	1	50%	3	3500	4	14%	28%	41,9%
20	0	1	1	40%	3	3500	4	14%	28%	37,9%
30	0	1	1	33%	3	3500	4	14%	28%	37,5%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	4	1	1	92%	3	3500	4	14%	28%	59,3%
10	10	1	1	100%	3	3500	4	14%	28%	64,2%
10	1	0	1	75%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	2	1	37%	3	3500	4	14%	28%	37,5%
10	1	1	0	28%	3	3500	4	14%	28%	36,3%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	7	3500	4	14%	18%	47,4%
10	1	1	1	71%	11	3500	4	14%	14%	44,7%
10	1	1	1	71%	3	1500	4	14%	12%	44,5%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	3	5000	4	14%	37%	54,2%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	3	3500	7	14%	40%	50,3%
10	1	1	1	71%	3	3500	12	14%	65%	61,8%
10	1	1	1	71%	3	3500	4	14%	28%	50,0%
10	1	1	1	71%	3	3500	4	60%	46%	54,0%
10	1	1	1	71%	3	3500	4	90%	64%	61,4%

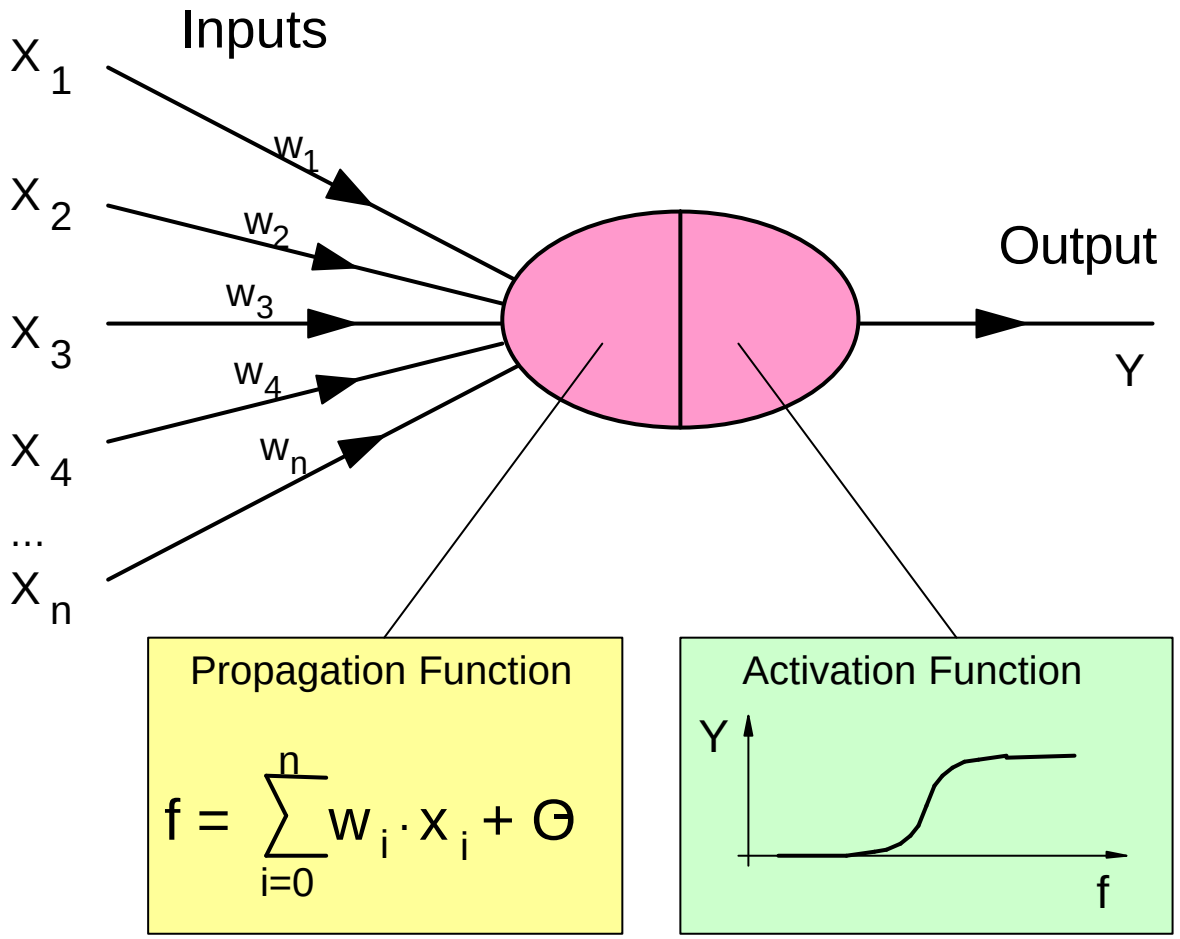
Attraverso questo semplice test, si può notare come il sistema si muova nella direzione indicata dall'esperto ogni volta che il valore di una variabile cambia nella direzione di un maggiore o minore rischio di frode.
Nella tabella riportata si possono esaminare i risultati della classificazione eseguita dal sistema esperto, in relazione al diverso grado di sospetto della richiesta di rimborso. Accanto all'indice di valutazione complessiva del caso esaminato (Indice Anomalia), il modello fornisce indicazioni su come il risultato finale venga a formarsi, attraverso la valutazione separata di differenti aspetti del caso (sospettabilità del contratto, modalità di incidente).
Attraverso la lettura di tutti i risultati (intermedi e finali) e delle regole di produzione attivate nel caso in esame (rule analizer), l'utente può valutare in modo trasparente le modalità e le ragioni della classificazione eseguita dal modello, correggendo eventuali errori o mancanze nel sistema di regole inserite.

Term	RB	IF	Aggr.	DoS	Res.A...
low = 0.50774	vea_inc	incRiscLuogo = low	1.00000	0.20313	0.20312
	vea_inc	Incmodalita = low	0.95996	0.20313	0.19498
	vea_inc	dannoVec = medio & Incmodalita = low & incRiscLuogo = low & veaEta = bassa	0.50000	0.10156	0.05076
	vea_inc	dannoVec = alto & Incmodalita = low & incRiscLuogo = low & veaEta = bassa	0.50000	0.10156	0.05076
	vea_inc	dannoVec = alto & Incmodalita = medium & incRiscLuogo = low & veaEta = bassa	0.04002	0.10156	0.00406
	vea_inc	dannoVec = medio & Incmodalita = medium & incRiscLuogo = low & veaEta = bassa	0.04002	0.10156	0.00406
high = 0.20310	vea_inc	veaEta = bassa	1.00000	0.10156	0.10154
	vea_inc	dannoVec = alto	0.50000	0.20313	0.10154

L'analisi del sistema prevede una fase di ottimizzazione, spesso svolta con il prototipo (off-line optimization) fino a quando gli esperti e l'ingegnere della conoscenza non verificano la completezza e l'accuratezza del modello.
Terminata la fase di *fine tuning* sul prototipo, il sistema esperto può essere compilato e integrato nel sistema informativo dell'utente. In caso vengano scoperti, successivamente all'integrazione, nuovi pattern di frode e/o cambino le procedure di valutazione da parte delle Compagnie, la logica decisionale può essere modificata "a caldo" semplicemente aggiornando e ricompilando la base di conoscenza.
Questa possibilità di aggiornare o modificare in tempo reale la logica di valutazione è tipica dei sistemi deduttivi e non può verificarsi con sistemi induttivi che devono "ricalcolare" la conoscenza sulla base dei dati più recenti.

2. Addestramento neurale di un sistema esperto

Le reti neurali (o ANN, artificial neural network) offrono un modello matematico dove le informazioni (inputs) sono rappresentate da un insieme di nodi (neuroni, in analogia ai neuroni del cervello) connessi in modo pesato con altri nodi collocati su diversi strati (layer).



I singoli nodi incorporano gli input ricevuti dai nodi connessi e utilizzano i pesi insieme ad una semplice funzione (sigmoide) per calcolare i valori di output. L'analogia col cervello umano è sinteticamente espressa dal fatto che l'attivazione (il maggior peso) di un input produce una attivazione degli input connessi in strati successivi e dell'output finale (e viceversa se il peso diminuisce). L'apprendimento viene effettuato tramite la modifica dei pesi di connessione mentre un insieme di input viene inserito nella rete.
Se la rete mantiene costante un certo vettore di input nella fase di addestramento viene definita unidirezionale (feed-forward). L'addestramento di backpropagation (il metodo più utilizzato), calcola i valori di output relativi a quel vettore di input e al valore dato dei pesi di connessione e li confronta ad un vettore di output desiderato. Calcolato l'errore tra i valori di output ottenuti e quelli desiderati, l'algoritmo di backpropagation cerca di minimizzare l'errore modificando il vettore dei pesi di connessione che generano quell'errore. Ottenuto un errore inferiore, il processo viene ripetuto e, dopo un certo numero di iterazioni, l'algoritmo fornisce un modello che converge producendo un risultato vicino a quello desiderato, a parità di input, attraverso la ripetuta modifica dell'insieme dei pesi.

Una ANN è dunque un metodo efficace per stabilire con quali pesi debbano essere aggregati ed elaborati gli input per ottenere un certo vettore di output.
La costruzione e l'addestramento di una rete neurale non pone alcun vincolo alle relazioni tra input e output, se non la presenza di un legame pesato.
In questa sede viene presentato l'addestramento di un sistema esperto rule based, dove gli stessi principi (rete feed forward) e la stessa tecnica di addestramento (backpropagation) vengono applicati ad un modello che contiene unicamente la definizione delle relazioni tra le variabili, senza però indicare il valore delle relazioni (cioè dei pesi di connessione o DoS, degree of support).
Si supponga ad es. che la procedura di valutazione precedentemente costruita indichi nel seguente blocco di regole un sostanziale equilibrio dell'importanza delle variabili utilizzate, con l'eccezione dell'età del veicolo assicurato, che ha un impatto inferiore (DoS = 10% anziché 20%) sia come fattore aggravante (iinc=high) che come fattore esimente (iinc=low). Questa differenza di valutazione è ovviamente relativa alla volontà del decisore che considera l'età del veicolo assicurato una variabile meno discriminante delle altre 3 inserite nel blocco di regole.

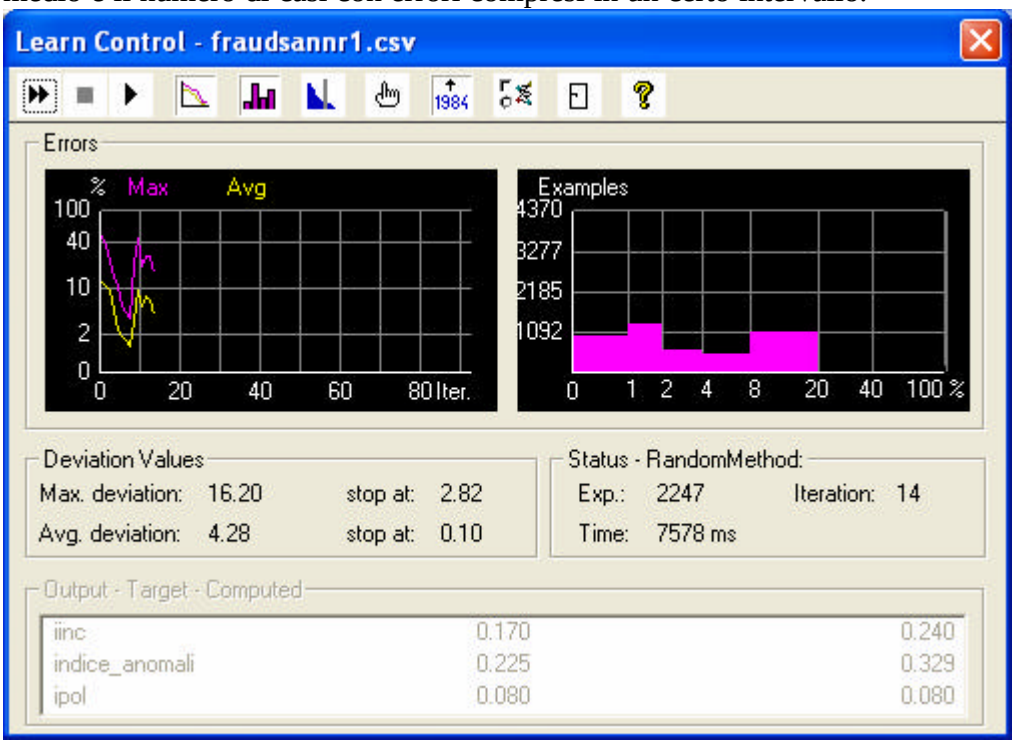
#	IF				THEN	
	dannoVec	Incmodalita	incRiscLuogo	veaEta	DoS	iinc
1	basso				0.20	low
2	alto				0.20	high
3		low			0.20	low
4		high			0.20	high
5			low		0.20	low
6			high		0.20	high
7				bassa	0.10	high
8				alta	0.10	low
9						

Nel caso si voglia addestrare il sistema per estrarre dai dati l'importanza delle singole regole (pesi o DoS), il vettore dei pesi viene imposto a 0 e lasciato libero di variare in modo del tutto equivalente alla procedura di addestramento di una ANN.

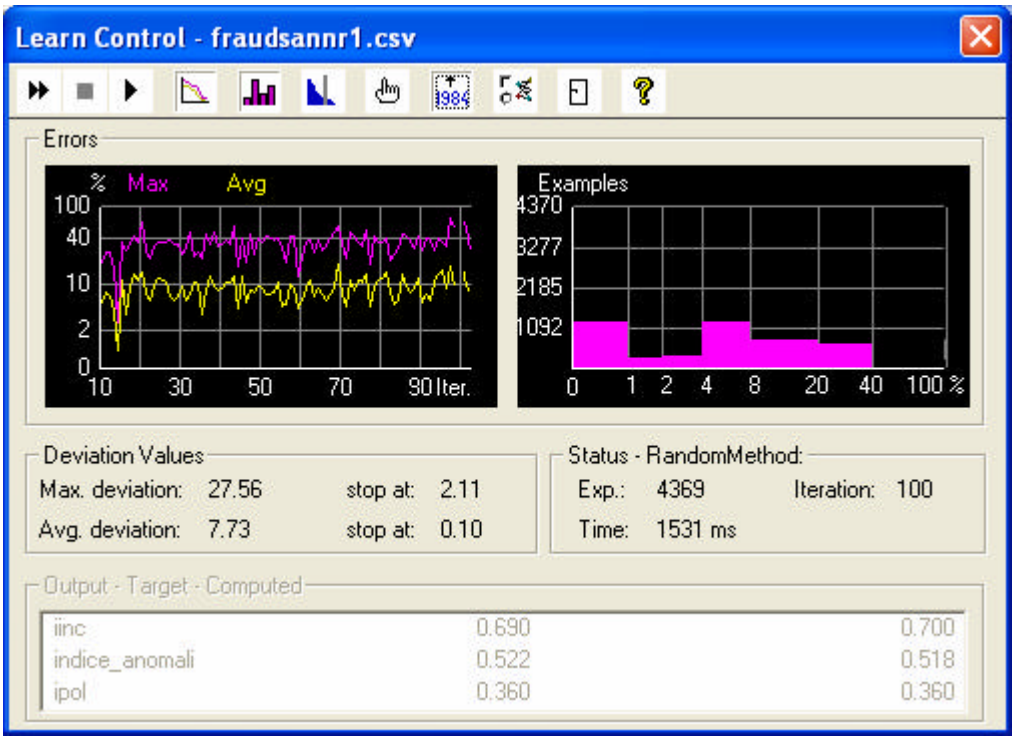
#	IF				THEN	
	dannoVec	Incmodalita	incRiscLuogo	veaEta	DoS	iinc
1	basso				0.00	low
2	alto				0.00	high
3		low			0.00	low
4		high			0.00	high
5			low		0.00	low
6			high		0.00	high
7				bassa	0.00	high
8				alta	0.00	low
9						

Così come nel caso generale, i dati di input vengono processati col primo vettore di pesi a disposizione (pari a 0) e l'output generato viene confrontato con un vettore di output desiderato o, nel caso di dati esistenti, reale e relativo a casi passati e già valutati.
L'algoritmo di backpropagation cercherà di minimizzare l'errore modificando ad ogni iterazione il vettore del grado di supporto delle regole (DoS, pesi di connessione) fino a un set di pesi ottimo che permette alla rete di convergere ad un valore minimo di errore totale (o al soddisfacimento di altre possibili condizioni terminali, ad es. un numero massimo di iterazioni).

Durante il processo di apprendimento, è possibile controllare il valore di errore massimo, di errore medio e il numero di casi con errori compresi in un certo intervallo.



Al termine dell'addestramento, l'algoritmo fornisce i risultati sulla qualità del processo e il set di pesi ottimi rispetto al problema.



Se l'algoritmo riesce a raggiungere risultati soddisfacenti, l'insieme dei pesi individuati è quello necessario a determinare un certo risultato (corretto, desiderato o passato), a partire da un certo vettore di input e dal sistema di relazioni tra le variabili indicate. (Nel caso considerato si può osservare la scarsa qualità del risultato ottenuto, in relazione alla natura poco coerente dei dati utilizzati per l'addestramento).

Nel caso il set di pesi individuato (in modo efficace) dall'algoritmo di addestramento diverga sensibilmente con quelli indicati dall'esperto, sono possibili due casi:

- la procedura di valutazione del processo decisionale di quella azienda è cambiata rispetto al passato
- la procedura di valutazione del processo decisionale di quella azienda è stata applicata, in passato, in modo diverso da quella dichiarata come ortodossa

#	IF				THEN	
	dannoVec	Incmodalita	incRiscLuogo	veaEta	DoS	iinc
1	basso				0.41	low
2	alto				0.23	high
3		low			0.22	low
4		high			0.29	high
5			low		0.19	low
6			high		0.11	high
7				bassa	0.09	high
8				alta	0.10	low
9						

Osservando i risultati del processo di apprendimento del caso di esempio (vedi figura), si può osservare che i pesi calcolati sulla base dei dati storici (che rappresentano le decisioni passate) sono diversi da quelli stabiliti dal decisore al momento della creazione della logica decisionale contenuta nel sistema esperto (che, in questo caso, si presume perfettamente fedele alle procedure).
Ad es., nel caso della regola 1:

Se Dannovec (danno al veicolo controparte) = basso allora iinc (rischiosità dell'incidente) =low

il peso di attivazione è sensibilmente maggiore (0,41) a quello indicato nel sistema non addestrato (0,2). Questo può significare una riduzione del significato esimente dell'evento nelle procedure più recenti di valutazione oppure una maggior importanza assegnata alla regola nei casi passati e precedentemente valutati dagli esperti anti frode. (ad es. i casi con danno "basso" al veicolo controparte sono stati considerati nella realtà più genuini di quanto indicato dalle procedure).
Un utilizzo non convenzionale dell'addestramento di un sistema rule based consiste nel processare i dati riferiti al passato con un sistema di regole che rappresenti fedelmente la procedura decisionale, ma senza indicazione dei pesi di attivazione delle regole.
Posto che il sistema esperto rispecchi fedelmente le procedure aziendali e il sistema di addestramento soddisfi le condizioni di convergenza, eventuali differenze possono essere discusse con il decisore nel senso di verificare quanto, in passato, tali procedure siano state effettivamente rispettate.
Come ultima possibilità (finora non analizzata), l'addestramento di un sistema rule based consente di evidenziare possibili errori o mancanze nella determinazione dei pesi delle regole contenute nella logica decisionale da parte dell'esperto o dell'ingegnere di conoscenza.

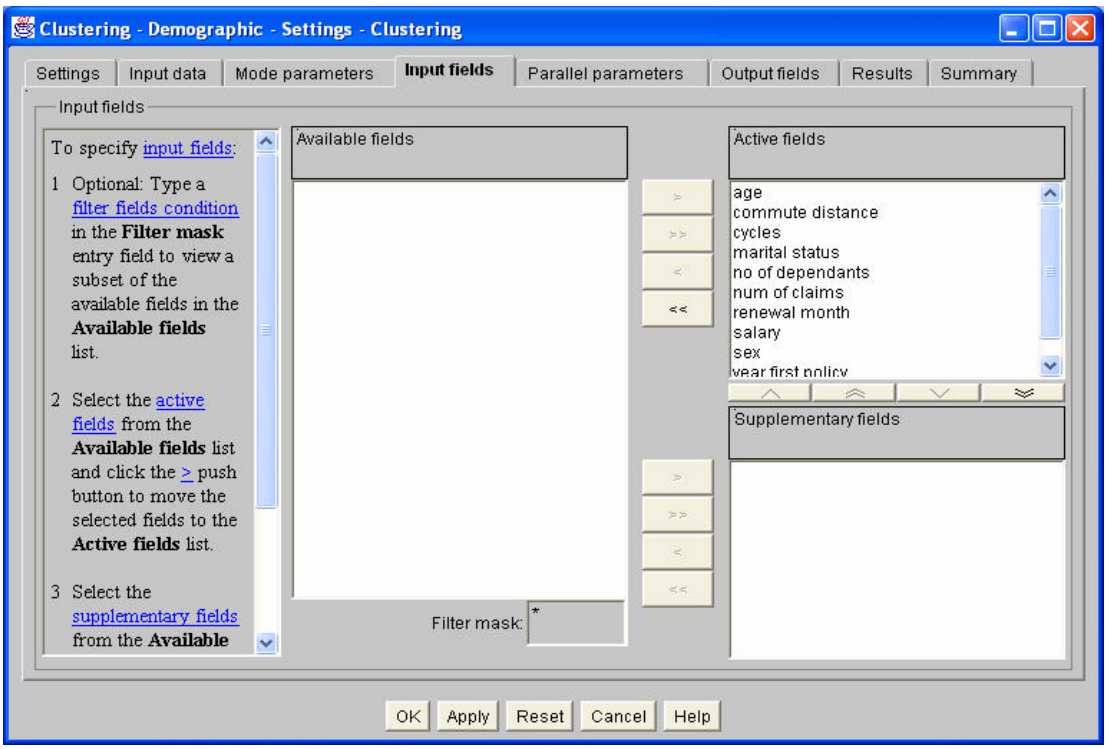
3. Demographic Clustering (Segmentazione)

L'obiettivo della segmentazione è quello di suddividere il database in segmenti di record simili, cioè in gruppi di record che condividono un certo numero di proprietà e possono essere considerati omogenei.
Per definizione, i record in differenti segmenti sono diversi per qualche aspetto. Calcolato il centro del segmento e le distanze tra questo e i record (omogeneità), i segmenti identificati nel processo di data mining debbono raggiungere un' alta omogeneità interna e un'alta eterogeneità tra i diversi segmenti.
La segmentazione di un Database viene effettuata per evidenziare sotto popolazioni omogenee che consenta di migliorare l'accuratezza dei profili (ad es. della clientela in un contesto di CRM). Una sotto popolazione che può essere "di individui maschi, anziani, in buona salute" o "donne, professioniste, a reddito elevato" può essere avvicinata con una politica di marketing specifica e maggiormente personalizzata.
L'algoritmo di segmentazione può contenere oppure no indicazioni dall'utente sul numero e le caratteristiche a priori dei segmenti da calcolare. Se non esistono indicazioni, il metodo viene denominato di apprendimento non supervisionato (unsupervised learning).
La demographic clustering opera principalmente su record aventi valori categorici. La tecnica di misurazione delle similarità tra i record (distanza euclidea o prossimità se il database comprende anche dati quantitativi) è basato su un metodo di voto denominato Condorset e produce un output di segmentazione non influenzato da forme gerarchiche predefinite.

Caso: **Fraudulent claims**
Tecnica: Clustering - demographic
Software: DB2 Intelligent Miner
Obiettivo: individuare segmenti ed assegnare i record contenuti nel DB a specifici cluster

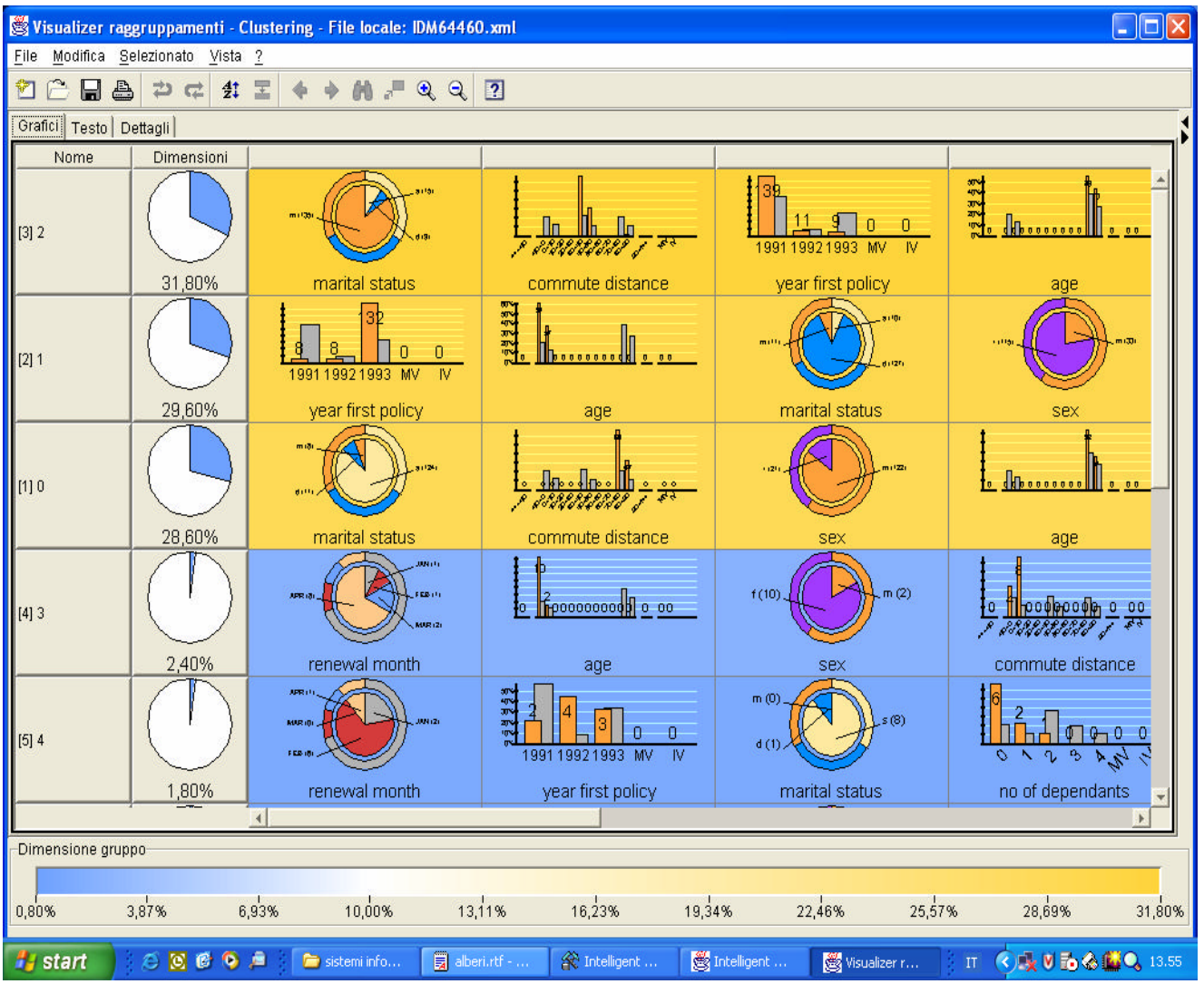
Database: flat, 500 record
Campi: age, commute distance, cycles, marital status, no of dependants, num of claims, renewal month, salary, sex, year first policy
Dati di esempio: Sample data for clustering

age	commute distance	cycles	marital status	no of dependants	num of claims	renewal month	salary	sex	year first policy
49.83	-3.28	1	"d"	3	3	"JAN"	39248.4	"f"	1991
49.85	4.13	8	"m"	2	6	"JAN"	39273.5	"f"	1991
50.07	1.16	5	"d"	4	1	"JAN"	40044	"m"	1993
95.57	40.19	1	"m"	0	4	"JAN"	39923.3	"m"	1991
95.82	78	1	"s"	2	8	"JAN"	40222	"f"	1991
52.47	-4.78	5	"d"	2	1	"APR"	39281.6	"m"	1992
49.31	-9.78	1	"d"	2	2	"JAN"	40066.4	"f"	1993
46.85	1.74	3	"m"	2	1	"JAN"	39002	"m"	1993
48.32	0.27	4	"d"	3	4	"JAN"	39038.1	"f"	1993
48.01	-4.23	5	"m"	3	1	"APR"	39733.5	"m"	1993
49.98	2.19	5	"d"	3	1	"JAN"	39374.2	"f"	1993
95.11	35.26	1	"m"	0	4	"JAN"	39402.4	"m"	1991

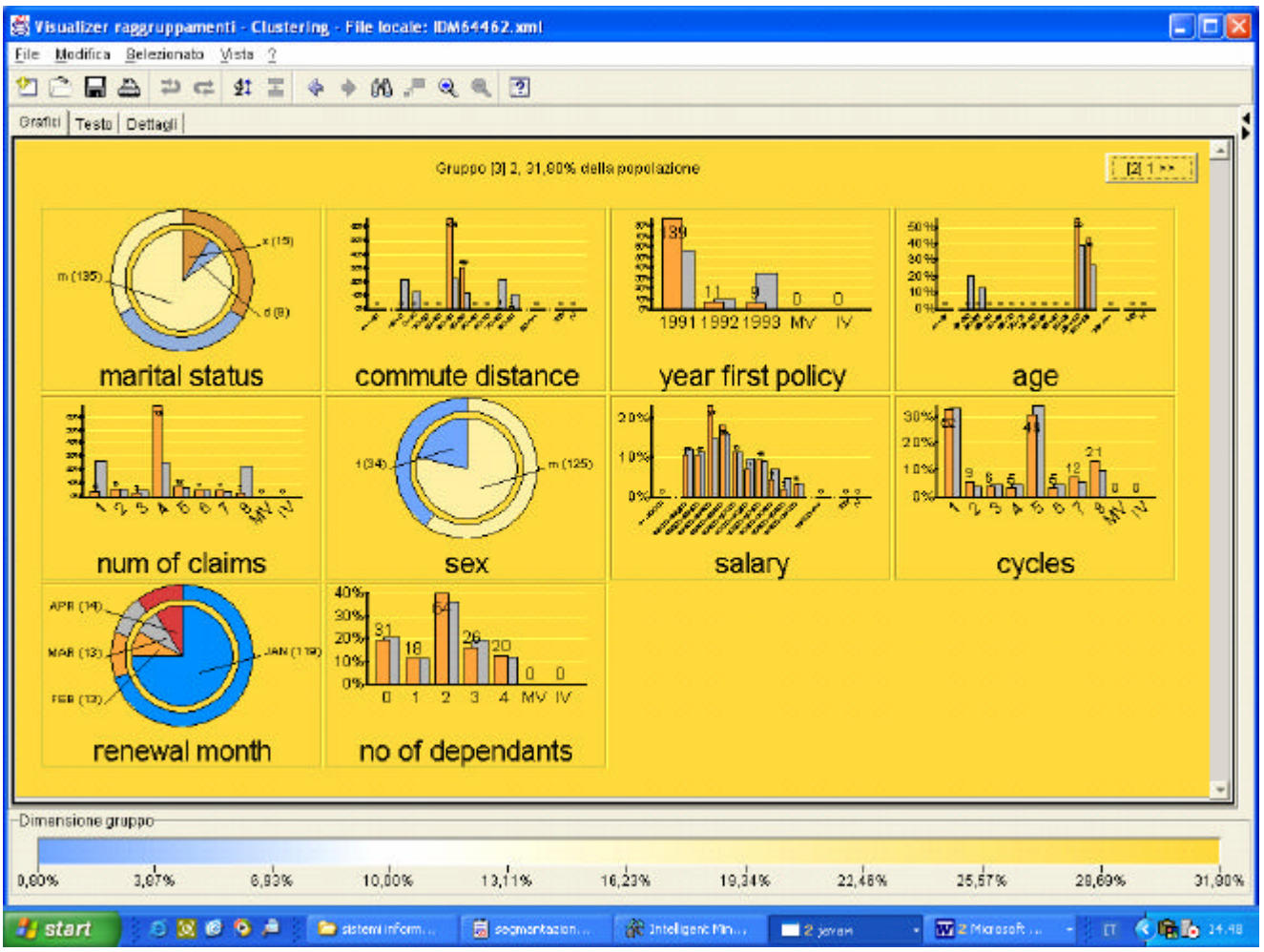


Caratteristiche dei Cluster identificati

- [3] 2 (31,80%)
  marital status è predominante m, ***commute distance*** è medio, ***year first policy*** è predominante 1991, ***age*** è alto, ***num of claims*** è predominante 4, ***sex*** è predominante m, ***salary*** è medio, ***cycles*** è predominante 1, ***renewal month*** è predominante JAN e ***no of dependants*** è predominante 2.
- [2] 1 (29,60%)
  year first policy è predominante 1993, ***age*** è basso, ***marital status*** è predominante d, ***sex*** è predominante f, ***commute distance*** è basso, ***num of claims*** è predominante 1, ***renewal month*** è predominante JAN, ***no of dependants*** è predominante 2, ***salary*** è medio e ***cycles*** è predominante 5.
- [1] 0 (28,60%)
  marital status è predominante s, ***commute distance*** è alto, ***sex*** è predominante m, ***age*** è alto, ***num of claims*** è predominante 8, ***year first policy*** è predominante 1991, ***salary*** è medio, ***no of dependants*** è predominante 2, ***renewal month*** è predominante JAN e ***cycles*** è predominante 5.
- [4] 3 (2,40%)
  renewal month è predominante APR, ***age*** è basso, ***sex*** è predominante f, ***commute distance*** è basso, ***marital status*** è predominante d, ***year first policy*** è predominante 1991, ***num of claims*** è predominante 1, ***cycles*** è predominante 1, ***salary*** è medio e ***no of dependants*** è predominante 2.
- [5] 4 (1,80%)
  renewal month è predominante FEB, ***year first policy*** è predominante 1992, ***marital status*** è predominante s, ***no of dependants*** è predominante 0, ***sex*** è predominante f, ***num of claims*** è predominante 4, ***cycles*** è predominante 1, ***commute distance*** è alto, ***age*** è alto e ***salary*** è basso.



Durante il processo di segmentazione, sono disponibili alcuni parametri di controllo dell'attività di mining sui dati, come ad es. il numero di segmenti in cui si desidera suddividere la popolazione, il numero di iterazioni sui dati che caratterizza la qualità del processo di segmentazione e le condizioni terminali (stopping condition) sul livello di accuratezza dei risultati.
La figura mostra i primi 5 maggiori segmenti identificati dall'algoritmo di data mining. Il numero sul lato sinistro rappresenta la dimensione del segmento, espressa in percentuale sul totale della popolazione. All'interno di ogni segmento, le variabili utilizzate (attive), sono rappresentate da sinistra a destra in ordine di importanza relativa, utilizzando l'indice di significatività statistica chi-quadrato. Ad esempio, per il segmento 2 la variabile "anno di prima polizza" e "età del contraente" sono indicate come le variabili più significative.
Ogni segmento ed ogni variabile al suo interno può essere visualizzata in un sotto grafico di tipo istogramma per le variabili quantitative o a torta per quelle categoriche. Ingrandendo ad es. lo studio relativo al primo segmento, si può notare come questo sia formato principalmente da persone anziane, soprattutto di sesso maschile, sposati, con polizze anziane ed un numero medio di sinistri.
Analizzando ancor più in dettaglio la distribuzione di una delle variabili indicate, si può vedere come la distanza dal luogo di lavoro sia, in questo segmento, particolarmente significativa ed alta rispetto alla distribuzione del totale della popolazione.



Statistiche dei gruppi

La tabella Statistiche fornisce una rapida panoramica delle seguenti informazioni:

- I nomi dei gruppi
- Le dimensioni assolute dei gruppi
- Le dimensioni dei gruppi in relazione al numero totale di record. Ad esempio, un valore del 20% nella colonna Dimensione vuol dire che il 20% di tutti i record appartengono a questo gruppo.
- L'omogeneità dei gruppi. Il fattore di omogeneità indica quanto sono simili i record che appartengono ad un gruppo. La tabella di similarità tra i gruppi mostra la similarità tra i gruppi. Ogni gruppo è confrontato a tutti gli altri gruppi. Il valore di similarità deve essere compreso tra 0 e 1,0.
    - 0 indica che i gruppi sono completamente diversi.
    - 1 indica che i gruppi sono identici.

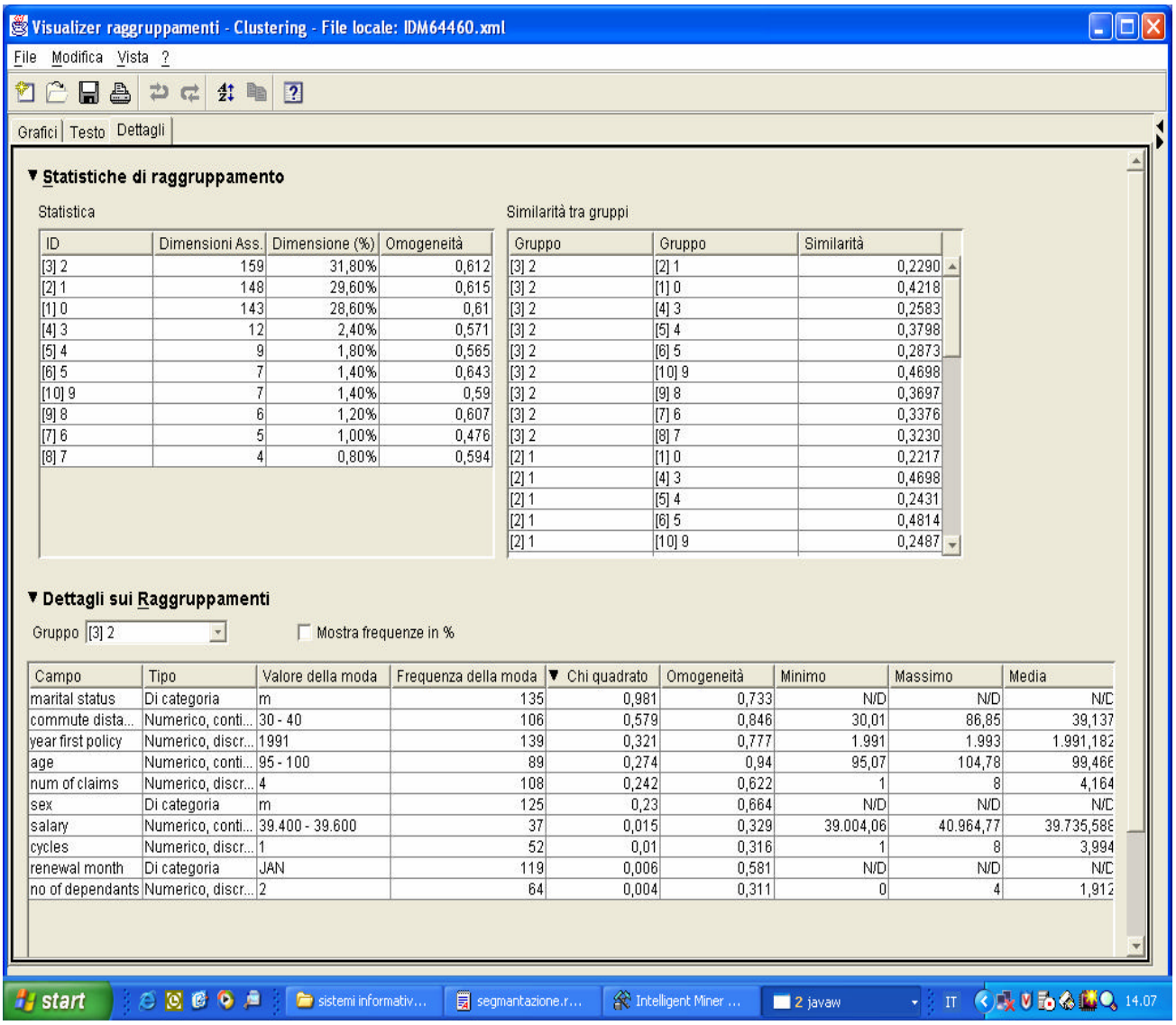
Dettagli sui gruppi

In questa sezione, sono visualizzate le seguenti informazioni per ciascun gruppo:

- Campo
- Tipo
- Valore della moda
- Frequenza della moda
- Chi quadrato
- Omogeneità

Se il campo non è numerico, o se le informazioni non sono disponibili nelle statistiche del modello, i seguenti campi non contengono alcun valore. Questo viene evidenziato con la stringa N/D (non disponibile).

- Minimo
- Massimo
- Media
- Deviazione standard
- Unità di distanza
- Valore aggregato



Dettagli sui campi

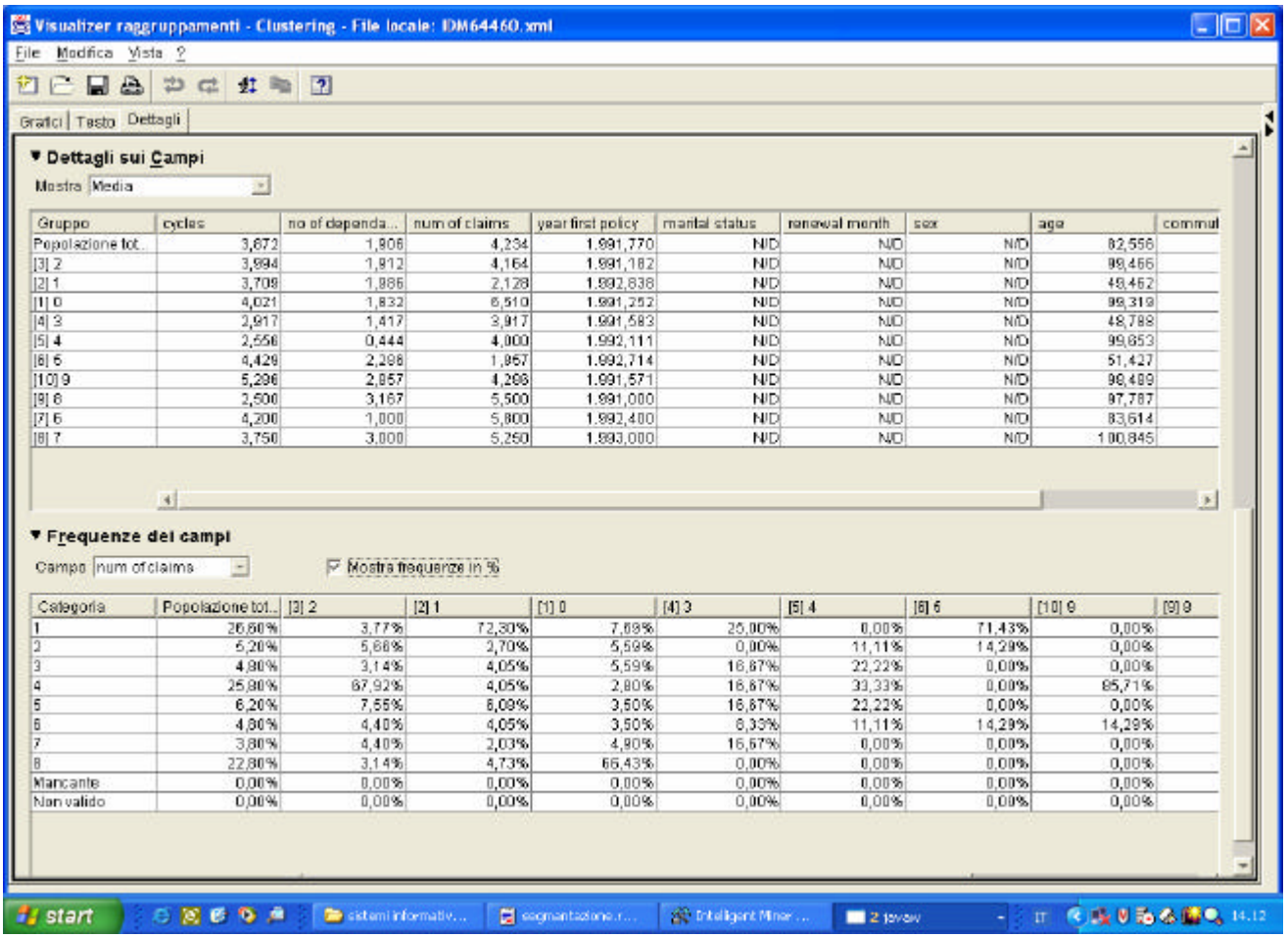
In questa sezione, è possibile confrontare i seguenti valori dei campi tra i diversi gruppi:

- Chi quadrato
- Omogeneità
- Massimo
- Media
- Minimo
- Frequenza della moda

- Valore della moda
- Deviazione standard
- Valore aggregato

Frequenza dei campi

In questa sezione, è possibile confrontare le frequenze dei singoli campi appartenenti ad un gruppo.



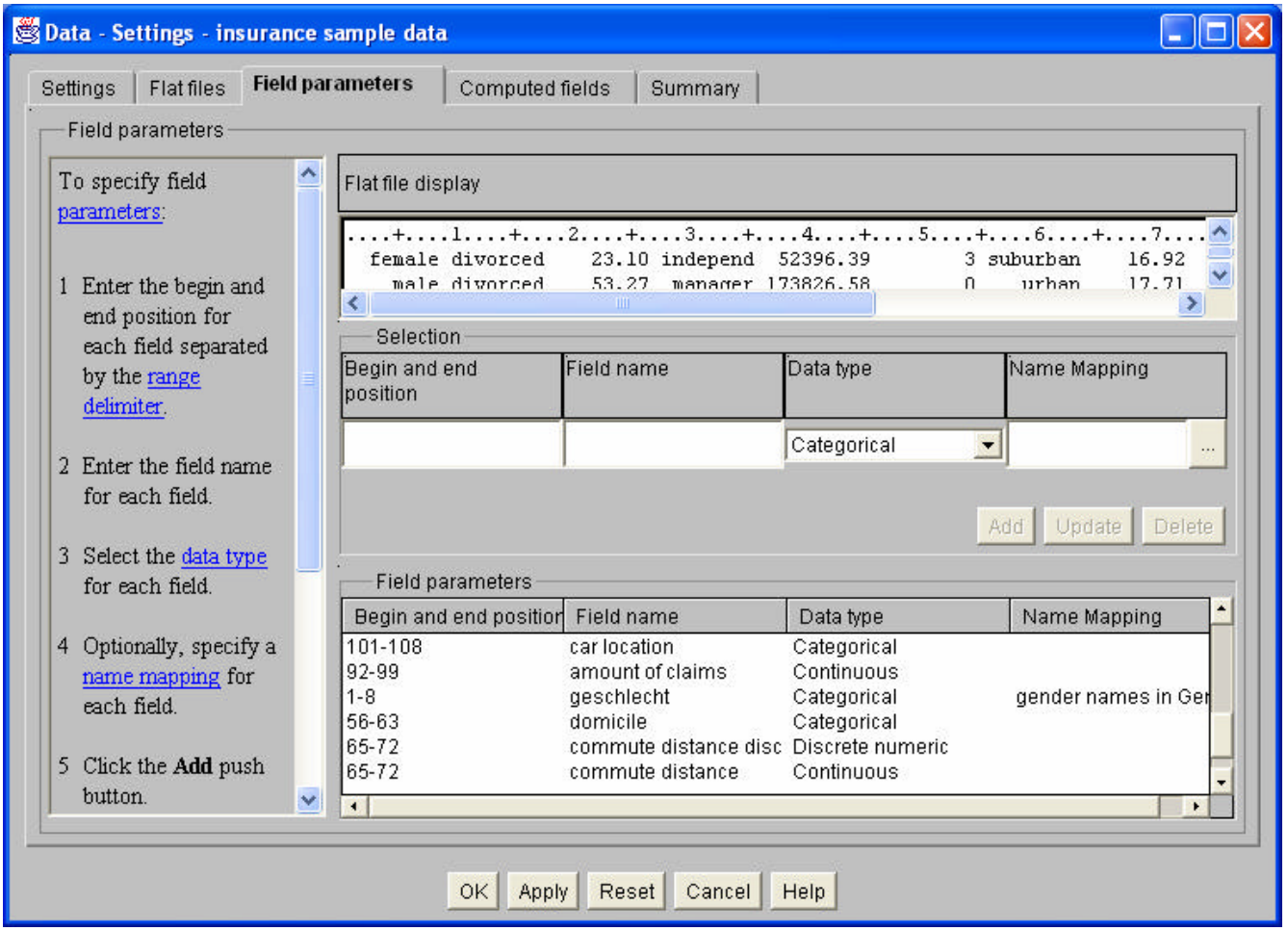
4. Alberi decisionali (predictive modelling)

Gli alberi decisionali sono basati su un processo supervisionato di apprendimento che permette di creare un modello di classificazione dei record che compongono un database, attraverso lo studio di un campione già classificato della popolazione, denominato training set.
Il modello indotto consiste nell'identificazione di schemi (pattern), essenzialmente generalizzazioni basate sull'osservazione dei record, che permettono di distinguere varie classi di record sulla base delle caratteristiche dei dati. Una volta costituito il modello, questo può essere utilizzato per definire la classe di record successivi e non classificati. In alternativa alla classificazione di una popolazione rispetto ad una variabile risultato, l'albero decisionale può essere utilizzato per comprendere come si forma un certo risultato analizzando la distribuzione dei dati. Nell'esempio presentato viene classificato un campione di record che contiene la variabile "numero di denuncie di incidenti" fatte ad un compagnia assicurativa. Elaborate le regole di classificazione, il modello può essere utilizzato per allocare correttamente set diversi (nuovi) di record o, alternativamente, per comprendere le relazioni di causa effetto che spiegano la formazione di un differente numero di denuncie.
Le tecniche di induzione supervisionata offrono vantaggi rispetto a modelli statistici, in quanto permettono di identificare fenomeni locali, mentre l'approccio statistico è basato su misurazioni dell'intera popolazione con distribuzione nota. Ad es. una variabile A potrebbe avere una scarsa correlazione diretta con una variabile "risultato" C di un'intera popolazione, ma essere particolarmente predittiva e significativa su un risultato intermedio (o un'altra variabile B) di una sotto popolazione, a sua volta fortemente correlata alla variabile risultato.
In questo caso i test statistici indicherebbero una relativa indipendenza tra le variabili A e C, data la scarsa significatività del sotto insieme dove si verifica la relazione tra le variabili A e B.
Il metodo di costruzione dell'albero decisionale prevede l'individuazione della variabile più determinante alla formazione del risultato in quello stadio di avanzamento. La scelta delle variabili determina la dimensione dell'albero e l'algoritmo (CART) deve quindi contenere criteri che minimizzino il numero di livelli e nodi dell'albero, massimizzando contemporaneamente il guadagno di informazioni. L'applicazione degli alberi decisionali è limitata ai problemi che possono essere risolti dividendo lo spazio delle soluzioni in rettangoli di dimensioni progressivamente inferiori.
Gli alberi decisionali sono utilizzati in problemi di classificazione (scoring, ranking), tipici di obiettivi di Risk management (customer retention, classificazione di clienti e fornitori, concessione del credito, individuazione di frodi, utilizzo anomalo di carte di credito…) e di Market management (cross e up selling, target marketing, segmentazione e profilazione della clientela…).

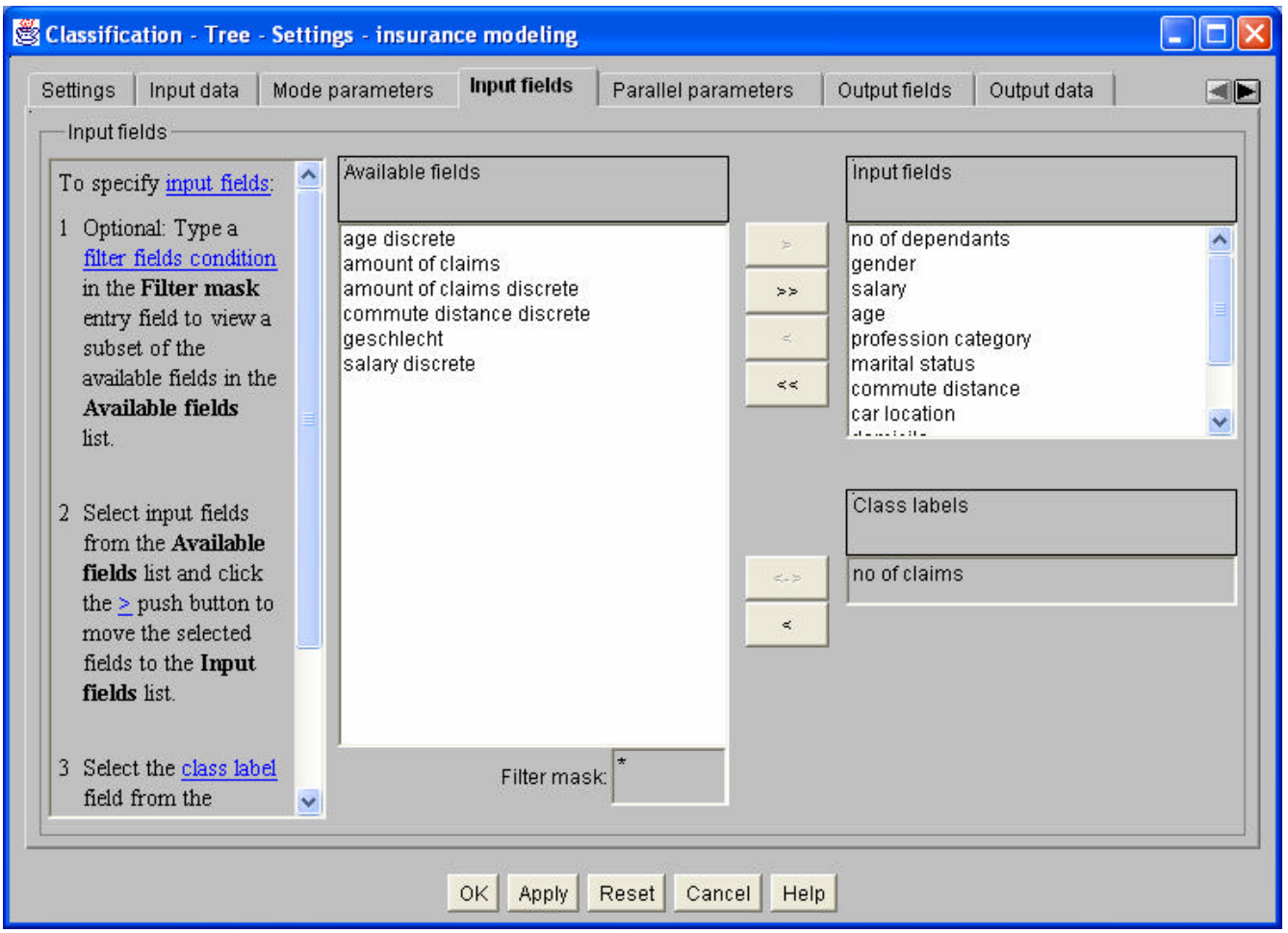
Caso: **Fraudulent claims**
Tecnica: Tree induction
Software: DB2 Intelligent Miner
Obiettivo: individuare un modello predittivo nella forma di albero decisionale che consenta di spiegare la formazione di un certo numero di richieste di rimborso assicurativo

Database: flat, 500 record
Campi: age, age of car, amount of claims, car location, car type, domicile, gender, profession, commute distance, marital status, no of dependants, num of claims, salary,
Dati di esempio: Insurance sample data

age	age of car	Amount of claims	car location	car type	commute distance	domicile	gender	marital status	no of claims	no of dependants	profession category	salary
23.1	3.24	760.59	" carpark"	" sedan"	16.92	"suburban"	" female"	"divorced"	" 0"	" 3"	"independ"	52396.4
53.27	2.24	2528.95	" garage"	" compact"	17.71	" urban"	" male"	"divorced"	" 0"	" 0"	" manager"	173827
49.4	3.52	1261.2	" carpark"	" compact"	18.72	" rural"	" female"	"divorced"	" 1"	" 0"	" worker"	51248.9
37.1	6.31	1969.96	" street"	" van"	17.97	"suburban"	" male"	" married"	" 1"	" 0"	" worker"	66357
30.88	3.18	1406.83	" carpark"	" sedan"	12.09	" rural"	" male"	" married"	" 3"	" 0"	" manager"	28984.3
27.36	1.47	410.03	" carpark"	" sedan"	16.4	" rural"	" female"	" married"	" 1"	" 2"	"employee"	48809.6
49.9	1.6	2940.89	" garage"	" sports"	9.63	" urban"	" male"	" single"	" 1"	" 1"	"independ"	152039



La tecnica prevede l'elaborazione dei campi (variabili di input) indicati come discriminanti per il riparto dei record in diverse categorie di output (risultato) ed evidenzia, in stadi successivi, le variabili più importanti nel determinare la classificazione (nodi) e le possibili regole di riparto dei dati che soddisfano la relazione input-output con una certa percentuale di casi sul totale (purezza). Il valore finale della regola (l'ultima variabile-nodo evidenziata) viene denominata foglia dell'albero.
Il software di Data Mining permette diverse visualizzazioni dei risultati:



Albero decisionale

La Vista albero visualizza i risultati dell'estrapolazione della classificazione ad albero in formato tabellare. L'albero è costituito da nodi differenti che corrispondono ad una riga della tabella. L'effetto di ogni riga è quello di evidenziare una regola di classificazione che indichi la classe di riparto prevista (cioè il punteggio) e il numero di casi in cui la regola funziona nel database esaminato (purezza della regola) rispetto al numero totale dei record. Le colonne della tabella visualizzano gli attributi dei nodi:

Albero

Ciascun nodo nella colonna Albero include un diagramma che mostra la distribuzione dei record nel nodo per le classi previste. Ognuna delle classi previste è rappresentata da un colore diverso. L'intensità di colore di un nodo rappresenta la percentuale di record che appartengono alla classe dipendente. Accanto al diagramma viene visualizzata la decisione dipendente.

ID nodo

L'ID del nodo mostra i livelli calcolati dell'albero. Ciascun nodo nell'albero è identificato da un ID nodo. L'ID nodo è creato aggiungendo all'ID del nodo di livello superiore ".x", dove x è la posizione del nodo rispetto agli altri nodi del livello. Il nodo "radice" ha l'ID 1.

Punteggio

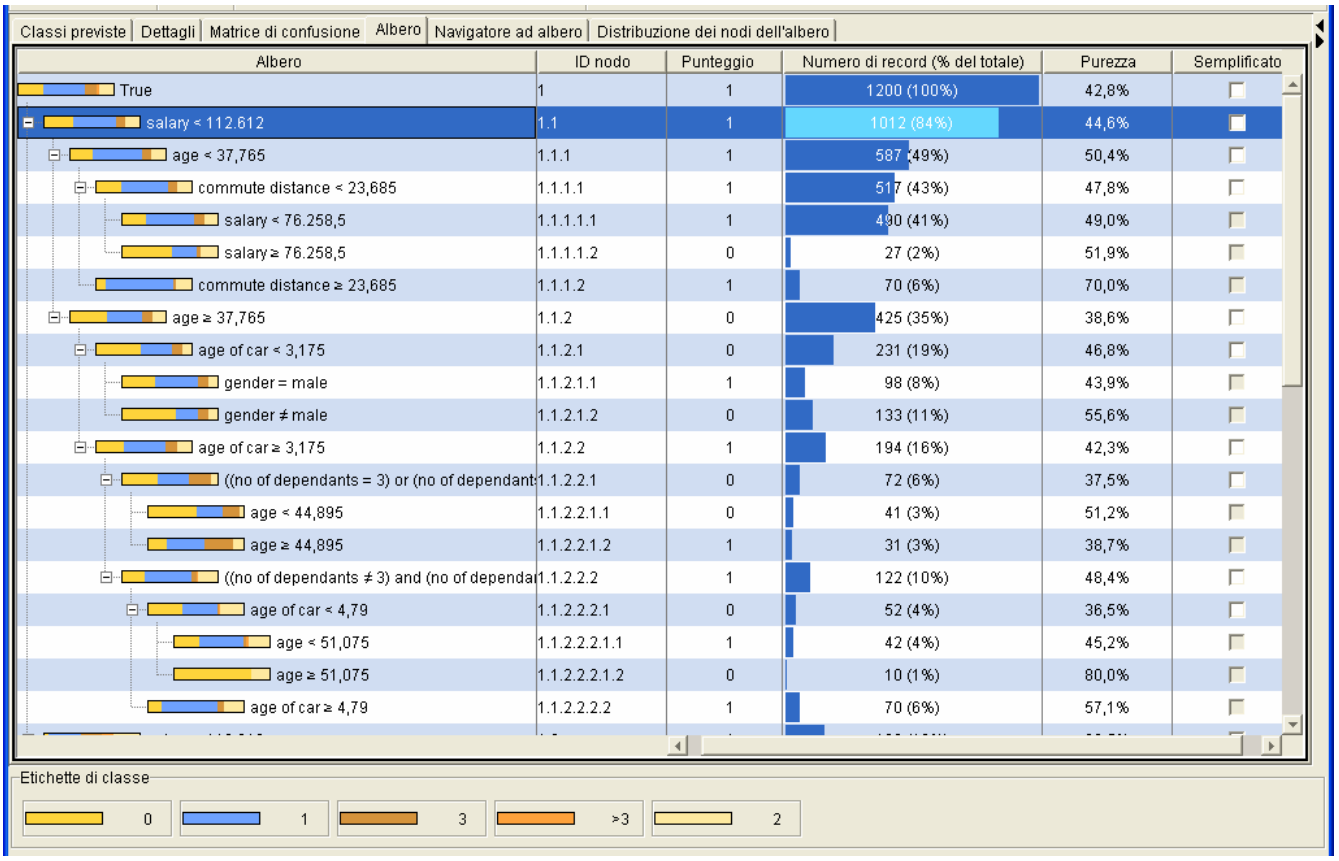
Il punteggio di un nodo mostra la classe in base alla quale sono stati previsti tutti i record del nodo. Ad esempio, si possono avere i punteggi 0, 1, 2, 3 o >3, corrispondenti al numero di denuncie effettuate.

Numero di record

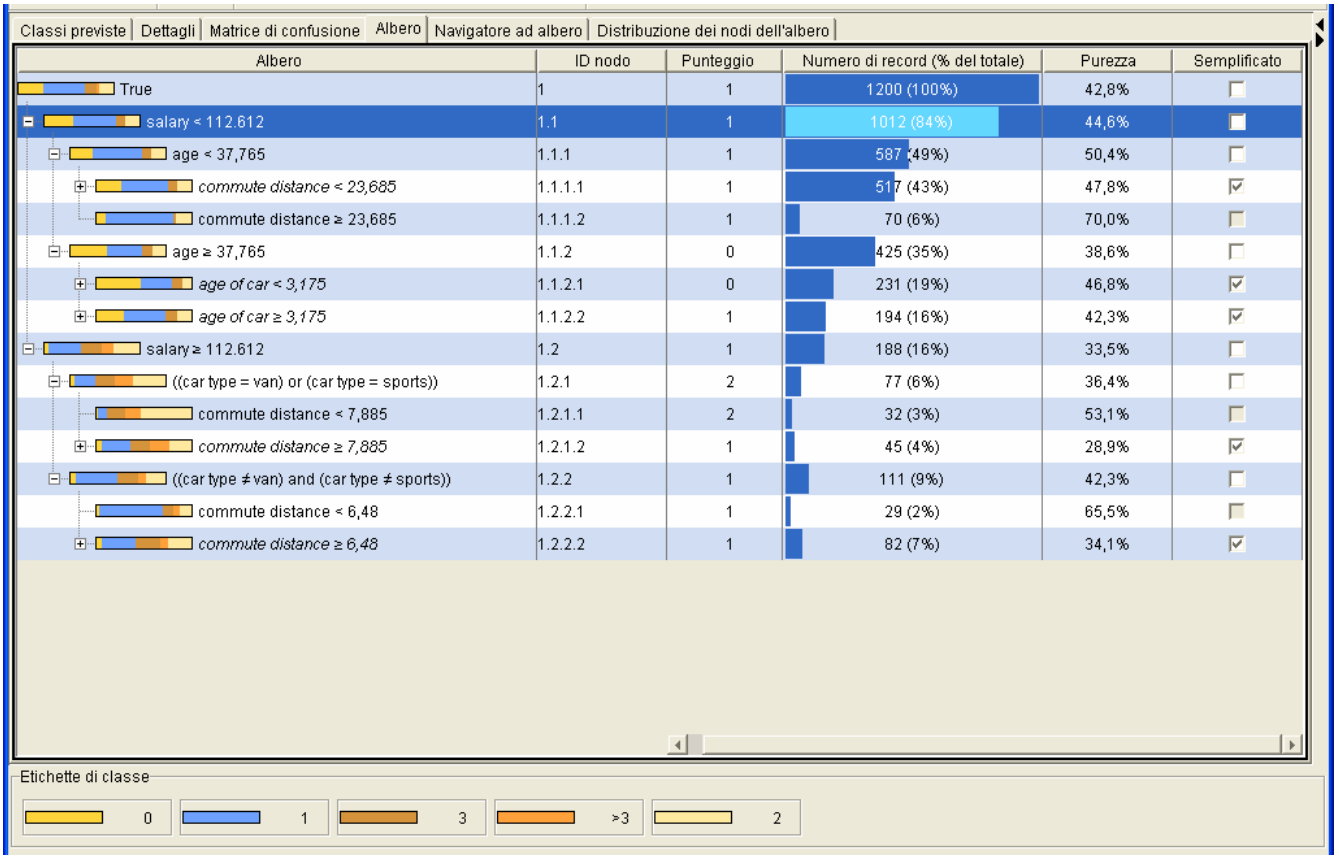
Il numero di record mostra il numero di record e la relativa percentuale in confronto con la popolazione totale. Inoltre, la percentuale del numero di record viene visualizzata da un istogramma.

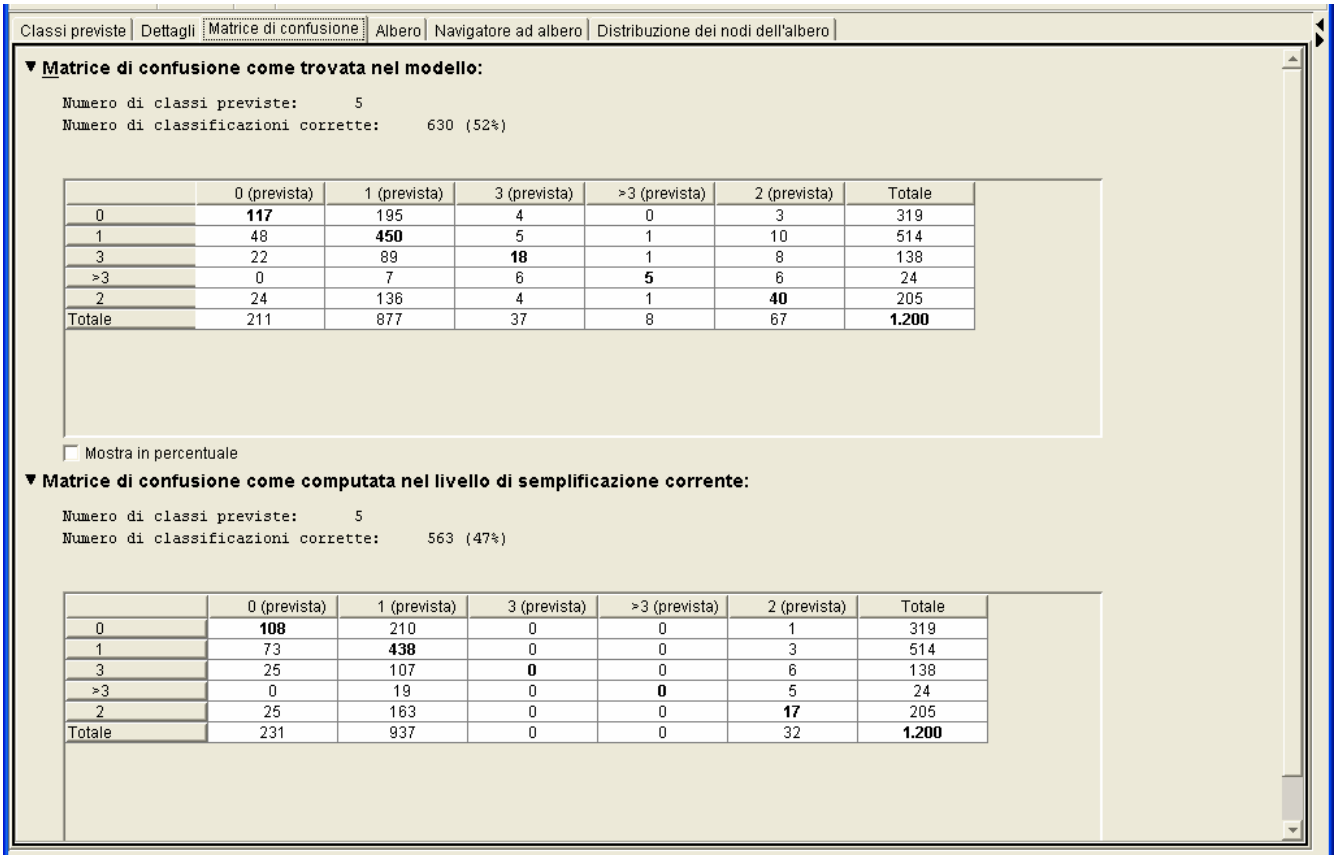
Purezza

La purezza viene elaborata dalla funzione di estrapolazione di Classificazione. Indica la percentuale dei record previsti correttamente nel nodo. Esempio: nel caso di claims > 3, l'unica regola evidenziata dal programma corrisponde alla sequenza:
se commute distance > 11,8 km e salary > 156.181 allora la classe (punteggio) è claims > 3 con una purezza del 62,5% (vedi Matrice di confusione). Cioè su 8 casi reali, 5 vengono correttamente assegnati secondo la regola definita dalla funzione di estrapolazione, mentre 3 record che soddisfano ugualmente la regola si distribuiscono in classi diverse.
Nel caso che regole diverse raggiungano lo stesso risultato (assegnino ai record la stessa classe con pattern diversi) con diversi gradi di purezza, la purezza espressa nella Matrice di confusione è data dalla media ponderata delle purezze. Esempio: se r1 ➔ 1 con purezza 50,4% su 587 record (49%) e r2➔1 con purezza 42,3% con 111 record (9%), il valore di purezza espresso dalla Matrice di confusione sarà =(50,4*587+42,3*111)/698



La Matrice di confusione viene calcolata dalla funzione di estrapolazione di Classificazione. Visualizza la distribuzione dei record in base alle classi attuali e alle loro classi previste. Fornisce indicazioni sulla qualità del modello corrente.
I modelli generati possono essere semplificati ad un certo livello di compressione (pruning).
In questo caso la funzione di estrapolazione calcola una diversa matrice di confusione, sulla base di regole semplificate (meno dettagliate, quindi meno precise) e la confronta con quella generata dal modello non semplificato.

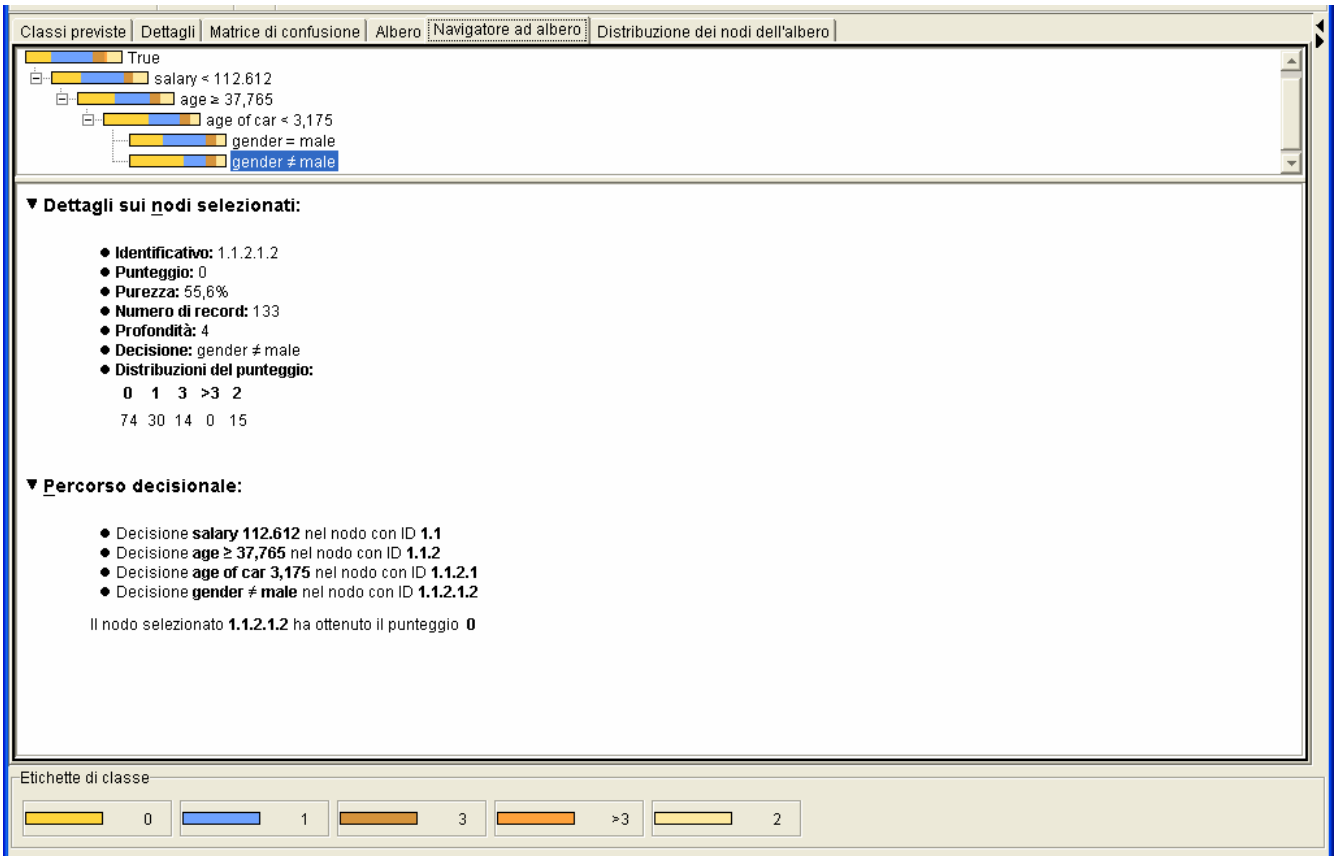


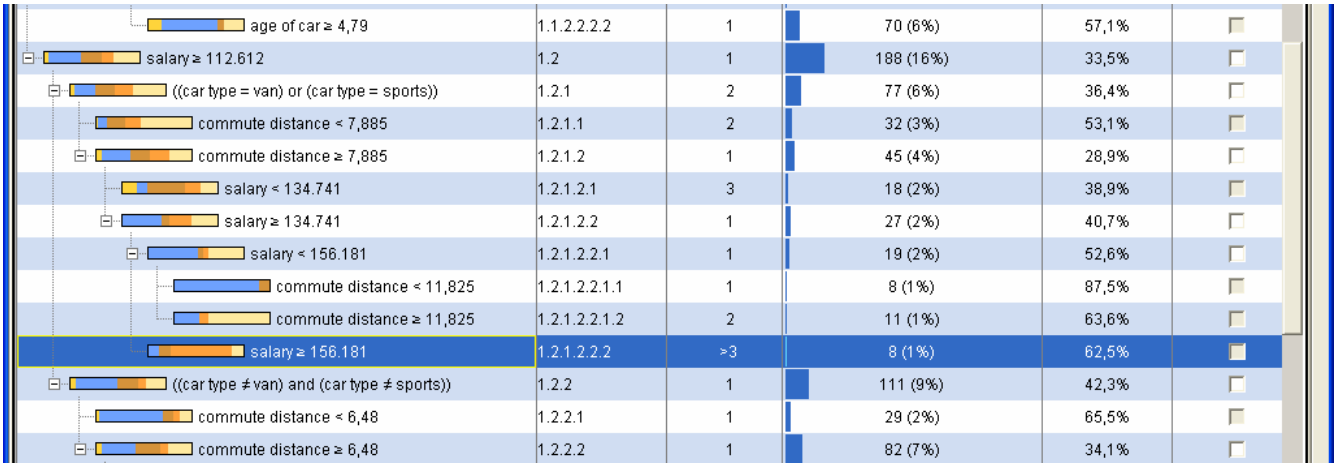


Nella Vista navigatore ad albero, sono presentate le informazioni di dettaglio sui nodi selezionati nella Vista albero o nella Vista distribuzione dei nodi dell'albero. Il percorso decisionale è visualizzato in formato testo nella sezione omonima. Viene anche visualizzato il punteggio del nodo

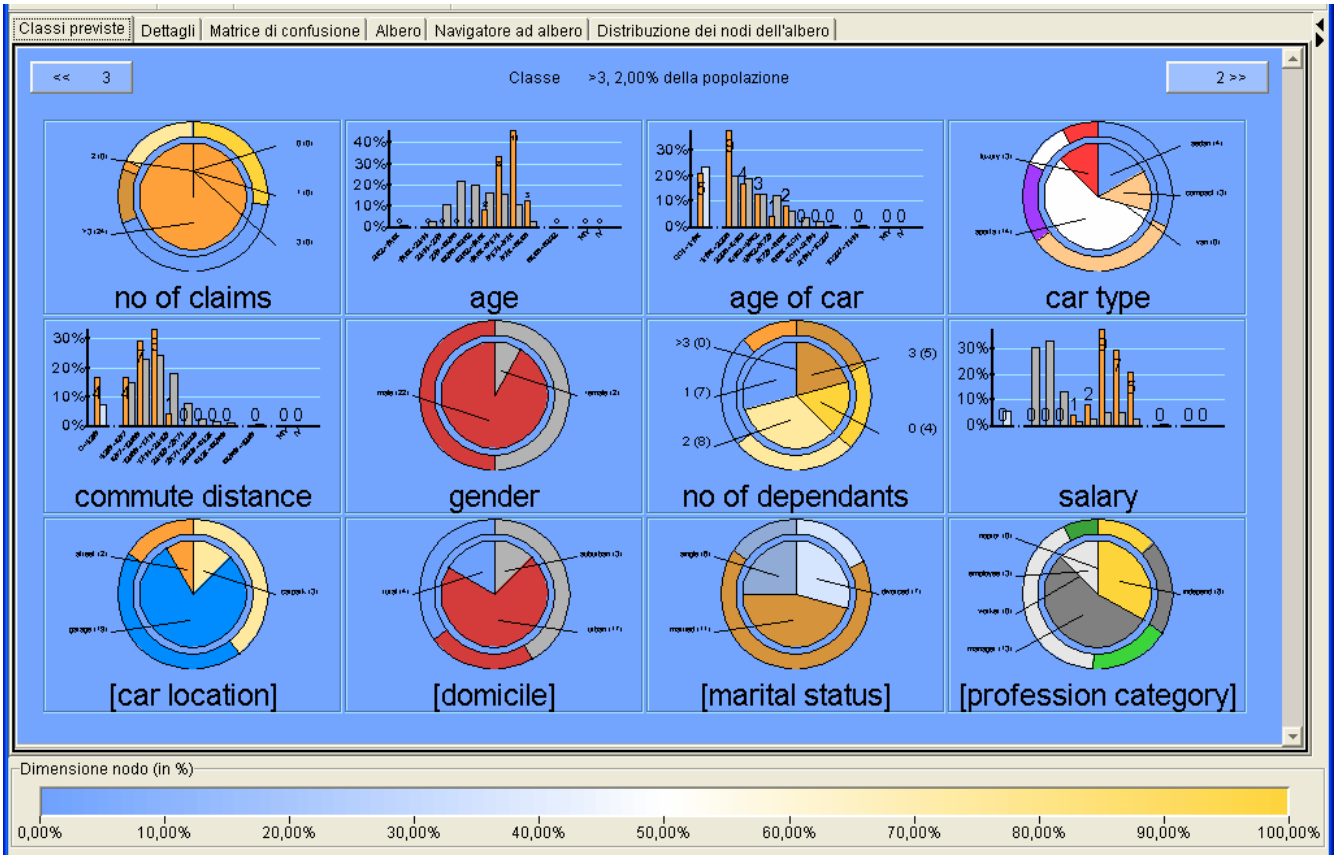
selezionato. Il percorso decisionale facilita la lettura della regola di classificazione calcolata dalla funzione di estrapolazione. Nell'esempio seguente si può osservare come il nodo selezionato esprima la regola:

se salary < 112.612 e age > 37, 7 e age of car < 3,1 e gender = female allora la classe è claims = 0 con purezza = 55,6%





L'elaborazione fornisce anche la rappresentazione grafica della distribuzione di ciascuna variabile all'interno delle classi di riparto del database, per una più efficace lettura della classificazioni effettuate dal procedimento di data mining.



Di fatto l'analisi dei cluster relativi alle possibili classi di riparto corrisponde ad un'analisi bivariata delle variabili usate per la classificazione, dove la classe di riparto funge da variabile dipendente.

In sintesi, la tecnica dell'albero decisionale consente di ricavare in modo induttivo le regole di classificazione (valutazione) dei dati (record) che nel caso dei Sistemi Esperti vengono acquisite in modo deduttivo intervistando gli esperti.
La classificazione induttiva di un albero decisionale è legata alla possibilità di poter recuperare informazioni (conoscenza) da un insieme di dati valido (completo e significativo), mentre quello di un sistema esperto si fonda sulla possibilità di recuperare la conoscenza da uno o più esperti.
Nonostante le differenze metodologiche le due tecniche hanno finalità analoghe (produzione in forma organizzata di un insieme di regole di classificazione) e possano essere utilizzate in modo alternativo o complementare a seconda del contesto e delle risorse attivabili dal decisore.

5. Link analysis

Association discovery

La link analysis cerca di stabilire relazioni tra i record o gruppi di record, anziché studiare il database come un unico soggetto (come ad es. la segmentazione o l'analisi predittiva). Le relazioni vengono spesso definite associazioni. Un'applicazione tipica della link analysis è l'Association discovery, che viene utilizzata per individuare combinazioni di articoli o servizi venduti all'interno delle stesse transazioni commerciali (Market Basket Analysis). Questa tecnica può essere usata efficacemente con obbiettivi di cross-selling, up-selling e targeting customers.
Durante il processo di data mining, la relazione viene evidenziata in forma di regola e misurata da due indicatori quantitativi :
confidenza: il numero di casi in cui si verifica la combinazione di vendita (rule head) rapportato al numero totale delle transazioni in cui si verifica la premessa (rule body). (es. la combinazione d'acquisto camicia-cravatta (rule head) si verifica nel 70% del totale delle transazioni in cui viene acquistata una camicia (rule body)
supporto: il numero di casi in cui la combinazione si verifica rispetto al numero totale delle transazioni (es. la R d'acquisto camicia➔cravatta si verifica il 13,5% del totale delle transazioni)

Sequential pattern discovery

La Sequential pattern discovery è una tecnica utilizzata per identificare associazioni di acquisto correlate nel tempo che rivelino la sequenza con la quale i clienti acquistano beni e servizi.
L'obbiettivo è la comprensione del comportamento di consumo di medio periodo per una corrispondente e tempestiva politica di promozioni (just in time).
Durante l'esecuzione della tecnica vengono individuate le sequenze di acquisto di ciascun cliente e calcolato il **supporto** della sequenza, come rapporto tra il numero di ricorrenze totali e il totale delle transazioni, reggruppate per cliente. E' una misura relativa che indica il numero dei clienti che supportano la sequenza rispetto al numero totale dei clienti.

Similar time sequence discovery

Nella ricerca di sequenze temporali di vendita la sequenza viene necessariamente riferita al tempo (e non soltanto alla loro sequenza ordinata, come nel caso precedente), in modo da stabilire una regola di comportamento riferite ad un'unità temporale. Serve ad ottimizzare gli acquisti e le scorte.

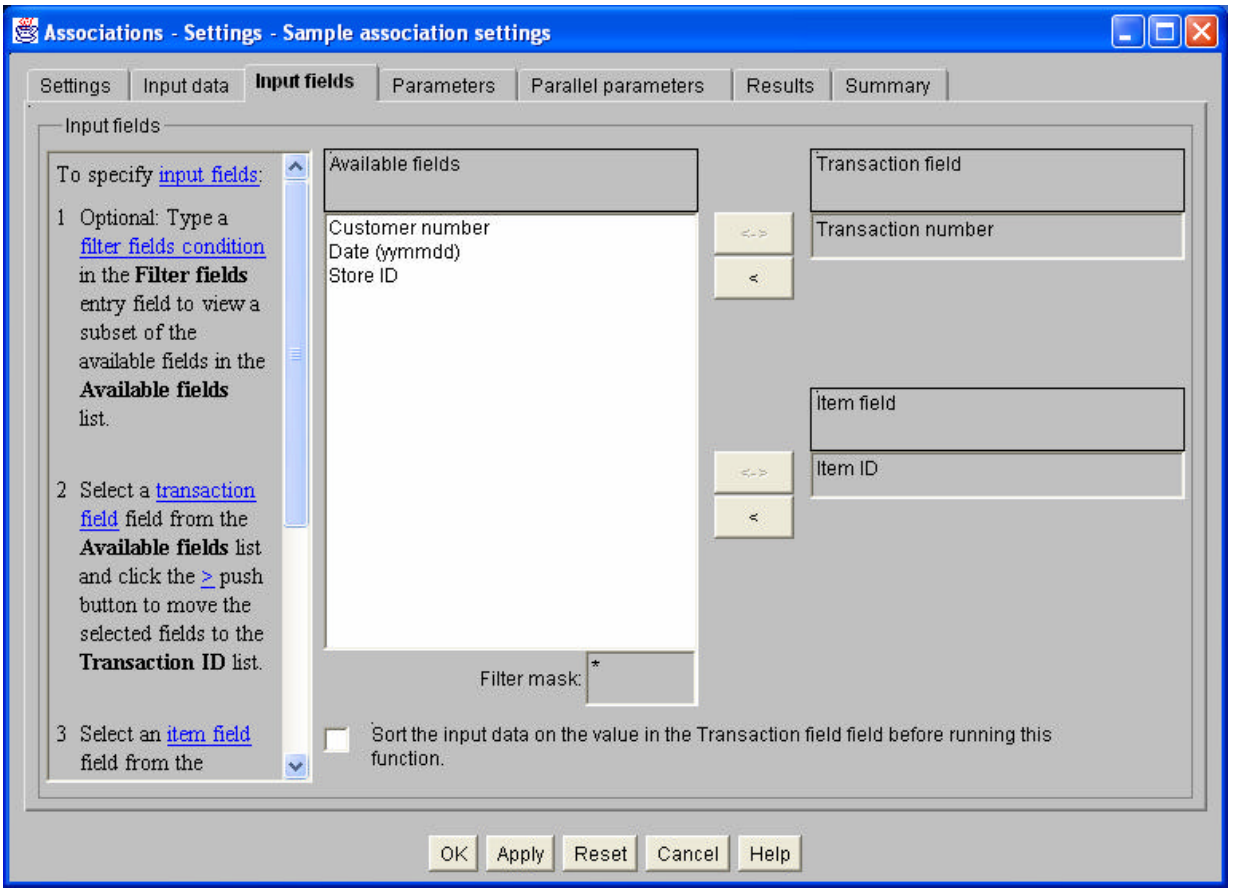
Caso: **MBA (Market Basket Analysis)**
Tecnica: Association discovery, Sequential pattern discovery, Similar time sequence discovery
Software: DB2 Intelligent Miner
Obiettivo: individuare combinazioni e sequenze nella vendita di articoli diversi

Database: flat, 1000 record
Campi: ID Customer, Date, Item ID, Store ID, Transaction ID
Dati di esempio:

Customer number	Date (yymmdd)	Item ID	Store ID	Transaction number
"0000001"	"950102"	"104"	"001"	"00001"
"0000001"	"950102"	"129"	"001"	"00001"
"0000001"	"950105"	"119"	"001"	"00077"
"0000001"	"950105"	"134"	"001"	"00077"
"0000001"	"950114"	"134"	"001"	"01077"
"0000002"	"950102"	"108"	"001"	"00002"
"0000002"	"950102"	"109"	"001"	"00002"

"0000002"	"950102"	"119"	"001"	"00002"
"0000003"	"950105"	"191"	"001"	"00079"
"0000003"	"950105"	"196"	"001"	"00079"
"0000003"	"950109"	"153"	"001"	"00983"
"0000003"	"950109"	"154"	"001"	"00983"

Configurazione dei parametri ed elaborazione dei dati:
Il DB viene necessariamente riordinato rispetto all'ID della transazione. La tecnica inizia contando le ricorrenze di vendita di un ID articolo. Viene così creato un vettore che indica le vendite del primo articolo rispetto a tutte le transazioni, raggruppate per ID transazione. Durante l'elaborazione vengono scartati i valori di vendita inferiori al supporto indicato.
Inserendo un secondo articolo, si crea una matrice a due colonne che conta le possibili ricorrenze di vendita dei due articoli raggruppate per ciascuna transazione. Il processo si ripete per gli n articoli considerati nell'elaborazione. Vengono così evidenziate regolarità nelle combinazioni di vendita (regole) per le quali vengono calcolate confidenza e supporto. Il supporto viene calcolato sia per il singolo articolo che per la relazione di vendita. Il rapporto tra la confidenza dell'associazione e il supporto dell'articolo premessa (la camicia) viene denominato lift. Se ad es. il supporto dell'articolo camicia (il numero di transazioni in cui una camicia viene acquistata) è del 20% sul totale delle transazioni e la confidenza camicia-cravatta è del 70%, il valore di lift della R è pari a 3,5. Cioè il valore atteso di vendita di una cravatta è di 3,5 volte se viene acquistata anche una camicia.

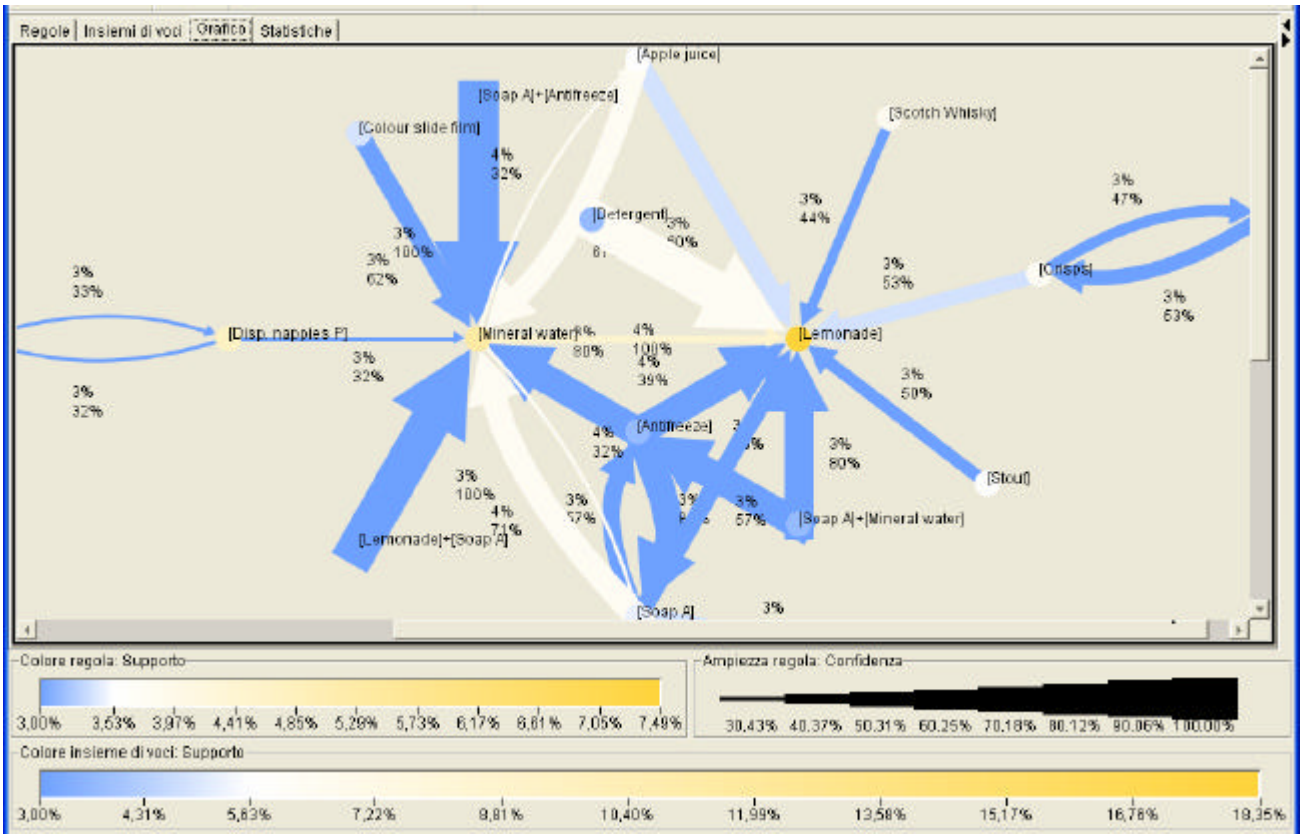


Esempio di Regole calcolate:

supporto	confidenza	corpo	testa
3,75%	100,00%	[Detergent]	[Lemonade]
3,00%	100,00%	[Soap A]+[Antifreeze]	[Mineral water]

3,00%	100,00%	[Antifreeze]+[Mineral water]	[Soap A]
3,00%	100,00%	[Lemonade]+[Soap A]	[Mineral water]
3,00%	80,00%	[Antifreeze]	[Soap A]
3,00%	80,00%	[Soap A]+[Mineral water]	[Antifreeze]
3,00%	80,00%	[Antifreeze]	[Lemonade]
3,00%	80,00%	[Antifreeze]	[Mineral water]
3,00%	80,00%	[Soap A]+[Mineral water]	[Lemonade]
3,00%	80,00%	[Gouda Cheese]	[Crackers]
3,75%	71,43%	[Soap A]	[Mineral water]
3,75%	66,67%	[Apple juice]	[Mineral water]
3,00%	66,67%	[Lemonade]+[Mineral water]	[Soap A]

Rappresentazione grafica:

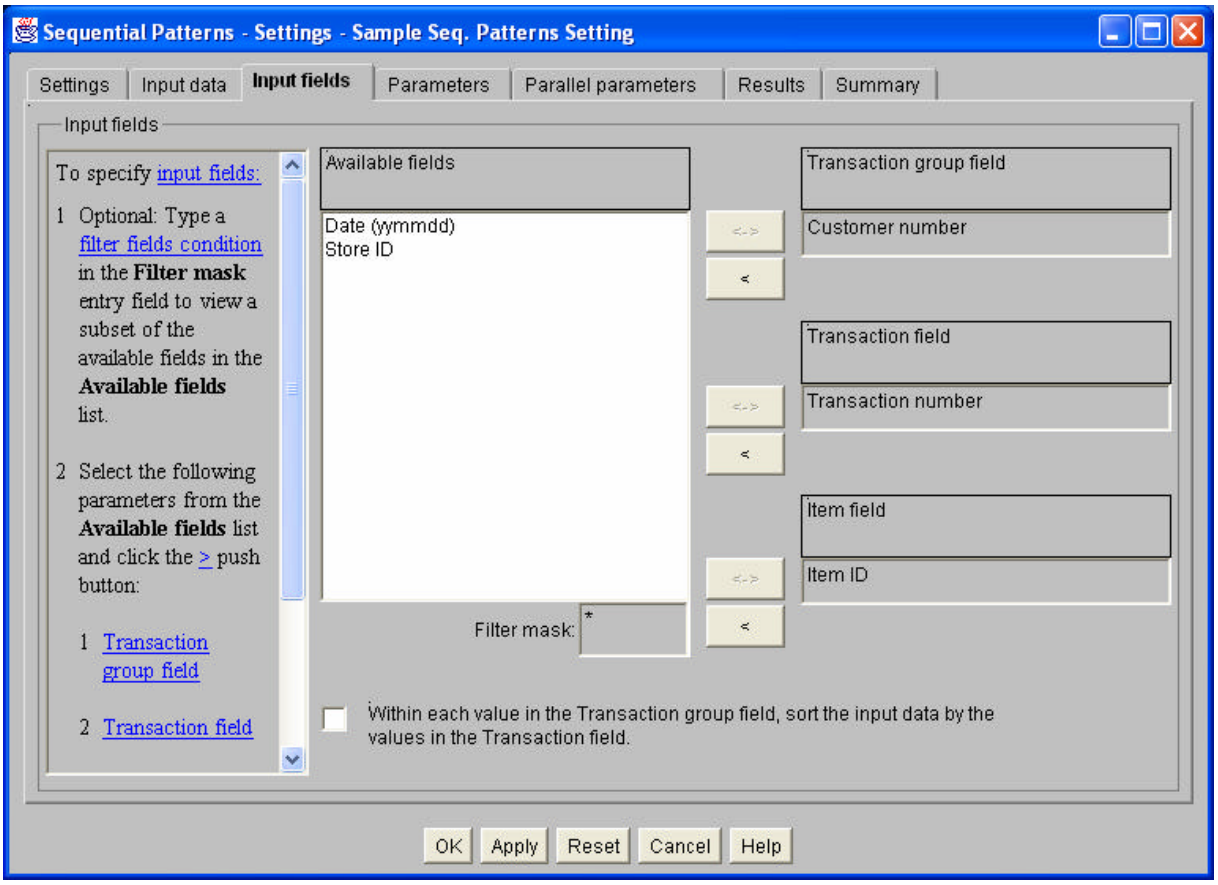


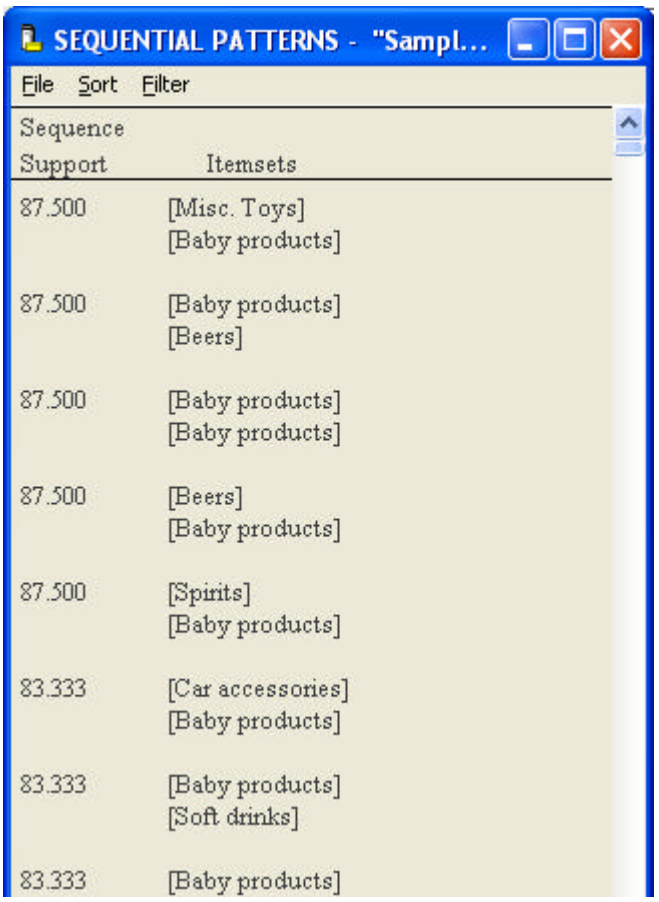
Il grafico evidenzia con delle frecce le relazioni (regole associative) nella direzione corpo testa, con colore corrispondente al valore di supporto ed ampiezza proporzionale alla confidenza della regola (vedi legenda). L'ampiezza e il colore dei nodi indica il supporto (frequenza di vendita sul totale transazioni) delle voci singole.
Nella tabella, così come nel grafico, si può ad esempio notare la relazione [Soap A]+[Antifreeze]➔ [Mineral water] con confidenza del 100% e supporto del 3%.

Statistiche
Il programma fornisce l'insieme delle informazioni di configurazione dell'elaborazione in una tabella di statistiche riassuntive:

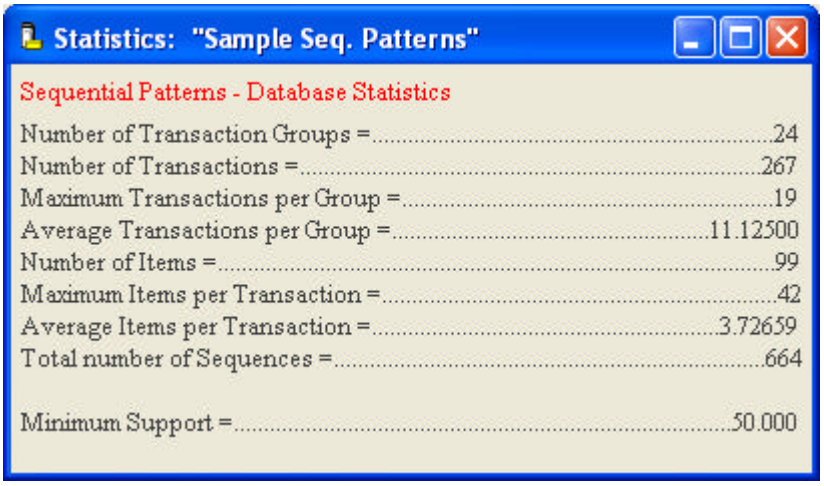
Confidenza minima della regola:	30,43%
Insiemi di voci visibili:	76
Lunghezza massima della regola :	--
Numero di insiemi di voci singole:	48
Numero di insiemi di voci usati nelle regole:	28
Numero di insiemi di voci:	76
Numero di transazioni:	267
Numero massimo di voci per transazione:	26
Numero medio di voci per transazione:	3,73
Regole visibili:	
Statistiche per oggetti visibili	39
Supporto minimo della regola:	3,00%

Sequential pattern discovery
Nella ricerca di sequenze di vendita il db viene riordinato per cliente e per ID transazione (in modo da rispettare la sequenza temporale), quindi raggruppato per cliente.



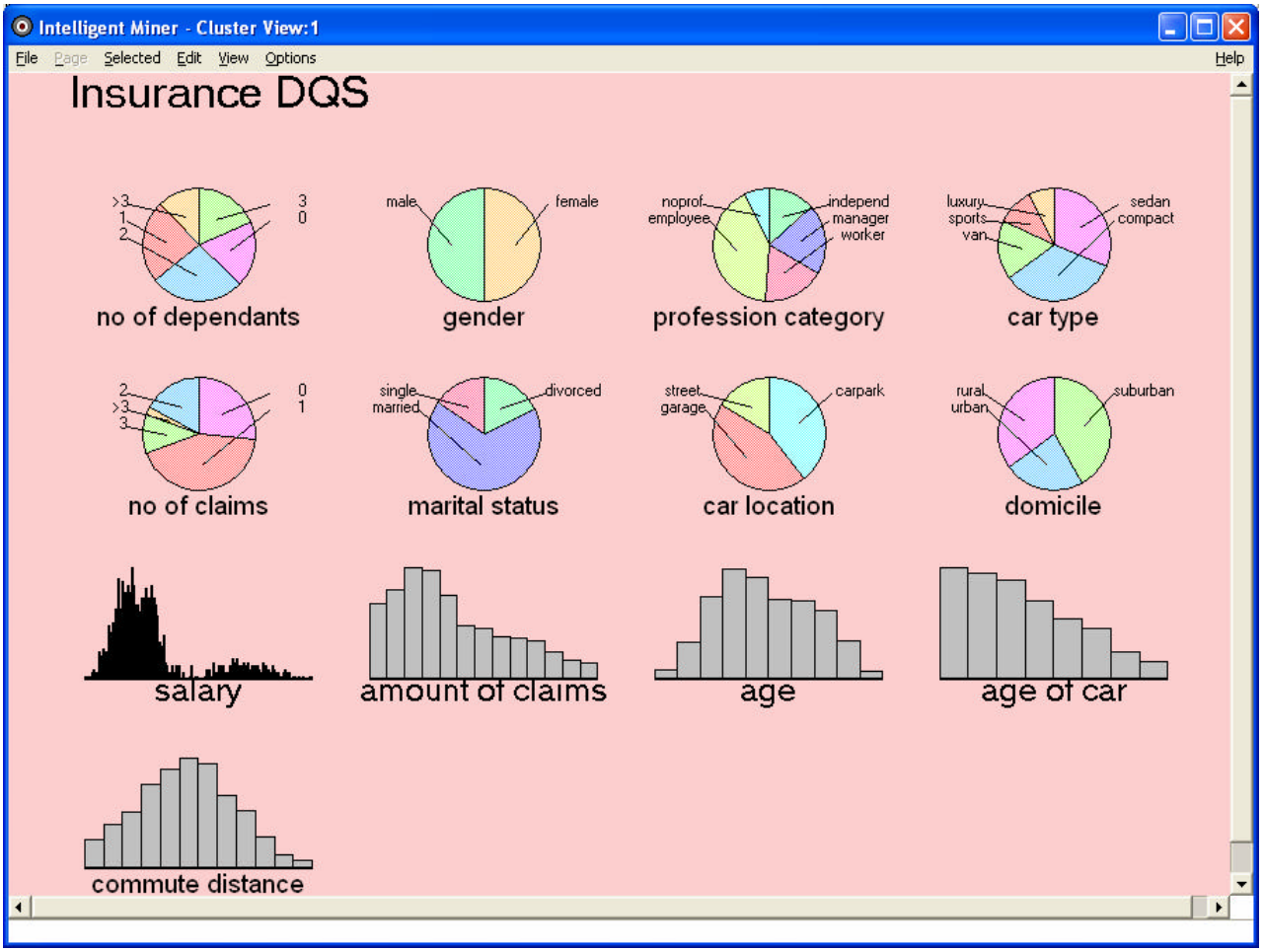


Nel caso in esame l'algoritmo recupera 664 sequenze di acquisto effettuate da 24 clienti. I risultati vengono riportati ordinati per supporto, cioè in relazione al numero o alla percentuale di clienti che rispettano la sequenza.



6. Analisi univariata

L'analisi univariata viene effettuata attraverso una serie di indici sintetici e con rappresentazioni grafiche univariate, in una fase di analisi preliminare o di pre processing dei dati, per una miglior comprensione del contenuto, della struttura e della qualità dei dati da analizzare e per individuare irregolarità e mancanze (invalid, noise, missing value, skew distribution).
Nel caso di variabili qualitative (categoriche), l'analisi della distribuzione di frequenza viene effettuata con grafici a torta, con suddivisioni proporzionali alle frequenze relative.
Nel caso di variabili quantitative è invece possibile calcolare, oltre al grafico, anche una serie di indici statistici (di posizione, di variabilità, di eterogeneità) che permettono una analisi più approfondita dei dati in esame.
Per i caratteri quantitativi continui, l'utilizzo di istogrammi per rappresentare la distribuzione di frequenza richiede la classificazione delle variabili in classi intervallari. Ciò comporta una perdita di informazioni, poiché si assume che le variabili si distribuiscano in modo uniforme all'interno delle



classi. In assenza di informazioni vincolanti, si assumono intervalli di ampiezza costante.
Nella tabella degli indici relativi alle variabili si possono osservare tra gli indici di posizione la media e la moda, cioè la classe o modalità a cui è associata la massima frequenza e i quantili (percentili), come generalizzazione della mediana, ovvero i valori che suddividono la distribuzione di frequenza in parti, con percentuali prefissate. Come misurazione della variabilità vengono evidenziati i valori estremi di della variabile (massimo e minimo), la deviazione standard (solo per variabili quantitative) e l'indice di entropia[6] (per variabili sia qualitative che quantitative), nella forma: $E = -\sum_{i=1}^{K} pi \log pi$ con p_i frequenze relative delle varie modalità o classi. L'indice

[6] P. Giudici "Data Mining", testo citato pp. 43-44

entropico risulta uguale a 0 in caso di perfetta omogeneità, mentre è uguale a log k in caso di massima eterogeneità.

Insurance DQS

Result created: 02/19/98 19:17:16

Result File : C:\DOCUME~1\STEFAN~1\IMPOST~1\Temp\6SCZDJNG.C2T

Cluster Characteristics :

Cluster Size		Id
Absolute	Relative(%)	
1200	100,00	0

Reference Field Characteristics(For All Field Types):

(Field Types : [ ]=Supplementary. CA=Categorical, CO=Continuous Numeric, DN=Discrete Numeric)

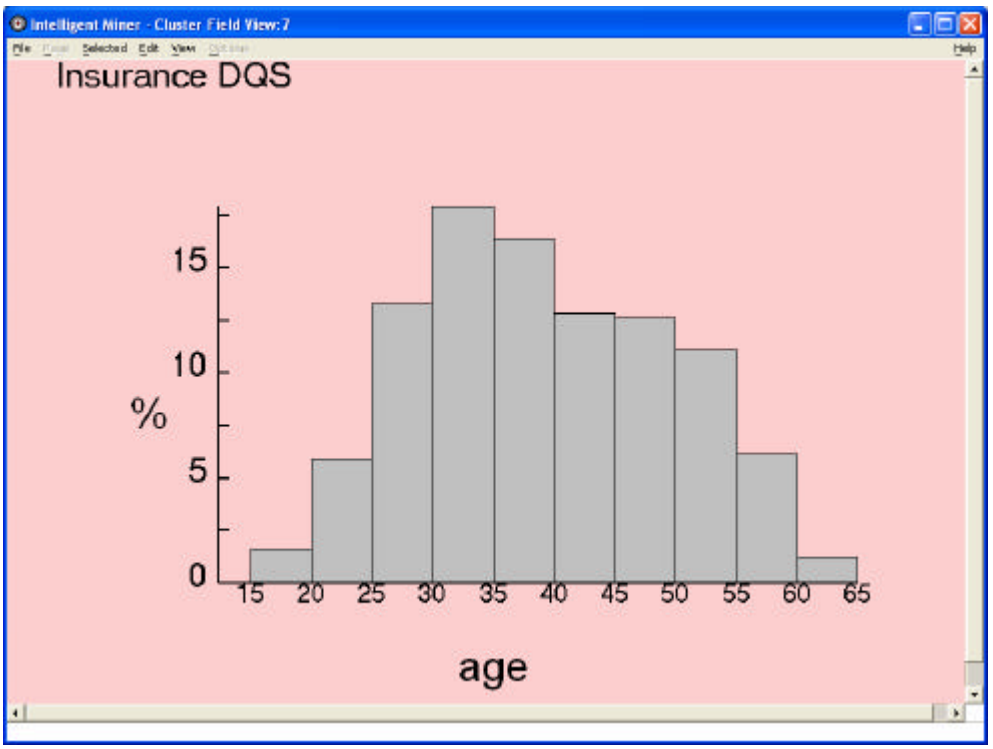
Id	Name	Type	Value Modal	Frequency Modal(%)	No. of Possible Values/Buckets
1	no of dependants	CA	2	26,75	5
2	gender	CA	female	50,08	2
3	profession category	CA	employee	41,33	5
4	car type	CA	compact	33,50	5
5	no of claims	CA	1	42,83	5
6	marital status	CA	married	67,08	3
7	car location	CA	garage	44,42	3
8	domicile	CA	suburban	41,75	3
9	salary	CO	52500	2,75	179
10	amount of claims	CO	625	14,00	13
11	age	CO	32,5	17,92	10
12	age of car	CO	0,5	19,50	8
13	commute distance	CO	13,75	15,17	12

Reference Field Characteristics(For Numeric Fields Only):

Id	Name	Value Minimum	Value Maximum	Mean	Deviation Standard
9	salary	15861,9	193000	68966,8	35729,5
10	amount of claims	1,89	5327,26	1291,68	954,436
11	age	9,32	69,92	39,0343	10,5327
12	age of car	0,01	11,44	3,07897	2,22862
13	commute distance	0	42,85	14,0441	7,06448

Quantile Information(For Numeric Fields Only):

Id	Name	Quantile	Threshold
9	salary	27252,28	2
		37683,96	10
		46010,07	25
		58670,07	50
		72251,15	75
		135023,92	90
		164392,05	98
11	age	19,55	2
		25,93	10
		31,05	25
		38	50
		47,65	75
		53,83	90
		58,78	98



Utilizzando il software di Data Mining, è possibile approfondire l'analisi su una singola partizione (variabile), ottenendo le dimensioni delle singole classi o modalità.

Insurance DQS Cluster 0 100,00% of population

Result created: 02/19/98 19:17:16

Result File : C:\DOCUME~1\STEFAN~1\IMPOST~1\Temp\6SCZDJNG.C2T
Number of Clusters : 1
Size of Cluster - 0 : Absolute: 1200, Relative: 100,00%

Cluster Field Characteristics(For All Field Types):

(Field Types : [ ]=Supplementary. CA=Categorical, CO=Continuous Numeric, DN=Discrete Numeric)

Id	Name	Type	Value Modal	Frequency Modal(%)	Chi-Squared	Entropy
11	age	CO	32,5	17,92	0,000	4,571

Cluster Field Characteristics(For Numeric Fields Only):

Id	Name	Value Minimum	Value Maximum	Mean	Deviation Standard
11	age	9,32	69,92	39,0343	10,5327

Field Details:

Field Name: age

Bound Lower	Bound Upper	Size %	Bound Lower	Bound Upper	Size %
9,32	15	00,5833	45	50	12,6667
15	20	01,6667	50	55	11,0833
20	25	05,9167	55	60	06,1667
25	30	13,3333	60	65	01,2500
30	35	17,9167	65	69,92	00,1667
35	40	16,4167	Missing Value		00,0000
40	45	12,8333			

7. Analisi bivariata

Nel caso bivariato e multivariato, gli indici permettono di calcolare l'esistenza di relazioni tra le variabili considerate, oltre a descrivere la distribuzione delle singole variabili.
L'esame della concordanza tra due variabili indica la tendenza ad associare valori elevati (poco elevati) della prima con valori elevati (poco elevati) della seconda. La discordanza è, viceversa, la tendenza ad associare modalità poco elevate di una variabile con modalità elevate dell'altra.
Il grado di concordanza o discordanza è misurato dalla Covarianza:

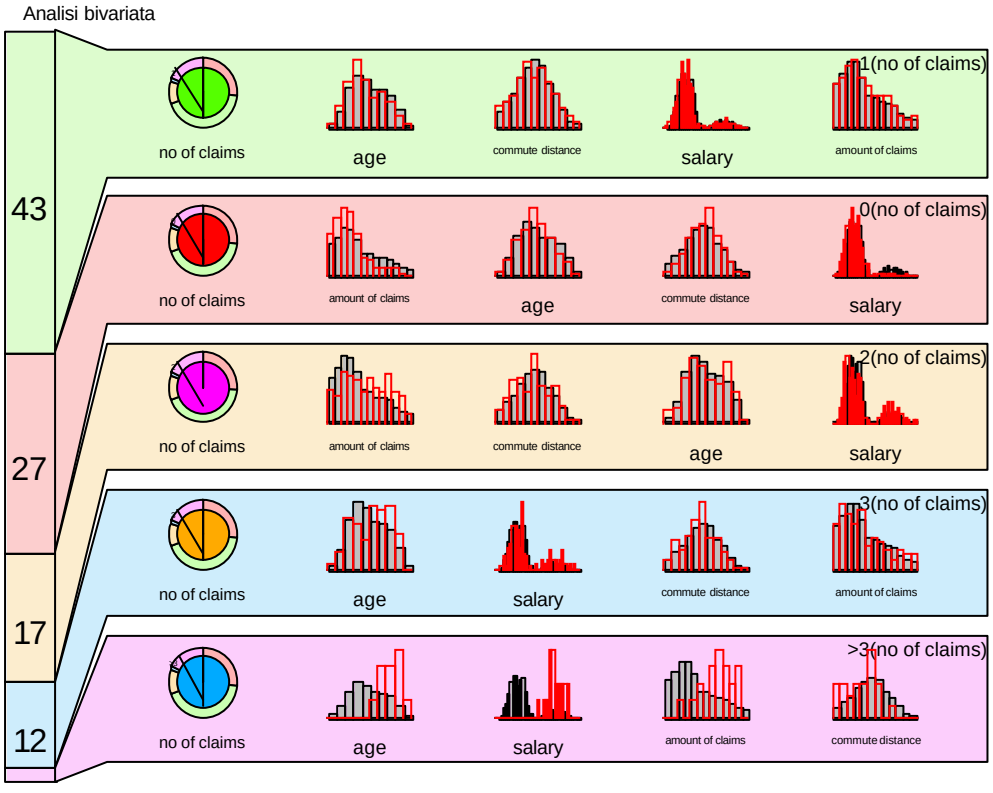
$$COV(X,Y) = \frac{1}{N}\sum_{i=1}^{N}[[x_i - m_X][y_i - m_Y]$$

La Covarianza è un indicatore assoluto e pertanto di difficile interpretazione. A questa viene solitamente preferito un indice relativo, denominato di correlazione lineare, nella forma:

$$r_{XY} = r(X,Y) = \frac{Cov(X,Y)}{s(X)s(Y)}$$

r(X,Y) assume valore 1 quando le due variabili sono totalmente e positivamente correlate, -1 quando le due variabili sono totalmente e negativamente correlate, 0 quando le due variabili non sono legate da alcun tipo di relazione lineare[7].
Utilizzando il software di Data Mining è possibile calcolare la matrice di correlazione tra le variabili, oltre al test chi-quadrato e F-statistico:

	amount of claims	commute distance	age of car	salary
amount of claims	1.0	-0.308140049489701	-0.0278550272855965	0.4750689008158
commute distance	-0.308140049489701	1.0	0.065974363363372	-0.1733463074645
age of car	-0.0278550272855965	0.065974363363372	1.0	-0.1854381594892
salary	0.475068900815828	-0.173346307464558	-0.185438159489292	1.0



[7] "Data Mining", P. Giudici, testo citato pp.50-54

Analisi bivariata

Result created: 06/16/04 10:18:11

Result File : C:\DOCUME~1\STEFAN~1\IMPOST~1\Temp\OHQHLRGC.C2T
Bivariate Statistics Field : no of claims

Split Values:

0(no of claims)
1(no of claims)
3(no of claims)
>3(no of claims)
2(no of claims)

Cluster Characteristics :

Cluster Size		Id
Absolute	Relative(%)	
319	26,58	0(no of claims)
514	42,83	1(no of claims)
138	11,50	3(no of claims)
24	2,00	>3(no of claims)
205	17,08	2(no of claims)

Reference Field Characteristics(For All Field Types):

(Field Types : []=Supplementary. CA=Categorical, CO=Continuous Numeric, DN=Discrete Numeric)

Id	Name	Type	Value Modal	Frequency Modal(%)	No. of Possible Values/Buckets
1	no of claims	CA	1	42,83	5
2	age	CO	32,5	17,92	10
3	amount of claims	CO	625	14,00	13
4	commute distance	CO	13,75	15,17	12
5	salary	CO	47500	10,58	37

Reference Field Characteristics(For Numeric Fields Only):

Id	Name	Value Minimum	Value Maximum	Mean	Deviation Standard
2	age	9,32	69,92	39,0343	10,5327
3	amount of claims	1,89	5327,26	1291,68	954,436
4	commute distance	0	42,85	14,0441	7,06448
5	salary	15861,9	193000	68966,8	35729,5

F-Test Results:

------- Numeric Field -------		Degree Of Freedom		F-Statistics	Probability
Numerator	Denonimator	Numerator	Denonimator		>F
amount of claims	age	1199	1199	8211,27	0
age	commute distance	1199	1199	2,22292	0
salary	age	1199	1199	1,15072E7	0
amount of claims	commute distance	1199	1199	18253	0
salary	amount of claims	1199	1199	1401,39	0
salary	commute distance	1199	1199	2,55795E7	0

Chi-Squared Test Results:

Input Field	Is Test Reliable?	Degree Of Freedom	Number Of Expected Values <5	Chi-Squared Statistics (C)	Probability >C
age	No	44	25	133,335	0
amount of claims	No	52	18	180,991	0
commute distance	No	48	19	77,1372	0,00482069
salary	No	140	126	368,583	0
no of claims	No	16	5	4800	0

8. Database relazionali e flat file (log file)

Un Database e' una collezione organizzata di informazioni. Qualsiasi sistema di gestione dei DB (DBMS, Database management system) permette di inserire, modificare, visualizzare e stampare le informazioni che vengono normalmente strutturate in una o più tabelle, divise in righe e colonne. Tutti i DBMS rispondono a esigenze comuni in termini di gestione dei dati, tra cui:

- La ridondanza dei dati
- L' uniformità dei dati
- L'indipendenza dalla piattaforma
- La sicurezza delle transazioni
- La possibilità di gestire correttamente un ambiente multiutente

I sistemi di gestione degli archivi vengono distinti per la loro capacità di collegare più tabelle di dati. Un Flat file è un archivio che non ha e non può avere collegamenti con altri archivi e coincide con un unica tabella in cui vengono registrate tutte le informazioni. Viceversa un RDBMS (Relational Database management system) permette di dividere il database in tante tabelle che contengono dati logicamente correlati attraverso un sistema di relazione tra campi chiave.
Le Relazioni tra le tabelle (ovvero tra campi delle tabelle) possono essere di diversi tipi:

- uno a uno: si ha quando un campo (es. cliente nella tabella Clienti) compare nella tabella primaria e in quella collegata una volta sola. In un DB che contiene un archivio Clienti e uno Fatture, per avere una relazione uno a uno nel campo cliente tra le tabelle clienti e fatture, dovremmo avere una sola fattura per ciascun cliente.
- uno a molti: avendo sicuramente più di una fattura per ogni cliente, lo stesso cliente può comparire nella tabella collegata n volte. La relazione è quindi tra un campo (cliente) che compare una volta nella prima tabella e molte volte nella seconda.
- molti a molti: si ha tra campi che contengono duplicati sia nella tabella primaria che in quella collegata.

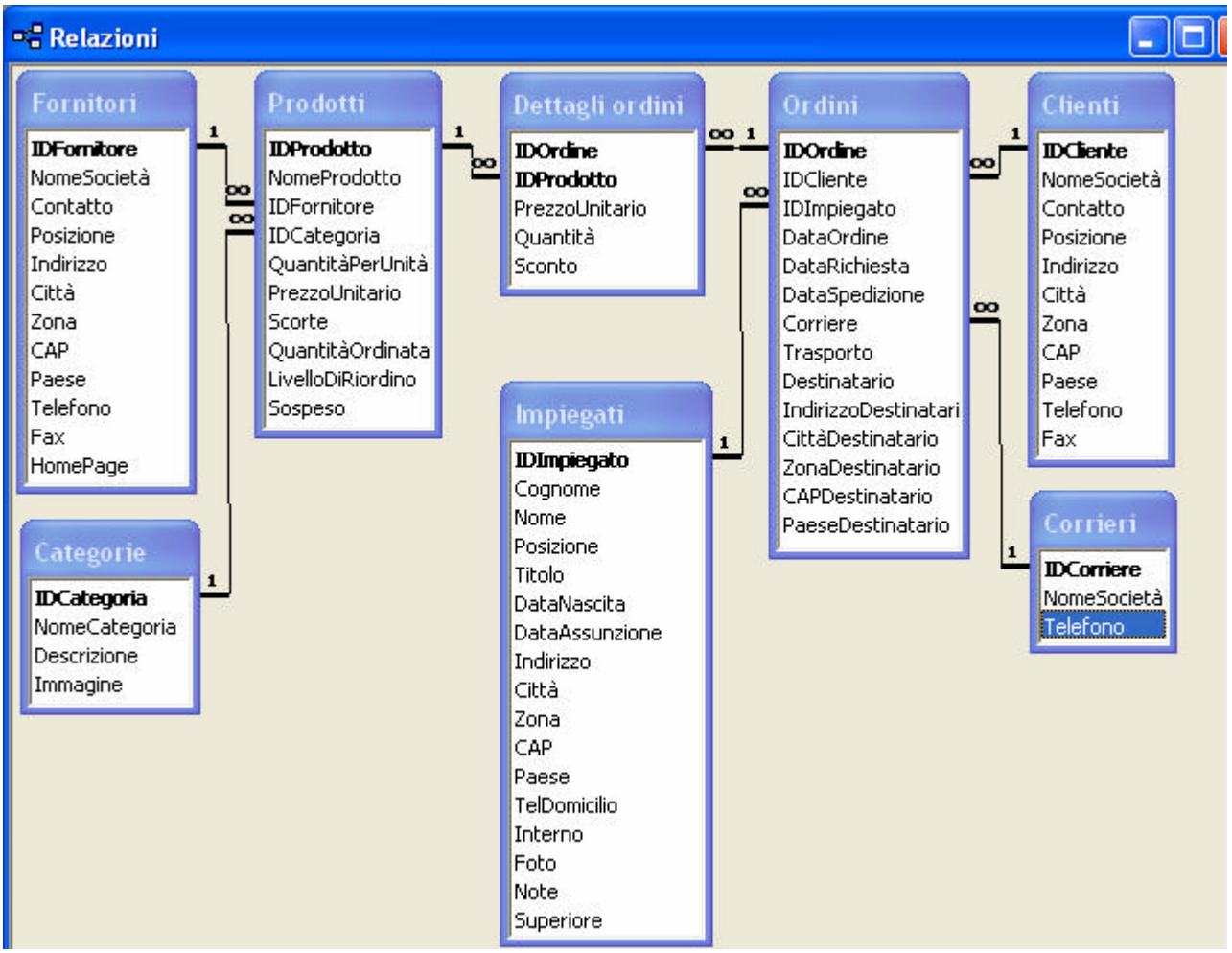
Una *Chiave Primaria* (o *primary key)* e' un campo (colonna di una tabella, es. cliente) o un gruppo di campi (colonne) che identificano in maniera univoca (singola) ogni record (riga) rispetto agli altri. E' cioè l'informazione (o l'insieme di informazioni) che permette di distinguere ogni singolo record da tutti gli altri. Viene usato come campo di collegamento nel lato 1 della relazione tra tabelle (tabella primaria).
Una *Chiave Esterna* (o *foreign key*) e' invece un campo presente in una tabella nella quale si registrano dati che coincidono con la chiave primaria di un' altra tabella (es. cliente nella tabella Fatture). Viene usato come campo di collegamento nel lato molti della relazione tra tabelle (supposto che la relazione si $1 \rightarrow \infty$).
La corrispondenza tra campi chiave viene denominata *Integrità Referenziale* e stabilisce le regole di alimentazione degli archivi collegati, nel senso che una tabella di lato molti non può registrare un valore che non sia stato precedentemente inserito nel campo chiave di lato 1. Questo sistema di controllo impedisce la formazione di record "orfani", cioè di record appartenenti a soggetti non registrati nell'archivio di quegli stessi soggetti.
La "*Violazione dell' integrità referenziale*" rende il database *inconsistente*, creando errori gravi, come ad es. l'emissione di una fattura ad un cliente non presente nell'archivio clienti dell'azienda.

Le relazioni possono essere definite in modi diversi, in relazione alle caratteristiche degli archivi:

equi join: visualizzano tutti i record di una tabella che hanno record corrispondenti in un'altra tabella. In altre parole una equi join tra clienti e fatture collega (e recupera) tutti i record dei clienti

che hanno ricevuto fattura, ma non quelli registrati nell'archivio clienti che non sono "movimentati" nell'archivio fatture.
join esterne: se l'interrogazione vuole verificare tutti gli clienti in archivio e le corrispondenti fatture, indipendentemente dall'esistenza di fatture, si ricorre a una join esterna.
self join: sono quelle in cui si mette in relazione la tabella con sé stessa, al fine di visualizzare eventuali anomalie dell'archivio
theta join: mettono in relazione le tabelle utilizzando operatori di confronto diversi da =

Si osservi ad es. il seguente sistema di relazioni, relativo al database di gestione Northwind.mdb (Prodotti, Fornitori, Clienti, Ordini, Fatture) fornito come esempio di Microsoft Access:

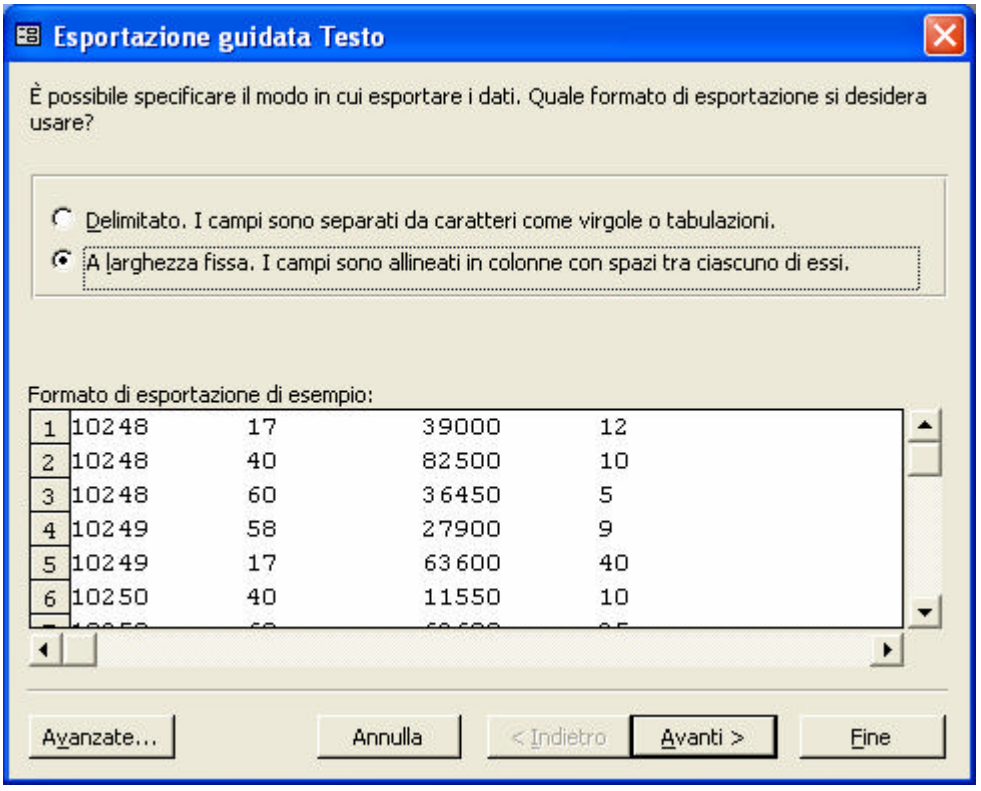


Il DB Northwind può essere utilizzato per costruire archivi su cui elaborare analisi di Business Intelligence. Anziché interrogare direttamente il sistema relazionale di archivi attraverso gli strumenti di analisi, è possibile ricavare semplici (flat) file di testo, denominati log file, sul tema che si desidera analizzare.
Un log file è solitamente un archivio che tiene traccia in modo automatico delle attività svolte dagli utenti (ad es. nei collegamenti ad un sito Web).
Riferito al normale e legittimo uso di un DB, può essere inteso come archivio storico di dati recuperati da una interrogazione (query) e contrassegnati in ordine temporale (timestamp) per analizzare parte dell'attività dell'utente.
Ad esempio i dati relativi agli ordini possono essere estratti dal DB per finalità di analisi, come ad esempio studiare eventuali regolarità nelle associazioni o nelle sequenze di acquisto. Ecco come si presenta l'archivio in formato tabellare:

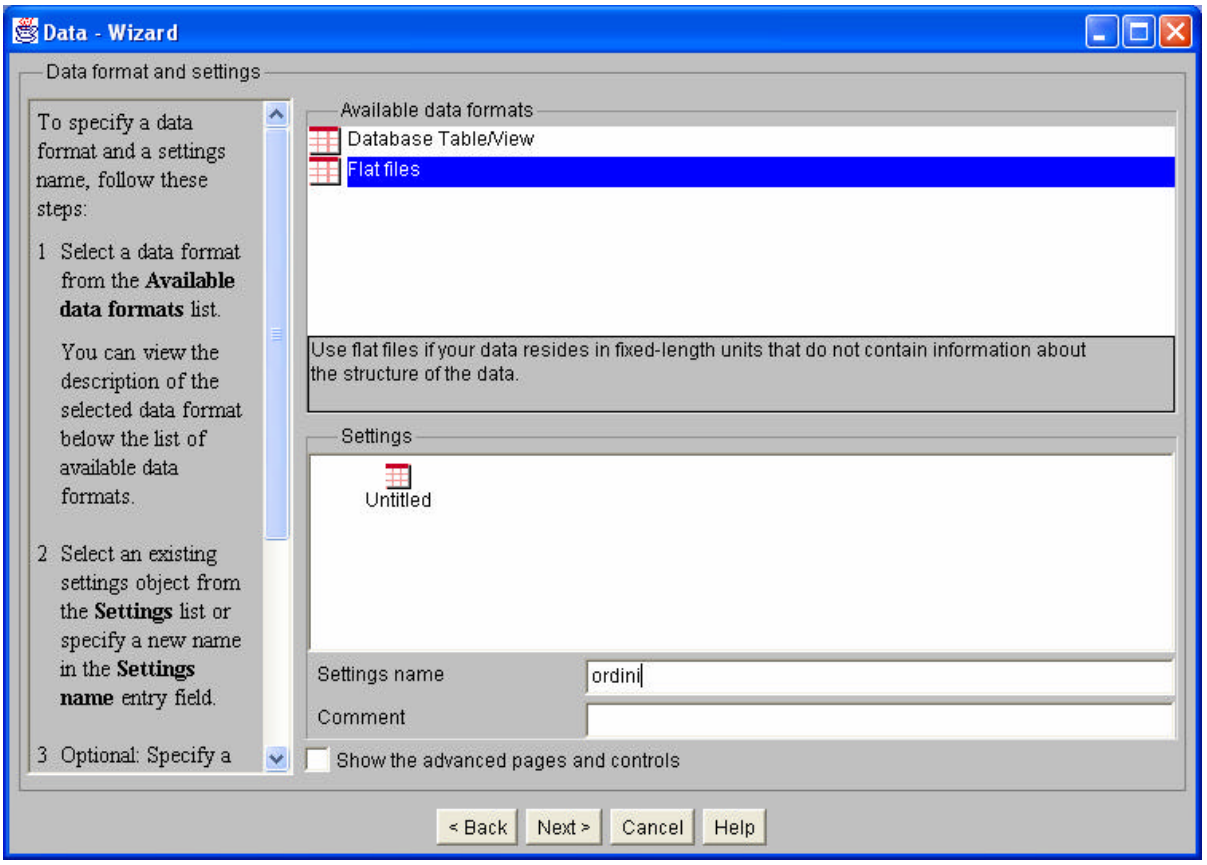
Dettagli ordini : Tabella

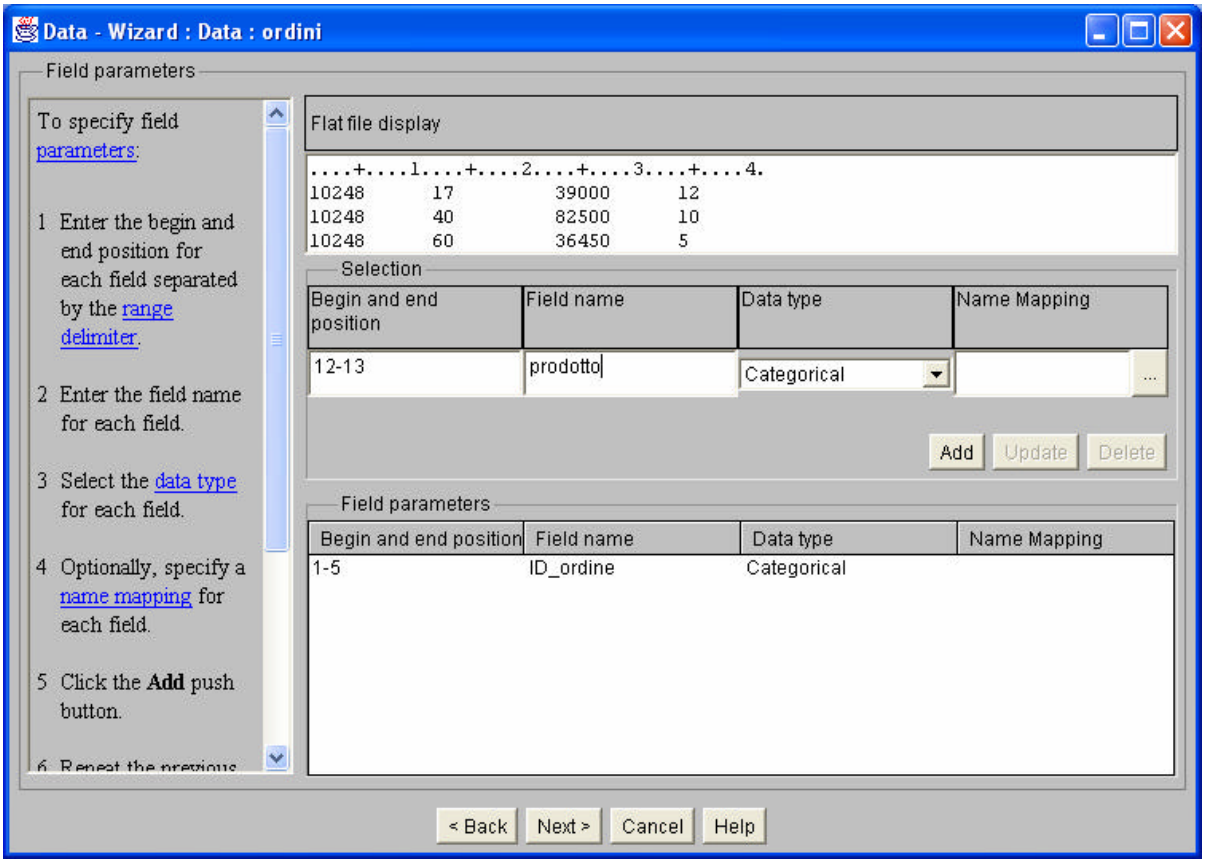
ID ordine	Prodotto	Prezzo unitario	Quantità	Sconto
10248	Queso Cabrales	L. 39.000	12	0%
10248	Singaporean Hokkien Fried Mee	L. 82.500	10	0%
10248	Mozzarella di Giovanni	L. 36.450	5	0%
10249	Tofu	L. 27.900	9	0%
10249	Manjimup Dried Apples	L. 63.600	40	0%
10250	Jack's New England Clam Chowder	L. 11.550	10	0%
10250	Manjimup Dried Apples	L. 63.600	35	15%
10250	Louisiana Fiery Hot Pepper Sauce	L. 25.200	15	15%
10251	Gustaf's Knäckebröd	L. 25.200	6	5%
10251	Ravioli Angelo	L. 23.400	15	5%
10251	Louisiana Fiery Hot Pepper Sauce	L. 25.200	20	0%
10252	Sir Rodney's Marmalade	L. 97.200	40	5%
10252	Geitost	L. 3.000	25	5%
10252	Camembert Pierrot	L. 40.800	40	0%
10253	Gorgonzola Telino	L. 15.000	20	0%
10253	Chartreuse verte	L. 21.600	42	0%
10253	Maxilaku	L. 24.000	40	0%
10254	Guaraná Fantástica	L. 5.400	15	15%
10254	Pâté chinois	L. 28.800	21	15%
10254	Longlife Tofu	L. 12.000	21	0%
10255	Chang	L. 22.800	20	0%

Il RDBMS consente quasi sempre l'esportazione dei dati in formato testo. Nel caso del programma MsAccess è sufficiente definire i parametri di esportazione salvando la tabella nel formato desiderato.



Una volta ottenuto il file ordini.txt, il programma di Data mining si occupa di ricostruire il tracciato record in modo manuale.

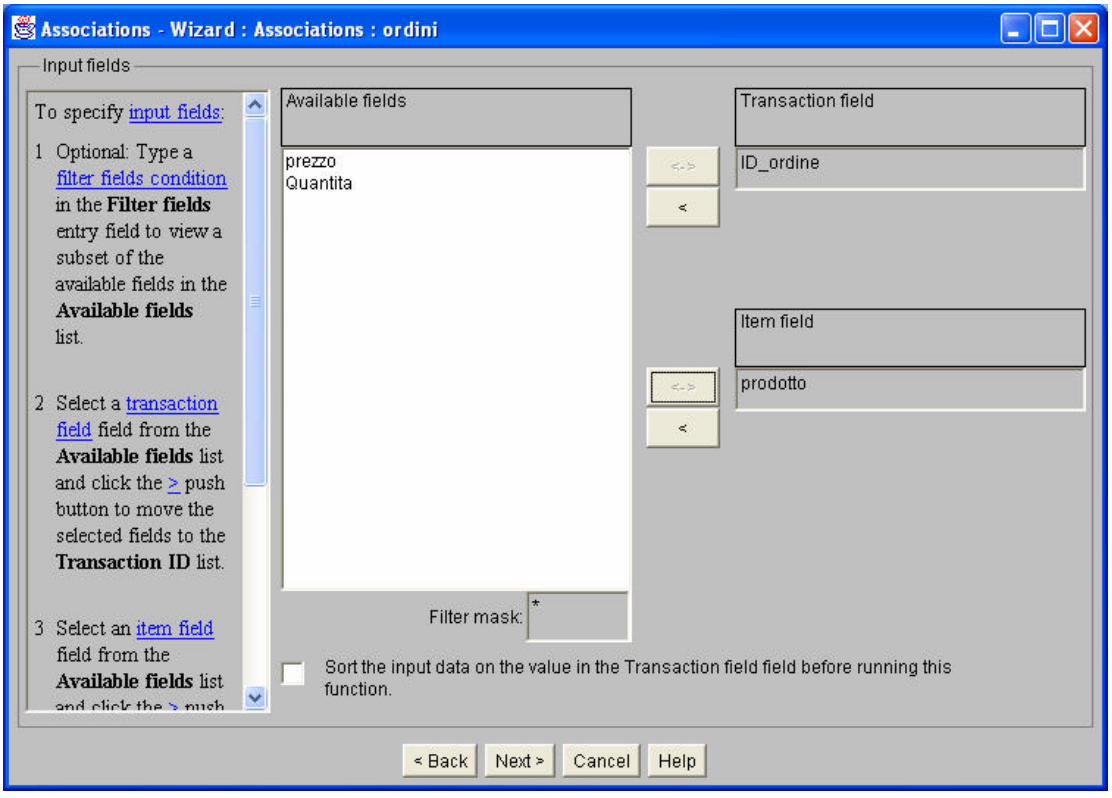


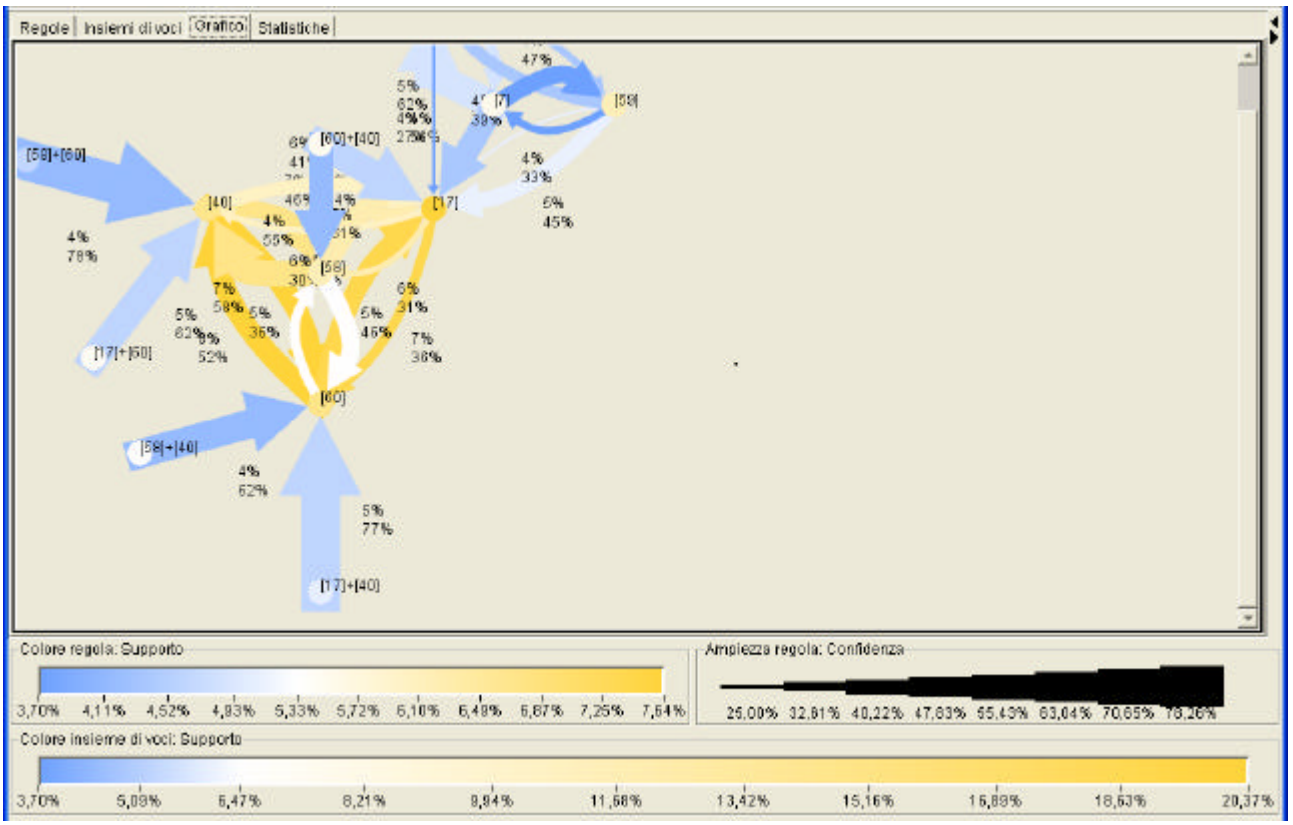


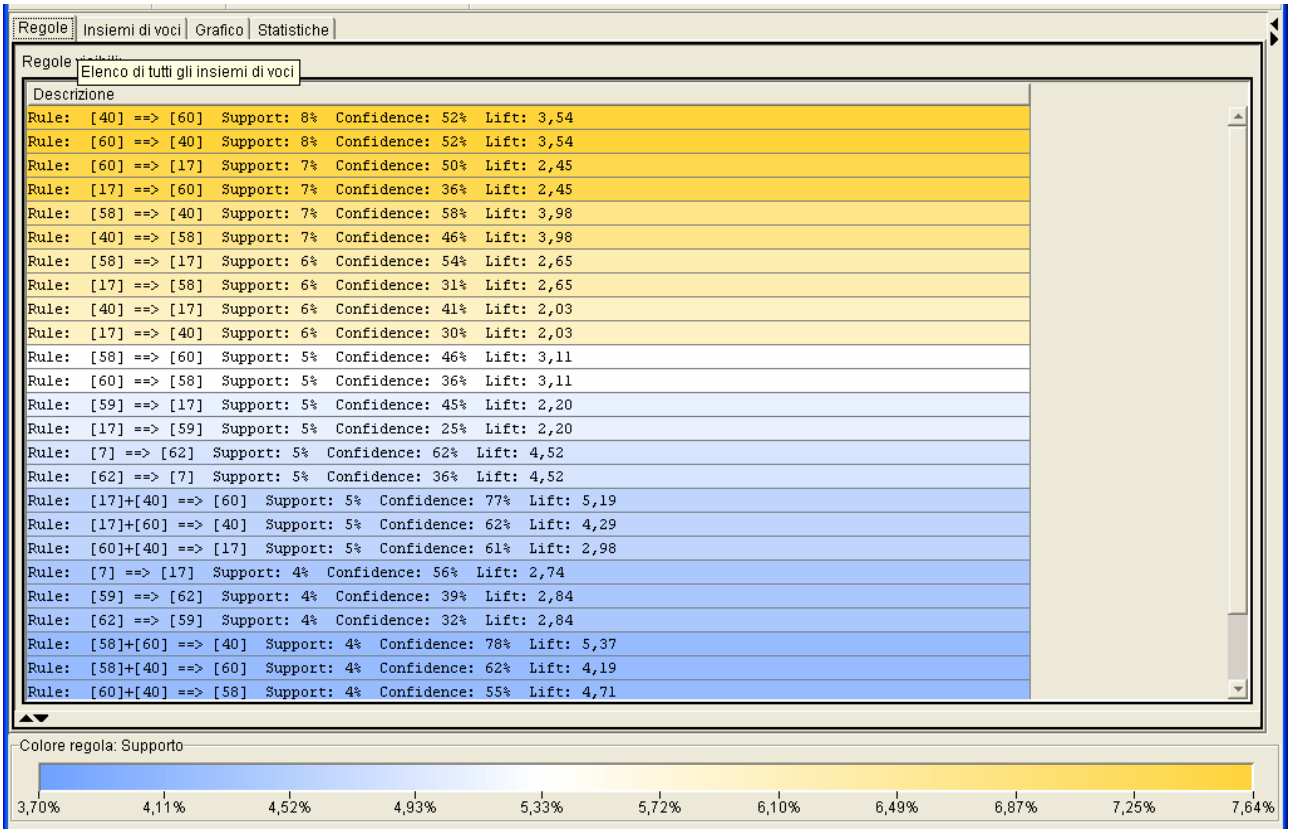
L’archivio importato nel client di Data Mining può essere aperto e verificato utilizzando il browser:

ID_ordine	Quantita	prezzo	prodotto
"10248"	12	39000	"17"
"10248"	10	82500	"40"
"10248"	5	36450	"60"
"10249"	40	63600	"17"
"10249"	9	27900	"58"
"10250"	10	11550	"40"
"10250"	15	25200	"58"
"10250"	35	63600	"60"
"10251"	6	25200	"17"
"10251"	15	23400	"40"
"10251"	20	25200	"60"

Una volta costruita la base di dati vengono applicate le normali procedure per la creazione di un’applicazione di mining. In questo caso vengono cercate eventuali combinazioni nelle associazioni di acquisto con la tecnica della Link analysis. La visualizzazione dei risultati consente di esaminare in modo grafico e tabellare le caratteristiche delle associazioni calcolate.







Utilizzando la stessa procedura, è possibile estrarre dal database relazionale Northwind altri archivi flat per applicare tecniche differenti coerenti ai diversi contesti decisionali (segmentazione dei clienti, dei fornitori, analisi delle sequenze di acquisto etc…)

Bibliografia

Customer Relationship Management – Farinet, Ploncher – Etas
The CRM handbook – Dyché – Addison Wesley
CRM – Tourniaire – McGraw-Hill
CRM – Greenberg - Apogeo
Introduzione al Data Mining – Poiger, Geatz – Mcgraw-Hill
Data Mining – P. Giudici – Mcgraw-Hill
Data Mining – Berry, Linoff – Apogeo
Miniere di dati – A. Ferrari – FrancoAngeli
Discovering Data Mining – IBM – Prentice Hall PTR
Fuzzy logic and Neurofuzzy applications in business and finance - Von Altrock C. - Prentice Hall PTR

Links

Algoritmi di classificazione http://hpc.isti.cnr.it/~palmeri/datam/doc/notes/notes.html
Automazione della forza vendita https://www.salesforce.com/
Metodo Condorcet http://www.genarts.com/karl/second-choice-voting.html
Test chi-quadrato http://www2.unipr.it/~bottarel/epi/assoc/chi_qua.htm
Software di data Mining http://www.kdnuggets.com/