UNIVERSITÀ DEGLI STUDI DI PISA

Facoltà di Scienze
Matematiche Fisiche e Naturali

Corso di Laurea Specialistica in Tecnologie Informatiche

Analisi dei Dati ed Estrazione di Conoscenza

Progetto "COOl Patterns"

Analisi delle vendite nella grande distribuzione

Federico Colla

Anno Accademico 2004/05

Indice

1 Introduzione

Il progetto è stato condotto seguendo le linee guida della metodologia CRISP-DM. Nella seguente relazione esiste una sezione per ogni fase del CRISP-DM e, per ciascuna di esse, vi sono diverse sottosezioni che corrispondono ad alcuni *task* generici propri della metodologia adottata. Non tutte le fasi/task della metodologia CRISP-DM sono state trattate nella loro interezza per via della natura didattica del progetto.

2 Strumenti di analisi

Per lo svolgimento del progetto sono stati utilizzati diversi strumenti software. In particolare:

- SPSS Clementine 8.1 (regole associative, alberi di decisione, esplorazione e preparazione dati)
- PrefixSpan_O: una implementazione freeware dell'algoritmo *Prefix Span* (pattern sequenziali)
- Microsoft Visual Studio .NET 2003
- 5 programmi di ausilio sviluppati appositamente per il progetto, descritti in dettaglio nelle sezioni seguenti
- EDXOR, Un editor di testi capace di gestire file di grosse dimensioni
- Alcuni script BATCH di supporto

Per quanto riguarda le risorse computazionali, tutte le analisi sono state effettuate utilizzando un computer portatile Pentium IV 2.56GHz con 512MB di RAM.

3 Business Understanding

3.1 Business Objectives

Lo studio svolto riguarda l'analisi delle vendite nella grande distribuzione. I dati disponibili sono relativi alle vendite di un ipermercato situato nella provincia di Livorno, effettuate nel trimestre Gennaio–Marzo del 2005.
Il progetto si prefigge due obiettivi. Il primo consiste nel trovare associazioni *interessanti* tra prodotti venduti insieme e sequenze tipiche di prodotti acquistati, a diversi livelli di astrazione.
Il secondo obiettivo riguarda l'estrazione di un profilo dei clienti che supportano le regole/sequenze di acquisto trovate nella fase precedente.
In particolare vogliamo essere capaci di rispondere alle seguenti domande:

1. Quali sono i prodotti il cui acquisto influenza l'acquisto di altri prodotti?

2. Quali sono i prodotti che vengono spesso comprati in sequenza nel tempo?

3. Qual'è l'identikit del cliente il cui comportamento di acquisto soddisfa un certo pattern?

3.1.1 Success Criteria

Non sono state date misure oggettive per la valutazione del successo dell'analisi condotta. L'unico requisito imposto riguarda il primo obiettivo dell'analisi. Si chiede infatti che le associazioni/sequenze di prodotti trovate siano *interessanti*. Non è stata data però una definizione di 'associazione/sequenza interessante'. L'interpretazione considerata durante lo svolgimento del progetto si riferisce a delle misure oggettive/soggettive spiegate in seguito.

3.2 Data mining goals

Gli obiettivi del progetto possono essere 'tradotti' in termini tecnici. Il primo obiettivo si riferisce alla scoperta di regole associative a singola dimensione e multilivello. La gerarchia considerata per i prodotti viene descritta nelle sezioni seguenti. Il secondo obiettivo riguarda la scoperta di pattern sequenziali di prodotti venduti nel tempo, anche in questo caso a diversi livelli di astrazione.
Per l'estrazione del profilo dei clienti l'idea è quella di creare, per ogni regola associativa/pattern sequenziale trovata, un albero di decisione che ha lo scopo di classificare i clienti rispetto ad un attributo target il quale è positivo se il cliente soddisfa la regola associativa/pattern sequenziale considerato, negativo altrimenti. L'analisi dell'albero risultante permette di estrarre il profilo di interesse.

3.3 Project plan

Il progetto puó essere articolato in diversi punti che saranno ripresi in dettaglio nelle sezioni seguenti.

- Esplorazione dei dati e preparazione
- Costruzione dataset per l'estrazione di regole associative
- Costruzione dataset per l'estrazione dei pattern sequenziali
- Estrazione delle regole associative
- Estrazione dei pattern sequenziali
- Tagging dei clienti rispetto alle regole associative selezionate
- Tagging dei clienti rispetto ai pattern sequenziali selezionati

- Creazione alberi di decisione
- Estrazione profili clienti

4 Data Understanding

4.1 Data collection

L'insieme dei dati iniziali è rappresentato da una serie di file di tipo e formato diverso, elencati (solo quelli rilevanti ai fini dell'analisi) nella seguente lista

- `aggregato_scontrini_iper_liv.lst` scontrini aggregati
- 78 file `iva_scontrini_AAAAMMGG.lst` che rappresentano gli scontrini di ogni giornata lavorativa del trimestre (AAAA anno, MM mese, GG giorno)
- `anag_articoli.lst` dati anagrafici relativi agli articoli
- `carte_liv.lst` dati anagrafici relativi ai soci
- `Classificazione Marketing.xls` classificazione gerarchica dei prodotti
- `tracciati.txt` descrizione del formato dei vari file

4.2 Data description

In questa sezione tutti i dati utili all'analisi sono descritti in dettaglio. In particolare per ogni sorgente di dati ne viene descritto il formato, la quantità (numero di record e di attributi) e l'insieme dei campi.
Se un attributo è segnato come 'filtrato' significa che è stato ritenuto irrilevante ai fini dell'analisi e quindi scartato in fase di loading (nodo *filtro* in

Clementine). Tali attributi non saranno contenuti nelle tabelle intermedie create durante le varie fasi della preparazione dei dati.

4.2.1 Scontrini dettagliati

Gli scontrini dettagliati delle vendite sono distribuiti su 78 file di testo, uno per ogni giorno lavorativo del semestre. In totale vi sono 6 720 858 record, ognuno dei quali ha lunghezza fissa di 63 caratteri e comprende i seguenti 8 campi:

- *data_scontrino*: char(10) con formato 'gg/mm/aaaa'
- *cassa*: number(3) numero cassa
- *scontrino*: number(4) numero scontrino
- *cod_categoria*: char(3) codice categoria (filtrato)
- *cod_art*: number(6) codice articolo
- *venduto_valore*: number(13) valore spesa (filtrato)
- *qta_pezzi*: number(10) quantità pezzi (filtrato)
- *qta_peso*: number(13) quantità peso (filtrato)

Esiste un record per ogni prodotto acquistato, e la tripla (*data_scontrino*, *cassa*, *scontrino*) identifica univocamente il carrello di appartenenza.
Il caricamento dei dati relativi agli scontrini dettagliati avviene attraverso un nodo di input per file a lunghezza fissa di Clementine.

4.2.2 Scontrini aggregati

Il file di testo `aggregato_scontrini_iper_liv.lst` contiene i dati aggregati relativi alle vendite del trimeste considerato, per un totale di 2 215 593 record.

Ogni record del file ha una lunghezza fissa di 69 caratteri e comprende i seguenti 8 campi:

- *data_scontrino*: char(10) con formato 'gg/mm/aaaa'
- *cassa*: number(3) numero cassa
- *scontrino*: number(4) numero scontrino
- *nro_carta*: number(8) numero carta socio
- *cod_aggregato*: number(4) giustapposizione settore/reparto (filtrato)
- *venduto*: number(13) valore spesa (filtrato)
- *venduto_promo*: number(13) valore spesa prodotti in promozione (filtrato)
- *venduto_pam*: number(13) valore spesa prodotti a marchio COOP (filtrato)

Le vendite aggregate sono costituite dai totali di vendita relativi ad ogni scontrino e divisi per settori e reparti. La tripla (*data_scontrino*, *cassa*, *scontrino*) identifica univocamente il carrello di appartenenza. Questi dati sono importanti per l'analisi perché permettono di legare gli scontrini con il numero di carta socio e quindi risalire alle informazioni riguardanti il cliente. Il caricamento dei dati relativi agli scontrini dettagliati avviene attraverso un nodo di input per file a lunghezza fissa di Clementine.

4.2.3 Anagrafica articoli

Il file di testo `anag_articoli.lst` contiene i dati relativi agli articoli in vendita nell'ipermercato. Il file contiene 37 954 records, ognuno dei quali ha una lunghezza fissa di 53 caratteri e comprende i seguenti 6 campi:

- *cod_art*: number(7) codice articolo
- *art_descr*: char(30) descrizione articolo
- *cod_presmkt*: char(1) vendibilità commerciale (filtrato)
- *cod_clmkt*: char(11) codice classificazione marketing
- *cod_clgest*: char(3) codice categoria marketing (filtrato)
- *cod_stat*: char(1) stato articolo (filtrato)

Il codice di classificazione marketing (*cod_clmkt*) permette di ricavare la descrizione dell'articolo a diversi livelli di astrazione (settore, reparto, categoria, subcategoria).
Tale codice è formato da una stringa avente formato 'SSRRCCCZZXX' dove SS è il codice del settore, RR è il codice del reparto, CCC è il codice della categoria, ZZ è il codice della subcategoria e XX non è utilizzato ai fini di questa analisi. Il caricamento dei dati relativi agli articoli avviene attraverso un nodo di input per file a lunghezza fissa di CLEMENTINE.

4.2.4 Gerarchia prodotti

La descrizione della gerarchia degli articoli è specificata nel file Excel `Classificazione Marketing.xls`. Da quest'ultimo è possibile estrarre 4 tabelle che descrivono ciascuna un livello della gerarchia.

La gerarchia considerata è quindi la seguente:

articoli < **subcategorie** < **categorie** < **reparti**< **settori**

Le nuove tabelle sono descritte nelle sezioni seguenti.

Settori

La tabella Settori (contenuta nel file `settori_descr.dat`) ha un totale di 9 record formati dai 2 campi separati da tabulazione *cod_settore* (codice del settore) e *descr_settore* (descrizione testuale del settore).

Reparti

La tabella Reparti (contenuta nel file `reparti.dat`) ha un totale di 54 record formati dai 3 campi *cod_settore*, *cod_reparto* (codice reparto) e *descr_reparto* (descrizione testuale del reparto).
Ogni reparto è identificato univocamente dalla giustapposizione del codice settore e del codice reparto.
A partire dalla tabella Reparti si è costruita un'altra tabella avente due campi: *key* (i cui valori sono la giustapposizione del codice settore e del codice reparto) e *descr_reparto*. La nuova tabella è memorizzata nel file di testo `reparti_descr.dat` in cui i valori dei campi sono separati da tabulazione.

Categorie

La tabella Categorie (contenuta nel file `categorie.dat`) ha un totale di 402 record formati dai 4 campi *cod_settore*, *cod_reparto*, *cod_categ* (codice categoria) e *descr_categ* (descrizione testuale della categoria).
Ogni categoria è identificata univocamente dal codice della categoria *cod_categ*.
A partire dalla tabella Categorie si è costruita un'altra tabella, in cui ogni record contiene solo il codice e la descrizione della categoria. La nuova tabella è memorizzata nel file di testo `categorie_descr.dat`, in cui i valori dei campi sono separati da tabulazione.

Subcategorie

La tabella Subcategorie (contenuta nel file `subcategorie.dat`) ha un totale di 1 516 record formati dai 5 campi *cod_settore*, *cod_reparto*, *cod_categ*, *cod_subcateg* (codice subcategoria) e *descr_subcateg* (descrizione testuale della subcategoria).
Ogni subcategoria è identificata univocamente dalla giustapposizione del codice categoria e del codice subcategoria.
A partire dalla tabella Subcategorie si è costruita un'altra tabella avente due campi: *key* (i cui valori sono la giustapposizione del codice categoria e del codice subcategoria) e *descr_subcateg*. La nuova tabella è memorizzata nel file di testo `subcategorie_descr.dat`, in cui i valori dei campi sono separati da tabulazione.

4.2.5 Anagrafica soci

Il file di testo `carte_liv.lst` contiene i dati relativi ai clienti. Il file contiene 97 814 records, ognuno dei quali ha una lunghezza fissa di 189 caratteri e comprende i seguenti 10 campi:

- *nro_carta*: number(10) numero carta socio
- *nro_socio*: number(5) numero socio (filtrato)
- *data_nasc*: char(10) con formato 'gg/mm/aaaa'
- *sesso*: char(1)
- *stato_civile*: char(1)
- *professione*: char(50)
- *titolo_studio*: char(50)

- *res_citta*: char(50) residenza (filtrato)
- *res_cap*: char(5) CAP residenza (filtrato)
- *socio_capofam*: number(5) (filtrato)

4.3 Data exploration

In questa sezione andiamo piú a fondo nella descrizione dei dati, in particolare quelli relativi ai clienti.
A partire dall'attributo *data_nasc* possiamo ricavare l'età del cliente, usando il nodo *nuovo campo* di CLEMENTINE. L'istogramma del nuovo attributo *age* è mostrato nella figura 1. In particolare, l'attributo *age* ha minimo 6, massimo 105 e media 51.

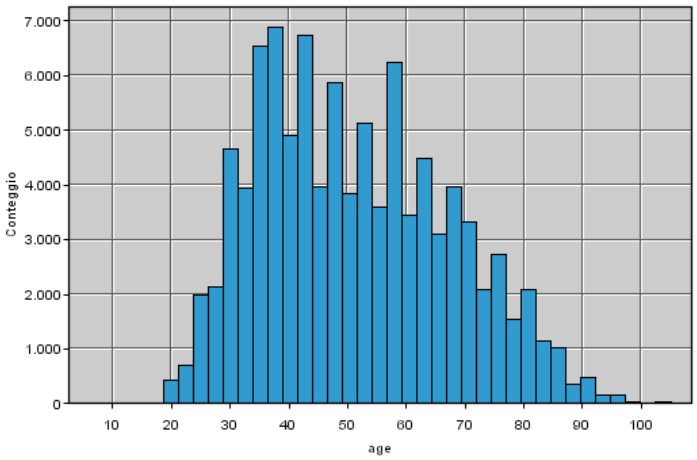

Figura 1: *Istogramma di age*

La distribuzione dell'attributo *sesso* è mostrata nella tabella 1. La moda di tale attributo è 'F'.

La distribuzione dell'attributo *stato_civile* è mostrata nella tabella 2. La

Valore	Proporzione %	Conteggio
F	61,544	60 198
M	38,456	37 616

Tabella 1: *Distribuzione di sesso*

moda di tale attributo è 'S', ovvero i clienti sposati sono la maggioranza.

Valore	Proporzione %	Conteggio
C	31,811	31 116
D	0,347	340
I	0,926	906
P	0,563	551
S	65,152	63 728
V	1,199	1 173

Tabella 2: *Distribuzione di stato_civile*

La distribuzione degli attributi *titolo_studio* e *professione* sono mostrate rispettivamente nelle tabelle 3 e 4.

5 Data Preparation — Obiettivo 1

In questa sezione viene descritto il processo di creazione dei dataset necessari alle analisi. Il dataset necessario all'estrazione dei profili dei clienti verrà discusso in seguito perché la sua creazione richiede dati derivanti dall'analisi relativa al primo obiettivo del progetto.

Valore	Proporzione %	Conteggio
AGRARIA	0,009	9
ALBERGHIERO	0,022	22
CLASSICO	0,106	104
ECONOMIA E COMM.	0,027	27
ELEMENTARE	2,961	2 897
GEOMETRA	0,259	254
GIURISPRUDENZA	0,038	38
INFORMATICA	0,0163	16
INGEGNERIA	0,077	76
LAUREA - ALTRO	3,151	3 083
LAUREA BREVE	0,001	1
MAGISTRALE	0,788	771
MATER. CLASSICHE	0,020	20
MATURITA' ALTRO	2,925	2 862
MEDIA INFERIORE	8,821	8 629
MEDIA SUPERIORE	6,438	6 298
NON INDICATA	72,1278	70 551
RAGIONERIA	1,487	1 455
SCIENTIFICO	0,359	352
SCIENZE ECONOM.	0,001	1
SCUOLA PROFESSIONALE	0,034	34
TECNICO AGRARIO	0,023	23
TECNICO INDUSTR.	0,276	270
TURISTICO	0,021	21

Tabella 3: *Distribuzione di titolo_studio*

Valore	**Proporzione %**	*Conteggio*
AGRICOLTORE	0,279	273
ALTRE PROFESSIONI	0,093	91
ARTIGIANO	2,190	2 143
CASALINGA	26,667	26 085
CLERO	0,0255	25
COMMERCIANTE	0,0368	36
COOP CONSUMO	4 0,239	234
DIRIGENTE DI AZIENDA	0,188	184
DISOCCUPATO	3,530	3 453
ENTE PRIVATO	0,905	886
ENTE PUBBLICO	5,877	5 749
IMPIEGATO	18,191	17 794
INSEGNANTE	0,047	46
LAVORATORE AUTONOMO	3,453	3 378
LIBERO PROFESSIONISTA	3,282	3 211
MILITARE DI CARRIERA	1,374	1 344
NON INDICATA	1,120	1 096
OPERAIO	16,2087	15 854
PENSIONATO	9,162	8 962
POSSIDENTE	0,008	8
STUDENTE	7.117	6 962

Tabella 4: *Distribuzione di professione*

5.1 Dataset construction

5.1.1 Dataset per l'estrazione di regole associative

L'obiettivo di questo task è creare il dataset in formato relazionale che contiene i dati delle le vendite di tutto il trimestre.
La tabella avrà i campi *key*, *nro_carta*, *cod_art*, *cod_subcateg*, *cod_categ*, *cod_reparto* e *cod_settore*. Il campo *key* ha lo scopo di identificare univocamente ogni transazione (carrello) e sarà descritto in dettaglio in seguito. La semantica degli altri attributi è quella descritta nelle sezioni precedenti.

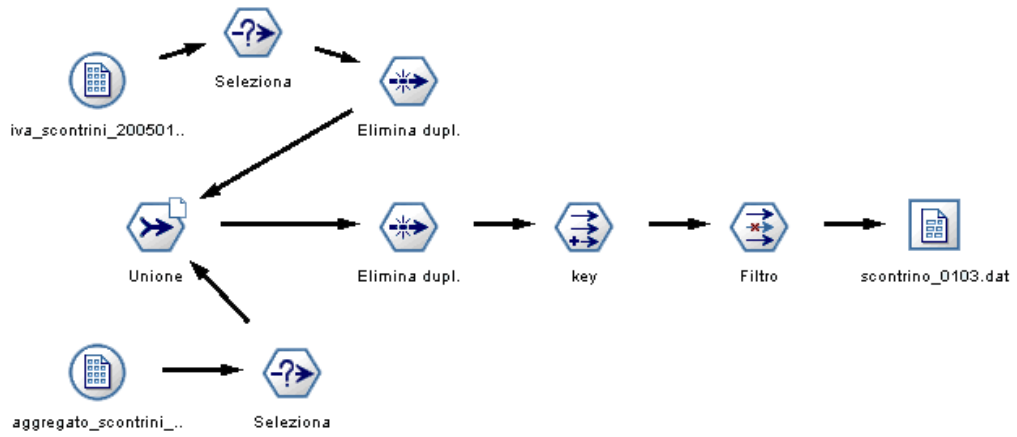

Figura 2: *Creazione dataset, processing scontrini*

Il processo di creazione del dataset è stato il seguente. Utilizzando lo *stream* di Clementine mostrato nella figura 2 sono stati processati tutti i 78 file relativi agli scontrini dettagliati, al fine di creare altrettanti 78 file contenenti le stesse informazioni iniziali piú il campo *nro_carta*, che mette in relazione ogni transazione con il cliente che l'ha effettuata.
Durante il caricamento dei dati, dallo scontrino dettagliato vengono eliminati i record aventi il codice articolo '174292', che corrisponde ad un articolo particolare, il sacchetto in plastica per i prodotti acquistato alle casse, presente in quasi tutte le transazioni e quindi ritenuto irrileventante ai fini delle

analisi successive.
Proseguendo nello stream, viene fatta una join con i dati degli scontrini aggregati, usando la tripla (*data_scontrino*, *cassa*, *scontrino*) come chiave. Tutti i valori duplicati sono stati rimossi, in modo tale da non avere piú record per lo stesso articolo nella stessa transazione.
Viene poi creato il campo stringa *key* i cui valori hanno il formato 'MMGGCCCSSSS' dove MM è il mese, GG è il giorno, CCC è il numero di cassa e SSSS è il numero dello scontrino. Si noti che ordinando lessicograficamente la chiave, si mantiene l'ordine temporale degli acquisti.
Infine vengono filtrati i campi *data_scontrino*, *cassa* e *scontrino* e il risultato scritto su un file di testo delimitato da tabulazione.

Dopo aver processato tutti gli scontrini dettagliati, i file risultanti sono stati concatenati, mantenendo l'ordine cronologico, per ottenere un unico file, `vendite_trimestre.dat`.

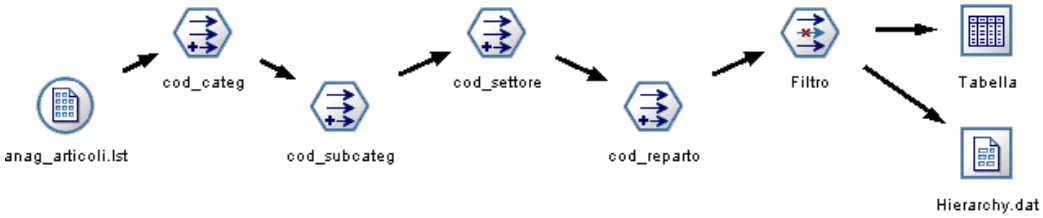

Figura 3: *Creazione gerarchia prodotti*

Per la costruzione della gerarchia degli articoli è stato usato lo stream Clementine mostrato in figura 3. A partire dall'anagrafica degli articoli, in particolare usando il campo codice di classificazione marketing (*cod_clmkt*), sono stati creati altri 4 campi che sono rispettivamente, il codice della subcategoria, della categoria, del reparto e del settore. Infine la descrizione e il codice iniziale di classificazione marketing sono stati filtrati e il risultato è

stato scritto nel file di testo `Hierarchy.dat`, i cui record sono delimitati da tabulazione.

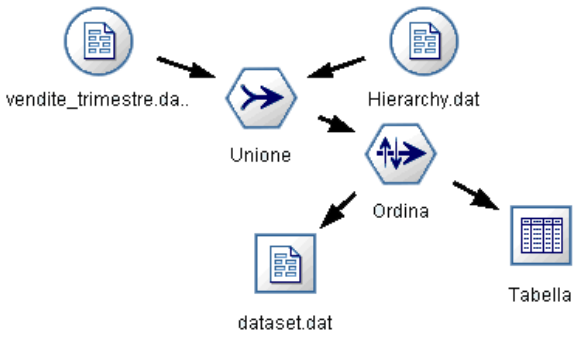

Figura 4: *Creazione dataset, fase finale*

Il dataset finale è stato quindi ottenuto attraverso lo stream di CLEMENTINE mostrato nella figura 4, in cui viene fatta una join delle due tabelle appena create usando come chiave l'attributo *cod_art*. Il risultato viene poi ordinato secondo gli attributi (*key* e *cod_art*) e scritto nel file di testo DATASET.DAT. Il dataset finale contiene 5 098 533 record ed un piccolo estratto è mostrato nella tabella 5. Il numero di record è ridotto rispetto a quello del file `vendite_trimestre.dat` per via dell'eliminazione dei duplicati, dei record relativi all'articolo '174292' e per via dell'*inner join* che elimina i record che non hanno corrispondenze.

key	nro_carta	cod_art	cod_categ	cod_subcateg	cod_reparto	cod_settore
01030011731	31403686	2561	009	01	01	01
01030011731	31403686	2545	009	01	01	01
01030011731	31403686	3393	009	03	01	01

Tabella 5: *Dataset per estrazione regole associative (estratto)*

5.1.2 Dataset per l'estrazione di pattern sequenziali

La creazione dei dataset per l'estrazione di pattern sequenziali ha richiesto uno sforzo maggiore, per via delle caratteristiche del software scelto per effettuare l'analisi, una implementazione freeware dell'algoritmo Prefix Span chiamata PrefixSpan_O.
Nel seguito, il termine *item* si riferisce ad un oggetto contenuto nella transazione, sia esso un codice di articolo, di categoria, di reparto, ecc.
Il software richiede che file di input sia binario ed abbia il formato qui descritto. Supponendo di avere due sequenze di transazioni

$$\{\ (0\ 1)\ (2)\ (1\ 3)\ \}, \{\ (1)\ (3)\ \}$$

dove 0, 1 e 2 sono gli identificativi degli item (articoli, categorie, ecc.) contenuti nelle transazioni, il formato del file di input è il seguente

$$0\ 1\ -1\ 2\ -1\ 1\ 3\ -1\ -2\ 1\ -1\ 3\ -1\ -2$$

in cui -1 delimita le transazioni dello stesso cliente e -2 delimita le sequenze di clienti diversi. Inoltre i codici identificativi degli item devono essere numeri interi di 32 bit, numerati da 0 a $|items| - 1$, dove $|items|$ è il numero totale di item presenti nel dataset.

Per creare il file di input è stato scritto un programma, DSBuilder, che dati la lista *ordinata* degli item che compaiono *effettivamente* nelle transazioni e il file `dataset.dat`, restituisce in output un file binario, formattato come descritto in precedenza, che contiene le sequenze di acquisto di tutti i clienti, al livello di astrazione desiderato. I file contenenti le liste degli item sono stati creati con Clementine, usando lo stream mostrato nella figura 5. I nuovi campi *item_ID* creati sono quelli specificati nella sezione 4.2.4 e vengono utilizzati come identificativi univoci delle subcategorie e dei reparti.

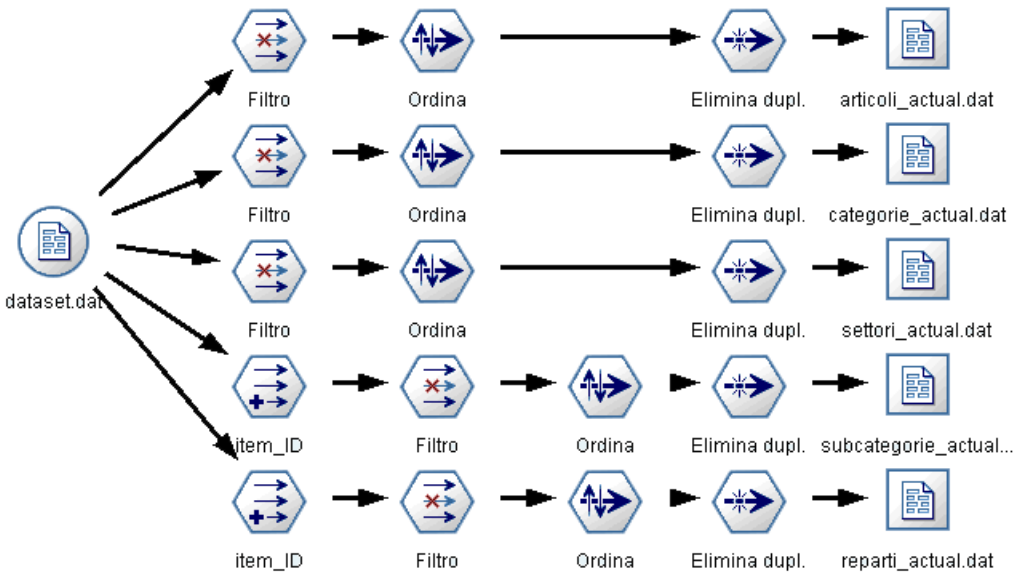

Figura 5: *Creazione liste item*

DSBuilder ricrea in memoria tutte le sequenze delle transazioni per cliente, e attraverso una funzione che mappa l'item nel nuovo identificativo, restituisce un file binario correttamente formattato. Inoltre DSBuilder crea un altro file di testo, chiamato *items*`_map.dat`, (ad esempio `articoli_map.dat`) che contiene la tabella di conversione dell'identificativo originale degli item, nell'identificativo usato per l'analisi con il software PrefixSpan_O.
Alla fine di questa fase avremo quindi 5 nuovi dataset chiamati `dataset_ps_`X`.dat`, dove $X \in \{articoli, subcategorie, categorie, reparti, settori\}$.

6 Modeling — Obiettivo 1

6.1 Estrazione regole associative

Il file `dataset.dat` contiene il dataset necessario per l'estrazione delle regole associative ad ogni livello di astrazione. Lo stream di Clementine usato per

lo scopo è quello mostrato nella figura 6. Per la creazione del modello è stato utilizzato il nodo *Apriori*. Anche in questo caso, I nuovi campi *item_ID* che appaiono nello stream sono quelli specificati nella sezione 4.2.4, e vengono usati come identificativo delle subcategorie e dei reparti. Il software ha permesso di effettuare l'analisi usando i dati in formato transazionale, usando l'attributo *key* come identificatore della transazione e, a seconda del livello di astrazione considerato, i codici di articolo, subcategoria, categoria, reparto e settore come attributi di input/output.

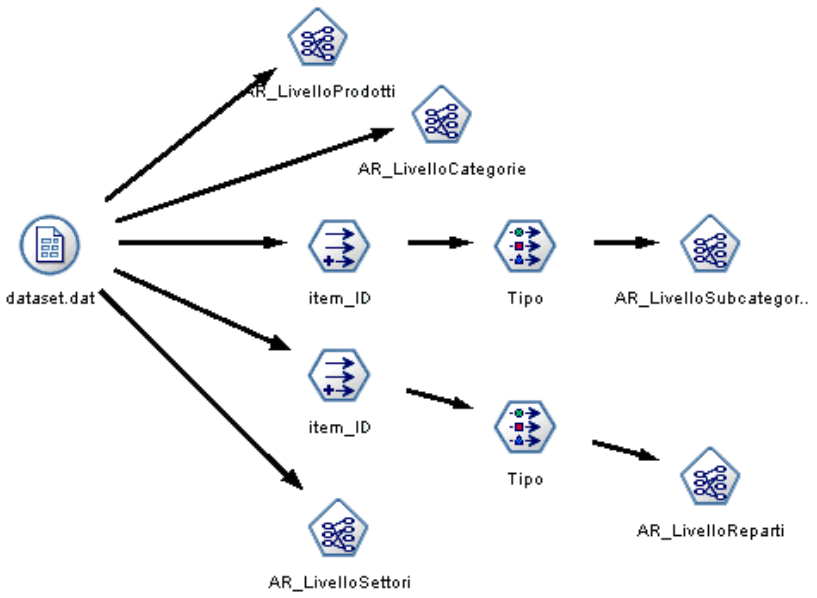

Figura 6: *Creazione modelli regole associative*

La strategia utilizzata per l'estrazione delle regole è quella del *reduced support*, ovvero ogni livello di astrazione ha la sua soglia di supporto minimo: piú basso è il livello nella gerarchia, piú piccola è la soglia di supporto minimo corrispondente. Le soglie di supporto e di confidenza minimo sono riportate nella tabella 6.

Livello	Supporto minimo	Confidenza minima
Articoli	0,01%	80%
Subcategorie	0,2%	75%
Categorie	0,7%	75%
Reparti	4%	75%
Settori	8%	80%

Tabella 6: *Soglie di supporto e confidenza per regole associative*

I risultati ottenuti sono presentati nella sezione 7.

6.2 Estrazione dei pattern sequenziali

Per l'estrazione dei pattern sequenziali è stato utilizzato il software PrefixSpan_O, reperibile all'indirizzo internet
`http://www-sal.cs.uiuc.edu/~hanj/software/prefixspan.htm`.
Anche in questo caso è stata adottata una strategia *reduced support*. Le soglie di supporto usate sono riportate nella tabella 7.

Livello	Supporto minimo
Articoli	2%
Subcategorie	10%
Categorie	20%
Reparti	30%
Settori	40%

Tabella 7: *Soglie di supporto per pattern sequenziali*

I risultati ottenuti sono presentati nella sezione 7.

7 Evaluation — Obiettivo 1

7.1 Selezione regole associative interessanti

Di seguito sono presentati i risultati ottenuti con l'algoritmo Apriori. Il modello ottenuto (nello specifico l'insieme di regole) è stato esportato in un file di testo in cui esiste un record per ogni regola e i cui valori sono delimitati da tabulazione. Un esempio di record presente nel file di output è mostrato nella tabella 8.

Istanze	Supporto	Confidenza	Lift	Conseguente	Antecedente 1
53	0.01	92.5	4237.263	283917	283920

Tabella 8: *Regole associative, output*

I record sono di lunghezza variabile in quanto ogni regola puó avere un numero di antecedenti diverso. Avendo applicato l'algoritmo sui *codici* degli articoli/categorie/ecc. le regole ottenute non sono direttamente interpretabili. Per ovviare a questo problema è stato scritto un programma, PrettyPrinterApriori, che dato in input l'insieme di regole 'grezze', restituisce un file di testo in cui i codici sono rimpiazzati dalle rispettive descrizioni testuali. Ad esempio la regola grezza

53 0.01 92.5 4237.263 283917 283920

viene trasformata in

[10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
Line: 70 Support: 0,01 Confidence: 92,5 Lift: 4237,263

dove *Line* indica la posizione della regola nel file di output originale.

Una regola associativa è stata ritenuta *interessante* se oltre ad essere valida, ha un valore di *lift* maggiore di 1, ovvero quando gli item sono correlati positivamente, e il contenuto devia da ció che viene considerata 'conoscenza

comune'. Per facilitare l'analisi, il programma PrettyPrinterApriori stampa solo le regole che hanno un lift maggiore o uguale ad 1.
Nelle sezioni seguenti sono elencate le regole interessanti trovate (una piccola selezione) per ogni livello di astrazione. Come ci si poteva aspettare, le regole interessanti sono concentrate nei livelli piú bassi della gerarchia.

Livello Articoli

[10 BICCH.CART.BIBO CIRC.200CC] → [PIATTI CART.BIBO CIRCUS D23X10]
Line: 70 Support: 0,01 Confidence: 92,5 Lift: 4237,263

[TELO 100X150 460 GR/MQ TU] [OSPITE 40X60 460 GR/MQ TU] → [ASCIUGAMANO 60X110 460 GR TU]
Line: 106 Support: 0,01 Confidence: 91,4 Lift: 965,993

[BOCC.CANI POLLO/TACCH.KG1.23] [BOC/NI GATTO VITELLO SIM.KG415] → [BOCC.GATTI CONIGLIO SIMBA G415]
Line: 107 Support: 0,01 Confidence: 91,4 Lift: 390,042

[PIATTO FRUTTA MAZIME B.CO CM21] [PIATTO F.DO MAXIME B.CO CM.17] → [PIATTO P.NO MAXIME B.CO CM.25]
Line: 161 Support: 0,01 Confidence: 90 Lift: 3052,386

[LENZUOLO PIANO 150X280 RIGHE] [LENZUOLO ANGOLI 90X200 TU] → [FEDERA 50X80 STAMPA RIGHE]
Line: 312 Support: 0,01 Confidence: 87,8 Lift: 809,222

[GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/MANZO G85] → [GOURMET PERLE FIL.CONIGLIO G85]
Line: 313 Support: 0,01 Confidence: 87,8 Lift: 492,757

[CUCCHIAIONE ACCIAIO INOX] [PALA FRITTO ACCIAIO INOX] [FORCHETTONE ACCIAIO INOX] → [SCHIUMAROLA IN ACCIAIO INOX]
Line: 510 Support: 0,01 Confidence: 85,7 Lift: 1912,523

[APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
Line: 745 Support: 0,01 Confidence: 83,3 Lift: 1314,55

[GASSOSA S. BENEDETTO LT.1.5] [CEDRATA SAN BENEDETTO LT.1.5] [ARANCIATA S.BENEDETTO LT.1,5] → [SPUMA BIONDA LT1.5 S.BENEDETTO]
Line: 778 Support: 0,01 Confidence: 83 Lift: 172,76

[BARAT.OVALE LT1,7 VTR COP.ACC.] [BARAT.OVALE LT0,84 VTR COP.ACC] → [BARAT.OVALE LT1,2 VTR COP.ACC.]
Line: 781 Support: 0,01 Confidence: 82,9 Lift: 1993,002

[MOUSSE GAT.COOP MANZ/FEGAT.G85] [MOUSSE GAT.COOP PES/TROTA G85] → [MOUSSE GATTO COOP POL/TAC.G85]
Line: 918 Support: 0,1 Confidence: 81,7 Lift: 712,617

Livello Subcategorie

[BIBITE-ARANCIATE] [SNACK SALATI-PATATINE] [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [BIBITE-COLE]
Line: 34 Support: 0,1 Confidence: 88,2 Lift: 11,084

[USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
Line: 881 Support: 0,1 Confidence: 84,7 Lift: 12,767

[USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-STOV. PLAST. COLORATA DECORATA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
Line: 5391 Support: 0,1 Confidence: 82,2 Lift: 12,391

[SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
Line: 5395 Support: 0,1 Confidence: 82,2 Lift: 10,34

[CARAMELLE/PROD. BASE ZUCCH.-ALTRE CARAMELLE] [CARAMELLE/PROD. BASE ZUCCH.-CARAM.NORMALI] [CARAMELLE/PROD. BASE ZUCCH.-GOMME DA MASTICARE] → [PRODOTTI BASE CIOCCOLATO-SNACK]
Line: 9938 Support: 0,1 Confidence: 81,2 Lift: 8,693

Livello Categorie

[UOVA] [OF PREPARATA] [VERDURA FRESCA] [LATTE] [FRUTTA FRESCA] → [ORTAGGI]
Line: 38 Support: 0,8 Confidence: 85,2 Lift: 1,893

[CAFFE] [UOVA] [VERDURA FRESCA] [FRUTTA FRESCA] → [ORTAGGI]
Line: 128 Support: 0,7 Confidence: 84,3 Lift: 1,871

[UOVA] [GRASSI] [VERDURA FRESCA] [AVICUNICOLO] → [ORTAGGI]
Line: 291 Support: 0,9 Confidence: 83,5 Lift: 1,854

[OLIO DI OLIVA] [UOVA] [SUINO] → [BOVINO]
Line: 4902 Support: 0,7 Confidence: 78,9 Lift: 1,757

[ZUCCHERO] [IGIENE CARTA] [DETERGENTI SUPERFICI] → [DETERGENZA TESSUTI]
Line: 11557 Support: 0,7 Confidence: 76,6 Lift: 2,247

Livello Reparti

[FRESCHI-CARNI BIANCHE] [FRESCHI-SURGELATI] [FRESCHI-GASTRONOMIA] → [FRESCHI-CARNI ROSSE]
Line: 4869 Support: 5,2 Confidence: 75,5 Lift: 1,217

Livello Settori

A questo livello di astrazione non sono state trovate regole aventi Lift maggiore di uno.

7.2 Selezione pattern sequenziali interessanti

Di seguito sono presentati i risultati ottenuti con l'algoritmo PrefixSpan. Nel file di output esiste una linea per ogni pattern sequenziale. Un esempio di record 'grezzo' presente nel file di output è il seguente

$$(0\ 1)\ (3)\ :\ 0.321787$$

in cui ogni gruppo di parentesi rappresenta un elemento della sequenza. Tra parentesi sono racchiusi gli identificativi degli item appartenenti alla stessa transazione. Infine viene stampato il supporto del pattern sequenziale.
Anche in questo caso il risultato non è direttamente interpretabile ed è stato scritto quindi un programma, PrettyPrinterPS, che ha lo scopo di rendere leggibili i pattern sequenziali trovati. In particolare, l'esempio di sopra viene trasformato in

{ [GENERI VARI] [FRESCHI] } { [TESSILE] }
Line: 175 Support: 0,331645

Nelle sezioni seguenti sono elencati i pattern sequenziali (una piccola selezione) trovati per ogni livello di astrazione.

Livello Articoli

{ [BANANE SFUSE] } { [KIWI SFUSI] }
Line: 793 Support: 0,02218

{ [PATATE P.GIALLA COOP VTB KG2,5] } { [PATATE BOLOGNA CAL.4/7 R.KG2.5] }
Line: 955 Support: 0,023268

{ [PIZZAIOLA TRIS LOCAT.GR.125X3] } { [MOZZAR.S.LUCIA TRIS GR.125X3] }
Line: 8547 Support: 0,010283

{ [FETTINE SCELTE VITELLONE] } { [PETTO POLLO COOP FETTE GL CF] }
Line: 36530 Support: 0,010928

{ [IMPASTO PIZZA] } { [IMPASTO PIZZA] } { [IMPASTO PIZZA] }
Line: 67093 Support: 0,007223

Livello Subcategorie

{ [AVICUNICOLO-POLLO] } { [BOVINO-VITELLONE] }
Line: 1296 Support: 0,148801

{ [FORMAGGI A SERV.ASSISTITO-SEMIDURI/DURI] } { [FORMAGGI LIBERO SERVIZIO-FORMAGGI FRESCHI] }
Line: 1056 Support: 0,100328

Livello Categorie

{ [BISCOTTI] } { [LATTE] }
Line: 70 Support: 0,222258

{ [FORMAGGI LIBERO SERVIZIO] } { [SALUMI LIBERO SERVIZIO] }
Line: 278 Support: 0,212621

{ [ORTAGGI] } { [VERDURA FRESCA] }
Line: 404 Support: 0,219369

Livello Reparti

{ [GENERI VARI-DROGHERIA ALIM. 1] } { [GENERI VARI-LIQUIDI] [FRESCHI-ORTOFRUTTA] }
Line: 714 Support: 0,381852

Livello Settori

{ [GENERI VARI] [FRESCHI] } { [TESSILE] }
Line: 175 Support: 0,331645

8 Data Preparation — Obiettivo 2

In questa e nelle sezioni che seguono è stato descritto il lavoro svolto per il raggiungimento del secondo obiettivo del progetto, ovvero l'estrazione di un

profilo dei clienti che supportano le regole/sequenze di acquisto trovate nella fase precedente.

8.1 Dataset construction

L'obiettivo di questo task è creare il dataset per l'estrazione dei profili dei clienti. Il processo adottato si avvale di due programmi scritti appositamente per lo scopo e di seguito descritti.
Il programma CustumerIDRetrieverApriori prende in input un file di testo contenente un insieme di regole associative 'grezze', come descritte nella sezione 7.1 e, per ciascuna di esse, restituisce un file di testo contenente una tabella i cui record di due campi (separati da virgole) contengono il numero della carta del cliente e un flag binario che indica se il cliente supporta o meno la regola e che sarà utilizzato come attributo *target* nella fase successiva. Esiste un record per ogni cliente che ha effettuato un acquisto nel trimestre considerato.
Il secondo programma, CustumerIDRetrieverPS, esegue la stessa operazione del precedente ma prende in input un file di testo contenente un insieme di pattern sequenziali 'grezzi', come descritti nella sezione 7.2. Il file prodotto in output ha la stessa struttura del precendente.

I dataset finali per l'estrazione del classificatore vengono costruiti facendo la join della tabella anagrafica clienti e queste nuove tabelle create (Si veda la figura 7). Il processo di join elimina i clienti che non hanno fatto acquisti nel trimestre in quanto non apportano nessuna informazione al fine della determinazione del profilo.

9 Modeling — Obiettivo 2

Lo scopo di questa fase è quello di creare dei modelli, degli alberi di decisione, per la variabile target definita in precedenza. L'algoritmo adottato per la costruzione dell'albero di decisione è il C5.0, fornito dal nodo *C5.0* di CLEMENTINE. E' stato creato un modello per ogni regola associativa/pattern sequenziale ritenuto interessante. Tutti i modelli sono stati costruiti utilizzando la tecnica *5-fold cross validation*.
Lo stream utilizzato per la costruzione dei modelli è mostrato nella figura 7.

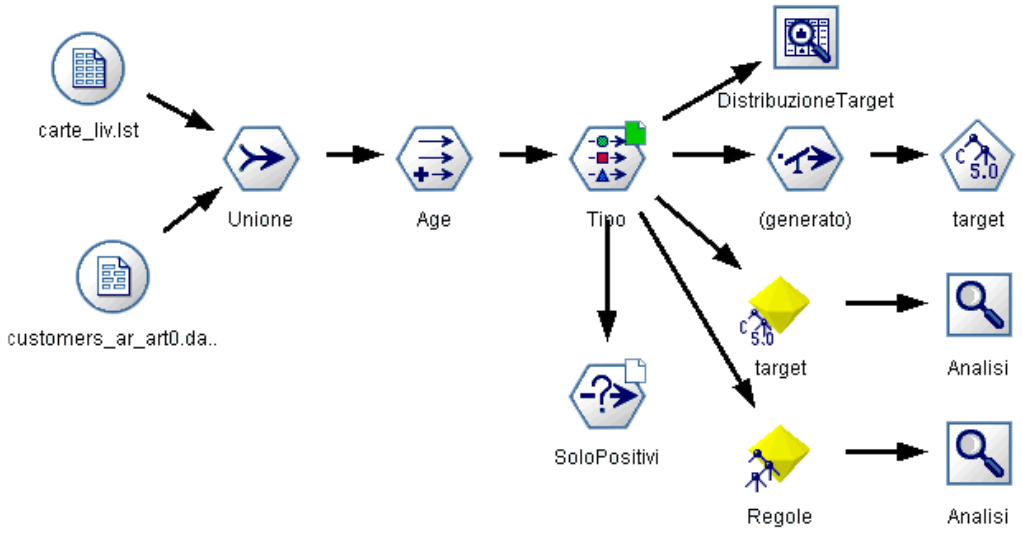

Figura 7: *Stream creazione profilo clienti*

Inizialmente viene fatto un join tra la tabella anagrafica clienti e una tabella creata con la tecnica descritta in precedenza. Il risultato è una nuova tabella che contiene i dati di tutti e soli i clienti che hanno effettuato acquisti nel trimestre preso in considerazione dall'analisi. Associato ad ogni cliente c'è un attributo binario (la variabile target) da predire.
Dalla data di nascita dei clienti è stato ricavato l'attributo *age*, ovvero l'età espressa in anni. Gli attributi presi in considerazione per l'analisi sono quelli

mostrati nella tabella 9.

Attributo	**Tipo**
sesso	flag
stato_civile	insieme discreto
professione	insieme discreto
titolo_studio	insieme discreto
age	intervallo

Tabella 9: *Attributi predittori*

Le prime analisi hanno mostrato che, essendo i clienti 'positivi' pochi rispetto al totale, l'algoritmo creava un albero composto da un solo nodo che classifica tutti gli esempi come negativi. Per risolvere il problema, per ogni dataset è stata calcolata la distribuzione della variabile target (nodo *esplora*) e grazie ad essa è stato creato uno nuovo nodo *bilanciamento* che ha permesso di avere un numero bilanciato di esempi positivi e negativi, per duplicazione degli esempi negativi.
A partire dall'albero di decisione ottenuti sono state generate le regole per la classificazione delle due classi in formato testuale. Per la creazione delle regole sono state impostati livelli di confidenza minimi del 95%. Per la determinazione del profilo cliente l'attenzione è stata posta sulle regole che classificano i clienti come positivi.
I risultati ottenuti sono discussi nella prossima sezione.

10 Evaluation — Obiettivo 2

Il profilo dei clienti che supportano le regole associative/pattern sequenziali, sono stati creati considerando le 3 regole di classificazione estratte dall'albero

di decisione piú significative in termini di supporto/confidenza. Un esempio di regola di classificazione è il seguente:
se $Age \leq 61$
e $Age > 57$
e *titolo_studio* = MEDIA INFERIORE
e *professione* in = CASALINGA
e *stato_civile* = C
allora T
dove 'T' è il flag che identifica i clienti 'positivi'.

10.1 Profili per Regole Associative

In questa sezione vengono discussi i risultati relativi alle regole associative trovate e i profili dei clienti corrispondenti.

Livello Articoli

La prima regola presa in considerazione è

[GOURM.GOLD DADINI GELLEE G85X8] [GOURMET PERLE FIL.C/MANZO G85] → [GOURMET PERLE FIL.CONIGLIO G85]
Support: 0,01 Confidence: 87,8 Lift: 492,757

Tale regola è supportata da 41 clienti. Il classificatore ottenuto ha una accuratezza del 96,07% su tutti i dati, e del 100% sui dati classificati come positivi.
Analizzando le regole di classificazione emerge che i clienti che supportano questa regola sono:

- Casalinghe, non sposate, di età tra i 57 e i 63 anni, che hanno la terza media inferiore come titolo di studio (confidenza: 99%)

- Ragazze single tra i 26 e i 29 anni che lavorano come impiegate (confidenza: 99.8%)
- Uomini pensionati di età minore di 51 anni (confidenza: 99%)

La regola

[APER.CAMPARI MIXX PEACH ML275] [APERIT.CAMPARI MIXX LIME ML275] [APERITIVO CAMP.GRADI 6,5 ML275] → [CAMPARI MIXX ORANGE ML275]
Support: 0,01 Confidence: 83,3 Lift: 1314,55

è supportata da 27 clienti. Il classificatore ottenuto ha una accuratezza del 97,6% su tutti i dati, e dell'85,19% sui dati classificati come positivi.
Analizzando le regole di classificazione emerge che i clienti che supportano questa regola sono:

- Ingegneri maschi aventi 26-27 anni (confidenza: 100%)
- Ragazze dai 26 ai 30 anni che hanno un lavoro autonomo e sono diplomate (confidenza: 99%)
- Impiegati single dai 26 ai 53 anni (confidenza: 96.5%)

Livello Subcategorie

La regola

[USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA] [USA E GETTA TAVOLA-ACCESSORI USA E GETTA] [USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA] → [USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI]
Support: 0,1 Confidence: 84,7 Lift: 12,767

è supportata da 158 clienti. Il classificatore ottenuto ha una accuratezza del 88,13% su tutti i dati, e del 100% sui dati classificati come positivi.
Analizzando le regole emerge che i clienti che supportano questa regola sono:

- Uomini celibi che lavorano per enti pubblici aventi 43-45 anni (confidenza: 98,1%)
- Liberi professionisti aventi titolo di studio media inferiore di 59-60 anni (confidenza: 99,4%)
- Militari di carriera sposati di aventi 38-41 anni (confidenza: 98,8%)

La regola

[SNACK SALATI-POP CORN/CEREALI] [SNACK SALATI-ESTRUSI] [BIBITE-ARANCIATE] → [BIBITE-COLE]
Support: 0,1 Confidence: 82,2 Lift: 10,34

è supportata da 155 clienti. Il classificatore ottenuto ha una accuratezza del 86,73% su tutti i dati, e del 100% sui dati classificati come positivi.
Analizzando le regole di classificazione emerge che i clienti che supportano questa regola sono:

- Ragazze disoccupate di 30-34 anni aventi un diploma magistrale (confidenza: 100%)
- Vedovi di 57-60 anni liberi professionisti (confidenza: 100%)
- Uomini/donne sposati di 32-40 anni e impiegati (confidenza: 100%)

Livello Categorie

La regola

[OLIO DI OLIVA] [UOVA] [SUINO] → [BOVINO]
Support: 0,7 Confidence: 78,9 Lift: 1,757

è supportata da 1 440 clienti. Il classificatore ottenuto ha una accuratezza del 64,83% su tutti i dati, e dell'89,24% sui dati classificati come positivi. Analizzando le regole di classificazione emerge che i clienti che supportano questa regola sono:

- Operai diplomati single aventi 33-34 anni (confidenza: 100%)
- Dirigenti di azienda sposati aventi 39-70 anni(confidenza: 100%)
- Casalinghe single diplomate aventi 33-39 anni (confidenza: 100%)

La regola

[UOVA] [OF PREPARATA] [VERDURA FRESCA] [LATTE] [FRUTTA FRESCA] → [ORTAGGI]
Support: 0,8 Confidence: 85,2 Lift: 1,893

è supportata da 1 543 clienti. Il classificatore ottenuto ha una accuratezza del 68,61% su tutti i dati, e dell'85,94% sui dati classificati come positivi. Analizzando le regole di classificazione emerge che i clienti che supportano questa regola sono:

- Uomini/donne di 44-50 anni che lavorano nel privato e hanno studiato materie scientifiche (confidenza: 100%)
- Donne single di 34-35 anni laureate che lavorano per enti pubblici (confidenza: 100%)
- Donne sposate aventi 20-27 anni (confidenza: 96.5%)

Livello Reparti

La regola

[FRESCHI-CARNI BIANCHE] [FRESCHI-SURGELATI] [FRESCHI-GASTRONOMIA] → [FRESCHI-CARNI ROSSE]
Support: 5,2 Confidence: 75,5 Lift: 1,217

è supportata da 7 241 clienti. Il classificatore ottenuto ha una accuratezza del 58,57% su tutti i dati, e del 73,61% sui dati classificati come positivi.
La generalità della regola associativa fa si che modello generi molte (circa 200) regole di classificazione per i soli casi positivi, ognuna delle quali ha confidenza 100% e un numero di istanze che la supportano molto piccolo (5/6 clienti per ogni regola).
Questo fatto rende impossibile creare un profilo sensato dei clienti che supportano tale regola associativa.

10.2 Profili per Pattern Sequenziali

In questa sezione vengono discussi i risultati relativi ai pattern sequenziali trovati e i profili dei clienti corrispondenti.

Livello Articoli

Il pattern sequenziale

{ [PIZZAIOLA TRIS LOCAT.GR.125X3] } { [MOZZAR.S.LUCIA TRIS GR.125X3] }
Support: 0,010283

è supportato da 599 clienti. Il classificatore ottenuto ha una accuratezza del 73,23% su tutti i dati, e del 94,16% sui dati classificati come positivi. Il profilo risultante è il seguente:

- Artigiani aventi 29-30 anni (confidenza: 100%)

- Single maschi e laureati aventi 60-70 anni (confidenza: 97%)
- Donne laureate di 30-31 anni che lavorano per enti privati (confidenza: 96.5%)

Il pattern sequenziale

{ [IMPASTO PIZZA] } { [IMPASTO PIZZA] } { [IMPASTO PIZZA] }
Support: 0,007223

è supportato da 426 clienti. Il classificatore ottenuto ha una accuratezza del 74,85% su tutti i dati, e del 99,59% sui dati classificati come positivi. Il profilo risultante è il seguente:

- Donne di 29-54 anni appartenenti al clero (confidenza: 100%)
- Uomoini single di 39-40 anni che lavorano per enti pubblici (confidenza: 100%)
- Donne sposate con laurea ma disoccupate di 33-35 anni (confidenza: 100%)

Livello Subcategorie

Il pattern sequenziale

{ [AVICUNICOLO-POLLO] } { [BOVINO-VITELLONE] }
Support: 0,148801

è supportato da 8 650 clienti. Il classificatore ottenuto ha una accuratezza del 56,79% su tutti i dati, e del 73,06% sui dati classificati come positivi. Il profilo risultante è il seguente:

- Donne diplomate casalinghe nubili di 32-36 anni (confidenza: 100%)

- Uomini sposati di 72-76 anni che non lavorano (confidenza: 100%)
- Impiegati sposati di 32-53 anni (confidenza: 100%)

Il pattern sequenziale

{ [FORMAGGI A SERV.ASSISTITO-SEMIDURI/DURI] } { [FORMAGGI LIBERO SERVIZIO-FORMAGGI FRESCHI] }
Support: 0,100328

è supportato da 5 805 clienti. Il classificatore ottenuto ha una accuratezza del 53,58% su tutti i dati, e dell'80,83% sui dati classificati come positivi. Il profilo risultante è il seguente:

- Liberi professionisti sposati di 34-38 anni (confidenza: 100%)
- Impiegati di 47-50 (confidenza: 96%)

Livello Categorie, Reparti e Settori

A questi livelli di astrazione, i pattern sequenziali sono risultati troppo generici e non hanno permesso la creazione di un modello accurato di classificazione e quindi l'estrazione dei profili corrispondenti.

11 Deployment

11.1 Final Report

In questo progetto sono state analizzate le vendite di un ipermercato relative al primo trimestre del 2005 al fine di estrarre regole associative e pattern sequenziali che descrivono comportamenti di acquisto dei clienti. Sono stati inoltre determinati i profili dei clienti che supportano alcune delle regole associative/pattern sequenziali rituenute interessanti. Purtroppo non è stato

possibile effettuare l'estrazione dei profili a tutti i livelli di astrazione, per via della generalità delle regole e pattern sequenziali ottenuti, e quindi dell'incapacità del classificatore di discriminare in modo soddisfacente le tipologie di clienti.