SCOPERTA DI REGOLE DI ASSOCIAZIONE

ATTIVITA' DEFINITA IN ORIGINE PER "MARKET BASKET DATA" [Agrawal '96]

ESEMPIO

#RECORD	LATTE	CAFFE'	BIRRA	PANE	BURRO	RISO	FAGIOLI
1	No	Si	No	Si	Si	No	No
2	Si	No	Si	Si	Si	No	No
3	No	Si	No	Si	Si	No	No
4	Si	Si	No	Si	Si	No	No
5	No	No	Si	No	No	No	No
6	No	No	No	No	Si	No	No
7	No	No	No	Si	No	No	No
8	No	No	No	No	No	No	Si
9	No	No	No	No	No	Si	Si
10	No	No	No	No	No	Si	No

SI VOGLIONO SCOPRIRE REGOLE R DEL TIPO:

R : SE X ALLORA Y

X E Y INSIEMI DI PRODOTTI, $X \cap Y = \varnothing$

SUPPORTO(R) = NUMERO DI RECORD CON X E Y / NUMERO TOTALE DI RECORD

CONFIDENZA(R) = NUMERO DI RECORD CON X E Y / NUMERO DI RECORD CON X

ESEMPIO

#RECORD	LATTE	CAFFE’	BIRRA	PANE	BURRO	RISO	FAGIOLI
1	No	Si	No	Si	Si	No	No
2	Si	No	Si	Si	Si	No	No
3	No	Si	No	Si	Si	No	No
4	Si	Si	No	Si	Si	No	No
5	No	No	Si	No	No	No	No
6	No	No	No	No	Si	No	No
7	No	No	No	Si	No	No	No
8	No	No	No	No	No	No	Si
9	No	No	No	No	No	Si	Si
10	No	No	No	No	No	Si	No

REGOLE SCOPERTE CON SUPPORTO MINIMO 0.3 E CONFIDENZA MINIMA 0.8:

- SE CAFFE’ ALLORA PANE — SUPP. = 0.3, CONF. = 1
- SE CAFFE’ ALLORA BURRO — SUPP. = 0.3, CONF. = 1
- SE PANE ALLORA BURRO — SUPP. = 0.4, CONF. = 0.8
- SE BURRO ALLORA PANE — SUPP. = 0.4, CONF. = 0.8
- SE (CAFFE’ & PANE) ALLORA BURRO — SUPP. = 0.3, CONF. = 1
- SE (CAFFE’ & BURRO) ALLORA PANE — SUPP. = 0.3, CONF. = 1
- SE CAFFE’ ALLORA (BURRO & PANE) — SUPP. = 0.3, CONF. = 1

PROPRIETA’ DELLE REGOLE DI ASSOCIAZIONE

LE REGOLE SONO ACCURATE (> SOGLIA MINIMA)
SONO ANCHE INTERESSANTI? NON SEMPRE!

DUE ESEMPI REALI:

1) **SE TOVAGLIONI DI CARTA ALLORA BIRRA**

   REGOLA INTERESSANTE (MOLTO SORPRENDENTE) E UTILE (METTERE LA BIRRA SULLA SCANSIA DI FIANCO AI TOVAGLIOLINI HA PRODOTTO UN INCREMENTO DI VENDITE DELLA BIRRA)

2) **SE BARBIE ALLORA BARRETTA DI CIOCCOLATA**

   REGOLA “SPURIA”: LA CIOCCOLATA E’ POCO COSTOSA E FACILMENTE REPERIBILE ALLA CASSA, QUINDI E’ ASSOCIATA A NUMEROSI PRODOTTI DIVERSI

CLASSIFICAZIONE

OGNI RECORD CONSISTE DI

- UN ATTRIBUTO TARGET (CLASSE)
- UN INSIEME DI ATTRIBUTI PREDITTIVI (CARATTERISTICHE O FEATURES)

Insieme di Training

F1	::::::::::	Fn	CLASSE
			A
			B
			A
			B
			C
			C

Insieme di Test

F1	::::::::::	Fn	CLASSE
			?
			?
			?
			?
			?
			?

SI VOGLIONO DEFINIRE DELLE REGOLE CHE LEGANO (I VALORI DE) GLI ATTRIBUTI PREDITTIVI E (I VALORI DE) L'ATTRIBUTO TARGET

LE REGOLE SERVONO PER PREDIRE IL VALORE (IGNOTO) DELL'ATTRIBUTO TARGET DI NUOVI RECORD, DI CUI SONO NOTI I VALORI DEGLI ATTRIBUTI PREDITTIVI

ESEMPIO

SI VUOLE UNA REGOLA PER PREDIRE SE UN DATO CLIENTE COMPRERA' UN PRODOTTO, NOTI IL SUO SESSO, LA SUA ETA' E IL PAESE DI PROVENIENZA

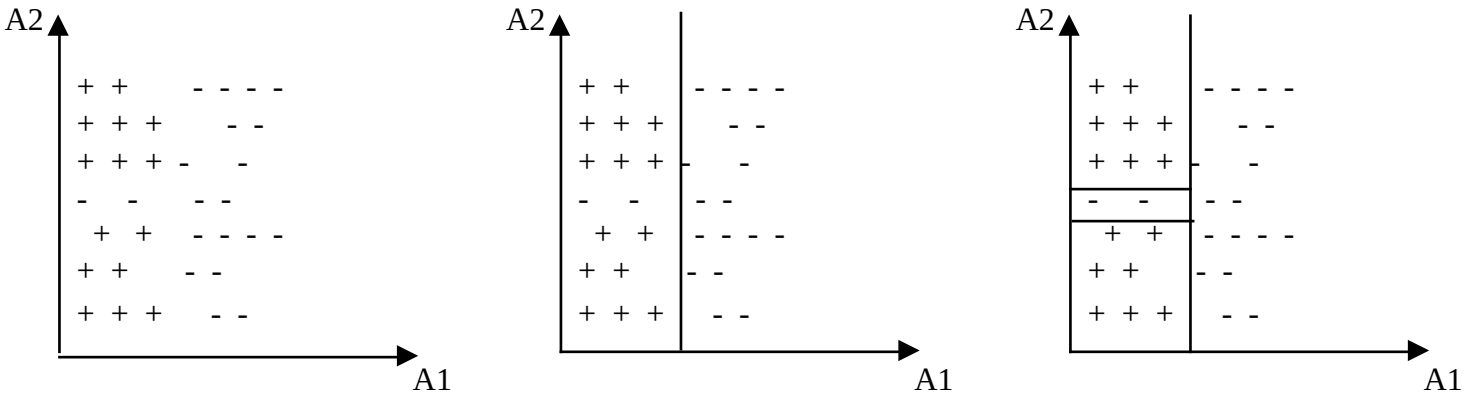
SESSO	**PAESE**	**ETA'**	**ACQUISTA(*)**
M	FRANCIA	25	Si
M	ITALIA	21	Si
F	FRANCIA	23	Si
F	ITALIA	34	Si
F	FRANCIA	30	No
M	GERMANIA	21	No
M	GERMANIA	20	No
F	GERMANIA	18	No
F	FRANCIA	34	No
M	FRANCIA	55	No

REGOLE CHE SI POSSONO INDURRE DAI DATI:

- SE (PAESE = GERMANIA) ALLORA (ACQUISTA = No)
- SE (PAESE = ITALIA) ALLORA (ACQUISTA = Si)
- (PAESE = FRANCIA & ETA' $\leq$ 25) ALLORA (ACQUISTA = Si)
- (PAESE = FRANCIA & ETA' $\geq$ 25) ALLORA (ACQUISTA = No)

PROPRIETA' DELLE REGOLE DI CLASSIFICAZIONE

ß PARTIZIONE DELLO SPAZIO DEI DATI IN REGIONI DI DECISIONE



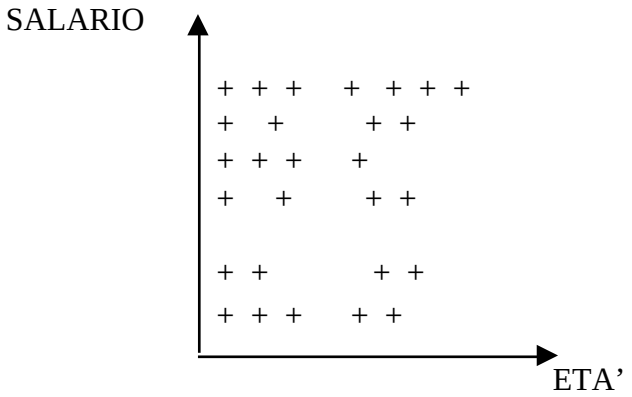
dati originali dati separati rispetto ad A1 dati separati rispetto ad A1 e A2

ß ACCURATEZZA E CAPACITA' DI GENERALIZZAZIONE

LA SECONDA REGOLA E' PIU' ACCURATA DELLA PRIMA (PERFETTA SEPARAZIONE) SUI DATI DI TRAINING.

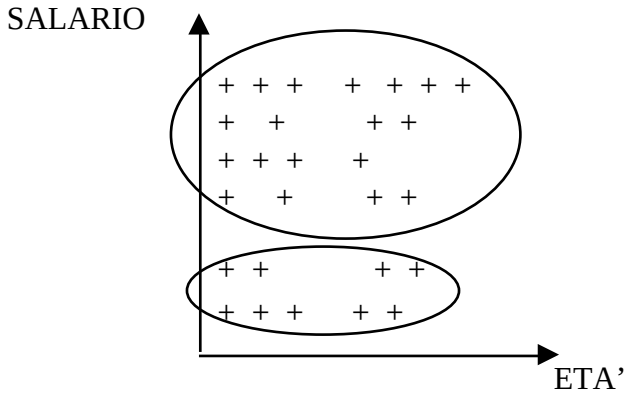
LA PRIMA PERO' POTREBBE RISULTARE MIGLIORE DELLA SECONDA SU DATI NON ANCORA VISTI

CLUSTERING



NON ESISTONO ATTRIBUTI TARGET

IL SISTEMA DEVE “SCOPRIRE” LE CLASSI, RAGGRUPPANDO RECORD SIMILI (= CON VALORI SIMILI DEGLI ATTRIBUTI) NELLA STESSA CLASSE



ESEMPIO

SI VUOLE CREARE UN INSIEME DI “PROFILI UTENTE” CHE DESCRIVANO LE TIPICITA’ DEI CLIENTI

SESSO	ETA’	SPESE VESTITI	SPESE CASA	SPESE LIBRI
M	25	3	0.2	0.07
M	21	1.5	2	0.07
F	23	0.5	1	0.03
F	34	2	1.3	0.1
F	30	0.7	3	0.06
M	21	1.5	1.9	0.2
M	20	2.1	0.1	0.3
F	18	2	3	0.01
F	34	2.3	3.1	0.15
M	55	1.2	3	0.05

COMPLESSIVAMENTE, IL CLIENTE MEDIO

- HA 41 ANNI
- SPENDE £1.680.000 IN VESTITI
- SPENDE £1.860.000 PER LA CASA
- SPENDE £104.000 IN LIBRI

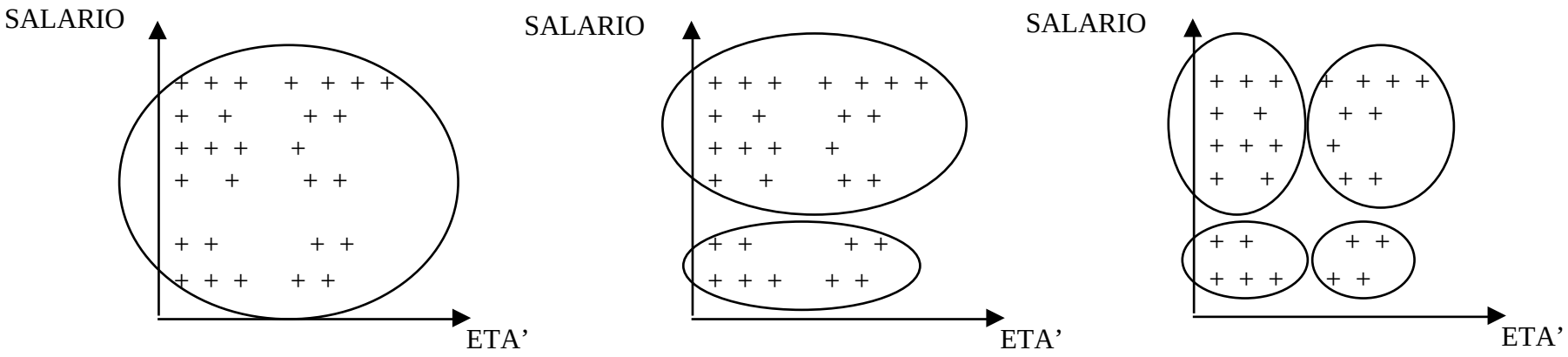
DOPO IL CLUSTERING, TROVIAMO DUE GRUPPI, IN CUI IL CLIENTE MEDIO

- HA 30 ANNI
- SPENDE £2.180.000 IN VESTITI
- SPENDE £1.320.000 PER LA CASA
- SPENDE £164.000 IN LIBRI

- HA 52 ANNI
- SPENDE £1.180.000 IN VESTITI
- SPENDE £2.400.000 PER LA CASA
- SPENDE £44.000 IN LIBRI

PROPRIETA’ DEL CLUSTERING

- E’ DIFFICILE VALUTARE LA QUALITA’ DI UNA CLUSTERIZZAZIONE (DIPENDENTE DAL DOMINIO)
-
- PROBLEMA: QUANTE CLASSI?



RELAZIONI CON IL MECCANISMO DI ASTRAZIONE