Dipartimento di Statistica
"Giuseppe Parenti"

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

WORKING PAPER 2004/06

Forme di aggregazione di dati e data-mining

Laura Grassini

Università degli Studi
di Firenze

Economic Statistics

FORME DI AGGREGAZIONE DI DATI E *DATA MINING*

Laura Grassini

Key words: Aggregazione, *summarization, smoothing.*

Abstract: This paper deals with the use of aggregated data in the analysis of very large data sets. This idea apparently contrasts with the proper nature of *data mining* which is mainly concerned with modelling and analysing data at micro level. However, aggregation, *smoothing* and *summarization* are approaches of data reduction that are widely used in *data mining*. After the definition of *aggregation, smoothing* and *summarization*, the paper is concerned with the problems and consequences connected with the use of aggregate data in two situations: (1) the computation of dependence measures(we considered the correlation index and a measure of monotonous dependence based on the mean difference); (2) the classification of units through linear discriminant analysis.

1 Introduzione

L'uso di dati aggregati è solo apparentemente in contrasto con la natura propria del *data mining*, che concerne essenzialmente la modellazione e l'analisi a livello di microdato. Nella pratica, infatti, aggregazione, *smoothing* e *summarization* (tradotto in italiano con *sommarizzazione*) sono metodi di riduzione dei dati che vengono frequentemente usati anche nel *data mining* [14].

Con riferimento a queste tematiche, dopo l'introduzione delle definizioni di aggregazione, *smoothing* e *summarization* (paragrafo 2), il lavoro esamina le conseguenze dell'impiego di dati aggregati su due analisi statistiche tipiche del *data mining*: (1) il calcolo di misure di dipendenza fra variabili; (2) la tecnica di classificazione di unità, basata sull'analisi discriminante lineare.

Il lavoro esamina, mediante un esercizio di simulazione, gli effetti dell'aggregazione dei dati sul il coefficiente di correlazione (paragrafo 3) e su una misura di relazione monotona fra le variabili, basata sulla differenza media (paragrafo 4). L'obiettivo di questa analisi è quello di valutare la sensibilità della misura di dipendenza al tipo di aggregazione, nell'ipotesi di corretta specificazione della microrelazione che lega le due variabili di interesse, variabili che sono assunte continue.

Per quanto concerne la classificazione di unità mediante la funzione discriminante (paragrafo 5), siamo interessati all'aggregazione di variabili nel senso della *summarization* e alle conseguenze sulla funzione discriminante lineare. L'analisi empirica opera su dati di bilancio di aziende dell'Italia del Nord-ovest e del Nord-est, e si propone di valutare la perdita di capacità discriminatoria con l'impiego di voci di conto aggregate (per altro di largo impiego nella pratica contabile), rispetto all'impiego di dati più dettagliati.

2 Aggregazione, *smoothing, summarization*

Nell'ambito del *data mining*, gli strumenti statistici vengono utilizzati per scoprire eventuali *pattern* nei dati, dati che hanno natura osservazionale e che sono caratterizzati da dimensionalità elevata [4]. In sintesi, gli aspetti tipici del data *data mining* sono [7]:

1. trattamento di grosse moli di dati;
2. uso intensivo di metodi di calcolo numerico (reti neurali, alberi decisionali, ecc.) e di algoritmi *time saving*;
3. interesse economico nella produzione di software specializzato per analizzare dati economico-aziendali;

4. rischio di una modesta qualità dei dati.

Nel *data mining*, l'analisi statistica è spesso caratterizzata da procedure automatiche di selezione, classificazione di unità e/o variabili e più raramente è basata sulla specificazione e stima di modelli con supervisione del ricercatore. L'approccio di ricerca è essenzialmente di tipo esplorativo, e la riduzione dei dati è un'operazione quasi sempre obbligata; essa si può operare rispetto alle unità, alle variabili, alle modalità o valori delle variabili [14].

Generalmente, col termine aggregazione si indica un metodo di riduzione del numero di unità statistiche. Si tratta di una funzione dei dati elementari che produce un dato aggregato o macrodato [10]. La costruzione di gruppi di unità (mediante metodi esplorativi o sulla base di ipotesi a priori) e il successivo calcolo dell'aggregato per i gruppi, produce nuove unità statistiche (i gruppi, per l'appunto) e i macrodati relativi (es. le medie, i totali di gruppo).

L'operazione di aggregazione può essere interpretata anche come una sorta di *data smoothing*. Infatti, se sostituiamo il valore di ogni unità con la media del gruppo di appartenenza, operiamo di fatto una approssimazione dei valori originari mediante un numero inferiore di valori o modalità.

Il termine *summarization*, infine, fa riferimento alla somma (o, più in generale, ad una combinazione lineare) di un certo numero di variabili relative alla singola unità statistica. Un esempio è rappresentato dalle poste contabili di un'impresa che sono una funzione lineare delle singole transazioni. Si noti che, talvolta, ([3], p. 217) questa operazione viene indicata col termine aggregazione.

Dopo queste brevi considerazioni sulla terminologia, veniamo ai temi affrontati nel presente lavoro, che si richiamano direttamente ai punti (1) e (2) del paragrafo introduttivo:

1. analisi dell'effetto dell' aggregazione/*smoothing* su misure di dipendenza fra variabili (paragrafi 3 e 4);
2. studio dell'effetto dell'operazione di *summarization* sull'analisi discriminante lineare quale metodo di classificazione delle unità (paragrafo 5).

3 Aggregazione e correlazione

3.1 Aspetti teorici

I primi studi degli effetti dell'aggregazione di unità sul legame di dipendenza fra due variabili, avevano prevalentemente interessi teorici ed erano concentrati sulla modellistica lineare [13]. In genere, si ipotizzava che il modello statistico (modello di regressione) fosse perfettamente specificato sui microdati e si ricavavano le conseguenze risultanti dall'aggregazione.

Fra i vari contributi, è senza dubbio da ricordare il lavoro di Grunfeld and Griliches del 1960 [6]. In particolare, i due autori dichiaravano che: (1) in pratica, non possediamo abbastanza informazioni sul comportamento a livello micro, per specificare in modo perfetto la microrelazione; (2) l'aggregazione di unità può ridurre gli errori di specificazione generando quello che può essere definito un *aggregation gain*. Ovviamente, se le microrelazioni sono mal specificate, ci potrà essere un significatuvo guadagno dall'aggregazione grazie all'eliminazione degli errori di specificazione ma anche una certa perdita dovuta all'aggregazione di differenti microrelazioni. In tutto ciò, la bassa qualità dei dati potrebbe essere proprio una fonte di *aggregation gain*.

Parlando di vantaggi che possono scaturire dall'aggregazione di unità, un ruolo fondamentale è giocato anche dal criterio di raggruppamento. Ad esempio, il raggruppamento delle famiglie in funzione del reddito percepito non è casuale rispetto a variabili economiche di interesse come la spesa in beni di consumo. In ogni gruppo, infatti, gli individui sono più omogenei rispetto a tale variabile che non rispetto ai disturbi accidentali [6]. Per questo motivo, tale raggruppamento tende ad esaltare il legame di dipendenza spesa-reddito.

Più recentemente, l'effetto dell'aggregazione sulla correlazione e su misure di dipendenza, è stato studiato per la messa a punto di metodologie volte alla garanzia della riservatezza nelle

indagini statistiche. Spesso i dati hanno un risoluzione disaggregata molto fine e, per limitare il rischio di *disclosure*, sono sottoposti ad una forma di aggregazione detta microaggregazione [5]. L'ottica della microaggregazione è quella di operare una sorta di *smoothing* dei dati, cercando di non distorcere troppo i legami di interdipendenza esistenti a livello di microrelazioni.

La filosofia della microaggregazione appare coerente con la quella del *data mining*, che si propone di lavorare a livello di microdato. Tuttavia la nostra ottica è parzialmente diversa in quanto possiamo anche ammettere che l'aggregazione modifichi i legami di dipendenza qualora siano presenti dati di bassa qualità.

La discussione sui vantaggi e sugli svantaggi derivanti dall'aggregazione di dati segue qui un'impostazione descrittiva piuttosto che inferenziale. Siamo interessati al potere esplicativo ottenuto tramite un modello e non agli errori nella stima dei parametri.

Si consideri il caso dell'aggregazione di N unità con riferimento a dati *cross section*. Si supponga di avere a disposizione due variabili e sia:

$$y_{ij} = \beta\, x_{ij} + e_{ij} \quad i = 1, 2, ..., K,\ j = 1, ..., n, \tag{1}$$

dove y_{ij} e x_{ij} sono misurate come scarti dalle rispettive medie, i indica il gruppo e j indica l'unità all'interno del gruppo. Per semplicità, consideriamo qui k gruppi (esaustivi e mutualmente escludentisi) della stessa numerosità . x_{ij} sono valori dati e e_{ij} la componente di disturbo tale che $E(e_{ij})=0$, $E(e_{ij}\ e_{sl})=\sigma^2$ se $i=s$ e $j=l$, $E(e_{ij}e_{sl})=0$ altrimenti.

A causa della struttura raggruppata dei dati, la devianza totale dei valori x_{ij} è scomposta nella somma di due componenti:

$$S_T^2 = S_W^2 + S_B^2 \tag{2}$$

dove, ricordando che i valori sono espressi come scarti dalla media generale, si ha:

$$S_T^2 = \sum_{ij} x_{ij}^2 \quad S_W^2 = \sum_{ij} (x_{ij} - \bar{x}_i)^2 \quad S_W^2 = \sum_{ij} \bar{x}_i^2 \tag{3}$$

Come si vede, S_T^2 è la devianza totale, S_W^2 è la devianza interna ai gruppi e S_B^2 è la devianza fra le medie di gruppo. La proporzione di varianza della variabile dipendente, spiegata dalla componente sistematica della (1) sarà, sui microdati:

$$\rho^2 = \frac{\beta^2\, S_T^2}{\beta^2\, S_T^2 + N\sigma^2} \tag{4}$$

La (4) esprime l'indice di determinazione lineare del modello teorico definito sui microdati.

Si consideri ora il modello che lega le medie di gruppo delle due variabili. Posto che valga la (1), si ha:

$$\bar{y}_i = \beta\, \bar{x}_i + \bar{e}_i \quad \bar{e}_i = \frac{1}{n} \sum_j e_{ij} \tag{5}$$

dove $E(\bar{e}_i) = 0,\ V(\bar{e}_i) = \sigma^2/n\ \ E(\bar{e}_i\ \bar{e}_s) = 0\ \ per\ \ i \neq s$

Sui dati aggregati (ovvero *smoothed*, se si considera che i valori originari x_{ij}, y_{ij} sono sostituiti con le medie di gruppo $\bar{y}_i\ \ e\ \ \bar{x}_i$) la proporzione della varianza della variabile dipendente, spiegata dalla parte sistematica del modello (5), è:

$$\rho_A^2 = \frac{\beta^2\, S_B^2}{\beta^2\, S_B^2 + N\sigma^2/n} = \frac{\beta^2\, S_B^2}{\beta^2\, S_B^2 + K\sigma^2} \tag{6}$$

dove $N=nK$ è il numero totale delle osservazioni.

Confrontando le espressioni (5) e (4), si ricava che:

$$\rho_A^2 \geq \rho^2 \ \ se \quad S_B^2 \geq S_T^2/n \tag{7}$$

Ponendo $\eta^2 = S_B^2/S_T^2$ si ha che, se il modello (1) è valido, l'aggregazione che soddisfa la relazione $\eta^2 > 1/n$ produce una correlazione più elevata di quella che si ottiene sui microdati. Si noti che l'espressione (7) rimane valida anche se i gruppi non hanno la medesima numerosità; in tal caso la (7) diventa:

$$\rho_A^2 \geq \rho^2 \quad se \quad S_B^2 \geq S_T^2/\bar{n} \tag{8}$$

dove $\bar{n}$ è la numerosità media dei gruppi.

Una aggregazione che verifica la condizione $\eta^2 > 1/n$, produce sui dati una correlazione più elevata. Si tratta di un tipo di aggregazione che tende a formare gruppi al loro interno omogenei in quanto nella (7)compare la devianza fra gruppi della variabile dipendente. Se esiste un legame fra x e y, l'aggregazione rispetto alla variabile x tenderà a produrre gruppi simili anche in termini dei valori di y [6]. In altre parole, una forma di aggregazione che non sia casuale rispetto alla variabile indipendente, determina una correlazione più elevata in confronto con quella ottenuta dai microdati.

Nel caso in cui la variabile indipendente è continua, può essere utile ricorrere ad un metodo ottimale di aggregazione. Poiché il coefficiente di correlazione è basato sul concetto di devianza, un criterio di raggruppamento particolarmente interessante è, ad esempio, quello che minimizza la devianza interna (ai gruppi) della variabile esplicativa x. Stabilito il numero K di gruppi, questo criterio conduce alla condizione [1]:

$$a_i = (\bar{x}_i + \bar{x}_{i+1})/2 \tag{9}$$

dove a_i è il limite superiore e $\bar{x}_i$ è la media della classe i-esima.

Da questa analisi vediamo quindi che, se la microequazione è perfettamente specificata, e si procede ad una aggregazione delle unità rispetto ai valori di x, è molto probabile che la macroequazione fornirà un più alto indice di determinazione lineare.

3.2 Risultati di un esercizio di simulazione

In questo paragrafo presentiamo i risultati di un esercizio di simulazione che ha preso in esame l'effetto dell'aggregazione di unità nell'ipotesi di modello lineare correttamente specificato a livello micro. Poiché nell'analisi di grosse moli di dati si scelgono spesso gruppi della medesima numerosità ([3],p.208), abbiamo optato per questa particolare struttura aggregativa secondo due diversi criteri di raggruppamento.

CRITERIO 1. I dati sono raggruppati in un prefissato numero di gruppi, rispetto a valori crescenti della variabile x. Ogni coppia dei microdati originari, viene sostituita (*smoothed*) con la rispettiva media di gruppo.

CRITERIO 2. I valori relativi alle due variabili sono aggregati separatamente in un numero prefissato di gruppi. Ne consegue che il singolo dato elementare sulla variabile x viene sostituito con la media del gruppo di appartenenza, che è definito rispetto a valori crescenti di x; il singolo dato elementare su y viene sostituito con la media del gruppo di appartenenza, che viene definito rispetto a valori crescenti di y.

I microdati sono stati ottenuti secondo due procedure:

1. generando 10000 valori di una variabile casuale normale bivariata con medie nulle, varianze unitarie e coefficiente di correlazione uguale a 0.4 e 0.3;

2. generando valori da variabili casuali χ_2^2 indipendenti, e poi trasformati ipotizzando una correlazione uguale a 0.4 e 0.3.

I risultati relativi al raggruppamento rispetto alla variabile x e a diverse numerosità di gruppi sono presentati nella Tabella 1. I risultati per il caso di dati disaggregati sono collocati nella prima riga della tabella.

Dalla Tabella 1, si può pertanto notare che l'effetto dell'aggregazione si fa sentire in modo rilevante già con *n=2*. La correlazione viene fortemente esaltata, tanto da raddoppiare se n=8. Si vede pertanto che, nell'ipotesi di corretta specificazione della microrelazione, questo tipo di aggregazione tende a modificare sensibilmente il valore della correlazione.

Tabella 1. Aggregazione rispetto a x (ipotesi di normalità)

Units per group	*Nr. groups*	r_{xy}=*0.3*	r_{xy}=*0.4*
1	10000	0.3072	0.3999
2	5000	0.4154	0.5234
4	2500	0.5411	0.6607
8	1250	0.6784	0.7843
10	1000	0.7243	0.8114
16	625	0.7965	0.8743
20	500	0.8371	0.8936
40	250	0.9105	0.9428
50	200	0.9238	0.9544
100	100	0.9583	0.9792

C'è da osservare che un sistema di aggregazione rispetto ad una determinata variabile può avere senso con due sole variabili sotto studio ma è meno utile in una situazione multivariata in cui si porrebbe il problema di scegliere la variabile classificatoria. Ci pare pertanto più interessante l'esame dei risultati della simulazione relativi al secondo criterio di raggruppamento, per il quale abbiamo anche considerato il caso di non normalità dei dati.

Come si può ricavare dalla Tabella 2, questo tipo di aggregazione non modifica l'associazione originale esistente fra le variabili, sia nel caso di normalità sia in quello di non normalità delle distribuzioni. E ciò accade soprattutto nel caso di gruppi di piccola dimensione, situazione questa più interessante per il *data mining*.

Tabella 2. Aggregazione separata di x e y

Unità	*Nr.*	*Normalità*		*Non normalità*	
per gruppo	*gruppi*	r_{xy}=0.3	r_{xy}=0.4	r_{xy}=0.3	r_{xy}=0.4
1	10000	0.3072	0.3999	0.3007	0.4044
2	5000	0.3072	0.3999	0.3010	0.4043
4	2500	0.3072	0.3998	0.3012	0.4046
8	1250	0.3071	0.3996	0.3020	0.4047
10	1000	0.3073	0.3998	0.3023	0.4045
16	625	0.3072	0.3995	0.3026	0.4051
20	500	0.3075	0.3999	0.3031	0.4055
40	250	0.3074	0.3995	0.3008	0.4040
50	200	0.3074	0.3991	0.2993	0.4029
100	100	0.3072	0.3991	0.2998	0.4025

Si deduce, quindi, che questo tipo di aggregazione appare preferibile al precedente, in quanto tende a preservare il valore 'vero' della correlazione nel caso in cui la microrelazione sia correttamente specificata.

4 Aggregazione e monotonicità del legame

In questo paragrafo, introduciamo una misura di monotonicità del legame di dipendenza di y da x e valutiamo l'impatto dell'aggregazione dei dati.

4.1 Definizioni e considerazioni teoriche

Si supponga che *(X,Y)* sia una variabile casuale non negativa, bivariata con media non nulla e finita *(E(X), E(Y))* e funzione di densità congiunta *f(x,y)*.

La codifferenza media di Y rispetto a X è definita come:

$$D_{y|x} = E_1E_2[(Y_1 - Y_2)\ \ sgn(X_1 - X_2)] \tag{10}$$

dove *sgn(w)* è 1 se $w > 0$, -1 se $w < 0$, 0 altrimenti; inoltre, (X_1, Y_1) e (X_2, Y_2) sono vettori casuali mutualmente indipendententi aventi la medesima distribuzione di *(X,Y)*. Un caso speciale di (10) è la differenza media di Y:

$$D_y = E_1E_2[(Y_1 - Y_2)\ \ sgn(Y_1 - Y_2)] \tag{11}$$

Analogamente si possono definire la codifferenza media $D_{x|y}$ di X rispetto a Y e la differenza media D_x di X. Questi indici di variabilità sono strettamente connessi con le curve di concentrazione (per dettagli cfr. [9]). In particolare:

$$G_x = \frac{D_x}{2E(X)} \qquad G_y = \frac{D_y}{2E(Y)} \tag{12}$$

sono, rispettivamente, l'indice di Gini per X e per Y. Inoltre, gli indici di concentrazione di X rispetto a Y e di Y rispetto a X (che costituiscono una generalizzazione dei corrispondenti indici del Gini) sono definiti come:

$$C_{x|y} = \frac{D_{x|y}}{2E(X)} \qquad C_{y|x} = \frac{D_{y|x}}{2E(Y)} \tag{13}$$

Da [9] e [11], deriviamo la seguente relazione:

$$-D_y \leq D_{y|x} \leq D_y \qquad -G_y \leq C_{y|x} \leq G_y \tag{14}$$

In particolare, se Y è costante, l'indice di concentrazione è zero; se $Y=kX$, dove k è una costante positiva, $D_{y|x}$ è uguale all'indice del Gini di X. In generale, se *Y=g(.)* e *g(X)* è una funzione crescente, X e *g(X)* (o Y) avranno esattamente lo stesso *ranking*. In tal caso, $D_{y|x}$ sarà uguale a D_y (e cioè, $C_{y|x}$ sarà uguale all'indice del Gini di Y). Analogamente accade per trasformazioni decrescenti.

Nel seguito, ci limiteremo a considerare situazioni di relazioni positive fra le variabili.

Da ([11],p. 76), se

$$E(Y|X = x) = a + bx \qquad E(X|Y = y) = a' + b'y \tag{15}$$

allora

$$b = \frac{D_{y|x}}{D_x} \qquad b' = \frac{D_{x|y}}{D_y} \qquad \rho^2_{xy} = \frac{D_{x|y}D_{y|x}}{D_xD_y} \tag{16}$$

Altri interessanti risultati emergono nel caso di normalità. Se *(X,Y)* è una normale bivariata con coefficiente di correlazione ρ_{xy} e deviazione standard σ_x, σ_y, si hanno le seguenti relazioni ([11],p. 71):

$$D_y = \frac{2\sigma_y}{\sqrt{\pi}} \qquad D_x = \frac{2\sigma_x}{\sqrt{\pi}} \qquad D_{y|x} = \frac{2\rho_{xy}\sigma_y}{\sqrt{\pi}} \qquad D_{x|y} = \frac{2\rho_{xy}\sigma_x}{\sqrt{\pi}} \tag{17}$$

Pertanto:

$$\frac{D_{y|x}}{D_y} = \frac{D_{x|y}}{D_x} = \rho_{xy} = R_{x|y} = R_{y|x} \tag{18}$$

Da questi risultati si deduce che il seguente indice può essere usato per valutare la forza della monotonicità del legame di dipendenza di Y da X

$$R_{y|x} = \frac{D_{y|x}}{D_y} = \frac{C_{y|x}}{G_y} \qquad -1 \leq R_{y|x} \leq 1. \tag{19}$$

4.2 Risultati di un esercizio di simulazione

Per analizzare il comportamento di $R_{y|x}$, abbiamo utilizzato gli stessi dati già elaborati nel paragrafo 3.2. I risultati sono presentati nella Tabella 3.

Possiamo osservare che, nel caso di microdati, i valori sono prossimi a quelli del coefficiente di correlazione nel caso normale (come atteso); tendono ad essere inferiori nel caso non normale. All'aumentare della dimensione dei gruppi, il valore di $R_{y|x}$ tende leggermente ad aumentare come ci si poteva aspettare, dato l'effetto perequativo dell'operazione di aggregazione.

Nel complesso, la misura della monotonicità del legame appare sensibilmente influenzata dalla forma della distribuzione, cosa del resto prevedibile, visto che l'indice tiene conto dell'ordinamento dei valori.

Tabella 3. Aggregazione separata e monotonicità

Unità	*Nr.*	*Normalità*		*Non normalità*	
per gruppo	*gruppi*	r_{xy}=0.3	r_{xy}=0.4	r_{xy}=0.3	r_{xy}=0.4
1	10000	0.3103	0.3980	0.2810	0.3619
2	5000	0.3103	0.3980	0.2811	0.3619
4	2500	0.3104	0.3981	0.2813	0.3621
8	1250	0.3107	0.3982	0.2818	0.3623
10	1000	0.3109	0.3984	0.2820	0.3623
16	625	0.3113	0.3986	0.2824	0.3626
20	500	0.3117	0.3989	0.2831	0.3630
40	250	0.3131	0.3998	0.2840	0.3636
50	200	0.3138	0.4001	0.2842	0.3636
100	100	0.3171	0.4025	0.2880	0.3656

5 *Summarization* e analisi discriminante lineare

5.1 Considerazioni teoriche

In questo paragrafo, esaminiamo l'impatto della *summarization* sulla capacità classificatoria di una funzione discriminante lineare. Si ipotizza di volere minimizzare il totale atteso degli errori di classificazione, secondo i principi dell'analisi discriminante. Il punto di riferimento è rappresentato da un vettore colonna x a p dimensioni, contenente i dati elementari originali; il vettore (colonna) dei dati sommarizzati è rappresentato da y, avente k elementi, dove $y{=}Ax$ e A è una matrice di k righe e p colonne, dove $k{<}p$.

In questo studio, la regola di *sommarizzazione* (e quindi la matrice A) si intende specificata esogenamente ai dati.

I principali benefici prodotti da una operazione di *sommarizzazione* sono: (1) rappresentazione più compatta dei dati, poiché $k{<}p$; (2) vantaggi se nei dati elementari sono presenti errori di misura; (3) produzione di nuova informazione (si pensi, ad esempio, ai saldi contabili o a voci contabili aggregate che comunemente vengono impiegate per l'interpretazione dei dati economico-finanziari di un'impresa).

Si supponga quindi di avere N osserazioni raggruppate in 2 gruppi che chiameremo G_1 e G_0. Il vettore colonna x_i di p elementi, osservato per l'unità i-esima, è considerato come la determinazione di una variabile casuale multinormale X dove:

$$X|G_0 \sim MN(\mu_0, \Sigma)$$

$$X|G_1 \sim MN(\mu_1, \Sigma)$$

e *MN* sta per *multinormale*; inoltre, μ_j è il vettore colonna delle p medie all'interno del gruppo *j (j=1,2)* e Σ è la matrice invertibile di varianza e covarianza, ipotizzata comune ai due gruppi.

Secondo la metodologia della discriminante lineare, la decisione classificatoria ottimale è quella che minimizza l'errore totale di classificazione rispetto alla verosimiglianza delle osservazioni [7]. Secondo questo criterio e sotto l'ipotesi di normalità, si ha quindi che:

se $(x_i - \mu_0)'\Sigma^{-1}(x_i - \mu_0) > (x_i - \mu_1)'\Sigma^{-1}(x_i - \mu_1)$

l'unità i-esima viene classificata in G_1, altrimenti in G_0.

Nell'ipotesi di multinormalità, il potere discriminante di questa regola classificatoria può essere espresso attraverso una funzione di distanza generalizzata fra i centroidi dei due gruppi. Tale distanza generalizzata è:

$d_x(G_0, G_1) = (\mu_0 - \mu_1)'\Sigma^{-1}(\mu_0 - \mu_1) = d_x'\Sigma^{-1}d_x$

dove $d_x = (\mu_0 - \mu_1)$. Quanto più grande è $d_x(G_0, G_1)$ tanto maggiore è il potere discriminante delle p variabili.

Si consideri ora l'operazione di *sommarizzazione* operata tramite la matrice A che produce il vettore $y_i = Ax_i$. Ovviamente, anche y_i è multinormale.

Ponendo

d_y=A $(\mu_0 - \mu_1)$=Ad_x

la distanza generalizzata fra i due gruppi, in termini del nuovo set di variabili, è:

$$d_y(G_0, G_1) = d_y'(A\Sigma A')^{-1}d_y = d_x'A'(A\Sigma A')^{-1}Ad_x \tag{20}$$

Per valutare quanta informazione utile ai fini dell'analisi discriminante viene preservata nel vettore y_i, si può usare il rapporto [2]:

$$R^2 = \frac{d_y(G_0, G_1)}{d_x(G_0, G_1)} = \frac{d_y'(A\Sigma A')^{-1}d_y}{d_x'\Sigma^{-1}d_x} \tag{21}$$

Si può dimostrare che la (21) assume valori compresi fra 0 e 1. Di conseguenza, c'è sempre la potenziale perdita di informazioni quando operiamo una trasformazione tipo Ax.

Per capire meglio quanto appena affermato, supponiamo che valga $\Sigma = I$ dove I è la matrice unitaria (o matrice identità) di dimensione p. Allora l'espressione (21) diventa:

$$\frac{d_y(G_0, G_1)}{d_x(G_0, G_1)} = \frac{d_y'(AA')^{-1}d_y}{d_x'd_x} \tag{22}$$

Immediatamente si può notare che la formula (22) è l'indice di determinazione lineare di un modello avente d_x come variabile dipendente e A come matrice delle covariate. Poiché l'indice di determinazione è inferiore o uguale a 1, la medesima formula mostra come ci sia sempre la potenziale perdita di informazioni quando operiamo una trasformazione del tipo Ax.

Quando l'espressione (22) è uguale a 1, non c'è alcuna perdita informativa dovuta alla trasformazione tramite A. E quindi, più grande è il valore della (22), minore informazione viene perduta con la *sommarizzazione* operata tramite la matrice A.

Questi risultati possono essere agevolmente estesi al caso di matrice non unitaria di covarianza Σ, mediante la diagonalizzazione di Σ che è ipotizzata essere invertibile. In tal caso, abbiamo che $\Sigma = \Gamma\Lambda\Gamma' = QQ'$ dove $Q = \Gamma\Lambda^{1/2}$ e $\Gamma'\Gamma = I$. Λ è la matrice diagonale dei p autovalori di Σ.

Definiamo quindi il vettore $z = Q^{-1}x$ che ha matrice di covarianza unitaria; si noti inoltre che $d_z = Q^{-1}d_x$. Infine, considerando che $y = Ax = AQz = Hz$ dove $H = AQ$, possiamo esprimere la (21) mediante le grandezze H e d_z anziché mediante A e d_x. Infatti abbiamo che:

$$\frac{d_y(G_0, G_1)}{d_x(G_0, G_1)} = \frac{d_y'(AQQ'A')^{-1}d_y}{d_x'(QQ')^{-1}d_x} = \frac{d_y'(HH')^{-1}d_y}{d_z'd_z} = \frac{d_y(G_0, G_1)}{d_z(G_0, G_1)} \tag{23}$$

Si noti come l'espressione (23), nella forma, sia analoga alla (22), dimostrando così che la (21) assume valori in[0,1].

5.2 Risultati di un'analisi di bilanci aziendali

Lo studio dell'effetto della *sommarizzazione* è stato condotto su dati di bilanci aziendali di 200 imprese operanti nel Nord-ovest (gruppo G_0) e 200 operanti nel Nord-est (gruppo G_1) d'Italia. I dati risalgono al 1997. Abbiamo condotto due tipi di analisi discriminante lineare: la prima sulle voci relative alla sezione attivo del conto patrimoniale; la seconda sulle voci della sezione passivo.

La Tabella 4 mostra le voci contabili relative alle attività. Le voci in grassetto rappresentano il risultato della *sommarizzazione*. La Tabella 5 riporta i risultati dell'analisi dettagliatamente per tipologia di aggregazione: abbiamo alternativamente aggregato le varie voci, partendo dalla situazione di completa disaggregazione (D,D,D) a quella di completa aggregazione dei dati nelle tre macrovoci (A,A,A). Per ogni situazione contemplata, abbiamo riportato il valore della grandezza (21) e la percentuale di corretta classificazione delle unità nei due gruppi, individuati dalle due ripartizioni territoriali del Nord Italia.

Tabella 4. Voci della sezione attivo

Cassa e Banche
Titoli e partecipazioni a breve
Liquidità immediate (LI)
Crediti commerciali a breve
Crediti a breve infra-gruppo
Crediti verso soci
Crediti diversi a breve
Ratei e risconti attivi
Liquidità differite (LD)
Magazzino materie prime
Magazzino prodotti in lav. e semilavorati
Lavori in corso su ordinazione
Magazzino prodotti finiti
Anticipi a fornitori
Disponibilità (DI)

Tabella 5. Sezione attivo

LI	LD	DI	Corretti %	R^2
D	D	D	59.0	1.000
D	D	A	56.0	0.871
D	A	D	57.3	0.746
D	A	A	55.5	0.632
A	D	D	58.3	0.970
A	D	A	54.8	0.857
A	A	D	57.3	0.746
A	A	A	56.0	0.631

D: disaggregata; A: aggregata

Le tabelle 6 e 7 riportano le voci della sezione passivo e i risultati relativi.

Tabella 6. Voci della sezione passivo

Debiti a breve verso banche
Obbligazioni a breve
Debiti a breve infra-gruppo
Debiti verso erario/enti previdenza
Fornitori, cambiali passive
Anticipi da clienti
Altre passività a breve
Fondo imposte e tasse
Utili da distribuire
Passivo corrente (PC)
Fondi per TFR
Debiti verso banche lungo termine
Obbligazioni a lungo termine
Debiti a lungo termine infra-gruppo
Debiti verso erario lungo termine
Fornitori e cambiali a lungo termine
Debiti diversi a lungo termine
Passivo consolidato (PML)
Capitale sociale
Riserve sovrapr. azioni
Riserve di rivalutazione
Riserva legale
Riserva per azioni proprie
Riserve statutarie
Altre riserve
- Perdite di esercizio
- Perdite esercizi precedenti
Utili d'esercizio a riserva
Utili esercizi precedenti
Capitale netto (CN)

Tabella 7. Sezione passivo

PC	PML	CN	Corretti %	R^2
D	D	D	61.8	1.000
D	D	A	62.3	0.659
D	A	D	63.3	0.739
D	A	A	60.5	0.376
A	D	D	62.3	0.676
A	D	A	57.3	0.351
A	A	D	63.0	0.527
A	A	A	58.3	0.241

D: disaggregata; A: aggregata

Nel complesso, nonostante la bassa capacità discriminatoria (vedi le percentuali di corretta classificazione) la *sommarizzazione* ha un impatto sensibile sui valori di R^2. Tuttavia, ad una diminuzione di R^2 non corrisponde necessariamente una perdita di capacità classificatoria (v. le percentuali di corretta classificazione), data la natura non normale dei dati di bilancio.

6 Conclusioni

I principali risultati ottenuti nelle analisi empiriche svolte sono i seguenti.

Riguardo alla aggregazione delle unità:

1. il raggruppamento separato delle due variaili sembra essere utile poichè nel caso di corretta specificazione della microrelazione non distorce significativamente la misura di associazione e di dipendenza fra le due variabili;
2. la misura di dipendenza monotona appare troppo influenzata dalla forma distributiva dei microdati tanto da distorcere il valore della dipendenza già sui microdati originali.

Per quanto concerne l'effetto della *sommarizzazione* dei dati elementari sulla capacitè discriminatoria e classificatoria dell'analisi discriminante lineare, si può dire che ciò dipende sensibilmente dalla forma distributiva dei microdati. Se questa è normale, la *sommarizzazione* tende a ridurre indubbiamente la capacità discriminatoria e anche quella classificatoria; nel caso di non normalità situazioni differenti possono verificarsi.

Bibliografia

[1] Aghevli B.B., F. Mehran (1981), *Optimal Grouping of Income Distribution Data*, Journal of the Americal Statistical Association, **373**, 22 – 26.

[2] Arya A., J.Fellingham, D.Schroeder (2000) *Accounting Information, Aggregation and Discriminant Analysis*, Management Science, **6**, 790 – 806.

[3] Berry M.J.A., G. S.Linoff (2001), *Data mining*, ed. Apogeo, Milano.

[4] Fayyad U.M., G. Pyatetsky-Shapiro, P.Smyth, R.Uthurusamy (1996), *Advances in Knowledge Discovery and Data Mining*, AAAI Press-The MIT Press.

[5] Franconi L., J. Stander (2002), *A model based method for disclosure limitation of business microdata*, Journal of the Royal Statistical Society, Series D, 51

[6] Grunfeld Y., Griliches Z. (1960), *Is Aggregation Necessarily Bad?*, The Review of Economics and Statistics, **XLII**, 1-13.

[7] Hand D. (1981) *Classification and Discrimination*, John Wiley and Sons.

[8] Hand D. (2000), *Methodological Issues in Data Mining*, Compstat 2000: Proceedings in Computational Statistics, Physica Verlag, 77-85.

[9] Kakwani N. C. (1980), *Income Inequality and Poverty*, Oxford University Press.

[10] SIS (1991), *Glossario dei principali termini su "La qualità dei dati statistici"*, a cura di A. Giommi, Bollettino n. 22, Società Italiana di Statistica (SIS), Roma.

[11] Taguchi T. (1981), *On a Multiple Gini's Coefficient and Some Concentrative Regressions*, Metron, **1**, 307-28.

[12] Taguchi T. (1988), *On the Structure of Multivariate Concentration. Some Relationships among Concentration Surface and Two Variate Mean Difference Regressions*, Computational Statistics and Data Analysis, **4**, 307-334.

[13] Theil H. (1957), *Specification Errors and the Estimation of Economic Relationships*, Review of the International Statistical Institute, 41-51.

[14] Weiss S.M., N. Indurkhya (1998), *Predictive Data Mining: a Practical Guide*, Morgan Kauffman Publishers.