***S**tatistica & **A**pplicazioni*
Vol. III, n. 1, 2005
pp. 59-79

Criteri ottimali e "quasi" ottimali di classificazione per caratteri continui nel caso della varianza fra i gruppi

Alessandro Zini[§]

Summary: *This work presents a continuous characters classification tecnique, which satisfies some optimality criterion, such as the maximisation of the variance "between" or the minimisation of the concentration "within" the groups, measured by some index which admits a decomposition. Such a tecnique is based on an algorithm which gives all the possible frequency distributions with fixed sum, the way that each class is not empty. As the algorithm is exponential, it is necessary to impose the minimum number of statistical units per class, which fact implies the construction of two classes of residual statistical units, to the left and to the right. In order to have indications about the optimal number of classes, the minimum number of statistical units per class, and about the imputation of the residual statistical units to the left and the right class, the tecnique will be applied to classify a 13868 individual income sample, drawn by Italian Bank in 1991.*

Keywords: *Exponential algorithm, Concentration within groups, Continous character, Optimality criterion, Variance between groups.*

1. Introduzione e presentazione del problema

Il problema di cui ci si occupa in questo lavoro è la classificazione di N intensità di un carattere continuo in un numero prefissato di gruppi k (non vuoti), in modo che venga ottimizzata una funzione di circostanza $V(x_1,\cdots,x_N)$. Si consideri il problema della massimizzazione della varianza fra i gruppi, oppure quello della minimizzazione della concentrazione nei gruppi, misurata da un qualche indice, ad esempio, l'indice di Herfindhal

[§] Dipartimento di Metodi Quantitativi per le Scienze Economiche ed Aziendali – Università degli Studi di Milano-Bicocca – Piazza dell'Ateneo Nuovo, 1, 20126 MILANO (e-mail: alessandro.zini@unimib.it).

(che ammette scomposizione in concentrazione "nei" e "fra" gruppi). Poiché la funzione *V* dipende da tutte le osservazioni, non è possibile in generale rintracciare criteri che permettano di identificare la soluzione ottimale in "pochi passi". Infatti, ai fini della soluzione del problema è necessario calcolare *V* per ogni distribuzione di frequenze su *k* gruppi (non vuoti). Il numero di tali distribuzioni è pari a $\binom{N-1}{k-1}$, numero che cresce esponenzialmente rispetto a *k*. Qualora *k* sia piccolo, è possibile, nonostante l'esponenzialità dell'algoritmo, pervenire alla soluzione ottimale, secondo la seguente

DEFINIZIONE 1 *(Criterio ottimale di classificazione)*
Si consideri una distribuzione per unità di un carattere quantitativo continuo ($x_1, x_2, ..., x_N$), dove $x_1 \le x_2 \le ... \le x_N$.

Sia $$A_N \equiv \left\{ (n_1, n_2, \cdots, n_k) : 1 \le n_1 \le n_2 \le \cdots \le n_k, \ \sum_{j=1}^{k} n_j = N \right\},$$
l'insieme di tutte le distribuzioni a frequenze ordinate in senso non decrescente. Risulta $card\ A_N \equiv |A_N| = \binom{N-1}{k-1}$.

Si definisca
$b_j \equiv \{x_{N_{j-1}+1}, x_{N_{j-1}+2}, \cdots, x_{N_j}\} \quad j = 1,2,\cdots,k$

$N_j \equiv \sum_{i=1}^{j} n_i, \quad j = 1,2,\cdots,k$ e $N_0 \equiv 0$

Sia $b \equiv (b_1, b_2, \cdots, b_k)$.
Dicesi *criterio ottimale di classificazione* ogni procedimento che permette di risolvere il seguente problema:

$$\max_{b \in A_N} V(b),$$

ove $V(b)$ è una misura di variabilità, di concentrazione o di altra natura come, per esempio, la varianza fra i gruppi.

◆

L'algoritmo efficiente che costruisce l'insieme *A* è esponenziale rispetto a *k.*[1]. Pertanto è necessario introdurre soluzioni sub-ottimali. A tal fine si introduce la seguente

DEFINIZIONE 2 *(criterio "quasi" ottimale di classificazione)*
Si consideri una distribuzione per unità di un carattere continuo $(x_1, x_2, ..., x_N)$, dove $x_1 \le x_2 \le ... \le x_N$.

Si considerino w gruppi di numerosità $n = \lfloor N/w \rfloor$ unità statistiche, dove $\lfloor \cdot \rfloor$ indica la parte intera, e si distribuiscano le rimanenti $d = N - w \cdot \lfloor N/w \rfloor$ unità statistiche in due classi denominate *residue*, sinistra e destra rispettivamente di numerosità d_1 e $d_2 (= d - d_1)$ Sotto tali condizioni.

$$\max_{c \in A_w} V(c), \qquad (1)$$

ove $c_j \equiv \{x_{w \cdot N_{j-1}+1}, x_{w \cdot N_{j-1}+2}, \cdots, x_{w \cdot N_j}\}$, $j = 1,2,\cdots,k$; $N_0 \equiv d_1$ e $c \equiv (c_1, c_2, \cdots, c_k)$. Ogni soluzione $S(w, d_1, d_2)$ del problema (1) si dice *criterio quasi ottimale di classificazione*. ◆

Dalla definizione 2 risulta chiaro che verrà scelta la soluzione $S^*(w, d_1, d_2)$ che, compatibilmente con il tempo macchina, al variare di (w, d_1, d_2) più si avvicina al valore massimo di *V*.

Nell'applicazione seguente, si considererà l'obiettivo della massimizzazione della varianza fra i gruppi.

2. I risultati di un'applicazione

In questo paragrafo si presenterà la parte applicativa, che affianca la parte teorica illustrata nei due paragrafi precedenti.
Si procede ad un'analisi dei risultati ottenuti, relativi al calcolo della varianza, confrontandoli tra loro attraverso tabelle riepilogative che mostrano come si distribuisce l'insieme dei dati[2] nelle k classi e nelle due classi residue.

---

[1] Ad esempio, con un processore di 2,5 GHz con $N = 100$ e $k = 8$ occorrono 12 ore-macchina, mentre con $N = 100$ e $k = 9$ occorreranno circa $12 \cdot 9$ ore-macchina, con $N = 100$ e $k = 10$ occorreranno circa $12 \cdot 9 \cdot 10$ ore-macchina.
[2] D'ora in avanti con il termine "residui" si indicheranno i "resti", così come sono stati definiti nei paragrafi 1 e 2.

I raggruppamenti effettuati sono stati creati in base ai parametri che si possono fissare a discrezione[3], cioè il numero di dati per gruppo (*n*) il numero di classi (*k*) in cui suddividere la distribuzione delle osservazioni, ed infine il numero di gruppi su cui l'algoritmo effettua tutte le permutazioni (*w*).
Dopo aver fissato i valori di queste variabili si è proceduto alla suddivisione degli eventuali residui all'interno delle due code, sinistra e destra, che si aggiungono, come avevamo già visto, al numero di classi (*k*) in cui suddividere i dati.
L'insieme di dati, su cui si sono basate le 265 prove effettuate, è composto da 13868 dati sui redditi individuali, che fanno riferimento ad un'indagine svolta dalla Banca d'Italia sulle famiglie dell'anno 1991.
La media totale è pari a 21.855,172 migliaia di lire, mentre la varianza totale è pari a 261.595.952 migliaia di lire al quadrato, con scarto quadratico medio pari a 16.173,928.
Le osservazioni sono state suddivise in k classi, dove k = 4, 5, 6, 7.
I vari tipi di suddivisioni effettuate sulla distribuzione dei dati sono quindi state di volta in volta ripetute, a parità di *n* e *w*, facendo variare il valore *k*.

2.1 Scelta di n (e di w)

Si può precisare che le 24 prove effettuate su k = 7 sono date proprio da queste coppie di valori, che possono essere riassunte per comodità in un'altra tabella.
Per quanto riguarda i risultati della varianza fra i gruppi ottenuti da tutte queste prove si rimanda alla sezione successiva 2.2, in cui si analizzerà come i valori della varianza fra i gruppi variano al variare di *k*, *n*, *w* e soprattutto come variano in base ad una diversa ripartizione dei residui nelle due code.

Tabella 1

Numero prove effettuate *	unità per gruppo (n) Numero gruppi (w) **	Residui Totali	residui a Sinistra	residui a Destra
3	n = 300 w = 46	68	30 48 50	38 20 18
3	n = 209 w = 66	74	30 48 50	44 26 24
3	n= 206 w = 67	66	30 48 50	36 18 16
3	n = 203 w = 68	64	30 48 50	34 16 14

[3]Compatibilmente con il numero di gruppi sui quali la macchina riesce a lavorare. Si ricorda che l'algoritmo è esponenziale.

3	n = 200 w = 69	68	30 48 50	38 28 18
3	n = 175 w = 79	43	20 25 30	23 18 13
3	n = 150 w = 92	68	30 48 50	38 28 18
3	n = 138 w = 100	68	30 48 50	38 28 18

* k =4, 5, 6, 7
** w = N/n ; N = 13868

Si analizzano ora i risultati della varianza fra i gruppi, si osserva in quali casi questa varianza si avvicina maggiormente alla varianza totale della distribuzione.
Il programma individua le 10 distribuzioni con varianza fra i gruppi più elevata.
Si riportano solo alcuni dei risultati (delle 265 prove effettuate), per mostrare come il numero delle classi *k* influisce sul valore della varianza fra i gruppi.
La Tabella 3 riporta i valori della varianza fra i gruppi per alcune coppie di *n* e *w*. Si riportano per maggiore immediatezza anche i valori percentuali, che in formula sono pari a: $\frac{\max Var_{FRA}}{Var_{TOT}} \cdot 100$.

Tabella 2

Unità per gruppo (n) Numero gruppi (w)	k = 4	K = 5	k = 6	k = 7
n = 1000 w = 13	213.796.176	217.069.568	218.554.656	
Valori percent.	81,727	82,798	83,546	
n = 575 w = 24	227.535.392	234.256.288	237.265.200	
Valori percent.	86,979	89,548	90,699	
n = 300 w = 46	227.855.712	236.795.696	240.749.776	243.168.064
Valori percent.	87,102	90,866	92,031	92,955
n = 283 w = 49	206.289.392	215.866.944	220.169.312	
Valori percent.	78,858	82,519	84,163	
n = 253 w = 59	215.705.712	225.793.632	227.260.624	
Valori percent.	82,457	86,313	86,874	
n = 212 w = 65	227.005.744	231.204.480	239.751.456	
Valori percent.	86,777	88,382	91,649	

n = 209 w = 66	227.471.696	235.888.568	240.468.944	243.002.304
Valori percent.	86,955	90,172	91,923	92,892

n = 206 w = 67	228.279.296	237.110.832	241.813.184	244.372.976
Valori percent.	87,264	90,640	92,437	93,416

n = 203 w = 68	228.364.768	237.332.336	242.064.848	244.648.192
Valori percent.	87,296	90,724	92,533	93,521

n = 200 w = 69	228.193.920	236.852.112	241.560.704	244.159.664
Valori percent.	87,231	90,541	92,341	93,334

Si può subito osservare che la varianza fra i gruppi aumenta sempre al crescere del numero di classi *k* in cui ripartire i dati[4].
In tutti i casi esaminati, la varianza fra i gruppi aumenta all'aumentare del numero delle classi.
In particolare, ad ogni k vi è un aumento della varianza che è mediamente dell'ordine di 2 punti percentuali. Ovviamente in base ai raggruppamenti scelti i valori cambiano molto, ma le "distanze" tra una classe e l'altra presentano delle regolarità.
Aumentando quindi il numero delle classi da 4 a 7 si ha un aumento della varianza fra i gruppi che è spesso intorno ai 6 punti percentuali.
Nella tabella 3 i risultati sono ordinati, all'interno di ogni classe *k*, secondo l'ordine decrescente della varianza fra i gruppi, avendo considerato solo quelle coppie di *n* e *w* che presentavano una varianza maggiore rispetto a quella di altri raggruppamenti.
Bisogna osservare che in questa analisi non si è prestata attenzione alla ripartizione dei resti nelle code.
Si è scelta una ripartizione dei resti casuale (all'incirca metà dei valori residui a sinistra e metà a destra) per k = 4 ed in seguito questa ripartizione è stata mantenuta anche per k = 5, 6, 7, altrimenti non sarebbe stato possibile confrontare i risultati.
Si riportano comunque i valori dei residui totali per ogni prova.
I risultati sono stati ordinati secondo il numero di gruppi *w*, ovvero seguendo l'ordine crescente dei gruppi (tabella 3).
In questo modo si può evidenziare meglio quale sia l'andamento della varianza, in corrispondenza di quale *w* aumenta oppure diminuisce, e se questo andamento è uguale in ogni classe *k*.

[4] Nella tabella 3 e nelle seguenti si è riportato solamente il primo valore della classifica, che è quello che, una volta fissati *k* e *w*, si avvicina maggiormente al valore della varianza totale.

Tabella 3

Unità per gruppo (n) Numero gruppi (w)	k = 5	k = 6	k = 7	residui totali
n = 300 w = 46	236.795.696	240.749.776	243.168.064	68
Valori percent.	90,519	92,031	92,955	
n = 206 w = 67	237.110.832	241.813.184	244.372.976	66
Valori percent.	90,640	92,437	93,416	
n = 203 w = 68	237.332.336	242.064.848	244.648.192	64
Valori percent.	90,724	92,533	93,521	
n = 200 w = 69	239.454.096	244.355.184	246.979.536	68
Valori percent.	91,535	93,409	94,412	
n = 175 w = 79	238.943.712	244.036.352	246.821.152	43
Valori percent.	91,340	93,287	93,352	
n = 150 w = 92	236.817.536	241.685.488		68
Valori percent.	90,527	92,388		
n = 138 w = 100	236.918.096	241.752.864		68
Valori percent.	90,566	92,414		

Da tabella 3, la varianza fra i gruppi aumenta gradualmente dal caso in cui il numero di gruppi è pari a 46 fino a 69, ma poi comincia a diminuire nuovamente fino a *w* pari a 100. Si noti però che i valori della varianza in questo ultimo gruppo sembrano aumentare nuovamente, anche se di poco (i valori in corrispondenza di *w* pari a 100 sono comunque inferiori al valore massimo che si ottiene, in corrispondenza di un numero di gruppi pari a 69).
La varianza quindi tende ad aumentare col numero di gruppi su cui si effettuano le permutazioni, ma in tutte e tre le classi l’aumento arriva solo fino ad un certo limite, dopo il quale la varianza torna nuovamente a diminuire.
Queste considerazioni sono però relative ad una ripartizione dei resti casuale; è opportuno, quindi, osservare come i risultati della varianza “fra” cambiano in relazione al criterio di ripartizione dei residui.
Per chiarire meglio si riporta una tabella (tabella 4) contenente i valori della varianza fra i gruppi, ordinati in senso decrescente, nel caso in cui i valori residui inseriti nella coda sinistra siano sempre pari a 30 osservazioni. In questa tabella si mostra qual è la classifica ordinata della varianza fra i gruppi per *k* = 6 e 7.

Si noti che, pur avendo un valore massimo della varianza in corrispondenza di un numero di gruppi *w* diverso nell'uno e nell'altro caso, per quanto riguarda le "posizioni" centrali, esse sono esattamente le stesse sia per $k = 6$ che per $k = 7$.

Tabella 4

	k = 6	residui tot.
n = 175 w = 79	243.393.632	43
Valori perc.	93,041	
n = 203 w = 68	242.064.848	64
Valori perc.	92,533	
n = 206 w = 67	241.813.184	66
Valori perc.	92,437	
n = 150 w = 92	241.752.912	68
Valori perc.	92,414	
n = 138 w = 100	241.685.488	68
Valori perc.	92,388	
n = 200 w = 69	241.560.704	68
Valori perc.	92,341	

	k = 7	residui tot.
n = 175 w = 79	246.224.464	43
valori perc.	94,123	
n = 203 w = 68	244.648.192	64
valori perc.	93,521	
n = 206 w = 67	244.372.976	66
valori perc.	93,416	
n = 200 w = 69	244.159.664	68
valori perc.	93,334	

	k = 6	residui tot.
n = 300 w = 46	240.749.776	68
Valori perc.	92,031	
n = 209 w= 66	240.468.944	74
Valori perc.	91,923	

	k = 7	residui tot.
n = 300 w = 46	243.168.064	68
valori perc.	92,955	
n = 209 w = 66	243.002.304	74
valori perc.	92,892	

Si può vedere che il valore massimo che la varianza fra i gruppi assume si verifica in corrispondenza di un numero di gruppi pari a 79 per $k = 6, 7$. Lo stesso si verifica anche per il secondo e il terzo valore immediatamente più elevati; in tutti e due i casi $k = 6, 7$ la varianza ha un andamento praticamente uguale: il valore più grande si ha con $w = 79$ e $w = 68$, mentre i valori più bassi si hanno quando $w = 66$.
Se si pone attenzione alla ripartizione dei resti nelle code, si nota che nel caso in cui la varianza assume il valore più elevato ($n = 175$) il numero di

residui totali è pari a 43, ed essendo 30 fra questi inseriti nella coda sinistra ne rimangono 13 a destra.
Si può, per il momento, osservare che quando si fissano 13 valori a destra (si inseriscono nella coda destra gli ultimi 13 valori) si ha sempre un valore elevato della varianza fra i gruppi.
Per comprendere meglio come la varianza assume valori differenti al variare della ripartizione dei resti nelle code, si possono riconsiderare esattamente le stesse prove viste prima, ma con i residui collocati diversamente nelle due code.
Ecco i risultati per k = 6, 7, con 48 residui a sinistra.

Tabella 5

	k = 6	residui tot.
n = 150 w = 92	244.900.912	68
Valori perc.	93,618	
n = 138 w = 100	244.899.568	68
Valori perc.	93,617	
n = 200 w = 69	244.402.816	68
Valori perc.	93,427	
n = 206 w = 67	244.252.848	66
Valori perc.	93,370	
n = 203 w = 68	244.013.872	64
Valori perc.	93,278	

	k = 7	residui tot.
n = 150 w = 92		68
valori perc.		
n = 138 w = 100		68
valori perc.		
n = 200 w = 69	247.018.336	68
valori perc.	94,427	
n = 206 w = 67	246.841.200	66
valori perc.	94,359	
n = 203 w = 68	246.628.320	64
valori perc.	94,278	

	k = 6	residui tot.
n = 209 w = 66	243.269.568	74
Valori perc.	92,994	
n = 300 w = 46	242.694.048	68
Valori perc.	92,774	

	k = 7	residui tot.
n = 209 w = 66	245.806.048	74
valori perc.	93,964	
n = 300 w = 46	245.166.928	68
valori perc.	93,719	

La varianza ha quindi uno stesso andamento sia per un numero di classi pari a 6 che per un numero di classi pari a 7.
Se si confrontano i risultati con quelli della tabella dove il numero dei residui a sinistra è pari a 30, si nota, ancora, che la varianza in questo caso è più grande, e che i valori migliori si hanno quando il numero dei gruppi è elevato (w = 92 o 100).

Si riportano ora le stesse prove cambiando ancora una volta il numero dei residui all'interno della coda sinistra.
Il numero dei residui a sinistra si fissa pari a 50.

Tabella 6

	k = 6	residui tot.
n = 138 w = 100	244.950.976	68
Valori perc.	93,637	
n = 150 w = 92	244.923.968	68
Valori perc.	93,626	
n = 200 w = 69	244.355.184	68
Valori perc.	93,409	
n = 206 w = 67	243.968.656	66
Valori perc.	93,261	
n = 209 w = 66	243.591.792	74
Valori perc.	93,117	
n = 203 w = 68	243.368.592	64
Valori perc.	93,032	
n = 300 w = 46	242.479.872	68
Valori perc.	92,692	

	k = 7	residui tot.
n = 138 w = 100		68
valori perc.		
n = 150 w = 92		68
valori perc.		
n = 200 w = 69	246.979.536	68
valori perc.	94,412	
n = 206 w = 67	246.565.056	66
valori perc.	94,254	
n = 209 w = 66	246.134.976	74
valori perc.	94,089	
n = 203 w = 68	245.992.272	64
valori perc.	94,035	
n = 300 w = 46	244.952.832	68
valori perc.	93,637	

Anche con questa ripartizione dei residui si ha una varianza fra i gruppi piuttosto elevata ed entrambe le classi (6 e 7) hanno uno stesso andamento della varianza. Se si analizzano nuovamente le unità a destra si vede che in questo caso i valori ottimali (cioè i raggruppamenti con varianza più elevata) si hanno quando i residui a destra sono pari a 18 (dato che il numero di residui totali per n = 150 e 138 è di 68 osservazioni, tolte 50 a sinistra ne restano 18 nella coda destra).
Considerando le tabelle relative a 48 e 50 residui a sinistra, si vede che i risultati migliori si hanno quando i residui a destra sono tra le 18 e le 20 unità.

Quando k = 5, 6, o 7 il *massimo* valore della varianza si ha quando le unità a destra sono pari a 18, per k = 4 invece si ha quando le unità residue a destra sono pari a 23.
Il risultato, invece, per cui le unità residue erano 30 a sinistra, e di conseguenza solo 13 nell’altra coda, risulta essere in qualsiasi classe il valore meno significativo, cio’ significa che anche per la coppia n = 175 e w = 79 si hanno valori migliori della varianza quando i residui a destra sono compresi tra 18 e 23 valori.
Per verificare se effettivamente il numero di residui a destra può essere così significativo, per il raggiungimento di un valore della varianza fra i gruppi ancora più elevato, si sono eseguite delle prove dove il numero di residui inseriti a sinistra è intorno alle 48 unità.
Si sono considerate diverse coppie di “numero di gruppi” e “numero di osservazioni per gruppo”, per k = 4, 5, 6. Si riporta sola la tabella riferita al caso k = 6.

Tabella 7

	K = 6	valori percent.	residui sin.	residui des.	residui tot.
n = 200 w = 69	244.029.824	93,285	46	22	68
n = 200 w = 69	244.245.664	93,367	47	21	68
n = 200 w = 69	244.373.120	93,416	49	19	68
n = 150 w = 92	244.481.360	93,457	46	22	68
n = 150 w = 92	244.717.904	93,548	47	21	68
n = 150 w = 92	244.906.016	93,619	49	19	68
n= 138 w = 100	244.481.360	93,457	47	21	68

I valori con 48 unità a sinistra erano pari a:

	k = 6	valori percent.	residui sin.	residui des.	residui tot.
n = 200 w = 69	244.402.816	93,427	48	20	68
n = 150 w = 92	244.900.912	93,618	48	20	68
n = 138 w = 100	244.899.568	93,617	48	20	68

In questo caso i valori della varianza si abbassano, fatta eccezione per n = 150 e w = 92 con 49 residui a sinistra.
Dal raffronto dei risultati della tabella con quelli dei casi k = 4, 5 (non riportati), si può concludere che *i valori della varianza fra i gruppi che tendono ad avvicinarsi il più possibile alla varianza totale della distribuzione si hanno in corrispondenza di un numero di residui nella coda destra che è tra le 18 e 20 unità, qualunque sia la coppia di n e w che si è scelta.*

2.2 Numero di unità residue fisse a destra

Sembra dall'analisi dei risultati che la coda destra assuma una notevole importanza. Pertanto, si riportano i risultati di alcune prove in cui si sono fissate il numero di unità da inserire nella coda destra, a parità di n e w.
La tabella 8 riporta la classifica ordinata della varianza nel caso in cui le unità a destra siano fisse a 20.

Tabella 8

	K = 5	residui tot.	residui sin.
n = 175 w = 79	239.524.480	43	23
valori perc.	91,562		
n = 200 w = 69	239.521.872	68	48
valori perc.	91,561		
n = 150 w = 92	239.492.960	68	48
valori perc.	91,550		
n = 209 w = 66	239.488.048	74	54
valori perc.	91,548		
n = 203 w = 68	239.483.872	64	44
valori perc.	91,547		
n = 206 w = 67	239.468.608	66	46
valori perc.	91,541		
n = 300 w = 46	238.667.104	68	48
valori perc.	91,235		
n = 138 w = 100	239.314.048	68	48
valori perc.	91,100		

	k = 6	residui tot.	residui sin.
n = 150	244.900.912	68	48
w = 92			
valori perc.	93,618		
n = 138	244.899.568	68	48
w = 100			
valori perc.	93,617		
n = 175	244.646.704	43	23
w = 79			
valori perc.	93,520		
n = 200	244.402.816	68	48
w = 69			
valori perc.	93,427		
n = 203	244.343.920	64	44
w = 68			
valori perc.	93,405		
n = 206	244.313.648	66	46
w = 67			
valori perc.	93,393		
n = 209	244.294.384	74	54
w = 66			
valori perc.	93,386		
n = 300	242.694.048	68	48
w = 46			
valori perc.	92,774		

Per $k = 5$, il valore più elevato si ha quando $n = 175$; se si confronta questo risultato con altri non riportati, a parità di n e k, si vede che il valore massimo si ha proprio quando il numero di unità residue a destra è pari a 20. Considerando, inoltre, i casi n = 209, 203 e 206 si osserva che i valori della varianza sono molto più elevati rispetto al caso di una ripartizione dei residui nella coda sinistra pari a 48. Infatti con questo raggruppamento il valore delle unità da inserire a destra era sempre inferiore o superiore alle 20 unità.
Per verificare se, data questa distribuzione di redditi, è davvero preferibile allocare almeno venti unità all’interno della coda destra, sono state effettuate altre prove, con un numero di residui a destra fisso a 40.
Si riportano nella tabella 13 i valori della varianza fra i gruppi, sempre per gli stessi raggruppamenti n e w analizzati fino ad ora.

Tabella 9

	K = 5	residui tot.	residui sin.
N = 138 w = 100	236.656.080	68	48

Valori perc.	90,466		
N = 209 w = 66	236.624.176	74	54
Valori perc.	90,454		
n = 150 w = 92	236.566.000	68	48
Valori perc.	90,431		
n = 206 w = 67	236.560.704	66	46
Valori perc.	90,429		
n = 200 w = 69	236.547.360	68	48
Valori perc.	90,424		
n = 300 w = 46	236.546.368	68	48
Valori perc.	90,424		
n = 203 w = 68	236.538.864	64	44
Valori perc.	90,421		
n = 175 w = 79	236.270.336	43	23
Valori perc.	90,318		
	k = 6	residui tot.	Residui sin.
n = 150	241.407.632	68	48
w = 92			
Valori perc.	92,282		
n = 138	241.339.584	68	48
w = 100			
Valori perc.	92,256		
n = 209	241.241.552	74	54
w = 66			
Valori perc.	92,219		
n = 200	241.237.824	68	48
w = 69			
Valori perc.	92,217		
n = 175	241.230.080	43	23
w = 79			
Valori perc.	92,214		
n = 206	241.228.096	66	46
w = 67			
Valori perc.	92,214		
n = 203	241.218.432	64	44
w = 68			

Valori perc.	92,210		
n = 300	240.492.960	68	48
w = 46			
Valori perc.	91,932		

Con questa ripartizione dei residui i valori diminuiscono ulteriormente.
Quindi, si conferma, mediante i risultati delle tabelle 8 e 9, che *una delle migliori ripartizioni che si possono effettuare con questo tipo di distribuzione è dato da un numero di residui a destra pari a 20 unità, a parità di n, k e w.*

2.3 Coda destra o sinistra nulle

Per allargare ulteriormente l’analisi si è inoltre proceduto ad una ripartizione dei dati in modo tale che i residui fossero concentrati tutti all’interno della coda sinistra o destra; *in ambo le tipologie di casi (per ogni valore di k esaminato) il valore della varianza fra i gruppi è diminuito drasticamente.*
Si riportano delle tabelle con qualcuno dei risultati delle prove effettuate con coda sinistra o destra nulla.

Nella tabella 10 i residui sono concentrati in una coda.

Tabella 10

Unità per gruppo (n) Numero gruppi (w)	k = 5	k = 6	residui Totali	Coda Sinistra	coda Destra
n = 300 w = 46	232.588.652	236.454.240	68	0	68
Valori percent.	88,911	90,389			
n = 209 w = 66	231.922.880	235.979.168	74	0	74
Valori percent.	88,656	90,207			
n = 206 w = 67	232.608.816	236.867.904	66	0	66
Valori percent.	88,919	90,547			
Unità per gruppo (n) Numero gruppi (w)	k = 5	k = 6	residui Totali	Coda Sinistra	coda Destra
n = 203 w = 68	232.796.144	237.098.960	64	0	64
Valori percent.	88,990	90,635			
n = 200 w = 69	232.459.120	236.618.752	68	0	68
Valori percent.	88,861	90,451			

Unità per gruppo (n)	k = 5	k = 6	residui	coda	Coda

Numero gruppi (w)			totali	sinistra	Destra
n = 300 w = 46	215.409.728	219.592.096	68	68	0
Valori percent.	82,344	83,943			
n = 209 w = 66	219.553.392	224.549.888	74	74	0
Valori percent.	83,928	85,838			
n = 206 w = 67	219.673.936	224.710.944	66	66	0
Valori percent.	83,974	85,900			
n= 203 w = 68	219.820.784	224.880.768	64	64	0
Valori percent.	84,030	85,964			
n = 200 w = 69	219.981.392	225.075.456	68	68	0
Valori percent.	84,092	86,039			

Qui, i valori della varianza sono ancora più bassi quando la coda sinistra è vuota. Questo è dovuto al fatto che i valori all'interno della coda destra corrispondono ai redditi più elevati.
Per mostrare ulteriormente come il fatto di avere la coda destra nulla, incida molto sul valore della varianza fra i gruppi, si riportano delle tabelle in cui si confrontano i risultati di alcune prove effettuate su k + 2 classi, con entrambe le code non vuote, con quelli delle prove calcolate su k* + 1 (dove una coda è vuota), dove k = k* - 1, sempre a parità di *n* e *w*.
In questo modo, come si può notare, i valori complessivi delle classi su cui si calcola la varianza rimangono invariati, l'unico elemento che varia è il numero delle classi su cui vengono calcolate le permutazioni: in un primo caso è k (come abbiamo sempre visto fino ad ora), nell'altro è k* (si ricorda infatti che i valori all'interno delle due code rimangono sempre fissi e non cambiano al permutare dell'algoritmo).
Si possono ad esempio raffrontare i valori della varianza nel caso in cui $k = 4$ e $k^* = 5$.
Per quanto riguarda i valori in corrispondenza di $k = 4$ si riportano, in tabella 11, i valori nel caso in cui la ripartizione dei residui sia con 48 unità fisse a sinistra.

Tabella 11

	k = 4	residui tot.
n = 300 w = 46	229.605.408	68
Valori perc.	87,771	
n = 209 w = 66	228.915.648	74

	k = 4	residui tot.
n = 200 w = 69	229.331.952	68
valori perc.	87,666	
n = 150 w = 92	229.642.768	68

Valori perc.	87,507	

valori perc.	87,785	

n = 206 w = 67	229.233.664	66
Valori perc.	87,628	

n = 138 w = 100	229.662.176	68
valori perc.	87,792	

n = 203 w = 68	228.891.504	64
Valori perc.	87,498	

Si osserva che questi valori sono inferiori a quelli riportati nella prima parte della tabella 10: ciò significa che, anche con la coda sinistra nulla, il numero di classi *k* incide molto sul valore della varianza fra i gruppi; a parità di *n* e *w*, avere un *k* elevato, anche in presenza di coda sinistra nulla, garantisce un valore più alto della varianza rispetto a quello che si avrebbe avuto con un numero di classi inferiore, ma con entrambe le code non vuote.
Se, invece, si raffrontano i risultati dei tabella 11 con quelli riportati nella parte inferiore di tabella 10, i valori della varianza sono di gran lunga superiori a quelli con coda destra nulla.
A parità di n e w, si preferisce, quindi, avere un numero di k inferiore, ma con entrambe le code non nulle, piuttosto che un k più elevato, ma con coda destra nulla.
Si osserva, inoltre, che i valori percentuali di tabella 11 sono maggiori di 5 punti percentuali rispetto a quelli della parte inferiore di tabella 10.
Pertanto, la coda destra assume una importanza notevole per il raggiungimento dell’obiettivo di rendere massima la varianza fra i gruppi.

3. Conclusioni

In questo lavoro si sono definiti criteri ottimali e “quasi” ottimali di classificazione per caratteri continui, basati sulla massimizzazione/minimizzazione di una funzione circostanza *V*.
Considerando il caso della massimizzazione della varianza fra i gruppi, l’analisi è stata accompagnata da una serie di prove, che hanno mostrato come sia possibile agire attraverso alcune variabili (numero di dati per gruppo *n*, numero di gruppi *w*, numero di classi *k* e ripartizione dei residui nelle due code) sul valore finale della varianza fra i gruppi, avvicinando il valore massimo assoluto.
Pur essendo i risultati delle prove in molti casi disomogenei, si possono comunque notare delle regolarità, e da queste è possibile trarre alcune considerazioni generali.
1. Molto importante è la ripartizione dei residui nelle due code, infatti, a seconda di come questi vengono ripartiti, il valore della varianza fra i gruppi si alza o si abbassa decisamente.

Nella distribuzione esaminata, cioè relativa ai redditi individuali di 13.868 redditieri, è proprio la ripartizione dei residui che risulta come l'elemento che in modo più significativo influisce sul valore della varianza fra i gruppi.
In base ad una diversa ripartizione dei residui, a parità di tutti gli altri fattori (*k*, *n* e *w*), vi sono delle variazioni nei valori della varianza fra i gruppi che raggiungono i 7 punti percentuali.
2. All'interno dell'analisi di come ripartire i residui, è soprattutto la coda destra ad assumere una notevole importanza.
Per verificare effettivamente l'esistenza di un valore "critico" in corrispondenza di un determinato numero di unità residue nella coda destra, sono state effettuate altre prove, tenendo fissi i valori dei residui all'interno della coda destra della distribuzione, a parità di *n*, *k* e *w*.
Si è osservato che questo valore critico esiste e che è dato da 20 unità residue a destra; questi 20 valori corrispondono ai valori più elevati dell'intera distribuzione.
Un'altra conferma dell'importanza della coda destra è data dalle prove effettuate con coda destra nulla, ovvero quando tutti i residui sono concentrati all'interno della coda sinistra: in questo caso i valori della varianza sono diminuiti drasticamente, molto di più rispetto al caso in cui era nulla la coda sinistra.
3. Bisogna notare che anche il numero delle classi *k* influisce sul valore della varianza, a parità di *n* e *w*. Si ha, infatti, un aumento di almeno due punti percentuali, se si considerano i valori della varianza fra i gruppi in relazione alla varianza totale della distribuzione, all'aumento di ogni *k*. Si noti che, nonostante il valore della varianza aumenti all'aumentare del numero delle classi, l'aumento è più limitato rispetto al caso in cui, a parità di *n* e *w*, si decide di agire sulla ripartizione dei resti nelle due code.
4. Infine, il numero minimo di osservazioni per gruppo *n*, risulta influenzare il valore della varianza fra i gruppi. Pur essendo i risultati non sempre regolari rispetto ai valori di *n* e *w*, bisogna comunque evidenziare che nella maggior parte dei casi, a parità di unità residue collocate a destra o a sinistra della distribuzione e a parità di numero di classi *k*, i valori della varianza fra i gruppi maggiori si hanno quando $n = 150$ o $n = 138$ (valori di *w* pari a 92 e 100).

Da quanto visto nel lavoro, l'utilizzazione dei criteri "quasi" ottimali, e il confronto fra essi, può risultare uno strumento utile ai fini della classificazione di caratteri continui in base alla massimizzazione/minimizzazione di una funzione di circostanza.

Riferimenti bibliografici

Banca d’Italia (1993), I bilanci delle famiglie italiane nell’anno 1991, *Suppl. al Bollettino Statistico. Anno III - Numero 404 – 14 Luglio 1993*.

Landenna G. (1984). *Fondamenti di statistica descrittiva*, Il Mulino.

Leti G. (1983). *Statistica descrittiva*, Il Mulino.

Takács L. (2001). Combinatorics *Handbook of statistics Vol. 4, 123 – 173, Edited by P. R. Krishnaiah and P. K. Sen, North Holland*.

Zenga M. (1989), *Introduzione alla statistica descrittiva*, Vita e Pensiero.

Zenga M.(2001). A multiplicative decomposition of Herfindahl concentration measure, *Metron*, Vol. LIX n. 1 – 2.