Revista Colombiana de Estadística
No. 31 y No. 32

CLASIFICACION NO JERARQUICA CON COORDENADAS FACTORIALES

LEONARDO BAUTISTA S.
Profesor Asociado
Universidad Nacional de Colombia
Departamento Administrativo
Nacional del Estado (DANE)

HERNÁN ABDÓN GARCÍA
Profesor Universidad de Nariño

RESUMEN Para un conjunto de n puntos p-dimensionales se desea obtener k clases disjuntas no vacías, en forma tal que la clasificación sea óptima, es decir que la inercia ENTRE sea máxima. La característica importante del conjunto de datos es que tanto n como p son muy grandes, lo cual implica que la utilización de los métodos clásicos de clasificación puede presentar serios inconvenientes. El análisis factorial se aplica como un paso intermedio que permite reducir la dimensión y establecer ejes factoriales (ortogonales) a utilizar en la tarea clasificatoria. El objetivo de este trabajo es analizar los efectos de tal procedimiento sobre los resultados finales de la clasificación. Puesto que los métodos de clasificación son de tipo algorítmo es imposible realizar una investigación de tipo analítico, por tal razón se planteó un estudio de tipo experimental.

Introducción

En los problemas prácticos más comunes de la clasificación, el investigador se encuentra ante una gran cantidad de variables y de observaciones. La "redundancia" de información sugiere la necesidad de reducir magnitudes a partir de un análisis de componentes principales. La gran cantidad de individuos obliga la aplicación de métodos de clasificación no jerárquica. En este artículo se analiza la utilización de coordenadas factoriales, producto del análisis normado de componentes principales, en los resultados clasificatorios obtenidos mediante la aplicación del algoritmo de K-Means.

Se planteó un diseño experimental en el que para cinco grupos de datos con diferente estructura inercial y para diferente cantidad de clases deseadas se reduce paulatina-

mente la cantidad de coordenadas factoriales a utilizar en la clasificación y se analiza su influencia en términos de la inercia explicada.

La utilización de los dos o tres primeros factores para realizar clasificaciones con gran cantidad de datos y un número alto de clases no producirá distorsiones importantes frente a la dispendiosa utilización de los datos originales. Sólo en casos de estructura inercial estrictamente esférica (poco frecuente en la práctica) la clasificación con las primeras coordenadas factoriales puede conllevar a errores importantes de confusión por pérdida de información.

1.Indicadores de comparación de clasificación.

Las coordenadas factoriales que se obtienen a partir de un análisis de componentes principales son el resultado de traslaciones y rotaciones, que conservan las distancias y los ángulos entre los puntos, razón por la cual el uso de todas las coordenadas factoriales para efectos clasificatorios no sólo no produce resultados diferentes a los obtenidos con las variables originales (clasificación completa) sino que no constituye reducción alguna. La reducción se encuentra cuando la actividad clasificatoria se realiza utilizando sólo los t primeros ejes factoriales (clasificación t-reducida), dejando de lado los restantes p menos t.

La inercia total de la nube dada por

$$I = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{t} \lambda_j + \sum_{j=t+1}^{p} \lambda_j$$

manteniendo fijas unas clases previas y notando con W la inercia "Dentro" y con B la inercia "Entre", se deja descomponer en:

$$I = \sum_{j=1}^{p} \lambda_j = \left\{ \sum_{j=1}^{t} W_j + \sum_{j=t+1}^{p} W_j \right\} + \left\{ \sum_{j=1}^{t} B_j + \sum_{j=t+1}^{p} B_j \right\}$$

que para efectos prácticos se notará:

$$I = \{W_t + W_p\} + \{B_t + B_p\}$$

y conduce a las particiones

$$I = B + W, \qquad I_t = B_t + W_t, \qquad I_p = B_p + W_p$$

La inercia explicada por la clasificación corresponde al término $B = B_t + B_p$ y la inercia de la nube proyectada sobre los primeros t ejes factoriales es $I_t = B_t + W_t$, ambas con B_t como parte común. W_t define la contribución factorial de la interpretación y B_p la contribución clasificatoria. [Lerman 1981]

Para comparar clasificaciones completas y t-reducidas se utilizará un coeficiente de estabilidad inercial, definido como la razón entre la tasa de inercia explicada t-reducida sobre la tasa de inercia explicada completa:

$$C = \frac{\frac{B_t}{I_t}}{\frac{B}{I}} \quad \text{haciendo} \quad D = \frac{B_t}{B} \quad \text{es claro que} \quad C = D \cdot \frac{I}{I_t}$$

Se toma como indicador de calidad, en el sentido de que la clasificación reducida explica tanta inercia como la clasificación completa el hecho de que C sea igual a uno. Se espera que también B sea próximo a uno, sin embargo es importante señalar que no están normados y aunque no es frecuente, se presentan valores superiores a la unidad. En principio un valor de D superior a la unidad implica un mejoramiento de la clasificación t-reducida frente a la clasificación completa, y se puede dar por

cuanto ésta última no es necesariamente un máximo absoluto de B, sin embargo se debe interpretar además como un mejoramiento debido en parte a la reducción de inercia total, reducción que implica contar con una menor cantidad de "ruido" en el conjunto a clasificar.

2. Plan de experimentación.

Se trabajó con cinco tipos de datos de acuerdo a su estructura inercial. El primer grupo está compuesto por series de datos en los que más del 90% de la inercia se explica en su primer factor. El segundo grupo de datos no son tan marcadamente "lineales" pero el primer factor recoge entre el 70% y el 90% de la inercia. El tercer grupo esta compuesto por casos, en los cuales la inercia se reparte sobre los dos primeros factores y supera el 70% de la inercia total. En el grupo cuatro se incluyen datos en los que se aprecia una diferencia importante entre las proporciones de inercia retenidas por el primero y el segundo factor. Finalmente el quinto grupo corresponde a los casos que se pueden llamar "esféricos" por cuanto la inercia total se reparte en proporciones similares entre todos los ejes factoriales.

De otro lado, se parte del supuesto que los resultados a obtener dependen también de la cantidad de clases que se requieran del proceso clasificatorio, por tal razón se consideran dentro del plan experimental los casos de clasificación en dos, tres y cuatro clases. Se trata entonces de 56 ejercicios dispuestos de la siguiente forma:

Cantidad de ejercicios clasificatorios por tipo de grupo factorial, según cantidad de clases requeridas.

	Grupo factorial tipo	Clases requeridas		
		Dos	Tres	Cuatro
1.	Más de 90% en F1		4	4
2.	Entre 70% y 90% en F1	6		6
3.	Más de 70% en F1 y F2	6		6
4.	Diferencias entre F1 y F2	5		5
5.	Inercia esférica	7		7
	TOTAL	24	4	28

Para estos 56 ejercicios se utilizan 36 conjuntos de datos, 13 datos reales provenientes de la literatura y 23 generados con modelos de simulación. En todos los casos se trabaja con el algoritmo K-Means convergente de la biblioteca MODULAD. Para cada conjunto de datos se mantienen el criterio de convergencia y los puntos que conforman los núcleos originales del algoritmo a lo largo del experimento, consistente en una clasificación completa y clasificaciones t-reducidas reduciendo el valor t desde p menos uno hasta uno.

3. Resultados.

Para el caso de datos del grupo factorial tipo uno, es decir con más del 90% de la inercia sobre el primer factor, los indicadores C_i y D_i (i= 1,2,3 la cantidad de factores utilizados) son todos muy cercanos a uno (cuadro de resultados Nro. 1), lo cual se interpreta como la viabilidad de utilizar un único factor sin pérdida de calidad en la clasificación. En el segundo grupo de datos la diferencia de inercia explicada cuando se utiliza un solo factor es un poco más acentuada. El indicador C_1 señala posibles ineficiencias que oscilan entre el 14% y el 35%. Los conjuntos 11 y 12 de datos

mostraron desviaciones hasta del 20% aún en el caso de trabajar con tres factores. Las clasificaciones con datos del tercer grupo factorial comienzan a evidenciar la tendencia consistente en que las ineficiencias son cercanas al 50% cuando se utiliza sólo el primer factor y se reducen paulatinamente a medida que se incorporan ejes factoriales. En este último caso la utilización de los tres primeros ejes condujo a un coeficiente de confusión todavía aceptable.

Cuadro Nro 1. Valores de los coeficientes de estabilidad inercial para cada uno de los experimentos realizados en los grupos experimentales de tipo uno, dos y tres.

Caso	Indiv.	Var.	Clas.	D_3	D_2	D_1	C_3	C_2	C_1
Tipo 1									
1a	16	4	3	0.9998	0.9997	0.9955	1.0004	1.0034	1.0145
1b	16	4	4	0.9999	0.9998	0.9966	1.0005	1.0034	1.0146
2a	20	5	3	1.0000	1.0000	0.9951	1.0029	1.0093	1.0352
2b	20	5	4	0.9995	0.9978	0.9761	1.0024	1.0072	1.0154
3	150	4	3	1.0000	0.9997	0.9827	1.0053	1.0225	1.0629
4a	20	4	3	0.9998	0.9999	0.9952	1.0039	1.0111	1.0773
4b	20	4	4	0.9993	0.9991	1.0260	1.0035	1.0113	1.1107
5	42	6	4	0.9994	0.9992	0.9829	1.0046	1.0159	1.0670
Tipo 2									
6	50	7	4	0.9993	0.9939	0.9962	1.0116	1.0393	1.1404
7	80	6	4	0.9989	0.9965	0.9956	1.0498	1.0917	1.1561
8	30	10	2	1.0000	1.0000	0.9951	1.0146	1.0338	1.1960
9a	100	4	2	0.9995	0.9994	0.9972	1.0282	1.1000	1.2405

Caso	Indiv.	Var.	Clas.	D_3	D_2	D_1	C_3	C_2	C_1
9b	100	4	4	0.9964	1.0071	0.9536	1.0251	1.1084	1.1862
10a	40	6	2	0.9995	0.9995	0.9652	1.0278	1.0558	1.2087
10b	40	6	4	0.9988	0.9985	0.8985	1.0271	1.0547	1.1252
11	56	10	2	1.0000	0.9995	0.9946	1.1179	1.1833	1.3456
12a	100	10	2	0.9992	0.9989	0.9986	1.1958	1.2525	1.3247
12b	100	10	4	0.9918	0.9934	0.9916	1.1869	1.2457	1.3154
13	100	9	2	1.0000	0.9999	0.9951	1.0078	1.0260	1.3365
14	20	5	4	0.9969	0.9954	0.9151	1.0166	1.1238	1.2598
Tipo 3									
15	88	5	4	0.9970	0.9992	0.9886	1.0978	1.2069	1.4293
16	23	9	2	0.9996	0.9996	0.9679	1.0073	1.0231	1.4104
17	88	10	2	0.9984	0.9975	0.9974	1.2887	1.3828	1.5043
18	80	6	4	0.9980	0.9989	0.8445	1.0194	1.1023	1.3285
19a	70	7	2	0.9995	0.9989	0.9964	1.1371	1.2935	1.5889
19b	70	7	4	0.9916	0.9828	0.9175	1.1282	1.2726	1.4631
20a	66	5	2	0.9996	0.9782	0.9748	1.1166	1.2204	1.5764
20b	66	5	4	0.9980	0.9933	0.9169	1.1148	1.2392	1.4827
21a	40	5	2	0.9897	1.1074	0.9769	1.1251	1.4808	1.7896
21b	40	5	4	0.9886	0.9502	0.8306	1.1238	1.2707	1.5174
22a	18	6	2	0.9923	0.9739	0.9698	1.0965	1.2361	1.8603
22b	18	6	4	1.0147	0.9265	0.7884	1.1213	1.1760	1.5123

El tratamiento de los datos del cuarto grupo reafirman la tendencia en el sentido de que se presenta una creciente ineficiencia a medida que la inercia de los datos

originales pierde concentración. El indicador C_1 toma valores mayores a dos y llega incluso a 3,27. El coeficiente C_3 se comporta alrededor de 1,5. En los resultados del grupo cuatro es característico el comportamiento decreciente de los coeficientes D_3, D_2, D_1. Los ejercicios con los datos del grupo cinco ratifican finalmente el hecho de que a medida que los datos asumen una forma más "esférica" se hace mayor la diferencia de inercias explicadas. El indicador C_1 muestra en este grupo valores de hasta 5,4. Resalta nuevamente que la calidad es muy sensible a la cantidad de clases, entre dos y cuatro el indicador C_1 presenta variaciones importantes.

Cuadro Nro 2. Valores de los coeficientes de estabilidad inercial para cada uno de los experimentos realizados en los grupos experimentales de tipo cuatro y cinco.

Caso	Indiv.	Var.	Clases	D_3	D_2	D_1	C_3	C_2	C_1
Tipo 4									
23	150	8	2	0.9986	0.9986	0.9983	1.5534	1.8238	2.2666
24	54	10	4	0.9879	0.9836	0.8929	1.5435	1.8058	2.0708
25a	140	10	2	0.9965	0.9965	0.9960	1.4939	1.7941	2.3333
25b	140	10	4	0.9863	0.9861	0.8905	1.4786	1.7754	2.0862
26	55	10	2	0.9976	0.9840	0.9837	1.2271	1.5408	2.3614
27	60	7	4	0.9784	0.9690	0.9515	1.3537	1.6924	2.3184
28a	60	7	2	0.9902	0.9876	0.9632	1.4687	1.8432	2.8698
28b	60	7	4	0.9819	0.9288	0.7141	1.4565	1.7335	2.1278
29a	90	10	2	0.9801	0.9759	0.9617	1.7929	2.2933	3.2768
29b	90	10	4	0.9972	0.9215	0.8486	1.8242	2.1656	2.8913

Caso	Indiv.	Var.	Clases	D_3	D_2	D_1	C_3	C_2	C_1
Tipo 5									
30a	56	6	2	0.9971	0.9641	0.9944	1.3350	1.7030	3.1781
30b	56	6	4	0.9893	0.8998	0.6546	1.3245	1.5894	2.0921
31a	60	6	2	0.9967	0.9905	0.9847	1.3319	1.8695	3.5232
31b	60	6	4	1.0134	0.9809	0.6719	1.3642	1.8512	2.4039
32a	100	10	2	1.0208	0.9762	1.2054	1.5453	1.9803	4.4590
32b	100	10	4	0.9746	0.9225	0.6637	1.4753	1.8714	2.4553
33a	90	10	2	0.9585	0.9190	0.9404	1.7158	2.2888	4.3322
33b	90	10	4	0.9654	0.8257	0.6529	1.7282	2.0565	3.0075
34a	100	10	2	0.9910	1.0141	1.0140	1.9338	2.7722	4.9370
34b	100	10	4	1.0022	0.8960	0.7052	1.9557	2.4495	3.4385
35a	90	8	2	1.1613	1.1427	1.1569	2.3272	3.2832	6.0603
35b	90	8	4	0.9501	0.8819	0.6071	1.9039	2.5341	3.1802
36a	100	9	2	1.0880	1.0671	0.9680	2.2844	3.1147	5.4390
36b	100	9	4	0.9345	0.8892	0.6113	1.9620	2.5954	3.4345

Por cuanto para el grupo con estructura inercial extremadamente esférica la clasificación con tres ejes factoriales arrojó un comportamiento particular a medida que aumentaba la cantidad de clases, se realizaron ensayos adicionales con conjuntos de cien datos, diez variables y clasificaciones de dos, cuatro y seis clases, estos resultados se muestran en el cuadro Nro. 3. En general el coeficiente D_i decrece a medida que se dejan de tener en cuenta factores, ($D_3 > D_2 > D_1$), y de otra parte, la variación de los D_i a lo largo de los cinco grupos factoriales es relativamente baja en comparación

con las variaciones de los valores C_i. Los valores de D_1 más distantes de la unidad se presentan para los conjuntos de datos del tipo cinco y lo hacen en forma moderada, se puede aventurar la conclusión de que estos valores decrecen tan levemente que para ciertos tipos de análisis pueden considerarse constantes. Es decir, parece existir una inercia "Entre", magnitud que se alcanza en la mayoría de las ocasiones independientemente de la estructura inercial de los datos. Esta conclusión implicaría dos hechos importantes: Primero, la clasificación con al menos un eje factorial, aún en los casos de mayor dispersión inercial, es siempre eficiente, en el sentido de que identifica la porción de B inherente a los datos. Segundo, el análisis de los C_i se refiere a la eficiencia clasificatoria pero es necesario relativizar las conclusiones. Teniendo en cuenta que el indicador C_i para valores constantes de D_i está determinado por $\frac{I}{I_i}$ es evidente que su marcado decrecimiento se debe básicamente a la reducción de inercia total I_i, que en casos de "ruido" es más una ventaja que un defecto.

Cuadro Nro 3. Valores de los coeficientes de estabilidad inercial para clasificaciones en el quinto grupo experimental con dos, cuatro y seis clases.

Caso	Clases	D_3	D_2	D_1	C_3	C_2	C_1
30a	2	0.9971	0.9641	0.9944	1.3350	1.7030	3.1781
31a	2	0.9967	0.9905	0.9847	1.3319	1.8695	3.5232
32a	2	1.0208	0.9762	1.2054	1.5453	1.9803	4.4590
33a	2	0.9585	0.9190	0.9404	1.7158	2.2888	4.3322
34a	2	0.9910	1.0141	1.0140	1.9338	2.7722	4.9370
35a	2	1.1613	1.1427	1.1569	2.3272	3.2832	6.0603
36a	2	1.0880	1.0671	0.9680	2.2844	3.1147	5.4390

Caso	Clases	D_3	D_2	D_1	C_3	C_2	C_1
30b	4	0.9893	0.8998	0.6546	1.3245	1.5894	2.0921
31b	4	1.0134	0.9809	0.6719	1.3542	1.8512	2.4039
32b	4	0.9746	0.9225	0.6637	1.4753	1.8714	2.4553
33b	4	0.9654	0.8257	0.6529	1.7282	2.0565	3.0075
34b	4	1.0022	0.8960	0.7062	1.9557	2.4495	3.4385
35b	4	0.9501	0.8819	0.6071	1.9039	2.5341	3.1802
36b	4	0.9345	0.8892	0.6113	1.9620	2.5954	3.4345
30c	6	0.9526	0.8376	0.5774	1.2755	1.4796	1.8456
31c	6	0.9830	0.8812	0.5585	1.3136	1.6630	1.9983
32c	6	1.0137	0.9000	0.5895	1.5345	1.8257	2.1806
33c	6	0.8827	0.7422	0.5212	1.5800	1.8485	2.4010
34c	6	0.9288	0.8186	0.5712	1.8125	2.2380	2.7811
35c	6	0.8571	0.7591	0.4983	1.7176	2.1810	2.6105
36c	6	0.8558	0.7793	0.4839	1.7968	2.2747	2.7187

Como conclusión final se puede afirmar que la clasificación t-reducida tiene muchas ventajas y muy pocas desventajas frente a la clasificación completa. La cantidad de ejes factoriales a utilizar en la clasificación depende de la forma del histograma de valores propios, pero cantidades razonables desde el punto de vista del computo de la clasificación conduce a muy buenos resultados.

Bibliografía

Anderberg, M. R.(1973), *Cluster Analysis for Aplications*, Academic Press, New York.

Bautista, L. y Ramos, J. (1988), *Análisis de datos de encuestas y tabulados*, Universidad Nacional, Bogotá.

Bock, H. H. (1987), *On the interface between Cluster Analysis, Principal Component Analysis, and Multidimentional Scaling. H.*.

Bozdogan and A. K. Gupta(eds.), *Multivariate Statistical Modeling and Data Analysis*, Reidel Publishing Company.

Escofier, B. et Pages, J. (1988), *Analyses Factorielles Simples et Multiples.*, Dunod, Paris.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley and Son, New York.

Jambu M. & Lebeaux M. O. (1983), *Cluster Analysis and Data Analysis*, North Holland.

Lebart, L., Morineau, A. and Warwick, K. (1984), *Multivariate Statistical Analysis.*, John Wiley, New York.

Lerman, J. C. (1981), *Classification et analyse ordinale des données*, Dunod, Paris.

Marriot, (1982) H., *Optimisation methods of Cluster Analysis,*, Vol. 69, Biometrika.

Ok-Sakun, Y. (1975), *Analyse Factorielle typologique et Lissage Typologique*, Tesis de Doctorado de tercer ciclo, Universidad de Paris, Paris.

Rao, C.R. (1964), *The use and inerpretation of Principal Component Analysis in Applied Research*, Sankhya A 26.

Rizzi, A. (1984), *Some Mathematical Properties of Cluster Methods, in Data Analysis and informatics III*, Eds. E.Diday & al., North Holland.

Takeuchi, K.; Yanai, H. and Mukherjee (1982), *The Foundations of Multivatiate Analysis*, Wiley Eastern Limited, India.

Volle, M. (1985), *Analyse des Données*, Economica, Paris.