Psicothema, 1996. Vol. 8, nº 3, pp. 709-715
ISSN 0214 - 9915 CODEN PSOTEG

CLASSIFICATION OF PRISON INMATES BASED ON HIERARCHICAL CLUSTER ANALYSIS

Albert Maydeu-Olivares
University of Illinois

Six clustering methods (Ward's method, average linkage clustering, complete linkage clustering, single linkage clustering, nearest neighbor, and *k*-means) were applied to the matrix of Euclidean distances among inmates in the 6-dimensional space defined by the following variables: 1) age, 2) length of the internment already fulfilled, 3) paranoid ideations and behavior, 4) hardness, 5) psychotic ideations and behavior, and 6) means-ends interpersonal problem solving skills. The classification determined by each of the clustering procedures was matched with the actual classification of the inmates into those who were and those who were not granted outside permits, by computing unconditional and conditional classification rates.

Ward's method, complete linkage and the k-means procedure were able to match satisfactorily the qualitative classification of the inmates. However, these methods were more efficient in determining which inmates would be denied an outside permit than in determining those that would be granted an outside permit. This also occurs in the qualitative classification: it is easy to determine which inmates will not make good use of an outside permit; it is much more difficult to determine which inmates will make good use of it.

Clasificación de internos penitenciarios basada en análisis jerárquico de clusters. Seis métodos de análisis jerárquico de clusters (el método de Ward, el de uniones ponderadas 'average linkage', el de uniones completas 'complete linkage', el de única union 'single linkage', el del vecino más próximo 'nearest neighbor', y el de k-medias) fueron aplicados a la matriz de distancias Euclídeas entre internos penitenciarios en el espacio de seis dimensiones definido por las siguientes variables: 1) edad, 2) condena cumplida hasta la fecha, 3) conductas e ideaciones paranoicas, 4) dureza de caracter, 5) conductas e ideaciones psicóticas, y 6) habilidades de resolución de problemas interpersonales. Las clasificaciones obtenidas mediante cada uno de estos métodos fueron comparadas con la clasificación realizada por los servicios penitenciarios de internos que disfrutan de permisos vs. los que no los disfrutan utilizando razones de clasificación condicionales e incondicionales.

Los métodos de Ward, de uniones completas, y k-medias fueron capaces de recuperar satisfactoriamente la clasificación de los internos realizado por los servicios penitenciarios. Sin embargo, estos métodos fueron mucho más eficaces en determinar a qué internos se les debian negar que a quienes se les debian conceder permisos. Esto no es sorprendente, ya que también ocurre utilizando los criterios cualitativos empleados por los servicios penitenciarios: es mucho más sencillo determinar qué internos no harán buen uso de los permisos que determinar quienes harán buen uso de los mismos.

Correspondencia: Albert Maydeu-Olivares
Dept. Statistics and Econometrics
Universidad Carlos III de Madrid
C/ Madrid 126-128
!8903 Getafe (Spain)

Many existing correctional systems include among its objectives not only the punishment of the offender, but also his/her rehabilitation. One of the most powerful ins-

truments that correctional systems can use towards this objective is the concession of permits to leave the prison. These can be weekend permits so that the inmate can visit his/her family, work permits so that the inmate can work outside the correctional institution, etc. Usually inmates who are granted this type of permits and make good use of them are eventually granted probation or conditional freedom. The decision of granting or denying a permit has extremely important individual and social consequences. Type I errors in the decision process (i.e., an inmate granted a permit who did not return) are generally accompanied by new criminal acts. Type II errors (i.e., an inmate denied a permit who would have made good use of it) have devastating consequences for the inmates. This decision process is almost invariably made from a qualitative perspective, and usually must take into consideration a set of legal guidelines that generally focus on variables such as type of crime, remaining time in prison, behavior during internment, etc. There is also a long tradition of research that has tried to show that psychological variables are causally related to performing criminal acts. Consequently, individual differences on certain psychological variables (e.g., social skills, psychopathy) have also been used to support the granting or denial of outside permits to correctional inmates.

Quantitative rules to assign outside permits to correctional inmates can be readily developed given a set of psychological and criminological predictors by means of discriminant analysis, or by more appropriate methods such as logistic or probit regression. These methods assume that the populations to be classified closely match a multivariate normal, or a multivariate logistic distribution. Then, by using the information contained in an actual classification of the subjects, these methods provide a classification rule. Since these methods use the actual classification of the subjects to derive the classification rule, the usefulness of such rule must be evaluated by cross-validation.

An alternative approach to classification can be obtained from cluster analytic methods. Very different approaches to cluster analysis exist (see Hartigan, 1975; Gordon, 1981; Jain & Dubes, 1988). The approach that will be used here involves the use of agglomerative hierarchical classification algorithms based on Euclidean distances among the subjects. Thus, Euclidean distances among the subjects in the n-dimensional space defined by a set of *n* predictor variables are obtained. Hierarchical clustering methods form nested groups of subjects by merging subjects or groups of subjects using a particular algorithm. Five agglomerative hierarchical clustering methods (Ward's method, average linkage clustering, complete linkage clustering, single linkage clustering, and nearest neighbor) and one non-hierarchical method (*k*-means) will be applied to the matrix of Euclidean distances among inmates in the 6-dimensional space defined by the following variables: 1) age, 2) length of the sentence already fulfilled, 3) paranoid ideations and behavior, 4) hardness, 5) psychotic ideations and behavior, and 6) means-ends interpersonal problem solving skills.

For each of the procedures, two clusters of subjects will be obtained based on their dissimilarities on these six variables. The classification determined by each of the clustering procedures will then be matched with the actual classification of the inmates into those who were and those who were not granted outside permits, by computing unconditional and conditional classification rates.

True positive and true negative rates are unconditional classification rates. In this case, the *true positive rate* is the proportion of subjects with outside permits correctly classified by the cluster analytic procedure. Si-

milarly, the *true negative rate* is the proportion of subjects without outside permits correctly classified by the cluster analytic procedure. However, for classification purposes, we are mostly interested in conditional classifications, rather than uncontidional ones (Widiger, Hurt, Frances, Clarkin, & Gilmore, 1984). Given a two-by-two classification table, we can obtain, for instance,

	Actual classification	
Clusters	**permit**	**no permit**
1	a	b
2	c	d

- *true positive rate* = a / (a + c)
- *negative positive rate* = d / (b + d)
- *positive predictive power* (PPP) = a / (a + b)
- *negative predictive power* (NPP) = d / (c + d)

Of these, PPP and NPP are conditional classification rates. In our case, the PPP rate is the probability that a subject would be granted an outside permit *given* that he/she has been assigned to cluster 1, whereas the NPP rate is the probability that a subject would be denied an outside permit *given* that he/she has been assigned to cluster 2.

Method

Subjects

A random sample of 108 male inmates from a medium security Spanish prison were tested on a large set of variables, including demographic, educational, personality, psychopathology, intelligence, and social competence variables. This sample represents approximately a fourth of the total population of the prison at the time of assessment. Details of the characteristics of this sample can be found in Guillén, Maydeu-Olivares, Pons, and Vigil (1989).

Measures

Six variables were selected from the pool of available variables by considering all variables that showed a correlation higher in magnitude than .20 with the actual classification of inmates with outside permits vs. inmates without outside permits, that is,

1) *age*;

2) *length of internment already fulfilled*;

3) *paranoic ideations and behaviors*, as measured by the Pa-6 scale of a Spanish adaptation (Roig-Fusté, 1986) of the Minimult version of the MMPI;

4) *hardness*, as measured by a Spanish adaptation of the Eysenck Personality Questionnaire's Psychoticism scale (EPQ-P);

5) *psychoticism*, as measured by the Sc-8 scale of a Spanish adaptation (Roig-Fusté, 1986) of the Minimult version of the MMPI;

6) *means-ends social problem solving*, as measured by a Spanish adaptation of the Means-Ends Problem Solving Questionnaire (Platt & Spivack, 1975).

To remove outliers from the sample, only those subjects up to 40 years of age and whose fulfilled length of internment was less than or equal to 150 months were included. The total available sample was 99 inmates. A further check for multidimensional outliers was performed using the method suggested by Bollen (1989: pp. 128-129). No outliers were pointed out by this procedure.

Procedure

The following clustering algorithms, as implemented in SAS 6 (SAS Inc., 1990), were applied to the Euclidean distances used as dissimilarity measures in the six-dimensional space defined by the predictor variables:

a) *Average linkage clustering* (Sokal & Michener, 1958), where the distance between two clusters is defined as the average dis-

tance between pairs of observations, one from each cluster.

b) *Complete linkage clustering* (Sorensen, 1948), where the distance between two clusters is defined as the maximum distance between an observation in one cluster and an observation in another cluster.

c) *Single linkage clustering* (Florek, Lukaszewicz, & Zubrzycki, 1951), where the distance between two clusters is defined as the minimum distance between an observation in one cluster and an observation in another cluster.

d) *Ward's method* (Ward, 1963), where the distance between two clusters is defined as the squared error criterion.

e) *Nearest neighbor method* (Wong & Lane, 1983). This method uses density estimates of a cluster of *k*-nearest observations around the cluster center to obtain modified dissimilarity measures that are subsequently used to perform single link clustering.

f) *K-means* (Hartigan, 1975; MacQueen, 1967). This method uses a fixed number of clusters for which a cluster seed is selected. Observations are assigned to clusters based on their distance to the cluster means.

Of these, the first five methods are hierarchical. *K* was arbitrarily set equal to four when using the nearest neighbor method. Since the *k*-means procedure may yield different solutions depending on the seed used to start the algorithm, three different random starting seeds were used with this procedure. All three produced the same solution. When using average linkage clustering and Ward's method, squared Euclidean distances were used, as required by these two methods. In all instances, the distances were computed from the raw data to incorporate the elevation, scatter, and shape of the subject's profiles (Cronbach & Goldine, 1953).

Results

In Table 1 the means, standard deviations, and inter-correlations among the six predictor variables are presented for each of the two groups of inmates: with and without outside permits. The correlations between each of the predictors and a dummy variable representing membership to one group of inmates or the other is also presented. As expected, older inmates, inmates that have been in prison for a long time, and inmates with better means-ends problem solving skills are more likely to be granted outside permits, whereas inmates with symptoms of paranoia or psychoticism, and those with a 'hard-core' personality are less likely to be granted outside permits. The larger differen-

Table 1
Means, standard deviations, and inter correlations among the predictor variables and the granting/denial of outside permits

a) Correlations between the predictors and the outcome variable (N= 99)

	age	time	MMPI-Pa6	MMPI-Sc8	EPQ-P	MEPS
Permit	.17	.37	–.20	–.19	–.19	.22

b) Inmates not allowed to leave the prison on permit (N= 88)

	age	time	MMPI-Pa6	MMPI-Sc8	EPQ-P	MEPS
Age		.23	–.02	–.38	–.29	.08
Time			.01	–.15	–.09	–.15
MMPI-Pa6				.42	.45	–.09
MMPI-Sc8					.44	.00
EPQ-P						–.04
Mean	27.33	46.72	4.30	4.67	3.68	.55
Std	4.51	23.07	1.88	1.83	2.92	.29

c) Inmates allowed to leave the prison on permit (N= 11)

	age	time	MMPI-Pa6	MMPI-Sc8	EPQ-P	MEPS
Age		.32	.26	–.42	.44	–.03
Time			.32	.15	–.20	.55
MMPI-Pa6				.15	.22	.29
MMPI-Sc8					.14	.40
EPQ-P						.11
Mean	29.82	76.91	3.10	3.55	1.91	.74
Std	4.21	31.31	1.38	1.63	2.02	.18

Notes. Permit is coded 0= no permit, 1= permit., time= length of internment fulfilled in months, EPQ-P= Eysenck's Personality Questionnaire-Psychoticism, MEPS= Means Ends Problem Solving questionnarire.

ces between both groups of inmates was found in their fulfilled length of internment. The base rate of outside permits is 11%.

In Table 2 I present a two-way table with the results of the clustering classifications vs. the actual classifications. Numerous ties were found. In Ward's method, ties were found at the following levels: {85, 83, 78, 75, and 73}; in average linkage at {85, 78, 74, and 71}; in complete linkage at {89, 85, 83, 78, 62, 58, and 56}; in single linkage at {86, 82, 78, 76, 70, 66, 65, 64, 61, 59, 56, 52, 30 and 24}; and in nearest neighbor linkage at {91, 60, 28, and 24}.

Table 2
Two-way classification table for several clustering algorithms

Clusters obtained	actual classification				
	outside permit		no outside permit		Total
1	K-means	9	K-means	37	46
	Ward	4	Ward	20	24
	Average Linkage	0	Average Linkage	1	1
	Complete Linkage	4	Complete Linkage	27	31
	Single Linkage	0	Single Linkage	1	1
	Nearest Neighbor	0	Nearest Neighbor	1	1
2	K-means	2	K-means	51	53
	Ward	7	Ward	68	75
	Average Linkage	11	Average Linkage	87	98
	Complete Linkage	7	Complete Linkage	61	68
	Single Linkage	11	Single Linkage	87	98
	Nearest Neighbor	11	Nearest Neighbor	87	98

In Table 3 I present the true positive and true negative rates, and the positive and negative predictive power for all six clustering methods. As it can be observed, three methods, average linkage, single linkage, and nearest neighbor gave exactly the same classifications. The results obtained using Ward's method and complete linkage are also very similar. Finally, the k-means procedure yielded a solution close to those offered by Ward's method and complete linkage.

Table 3
True positive, true negative, and conditional classification rates of outside permit assignment

Clustering algorithm	true positives	true negatives	PPP	NPP
K-means	.81	.58	.20	.96
Ward's	.36	.77	.17	.91
Average Linkage	0	.99	0	.89
Complete Linkage	.36	.69	.13	.90
Single Linkage	0	.99	0	.89
Nearest Neighbor	0	.99	0.	89

Notes. PPP= positive predictive power; NPP= negative predictive power

The solution obtained by using average linkage, single linkage, and nearest neighbor is very poor. Clearly, these methods failed to form two groups. That is, they assigned 98 out of the 99 subjects in the study to the no-outside-permit group. They give a *zero true positive rate!*, and a 99% true negative rate. They also give a zero PPP rate, and a 89% NPP rate.

Ward's method and complete linkage yielded much better solutions. Of these two methods, Ward's method was slightly superior. The best recovery of the actual classification of the inmates was obtained by the k-means procedure. This procedure gave a 81% true positive rate, but only a 58% true negative rate. The k-means procedure also gave the better conditional classification rates. If an inmate is in cluster 2, then there is a 96% chance that he will be denied an outside permit. However, if he is classified into cluster 1, then there is only a 20% chance that he will granted an outside permit.

Conclusions

Cluster analysis methods based on a small number of predictors are able to match satisfactorily a qualitative classification of the inmates. However, cluster analysis are more efficient in determining which

inmates will be denied an outside permit than in determining those that will be granted an outside permit. It is important to realize, that this also occurs in the qualitative classification: it is easy to determine which inmates will not make good use of an outside permit; it is much more difficult to determine which inmates will make good use of it. For this reason, it would be desirable to match the cluster analytic classification not with the decision of granting or denying an outside permit, but with the performance of the inmates during their outside permits, and eventually, with their long term performance after being released.

It is important to notice that some clustering procedures failed to classify the inmates into two groups. However, when only two groups are requested the influence of outliers may cause the clustering algorithms to fail and although a method of detection of multidimensional outliers was applied to this data, no completely satisfactory method of detecting multidimensional outliers exist. Therefore, when using cluster analysis, it is important to use several clustering methods and compare their results.

Acknowledgements

This research was supported by a Fulbright-La Caixa scholarship to the author, who is now at Dept. of Statistics and Econometrics, Universidad Carlos III de Madrid, C/ Madrid 126-128, 28903 Getafe (Spain). E-mail: amaydeu@est-econ.uc3m.es

Referencias

Bollen, K.A. (1989). *Structural equations with latent variables*. New York: Wiley.

Cronbach, L.J., and Gleser, G. (1953). Assessing similarity among profiles. *Psychological Bulletin, 50*, 456-472.

Florek, K., Lukaskzewicz, J., Perkal, J., and Zubrycki, S. (1951). Sur la liason et la division des points d'un ensemble fini. *Colloquium Mathematicae, 2*, 282-285.

Gordon, A.D. (1981). *Classification*. London: Chapman and Hall.

Guillén, A., Maydeu-Olivares, A., Pons, J., and Vigil, A. (1989). *Resolución de problemas interpersonales y delincuencia: Un estudio comparativo*. [Interpersonal problem solving and delinquency: A comparative study]. Unpublished document. Departament d'Educació i Psicologia. Universitat de Barcelona.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Jain, A.K., and Dubes, R.C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

MacQueen, J.B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkely Symposium on Mathematical Statistics and Probability, 1*, 281-297.

Platt, J.J., and Spivack, G. (1975). *Manual for the Means-Ends Problem Solving Procedure (MEPS). A measure of interpersonal cognitive problem-solving skill*. Unpublished document. Philadelphia: Hanhemann Community Mental Health/Mental Retardation Center.

Roig-Fusté, J.M. (1986). *Exploración objetiva de la personalidad normal y anormal (a través del MMPI)* [Objective assessment of normal and abnormal personality]. Barcelona: Tesys.

SAS Institute (1990). *SAS/STAT User's Guide*, Version 6, Fourth Edition. Cary, NC: SAS Institute.

Sokal, R.R., and Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38*, 1409-1438.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in analyses of the vegetation on Danish commons. *Biologiske Skrifter, 5*, 1-34.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *77*, 841-847.

Widiger, T.A., Hurt, S.W., Frances, A., Clarkin, J.F., and Gilmore, M. (1984). Diagnostic efficiency and DSM-III. *Archives of General Psychiatry*, 41, 1005-1012.

Wong, M.A., and Lane, T. (1983). A k-th nearest neighbor clustering procedure. *Journal of the Royal Statistical Society*, *Series B*, *45*, 362-368.

Aceptado el 23 de marzo de 1996