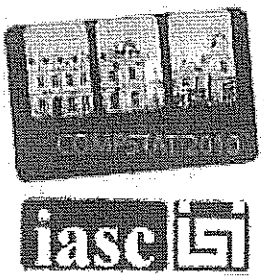
Proceedings in Computational Statistics

2010

Edited by
Yves Lechevallier
Gilbert Saporta

Physica-Verlag
A Springer Company



http://www.springer.com/978-3-7908-2603-6

Proceedings of COMPSTAT'2010
19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers
Lechevallier, Y.; Saporta, G. (Eds.)
2010, XXX, 621 p. 105 illus., Softcover
ISBN: 978-3-7908-2603-6
A product of Physica-Verlag Heidelberg

Contents

A Comparison between Beale Test and Some Heuristic Criteria to Establish Clusters Number

Angela Alibrandi[1] and Massimiliano Giacalone[2]

[1] Department of Economical, Financial, Social, Environmental, Statistical and Territorial Sciences (S.E.FI.S.A.S.T.), University of Messina, Via dei Verdi 75, 98122 Messina, *aalibrandi@unime.it*

[2] Department of Public Organization Law, Economy and Society (D.O.P.E.S), University of Catanzaro "Magna Graecia", Campus of Germaneto, 88100 Catanzaro, *maxgiacit@yahoo.it*

Abstract. In cluster analysis the individualization of the adequate clusters number represents a fundamental decision to be taken into correctly assigning the units to not-previously defined groups of observations. Many criteria have been proposed in literature in order to establish the best clusters number. Purpose of this paper is to examine the theoretical bases of some most common criteria: Beale test, based on the significance logic and two heuristic methods as Pseudo T^2 Hotelling and Cubic Clustering Criterion. Moreover, we want to compare them in terms of flexibility and applicability, taking in account the assumptions on which they are based; finally we apply all these criteria on real data and we compare the obtained results.

Keywords: clusters number, grouping structure, Beale test, T^2Hotelling, Cubic Clustering Criterion

1 Introduction

As it is known in literature, cluster analysis (Johnson, 1998) is a multivariate technique that aims to assign statistical units in not-previously defined categories, creating groups of observations in order to be homogeneous within them and heterogeneous among them. So, purpose of the clustering is to synthesize statistical units in an inferior number of clusters, maximizing the infra-groups distance and minimizing, in the same time, the intra-groups variability. Cluster analysis represents an ideal data-mining tool because the classes or groups that the data form are unknown, especially as the state definition is expanded to include an increasing number of variables. Cluster analysis uncovers these underlying patterns in the data and assigns each case to a cluster. Unlike the discriminant analysis, in the cluster analysis there isn't information about the number of cluster and the characteristics of the groups in the population. The individualization of the grouping structure constitutes, therefore, a fundamental decision to be taken. In literature various criteria (Milligan and Cooper, 1985) have been proposed to individualize

the best structure and to assess the cluster validity ((Xu Rui et al., 2008; Halkidi, 2002). In this context our paper aims to examine three of the most utilized criteria: Beale test (Gordon, 1999), based on the significance logic and two heuristic methods as Pseudo T^2 Hotelling (Halkidi, 2002) and Cubic Clustering Criterion (Sarle, 1983). Moreover, we want to compare them in terms of flexibility and applicability, taking in account the assumptions on which they are based; finally we apply all these criteria on real data and we compare the obtained results. In particular, the paper is so structured:

- in paragraph 2 the theoretical bases of the three considered criteria are exposed;
- in paragraph 3 the application of the examined criteria is shown and a comparison between the obtained rsults is performed;
- in paragraph 4 some final remarks and a discussion conclude the paper.

2 Theoretical bases of Beale test, CCC and PST2

2.1 Beale's probabilistic algorithm

Beale's probabilistic algorithm (Beale, 1969) replies to the exigency of choosing the suitable number of clusters, allowing to verify the significance of grouping. Such as it is reported in Gordon (1999), Beale test is based on a F-type statistic and allows to compare goodness of clustering with r clusters compared to $r-1$ clusters, capturing the tightness of clusters. The criterion refers to a matrix of Euclidean distances. Let's suppose to have k quantitative modalities on each of n statistic units of a population and we need to individualize a grouping of n units in r groups, with $r < n$. Let's indicate by $W(r)$ the residual sum of squares (within group), relative to a partition in r clusters. Beale test allows to verify the hypothesis according to which, proceeding from $r-1$ to r clusters, there is a significant reduction of within-groups deviance. In order to decide if a partition with r clusters has to be preferred to another with $r-1$, we can employ:

$$F = \frac{W_{r-1} - W_r}{W_r} \tag{1}$$

whose asymptotic critical region is the right tail of F distribution, with $\nu_{num} = k$ and $\nu_{den} = k(n-r)$ degrees of freedom.

Beale proposed a correction factor, indicated with C that, for large samples, is function of the attended diminution of W_r to the increasing of r

$$C = \frac{n-(r-1)}{n-r}\left(\frac{r}{r-1}\right)^{\frac{2}{k}} - 1 \tag{2}$$

The adjusted statistic test is given by the ratio between F and the correction factor C:

$$F' = \frac{F}{C} \tag{3}$$

F' keeps the same degrees of freedom [$\nu_{num} = k$ and $\nu_{den} = k(n-r)$]. This test has to be calculated for every couple $(r-1)$ and r, until we reach the significance of the test. In order to verify the above-mentioned significance, the test has to be compared with the critical value of the Snedecor -Fisher F test with $\nu_{num} = k$ and $\nu_{den} = k(n-r)$ degrees of freedom, at a fixed α level of significance.

Large values of the test indicate a better clustering solution. So, if the empirical F is greater than the critical F (i.e. the associated p-value is less than $\alpha = 0.05$), we can say that the change from $r-1$ to r clusters yields the reduction of a significant quantity of within-groups deviation and so r can be considered the optimal number of groups; this result indicates a stopping point. Otherwise, the solution with the smaller number of clusters has to be preferred.

2.2 The Pseudo T^2 statistics

Another method of judging the number of clusters is the $Pseudo-T^2$ statistic (PST^2), that is a variant of Hotelling's T^2 (Halkidi, 2002), based on the assumption that two clusters are drawn from two independent multivariate normal distributions with the same mean and covariance. PST^2 computed to compare the means of two aggregated clusters in hierarchical models and is expressed as follows:

$$PST^2 = \frac{n_1 - n_2}{n_1 + n_2}(\bar{x}_1 - \bar{x}_2)^T W^{-1}(\bar{x}_1 - \bar{x}_2) \tag{4}$$

where n_1 and n_2 indicate the number of observations into the two clusters, $\bar{x}_1$ and $\bar{x}_2$ are the mean values, respectively, and W is the unbiased pooled covariance matrix estimate. In particular, PST^2 measures the degree of separation between the two last aggregated clusters; it can't be considered as significance test because it isn't distributed exactly as t random variable. So, for its interpretation we have to examine the values that it assumes: elevated values suggest to arrest the clustering to the previous level. If the $Pseudo-T^2$ statistic value is large, the means are significantly different and so the considered clusters should not be combined; if the value is small, instead, the clusters can safely be combined. A general rule for interpreting the PST^2 is to observe all values of statistics, calculated for each number of clusters,

until a value results markedly larger than the previous one, establishing the acceptability of the partition. The choice of groups is based on the analysis of the peaks achieved by that index, typically we have to prefer a group with k +1 classes if at the k classes, the index assumes high values, followed by a sharp fall.

2.3 The Cubic Clustering Criterion

Another criterion proposed in literature in the choice of the optimal number of cluster is the Cubic Clustering Criterion or CCC (Sarle, 1983). It's based on the assumption that an uniform distribution on a hyperrectangle will be divided into clusters shaped roughly like hypercubes. The CCC aims to verify the null hypothesis according to which the clusters are hypercubes (obtained from a uniform distribution on a hyperbox) against the alternative one for which the data have been sampled from a mixture of spherical multivariate normal distributions, with equal variances and sampling probabilities. In other words, CCC is a comparative measure of the deviation of the clusters from the expected distribution, if data points were obtained from an uniform distribution. The criterion is calculated as:

$$CCC = \ln\left(\frac{1 - E(R^2)}{E(R^2)}\right)k \tag{5}$$

where $E(R^2)$ is the expected R^2 value, R^2 is the observed R^2 (ratio of "between group variance" on "within group variance", that furnishes a measure of the clustering quality) and k is the variance-stabilizing transformation.

For its interpretation, we have to consider that:

- large positive values of CCC (>3) indicate good clustering, showing a larger difference from an uniform (no clusters) distribution;
- if CCC continues to increase with the number of clusters, it may be an indication of formation of pockets of sub-clusters in more clusters;
- values between 0 and 2 indicate potential clusters; negative values indicate that grouping structure isn't appropriate;
- large negative values can indicate outliers.

This criterion seems to give good results especially on samples of high abundance, while its strength could be lower if the number of observations in each group is low. In any case, it may not be used in probabilistic terms, choosing the number of groups in correspondence to an absolute maximum or relative maximum possible. For these reasons, it's helpful to plot the CCC values calculated for each number of clusters and to look for the peaks where CCC >3. However, the CCC may be incorrect if variables are highly correlated.

For the last two criteria, which are heurustic, the information deducible from the construction of these indicators should not be interpreted in probabilistic terms, they may be used, noting the trend, so as to identify potential "natural" clusters of the considered units.

2.4 Main differences among the three examined criteria

From a methodological point of view, we can compared the examined criteria. Beale test is based on the significance logic and it is characterized by large flexibility and applicability because it's released by restrictions on assumption about the distribution of the studied variables. PST^2 statistic is based on the assumption of normality and CCC index is based on the assumption of uniformity. For both criteria these assumptions are often hardly realizable.

Moreover, either PST^2 or CCC must be calculated to every hierarchical level, whilst Beale test has to be carried until the reaching of the significance.

Finally, CCC and PST^2 are heuristic methods and they can't be analytically considered because the sampling distributions for these indexes are unknown; on the contrary, Beale test is based on F test and follows the same distribution.

3 An application to real data

In order to illustrate the utility of the above - mentioned criteria we have applied them on a real dataset. We have examined monthly data, referred to building abusiveness phenomenon in Messina, noticed by the department of Environmental Police of Messina in the year 2007, for each of fourteen districts in which the city was divided.

Our variables are represented by the count of violations to some articles of the Regional Law 10 August 1985, n. 37 "New norms in subject of urbanistic control of the activity - house building, rearranges urbanistic and confirmation of the unauthorized works":

- the first concerned the articles 5 and 9 (Administrative Sanctions);
- the second referred to article 20 (Urbanistic Law);
- the third related to sequestration (Building Sequestrations).

Before data analysis (carried on using the centroid-method, the Euclidean metric and the hierarchical classification) each variable has been standardized.

Tables 1 and 2 report the results of Beale test and the application of the two heuristic methods PST^2 and CCC, respectively.

Comparison	F	ν_{num}	ν_{den}	p-value
Cluster 3 vs 2	F=0.143	36	396	0.984
Cluster 4 vs 3	F=0.647	36	360	0.973
Cluster 5 vs 4	F=1.384	36	324	0.022

Table 1. Results of Application of Beale test

Clusters	PST^2	CCC
1	4.10	0
2	5.8	1.86
3	3.5	2.16
4	4.5	2.66
5	4.7	3.9
6	12.7	1.36
7	3.1	1.52
8	2.0	3.06
9	3.7	4.74
10	5.6	7.32
11	1.7	8.62
12	1.7	10.72
13	4.0	11.42
14	2.8	11.52

Table 2. Results of Application of CCC and PST^2 statistics

It's evident that the suitable number of clusters in this framework can be determined by the comparison between five and four clusters, as we can see in the last column of Table 1. Finally, in the last row of last column in Table 1 we note a significant p-value ($\alpha = 0.05$), that suggests us to choose five clusters as optimal partition.

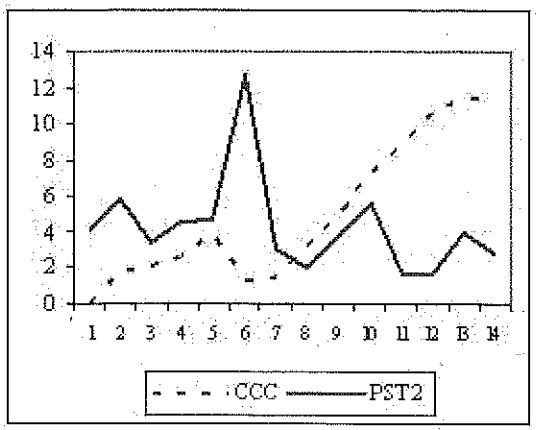


Fig. 1. Results of CCC and PST2 statistics application

Examining Table 2 and Figure 1 we can note that the largest value of PST^2 is reached for six clusters and this criterion suggests that the optimal choice is the previous level of clustering. Also CCC indicate five cluster as

best partition, because of the presence of a peach. The fourteen districts are grouped in five clusters as it is illustrated in Table 3.

Clusters	Districts
1	III - XIV - II - IV - XIII
2	V - X - VII
3	I - VIII - XII
4	VI - XI
5	IX

Table 3. Allocation of the fourteen Districts in the five Clusters

We are aware that the potential corrected solutions could be more than one. In our application, on the bases of the obtained results, we considered appropriate to use the partition into five groups, because all the applied criteria lead to the choice of this optimal clusters number.

4 Final Remarks and discussion

The ability to identify the appropriate number of clusters for a given set of data is one of the most fundamental shortcomings of non-hierarchical techniques. While local knowledge and experience can play a role in data analysis, user defined parameters such as k groups builds significant subjectivity into analysis. Furthermore, implicit to most discussions of the k-means approach there are no established methods for determining the optimal number of clusters (Levine, 1999). In fact, there are many and many methods outlined in the statistics literature detailing potential methods for detecting the appropriate number of clusters (Everitt 1979; Gordon 1998; Gubresic, 2006; Lozano J.A. et al (1986), Milligan and Cooper, 1985). Three of the more effective procedures for determining the number of clusters are examined in our data set: the CCC (test statistic provided by the SAS package), the PST^2 statistic and the Beale test. The CCC column of SAS output has been analyzed from n groups to 1 group. These inflection points are indicative of appropriate cluster groupings for the data. Moreover, we observed more than a single inflection point. Alternatively, graphic plots of CCC values have been utilized for our analysis. The CCC values were also used in conjunction with pseudo F (PSF) and t^2 statistics in our application. Both measures provide additional information, with large PSF values suggesting a good stopping point. Inflections in the t^2 statistic also suggest possible cluster stops. On the bases of the jointly use of the three examined criteria (CCC, PST^2 statistics, and Beale test), we can note that all of them confirm that the best partition of our observations can be reached by grouping in five clusters.
Comparing the three criteria, we can retain that Beale test, based on the significance logic, represents the most flexible and applicable compared to the

others: in applications the assumptions of normality of PST^2 statistic and uniformity of CCC index are often hardly realizable, so Beale test is more extendible because it's released by restrictions on assumption about the distribution of the studied variables. Both PST^2 and CCC must be calculated to every hierarchical level, while Beale test has to be carried on until the reaching of the significance, that represents the stopping point.
Moreover, CCC and PST^2 are heuristic methods and they are rather lacking because the sampling distribution for these indexes are unknown; on the contrary, Beale test is based on F test, assumes the same hypothesis and follows the same distribution, so it's a significance test that inferentially has to be preferred than the other criteria of choice.

References

BEALE E. M. L.(1969): Euclidean Cluster Analysis. Bulletin of International Statistical Institute. In: *Proceedings j of the 37th Session.*

EVERITT B.S.(1979): Unresolved problems in cluster analysis. *Biometrics, 35 (1), 169-181.*

GORDON A.D. (1998): *How Many Clusters? An Investigation of Five Procedures for Detecting Nested Cluster Structure.* Data Science, Classification and Related Methods. Springer-Verlag.

GORDON A.D. (1999): *Classification.* 2nd edition. Chapman and Hall, 60-65.

GRUBESIC T.H. (2006): On the application of fuzzy clustering for crime hot spot detection. *Journal of Quantitative Criminology Springer Netherlands 22(1),77-105.*

HALKIDI M., BATISTAKIS Y. and VAZIRGIANNIS M.(2002): On clustering validation techniques. *Journal of Intelligent Information System, 17, (2), 107-145.*

JOHNSON D.E.(1998): Cluster analysis. *Applied Multivariate Methods for Data Analysis,* Duxbury Press, *319-396.*

LEE K. M., HERRMANA T. J., LINGENFELSERA J. and JACKSON D. S. (2005): Classification and prediction of maize hardness-associated properties using multivariate statistical analyses. *Journal of Cereal Science (41),85-93.*

LEVINE N. (1999): *CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations.* Washington DC: Ned Levine and Associates; National Institute of Justice.

LOZANO J.A., LARRANAGA P. and GRANA M. (1996): Partitional cluster analysis with genetic algorithms: searching for the number of clusters. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba: *Data Science, Classification and Related Methods,* Tokyo, Springer-Verlag.

MILLIGAN G.W. and COOPER M.C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika (50), 159-179.*

SARLE W.S. (1983): Cubic Clustering Criterion. *SAS Technical Report,(1), p.108.*

XU RUI, II DONALD C. WUNSCH (2008): Recent advances in cluster analysis, *International Journal of Intelligent Computing and Cybernetics,* Emerald Group Publishing Limited.