Iterative Partitioning of a Set of Judges

Disaggregazione Iterativa di un Insieme di Graduatorie

Agostino Tarsitano
Dipartimento di Economia e Statistica - Università degli Studi della Calabria
agotar@unical.it

Summary: iterative relocation is used to partition a sample of *n* observers (for each of which the ranking of the same set of *m* items is considered) into g homogeneous and nonoverlapping clusters. The classification is obtained by using the global coefficient of concordance as a quality evaluator and trasfers director.

Keywords: nonhierachical classification, coefficient of concordance, transfer algorithm

1. Introduction

Consider a fixed set of of *m* object $\boldsymbol{a}_t=(a_{t1},a_{t2},\ldots,a_{tm})'$ ordered according to the different degree in which they possess a common attribute. Let the object be thoroughly mixed and then rearranged by an observer in order of merit according to his/her judgement and let *n* be the number of judges and assume that each judge ranks indipendently (at least, stochastically) of the other judges. There are cases in which it is more important to rank some entity correctly than to correctly rank others (for example, it is more satisfactory the placing of the winner in a race in the first rank than the placing of the worst contestant last). In other cases would be useful obtain explicit information about subsequences of objects on which judges have similar views. The purpose of the present paper is to outline a new algorithm for clustering multivariate ordinal data which is able to discover wether all the judges can be considered as belonging to the same population, if there exist judges markedly remote from the others or different subgroups among the judges.

2. Assessment of cluster homogeneity using Kendall's coefficient

It has been observed (e.g. Gibbons, 1976, p. 303) that there is no such thing as perfect disagreement for more than two sets of ranking. If, say, the first set of ranks is arranged in natural order and the corresponding ranks of the second set appear in reverse natural order, there is no way that the third set can be in complete disagreement with both of the first two set of rankings. In this sense, the criterion guiding the new classification procedure is oriented toward cluster homogeneity rather the distance between centroids. The most popular measure of the agreement is the Kendall coefficient of concordance

$$K=\alpha\frac{\sum_{r=1}^{m}S_r^2}{n}-\beta;\qquad S_r=\sum_{t=1}^{n}a_t;\quad \alpha=\frac{12}{m\left(m^2-1\right)};\ \beta=\frac{3(m+1)^2}{m^2-1}$$

which is defined as the sum of squares of then *m* totals of the *n* rank values for each object about the mean *n(m+1)/2* divided by the maximum of this expression: $n^2(m^3-m)/12$, which occurs if all the *n* rankings agree; *K* would equal one for perfect agreement and tend toward zero for no agreement, but it can never be negative. Kendall noted that , if $\bar{\rho}$ stands for the mean of the Spearman's rank correlation coefficient between the *n(n-1)/2* pairs of rankings, then $\bar{\rho} = (nK-1)/(n-1)$. Iterative schemes for cluster analysis are concerned with making membership changes which optimize a numerical criterion. In this paper we quantify the overall resemblance of the judges with

$$W^{(q)} = \sum_{i=1}^{g}\left(\frac{n_i}{n}\right)K_i^{(q)} \qquad where \qquad K_i^{(q)} = \alpha\frac{\sum_{r=1}^{m} {}_q S_{ir}^2}{n_i^q} - \beta; \quad n = \sum_{i=1}^{g} n_i$$

$K_i^{(q)}$ accounts for the concondance among the judges belonging to cluster *i* at the q-th iteration. Since *W* is a weighted average of the $\{K_i\}$ it ranges from the lowest to the highest group concordance coefficients. If $g_2 > g_1$, the final value of criterion for g_2 groups $W(g_2)$ is greater than $W(g_1)$ computed for g_1 groups; however, violation have been encountered in some data sets.The goal of the algorithm is to obtain homogenous groups of judges for which all group members exhibit significant correlation among themselves and low correlation with judges in other groups. Ehrenberg (1952) is skeptical on the Spearman approach and develops an alternative measure of the agreement for multivariate ordinal data. However, Tarsitano (2001) has found the Spearman's *r* the most flexible among several measures of pairwise concordance.

3. A new algorithm for clustering ordinal data sets

To determine a clustering of the judges $\boldsymbol{g}=(g_1, g_2, \ldots, g_n)'$ with $g_i=j$ if $\mathbf{a}_i$ belongs to the j-th cluster I propose an heuristic method which, as it is customary in the iterative partitioning framework, has three essential features: starting the process, reassigning entities, overcoming local optima.

There are a wide variety of techniques available for choosing the first *g* leaders where *g* is assumed to be known. In the present study we use the method suggested by Kennard and Stone (1969). The first two leaders are selected by choosing the two judges that are farthest apart. Let $P=\{P_1, P_2, \ldots, P_g\}$ be the set of the *k* judges that have been chosen as leaders; the *(k+1)-th* leader is chosen from among the remaining (*n-k)* candidates using

$$\Delta_{k+1}^2 = \underset{i\notin P}{Max}\left\{\underset{s=1,2,\ldots,g}{Min}\left[\sqrt{\sum_{r=1}^{m}\left|a_{ir}-p_{sr}\right|^2}\right]\right\}$$

The first initialization of the membership vector $g^{(0)}$ is derived according to the rule

$$\gamma_i^{(0)} = j \;\; if \;\; \sqrt{\sum_{r=1}^{m}\left|a_{ir}-p_{jr}\right|^2} = \underset{s=1,2,\ldots,g}{Min}\left\{\sqrt{\sum_{r=1}^{m}\left|a_{ir}-p_{sr}\right|^2}\right\};$$

The central step of an iterative relocation scheme is the way in which entities are moved from one cluster to another. Let $W^{(q+1)}$ the pooled concordance coefficient after that transfer of a_t from cluster j to cluster i has taken place. The effect is

$$\Delta W = \frac{\alpha}{n}\left\{\frac{\sum_{r=1}^{m} S_{ir}^2}{n_i(n_i-1)} - \frac{\sum_{r=1}^{m} S_{jr}^2}{n_j(n_j+1)} + \frac{(n_i+n_j)\sum_{r=1}^{m} a_{tr}^2}{(n_i-1)(n_j+1)} + \frac{2\sum_{r=1}^{m} a_{tr}\left[(n_i-1)S_{jr} - (n_j+1)S_{ir}\right]}{(n_i-1)(n_j+1)}\right\}$$

If $\Delta W>0$ then $W^{(q+1)} > W^{(q)}$. It seems reasonable exhamining the potential effects of switching a_t into every cluster (except the one it is in) and finding the greatest value satisfying $\Delta W>0$. Thus each entity is tranferred (if transferred) to the cluster which maximizes the impact of the transfer. This step would have the drawback of preventing more effective transfers involving cluster i and j; moreover, the results depend on the order the entities are considered one by one. A better procedure would be performing a complete scanning of the entities and all candidate transfers for which $\Delta W>0$ are retained. They are then sorted in ascending order of magnitude and executed starting with the first, but excluding those affecting clusters already interested in a transfer. Bansfield and Bassill (1977) have suggested that a local maximum might be circumvented by swopping two entities a_t and a_s with $g_t=i$ and $g_s=j$, $i\neq j$. The effect on our criterion is

$$\partial W = \frac{\alpha}{n}\left[\left(\frac{n_i+n_j}{n_i n_j}\right)\sum_{r=1}^{m}(a_{tr}-a_{sr})^2 + \left(\frac{2}{n_i n_j}\right)\sum_{r=1}^{m}(a_{sr}-a_{tr})(n_j S_{ir} - n_i S_{jr})\right]$$

All the candidate swaps are tested in turn and those for which $\partial W>0$ are executed with the selective procedure used for the transfers. The algorithm produces a clustering which is only locally optimal, given the starting configuration. The aggregate Kendall concordace coefficient may not be increased neither by trasferring a judge from one cluster to another nor by exchanging the cluster to which a pairs of judges belong; however, different partitions may have the same or greater values of W.

4. Experimental results

A few experitments have been made in a preliminary attempt at validation of the new algorithm (in the following it should be remembered that we are investigating the rankers and not the items ranked). First, I re-analyze the data presented in Rizzi and Badaloni (1972). The first data set is an artificial example including 35 complete rankins of 10 objects supposed to have 6 groups according to a density search method. Fig.1 shows the quantity $Ln(W_{i+1}/W_i)$ plotted against the number of clusters (also known as log-scree plot). A pronounced peak in the graph indicates that $g=i+1$ is a good candidate value for the correct number of clusters. This would suggest $g=5$ or $g=7$ for the first data set; in particular, the structure for $g=7$ is very similar to the classification found by the two authors. The second example is a study involving 121 students who rated 14 profession according to their social prestige The solution g=10 proposed by Rizzi and Badaloni is compatible with the graph although a lower number of clusters seems more plausible.

Figure 1: *Log-scree plot for the Rizzi Badaloni 1st data set*

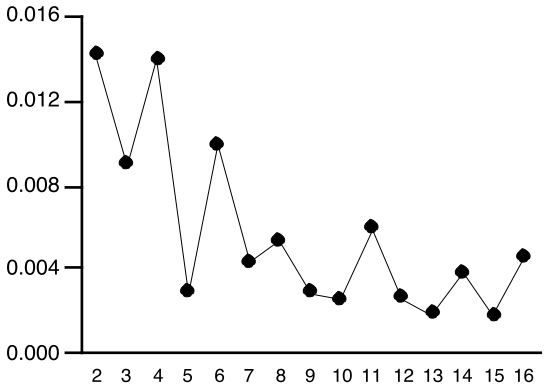


Figure 2: *Log-scree plot for the Rizzi Badaloni 2nd data set*

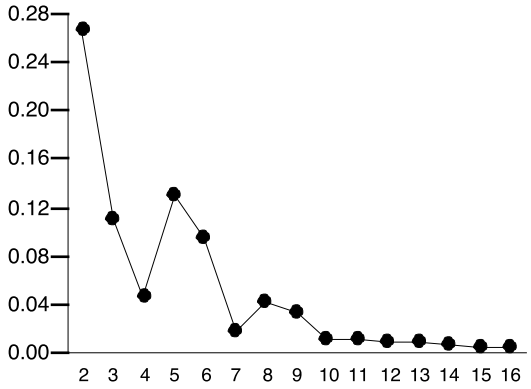


Figure 3: *Log-scree plot for required courses data set*

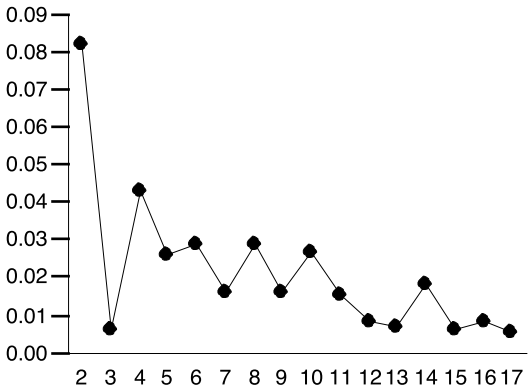
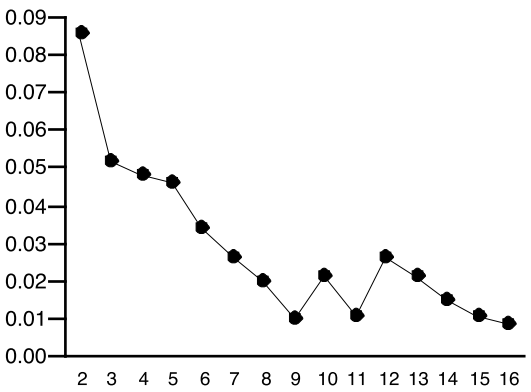


Figure 4: *Log-scree plot for a four group artificial data set*



The third data set includes 67 rankings of the 14 required courses for a degree in Economics according to the student's satisfaction with the grade received. The appropriate number of clusters is not known; however g=2 appears to be consistent with the study conducted by Tarsitano (2001) in which the student evaluations are divided according to their mean grade (lower than 26/30 and equal or greater than 26/30). The fourth examples is a concocted data set presenting a strong four group structure fairly well exposed in Fig. 4.

References

Bansfield C.F. Bassill L.C.(1977) Algorithm AS113: a transfer algorithm for non-hierarchical classification. *Appl. Statist*, 26, 206-210.

Dickinson J.D.(1976) *Nonparametric methods for quantitative analysis*, Holt, Rinehart and Winston, New York.

Ehrenberg A.S.C.(1952) On sampling from a population of rankers. *Biometrika*, 39,82-87.

Kennard R.W. Stone L.A.(1969) Computer aided design of experiments. *Technometrics*, 11,137-148.

Rizzi A. Badaloni M.(1972) Contributi alla cluster analysis. *Metron*. 30, 154-208.

Tarsitano A. (2001) Clustering gerarchica dei corsi obbligatori ad Economia. Submitted for pubblication.