Weighted rank correlation and hierarchical clustering

Raggruppamento gerarchico basato su correlazioni di ranghi pesati

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

Riassunto: si adopera la classificazione gerarchica aggregativa per classificare *n* giudici (per ognuno dei quali si considera la graduatoria completa proposta su *m* item) in gruppi disgiunti ed omogenei. Lo schema si basa su matrici di dissimilarità ricavate da coefficienti di correlazione ponderata che sono in grado di cogliere meglio di quelli tradizionali certi profili di giudizio e consentono la definizione di gruppi più significativi di giudici.

Keywords: agglomerative classification, weighting attributes

1. Introduction

Measuring agreement between two set of rankings is an issue frequently encountered in many research studies. Classic fields where rank data occur are market segmentation, consensus formation, scales of symptoms and feelings, information retrieval. The situation considered this paper is as follows: a fixed set of *m* item is arranged in order according to the different degree in which they possess a common attribute with 1 assigned to the most preferred item, 2 to the next-to-most preferred and so forth by *n* judges $\{O_1, O_2, \ldots, O_n\}$. Gaps or ties are not allowed: a ranking is simply a permutation of the integers *1* through *n*. Each judge ranks independently of the other judges, but there is reason to believe that there exist different subgroups among the *n* judges.

Cluster analysis (CA) is the procedure by which we objectively group together rankings on the basis of their differences and similarities. A crucial issue of CA is to decide wether rankings should be described by either a pattern matrix or a dissimilarity matrix. A pattern matrix is a *(nxm)* matrix $\boldsymbol{S}=(s_{ij})$ where s_{ij} denotes the rank given by the i-th judge to the j-th object; a dissimilarity is a *(nxn)* matrix $\boldsymbol{D}=(d_{ij})$ where d_{ij} measures the degree of the interjudge closeness (usually, the latter is derived from the former). The most common outputs of CA are a partition of rankings (which is applied to $\boldsymbol{S}$) and a hierarchical classification (which is applied to $\boldsymbol{D}$). The problem of iterative partitioning a set of judges into disjoint types may be investigated by using a k-means algorithm based on a simple function of the average Spearman's ρ. Alternatively, the intergroup agreement can be evaluated by the average Kendall's τ.

The purpose of the present study is to obtain a successively-nested set of partitions based on a dissimilarity matrix compatible with ranked data. The contents of the various sections are as follows. Section 2 reviews the general formulation of weighted rank correlations in which the incorporation of a weight function allows more flexibility in the classification. In Section 3 rank correlations are transformed in dissimilarity coefficients. In addition, the choice of an agglomerative algorithm is briefly discussed. Finally, in Section 4, we show experimental results and summarize our conclusions.

2. Weighted rank correlations

Researchers in many fields have become increasingly aware that there are certain patterns of resemblance which may reflect significant facets of the association between two rankings. In evaluating these relationships one would require a classification method that captures the specific pattern and is not unduly affected by few possibly incongruous data. Quade and Salama (1992) showed that statistical methods for measuring association when the magnitude of intercategory distances cannot be ignored, group naturally in two classes: weighted rank correlation (w.r.c.) and correlation of scores. In this note, we pursue the first approach. In particular, we analyze a weighted version of the Spearman's ρ

$$r_{ij}(w) = \frac{E(pq) - E(p)E(q)}{\sqrt{E\left[p^2 - E(p)\right]E\left[q^2 - E(q)\right]}}; \quad p = \sum_{k=1}^{m} w_k x_k, \quad q = \sum_{k=1}^{m} w_k^2 x_k \tag{1}$$

where x_k is the k-th rank given by judge O_i after that the ranks given by judge O_j have been arranged in their natural order. Vector $\boldsymbol{w}=\{w_1, w_2,\ldots,w_m\}$ is a monotone system of weights. The expectations *E(.)* is taken over all *m!* permutations. Two special cases of (1) are known in the literature. The Mango index (mango, 1997) is obtained by using $w_i=i^2$, $i=1,\ldots, m$

$$r_M = 1 - \frac{3\left\lfloor m^2(m+1)^2 - 4\sum_{k=1}^{m} k^2 x_k \right\rfloor}{m(m-1)(m+1)^2} \tag{2}$$

This measure is based on the sum of the second order minors extracted from the *(mx2)* matrix having **x** as first column and the natural ordering as second column. The Blest index (Blest, 2000) is obtained by using $w_i=(m-i+1)^2$, $i=1,\ldots,m$

$$r_B = 1 - \frac{\left\lfloor 12\sum_{k=1}^{m}(m+1-k)^2 x_k - m(m+1)^2(m+2) \right\rfloor}{m(m-1)(m+1)^2} \tag{3}$$

Which is based on the differences between the accumulated ranks of the two judges. Although the two procedures appear to be entirely different, each of r_M and r_B can be derived from the other. Furthermore, $(r_m + r_b)/2=\rho$ *i.e.* the Spearman's coefficient. The values of r_M and r_B range between -1 and 1. The former occurs between s_i and $=m-s_i+1$ for $i=1,\ldots, m$ whereas +1 occurs only in the case of maximum concordance. A value near to zero indicates no association between the rankings. The index of Mango and Blest act as complementary rank order statistics in cases of limited resource allocation because ascribing higher importance to one item reduces the importance of another. For example, it is more satisfactory the placing of the winner in a race in the first position than the placing of the worst contestant last (ceiling effect). In other cases differences in low ranks would seem more critical (floor effect). For example, when an admission office expunges the less qualified candidates. The coefficient r_B is more sensitive to floor effects because weighs the front ranks heavier than the back ranks. The r_m reacts more to ceiling effects. Another interesting feature is that their values are synchronized to quantify bipolarity condition that is comparisons in which the top-down and the bottom-up process simultaneously affect the same attribute giving rise to a bidirectional effect.

3. Correlations, distance measures and agglomerative clustering

The basic assumption of this paper is that a rank correlation between judge *i* and the *j* can be used to quantify the similarity/dissimilarity between them. Since $-1 \leq r_M, r_B \leq 1$ these coefficients have to be transformed into dissimilarities in the interval (0,1). Anderberg (1973) suggested a linear transformation of the Spearman's footrule which also corresponds to the general coefficient of similarity proposed by Gower (1971) for rank data. Kaufman and Rouseeuw (1990) proposed a linear transformation of the of the Spearman's ρ. Rank correlations can be mapped to distances using the fact that the matrix $\boldsymbol{R}=(r_{ij})$ is positive definite or positive semi-definite. Consequently, the transformation $d_{ij}(w)=[1-(r_{ij}(w)+1)/2]^{0.5}$ result in metric dissimilarity matrices for each weight function. The use of correlations in CA is in general controversial, but appears legitimate for rankings since each judgement is being averaged over homogeneous attributes.

An agglomerative clustering starts with each judge forming a separate group. It successively nests the group close to one another until all of the groups are merged into one or until a stopping rule is satisfied. Numerous agglomerative algorithms are reported in exploratory data analysis. They differ mainly in their definition of intergroup dissimilarity (link). The focus of this section is to use monotone invariant procedures in which the construction of the particular hierarchy depends solely on the rank order of the dissimilarity. While other links are possible, the type of data we are using precludes somewhat methods using average of correlations (UPGMA, UPGMC, WPGMC, Ward). In particular, we have employed the complete-link method because it is expected to identify stereotypal judge types and avoids chaining effect.

4. Experimental results and conclusions

Test data were generated in two steps. In the first, we have defined six pivot permutations

Natural order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Inverse order	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
Floor effect	1	2	3	4	5	15	14	13	12	11	10	9	8	7	6
Ceiling effect	10	9	8	7	6	5	4	3	2	1	11	12	13	14	15
Direct bipolarity	1	2	3	4	11	10	9	8	7	6	5	12	13	14	15
Reflected bipolarity	15	14	13	12	5	6	7	8	9	10	11	4	3	2	1

Then, we have generated 14 rankings by swapping the ranks of two randomly selected positions of each of the pivot permutations. A total of n=90 "judges" have been obtained. The following table reports the alternative classifications of rankings into six groups, assuming dissimilarities based on Mango and Blest indices and two other commonly used rank correlation coefficients: Spearman's ρ and Kendall's τ. Tηε last row shows the corrected Rand index of proximity between the "true " partition and that achieved by the method. Cell entries are the rankings which have been moved from one group to another.

Table 1: alternative classifications of n=90 rankings

Type	Cluster	Mango	Blest	Spearman	Kendall
Natural order	1-15	+62	+62	+62	+62
		-6,-9,-13	-6,-12	-6,-12	-12
Inverse order	16-30	+79,+90	+81	+81	+81,+82
		-26,-29	-17,-21	-17,-21	
Floor effect	31-45		+48,+87		
Ceiling effect	46-60	+9,+13,+67	+12	+12	+12
		+70	-48	-48	
Direct Bipolarity	61-75	+6	+6	+6	
		-62,-67,-70	-62	-62	-62
Reflected Bipolarity	76-90	+26,+29	+17,+21	+17,+21+48	
		-79,-90	-81,-87	-81	-81,-82
C.Rand index		0.7495	0.7941	0.8186	0.8939

It is full evident that ρ and τ show scarce discriminating power over the simulated rankings (here, high values of the Rand index have a negative interpretation). In contrast, r_M and r_B allow reassigning several judges. For instance, according to the Mango index, judge$O_{79}=(15,14,6,12,5,13,7,8,9,10,11,4,3,2,1)$ and judge $O_{90}=(15,14,6,12,5,13,7,8,9,10,11,4,3,2,1)$ supposed to be affected by reflected bipolarity, have been moved to a cluster characterized by inverse ordering because r_M gives more weight to agreement between back ranks than to agreement between front ranks. It is important to realize that the choice of a w.r.c. coefficient presupposes the existence of particular type of clusters. Of course, there are no standard rules as to how the weights of $r_{ij}(w)$ should be chosen. However, the characteristics of a w.r.c. would simplify the selection of dissimilarity measures and increase the generality of research findings.

References

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

Blest D.C. (2000). Rank correlation - An Alternative measure. Australian and New Zealand Journal of Statistics. 42, 101-111.

Gower J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics, 27, 857-854.

Kaufman L., Rousseeuw P.J. (1990). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons, New York.

Mango, A. (1997), Rank correlation coefficients: A new approach. Computing Science and Statistics. Computational Statistics and Data Analysis on the Eve of the 21st Century. Proceedings of the Second World Congress of the IASC, 29, 471-476.

Quade D., Salama I.A. (1992). A survey of weighted rank correlation. P.K. Sen and I.A. Salama (Eds.). Order statistics and nonparametrics: theory and applications. Elsevier Science Publishers B.V., Amsterdam, 213-225.