A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 123, 1992

*Uwe Partzsch**

ABOUT THE EVALUATION OF PARTITIONS IN CLUSTER ANALYSIS

1. INTRODUCTION

It is often required in cluster analysis to compare independently determined partitions of the same set of data. Empirical measures for comparing the solutions of cluster algorithms were introduced mainly in the past twenty years. An excellent summary has been given by K r a u t h (1983). Most of these measures are only empirical ones e.g. the coefficient according to R a n d (1971), K l a s t o r i n (1980), R o h l f (1971), F o w l k e s (1980). In a large number of studies, alternative classifications are compared simply by subjective visual examination.

In this paper, statistical methods will be outlined for comparing classifications in which the investigated objects are fitted into nonoverlapping classes. A statistical method based on the general nonparametric test strategy worked out by M a n t e l (1968) is presented for testing the agreement of two cluster partitions of one and the same set of observed objects $S = \{o_1, \ldots, o_n\} \subset R^p$ with the usual assumptions for disjunct partitions

$$\mathcal{P}_m = \{P_{m.1}, \ldots, P_{m.k}\}, \qquad m = 1, 2$$

with k_m clusters $P_{m.1}, \ldots, P_{m.k}$ in the partitions $\mathcal{P}_m$ $(m = 1, 2)$ (see B o c k, 1974).

Contrary to the generally used heuristic measures of simila-

* Lecturer at the High School of Economics, Berlin.

rity, we examine here the possibility of testing the hypothesis of correspondence by chance for the two partitions. The inference procedure depends on some common nonparametric notions. Appropriate mean and variance formulae are presented for two special measures of similarity. Furthermore, some aspects of the asymptotic behaviour are discussed for the distribution of the proposed statistics under the null hypothesis. For lack of the general result of normal approximation for the probability distribution of the statistic under the null hypothesis, some conservative approximate tests for the index of similarity are introduced on the basis of known probability inequalities. This is illustrated by a simple example.

2. THE GENERAL TEST STRATEGY

To develope the comparison procedure we have to proceed from several assumptions:

At first, we assume that the two independently determined partitions $\mathcal{P}_m$ are both disjunct partitions of the same set of data. That means

$$\bigcup_{f=1}^{k_m} P_{m,f} = S \text{ with } P_{m,j} \cap P_{m,l} = \emptyset \qquad j \neq l; \quad j, l = 1, \ldots, k_m, \quad m = 1, 2.$$

Furthermore, it is assumed that the information about the classification of the n objects may be reduced to single numerical values defined for each pair of objects in the form of the elements in a so-called structure matrix. That means for each pair of objects it is only important to know something about the membership of the objects in both partitions. So we assume that there are defined structure matrices for the two partitions

$$A(P_1) = [a_{ij}] \text{ and } B(P_2) = [b_{ij}], \; A(P_1), B(P_2) \in R^{nxn}$$

which are constructed on the relationships of the objects $\{O_1, \ldots, O_n\}$ in the two cluster partitions P_1 and P_2, respectively.

Finally it is assumed, as one of several obvious possibilities, that all possible partitions with the same cluster structure (the same number of clusters and the same number of objects in each cluster) as the two given partitions P_1 and P_2 are equally pro-

bable. This uniform distribution over the partitions with equal number of clusters and equal number of objects in these k_m ($m = 1, 2$) clusters has certain advantages. With these assumptions some possibilities of approximation can be found for carrying out the desired significance test. For the statistic

$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot b_{ij} \qquad (1)$$

which has been defined on the basis of the structure matrices $A(P_1)$ and $B(P_2)$ we construct then a significance test for the null hypothesis of correspondence by chance vs. the alternative hypothesis of agreement between the two partitions.

Therefore, we have to know the distribution of the statistic L under the null hypothesis. For calculating this distribution we take the structure matrix $A(P_1)$ as fixed and permutate the rows and columns of $B(P_2)$ simultanously. So we get n! equally probable structure matrices $B_p(P_2)$, $p = 1, \ldots, n!$, which have the same elements as $B(P_2)$. These values stand only in other places in the matrices $B_p(P_2)$.

From these n! pairs $((A(P_1), B_p(P_2)), p = 1, \ldots, n!)$ we compute the probability distribution of L under the null hypothesis. On the obtained distribution of L under the hypothesis that all relabellings have the same chance of occurrence (i.e. under the hypothesis of randomness) we found the critical value for rejecting the hypothesis of correspondence by chance. Consequently, if the original calculated value of L is sufficiently extreme with respect to the probability distribution of the statistic L under the hypothesis of random relabelling, the hypothesis is rejected and the value of L is assumed to denote a significant degree of common structure in the two cluster partitions, which is reflected in the structure matrices $A(P_1)$ and $B(P_2)$. So, one can adopt this approach worked out by Mantel as a simple statistical procedure for evaluating two given cluster partitions. Unfortunately, in practice it is often difficult to compute the permutation distribution of the statistic L. For instance, the number of objects n is too large for computing any resultant values of L for building up the permutation distribution in many fields of application.

3. SOME ASPECTS OF THE DISTRIBUTION OF THE STATISTIC L

For computation of the distribution of the statistic L under the null hypothesis we have to calculate n! values of L in the form of term (1) from the n! structure matrices $((A(P_1), B_p(P_2)), p = 1, \ldots, n!)$. For instance, it will be necessary for only $n = 20$ objects to compute more than 2.43×10^{18} values of L, and therefore relabel the elements of matrix $B(P_2)$. Consequently some possibilities of approximation have to be found for carrying out the desired significance test. Application of the central limit theorems may be considered as a loophole to get information about the asymptotic distribution of L in the case of large sample numbers. But according to P a r t z s c h (1985), the usual assumption of independence of the random variables defined in (1) in general is not true. So the use of the central limit theorems which is based on the independence of these variables in the sum is not possible.

The use of central limit theorems with weakly dependent random variables, as introduced by I b r a g i m o v and L i n n i k (1971) and by B i l l i n g s l e y (1968), has to be examined with respect to the degree of dependence between the variables defined by the structure matrices. In other words, the application of these theoretical tools depends on the special definition of the matrices $A(P_1)$ and $B(P_2)$ and on the structure of the two partitions (number of clusters, number of observations in general and in each cluster).

As a second obvious alternative, a random sample of the n! permutation matrices may be drawn (with replacement) and, on the basis of these n_a random samples, an approximation to the exact permutation distribution may be constructed by using the observed values of the statistic L in this random sample of size n_a. To be sure of getting a good approximation, we use the lemma proved by P a r t z s c h (1985):

LEMMA. Let $(1 - \alpha_o)$ be an arbitrary significance level. For α_a we assume

$$\alpha_a = \frac{\text{int}[n_a \cdot \alpha_o + 2.33 \cdot \sqrt{n_a \cdot \alpha_o \cdot (1 - \alpha_o)}] + 1}{n_a} \qquad (2)$$

where $n_a = f(\alpha_o, \alpha_a)$ is an integer one. Furthermore, let $1_{(1-\alpha_o)}$ and $1^{n_a}_{(1-\alpha_a)}$ be the $(1 - \alpha_o)$ - and $(1 - \alpha_a)$ quantiles of the permutation and the approximation distribution under the null hypothesis, respectively. Then

$$\Pr(L \geq 1^{n_a}_{(1-\alpha_a)} \;/\; L \geq 1_{(1-\alpha_o)}) \approx 0.99 \qquad (3)$$

That means, if we assume L to be significant at the level $(1 - \alpha_o)$ in the original permutation distribution, then it is true with 0.99 probability that the statistic L is also significant at a level of not less than $(1 - \alpha_a)$ in the approximated distribution based on the random sample of size n_a.

In our computations we chose $n_a = 1000$. We then get, for instance, the relationship

$$\Pr(L \geq 1^{1000}_{0.982} \;/\; L \geq 1_{0.99}) \approx 0.99.$$

That means: if L is significant at the level of 0.99, then the significance level of L in the approximation distribution is not less than 0.982 with a probability of 0.99. Computation with the approximation distribution has proved that great caution should be taken by any researcher wishing to rely on the adequateness of a normal approximation, especially if the sample size n is small. Simulation studies $(n \leq 40)$ often yield a nonnormal approximation of the distribution of L using structure matrices definitions as in equations (9), (10), and (11), (12), respectively.

With the lack of the general result of a normal approximation for probability distribution, simple inequalities according to Cantelli or Csebyshev may be used to provide a significance level for the proposed hypothesis test.

Thus, we get a parameterfree conservative test by using the inequality according to Cantelli

$$Pr((L - E(L)) \geq z \cdot D(L)) \leq \frac{1}{(1+z^2)} \tag{4}$$

or the inequality according to Chebyshev

$$Pr(|L - E(L)| \geq z \cdot D(L)) \leq \frac{1}{z^2} \tag{5}$$

for the one-tailed or the two-tailed test, respectively. The computed probabilities can be compared then with an arbitrarily chosen significance level.

The mean and variance formulas under the null hypothesis are obtained on the basis of the randomization model:

$$E(L) = [n(n-1)]^{-1} \cdot a_1 \cdot b_1 \tag{6}$$

$$\begin{aligned} D^2(L) = {} & [n(n-1)(n-2)(n-3)]^{-1}(a_1 - 4a_2 + 2a_3) \cdot \\ & \cdot (b_1 - 4b_2 + 2b_3) + 4[n(n-1)(n-2)]^{-1} \cdot \\ & \cdot (a_2 - a_3) \cdot (b_2 - b_3) + 2[n(n-1)]^{-1} \cdot a_3 b_3 - \\ & - (E(L))^2 \end{aligned} \tag{7}$$

where the parameters a_1, a_2, a_3 and the corresponding b_1, b_2, b_3 (from the structure matrix $B(P_2)$) are computed in the following way:

$$\begin{aligned} a_1 &= \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\Big)^2, & b_1 &= \Big(\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\Big)^2 \\ a_2 &= \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} a_{ij}\Big)^2, & b_2 &= \sum_{i=1}^{n}\Big(\sum_{j=1}^{n} b_{ij}\Big)^2 \\ a_3 &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2, & b_3 &= \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2 \end{aligned} \tag{8}$$

4. SPECIFICATION OF L FOR TWO SPECIAL DEFINITIONS OF THE STRUCTURE MATRICES

From (6), (7) and (8) it follows that the distribution parameters of L are functions of the special defined structure matrices $A(P_1)$ and $B(P_2)$. The specification of L will be provided now for two special definitions of structure matrices. Contrary to the measure, provided by Rand, we get a measure in which the

agreement between clusters of small size has a greater effect on the size of the index of correspondence if the elements of the structure matrices are defined in the following way:

$$a_{i,j} = \begin{cases} \left[|P_1^*| \cdot \binom{|P_{1,1}^*|}{2} \right]^{-1} & \text{if } (\exists 1 \leq |P_{1,}| : o_i, o_j \in P_{1,1}^* \in P_1^*) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

$$b_{i,j} = \begin{cases} \left[|P_2^*| \cdot \binom{|P_{2,1}^*|}{2} \right]^{-1} & \text{if } (\exists 1 \leq |P_2| : o_i, o_j \in P_{2,1}^* \in P_2^*) \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where P_m^* is the set of clusters in P_m with not less then two objects ($P_m^* = \{P_{m,1}^* : |P_{m,1}| \geq 2\}$, $m = 1, 2$) and $|A|$ denotes the number of elements of set A.

With an appropriate definition of the structure matrices, especially as

$$a_{ij} = \begin{cases} 1, \text{ if } (\exists 1 \leq |P_1| : o_i, o_j \in P_{1,1}, P_{1,1} \in P_1) \\ 0 \text{ - otherwise} \end{cases} \quad (11)$$

$$b_{ij} = \begin{cases} 1, \text{ if } (\exists 1 \leq |P_2| : o_i, o_j \in P_{1,2}, P_{1,2} \in P_2) \\ 0 \text{ - otherwise} \end{cases} \quad (12)$$

one gets another special form of index 1 which is identical with the simple measure C_{SR} provided by Rand (see also (15)) and which improves the possibility of testing this simple measure of correspondence with the procedure presented here.

The following lemma (see also P a r t z s c h, 1985) may be used to simplify the computations:

LEMMA. If there does not exist a trivial partition (i.e. $|P_1| < n$ and $|P_2| < n$), the parameters in (8) can be expressed by the following formulae:

a) with the assumption of definition for the structure matrices as in (9) and (10)

$a_1 = 4$,

$$a_2 = \frac{4}{(k_1^*)^2} \sum_{l=1}^{k_1^*} |P^*_{1,l}|^{-1}$$

$$a_3 = \frac{2}{(k_1^*)^2} \sum_{l=1}^{k_1^*} \binom{|P^*_{1,l}|}{2}^{-1} \qquad (13)$$

where $k_1^* = |P_1^*|$ and

h) with the assumption of definition for the structure matrices as in (11) and (12):

$$a_1 = (\sum_{l=1}^{k_1} |P_{1,l}|^2 - n)^2,$$

$$a_2 = \sum_{l=1}^{k_1} |P_{1,l}|^3 - 2 \sum_{l=1}^{k_1} |P_{1,l}|^2 + n,$$

$$a_3 = \sum_{l=1}^{k_1} |P_{1,l}|^2 - n \qquad (14)$$

where $k_1 = |P_1|$.

Analogous formulas for parameters b_1, b_2 and b_3 are given if the notation from the second cluster partition is used in (13) and (14), respectively. In this way it is very simple to compute the moments of the statistic L in dependence on the defined structure matrices, and, with the probability in equations according to Cantelli or Chebyshev, we get the desired result for the proposed significance test. This procedure may also be applied to other heuristic measures of correspondence if appropriate structure matrices can be defined in adequate form.

5. AN EXAMPLE

As a very simple example to explain the generality of the concept introduced above, let us take $S = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8, O_9, O_{10}, O_{11}, O_{12}\}$ and consider the following two cluster partitions (n = 12):

$P_1 = (1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3)$,

$P_2 = \{1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5\}$.

The second partition is a more detailed partition of the first one. For the empirical coefficients introduced on the basis of the simple contingency table, i.e.

Scheme 1

Contingency table

P_1 \ P_2	$ind_2(O_j) = ind_2(O_i)$	$ind_2(O_j) \neq ind_2(O_i)$
$ind_1(O_j) = ind_1(O_i)$	a	b
$ind_2(O_j) \neq ind_1(O_i)$	c	d

with $ind_m(O_j) = 1$ if $O_j \in P_{m,1} \in P_m$, $(m = 1, 2)$ in the form

$$C_R = (a + d) / (a + b + c + d) \quad \text{(Rand, 1971)} \quad (15)$$

$$C_{SR} = a \quad \text{(Rand, 1971)} \quad (16)$$

$$C_J = a / (a + b + c) \quad \text{(Rohlf, 1971)} \quad (17)$$

$$C_I = a / \sqrt{(a + b)(a + c)} \quad \text{(Fowlkes, 1980)} \quad (18)$$

the associated values of these measures of similarity are given as

$C_R = 0.84$,

$C_{SR} = 9.0$,

$C_J = 0.47$,

$C_I = 0.69$.

But there is no possibility for deciding about the significance of correspondence between the two observed cluster partitions. Subjective decision can only be made if we know the limits of these heuristic indices of similarity. For the proposed statistic L, the associated values are given with the structure matrices defined in (9) and (10) as

$a_1 = 4.000$, $b_1 = 4.000$, $E(L_w) = 0.03030$,

$a_2 = 0.348$, $b_2 = 0.346$, $D(L_w) = 0.02198$,

$a_3 = 0.133$, $b_3 = 0.293$, $L_w = 0.11560$,

and with the structure matrices defined in (11) and (12) as

$a_1 = 1440$, $b_1 = 324$, $E(L_{SR}) = 5.18$,

$a_2 = 128, \quad b_2 = 30, \quad D(L_{SR}) = 2.67,$

$a_3 = 38, \quad b_3 = 18, \quad L_{SR} = 18.000,$

respectively. Hence it is shown with

$$Pr(|E(L_w) - L_w| \geq z \cdot D(L_w)) \leq 0.049,$$

$$Pr(|E(L_{SR}) - L_{SR}| \geq z \cdot D(L_{SR})) \leq 0.042$$

that there is a real agreement between the two partitions at least at the significance level of 0.95. So the hypothesis of correspondence by chance is rejected at the level of 0.95.

6. DISCUSSION

The presented procedure is suitable for solving many theoretical problems of cluster analysis and practical problems, too. Here are some examples:

- finding an "optimal" cluster algorithm for special structures of the object space, defined by a mixture of distribution functions;
- obtaining more information about special properties of several cluster algorithms in simulation studies;
- controlling the stability and error robustness of cluster partitions;
- comparing two classification schemes for objects obtained from different data sets;
- comparing two classification schemes for objects obtained from the same set of data but from diverse point of view.

Further generalization and modification of the proposed procedure are available. For example, the statistic L may be defined with a nonuniform distribution on the space of the cluster partitions. Furthermore, other measures of correspondence may be defined on the basis of available structure matrices, and their behaviour has to be examined.

In future it will be interesting to examine in which cases the central limit theorems may be used and whether some improvements may be made by using other probability inequalities. The main outcome is, however, that this is an objective statistical method to compare several independently determined cluster parti-

tions. It is suitable for solving many theoretical problems in the field of cluster analysis and other applications.

REFERENCES

Billingsley P. (1968), *Convergence of Probability Measures*, J. Wiley and Sons, New York.

Bock H. H. (1974), *Automatische Klassifikation*, Vandenhoeck and Ruprecht, Goettingen.

Folwkes E. B. (1980), *A New Measure of Similarity Between Two Hierarchical Clusterings*, Meeting of the Classification Society, Boulder.

Ibragimov I. A., Linnik Yu. V. (1971), *Independent and Stationary Sequences of Random Variables*, Wolters Noordhoff, Groningen.

Klastorin T. D. (1980), *Merging Groups to Maximize Object Partition Comparison*, "Psychometrika", No. 45, p. 425-433.

Krauth J. (1983), *Evaluation von Verfahren der Automatischen Klassifikation*, "Studien zur Klassifikation", Nr. 13, p. 203-212.

Mantel N. (1968), *The Detection of Disease Clustering and Generalized Regression Approach*, "Cancer Research", No. 27, p. 209-220.

Partzsch U. (1985), *Anwendung multivariater Klassifikationsverfahren zur Erstellung kulturpflanzenartenbezogener Witterungstypen*, Berlin.

Rand W. J. (1971), *Objective Criteria for the Evaluating of Clustering Methods*, "Journal of the American Statistical Association", No. 66, p. 846--850.

Rohlf F. J. (1971), *Methods of Comparing Classifications*, "Annals of Ecology and Systematics", No. 5, p. 101-113.

Uwe Partzsch

OCENA PODZIAŁÓW W ANALIZIE WIĄZEK

W analizie wiązek często konieczne jest porównywanie wyznaczonych niezależnie podziałów w tym samym zbiorze danych. W artykule opisano statystyczną metodę testowania odpowiedniości dwóch podziałów wiązkowych przy zastosowaniu ogólnego nieparametrycznego testu Mantela (1968). Dla dwóch szczególnych miar podobieństw podane są wzory na średnią i wariancję.

Z powodu braku ogólnych wyników normalnej aproksymacji rozkładu prawdopodobieństwa zaproponowanej statystyki wprowadza się pewne tradycyjne testy przybliżone oparte na znanych w rachunku prawdopodobieństwa nierównościach. Ponadto omawia się pewne asymptotyczne własności statystyki przy założeniu prawdziwości hipotezy zerowej.