COENOSES 4 (1): 1-10, 1989

A METHOD FOR GENERATING CONSENSUS PARTITIONS AND ITS APPLICATION TO COMMUNITY CLASSIFICATION

J. Podani, Research Institute for Ecology and Botany, Hungarian Academy of Sciences, Vácrátót, H-2163
and Dept. of Plant Taxonomy and Ecology, L. Eötvös University, Budapest, H-1083, Hungary

Keywords: Agglomerative clustering; Complete linkage; Consensus index; Consensus interval; Iterative relocation; Single linkage

Abstract. An efficient heuristic method called MINGFC (for MINimization of Global Fusion Criterion) is proposed to select approximately optimal consensus partitions from the consensus interval defined by Neumann and Norton. The method utilizes a dissimilarity measure, the number of partitions in which two objects belong to different classes. A new consensus index is defined as the ratio of the average of all within-class dissimilarities to the average of all between-class dissimilarities. The lower this ratio, the more appropriate a given partition is as a consensus of the alternative classifications. This consensus index serves as the fusion criterion in the agglomerative clustering algorithm of MINGFC which generates a series of consensus partitions. The result is represented by a set of at least two trees, in graph theoretical terms: a consensus forest. To obtain a unique solution for any consensus problem, two procedures are suggested to resolve ties encountered during the clustering process. If a particular level of the consensus forest is of primary concern, the partition into the given number of classes may be further improved by an iterative relocation procedure. Artificial partitions and actual vegetation data provide the basis for illustrating MINGFC and iterative relocation, for evaluating the tie-breaking procedures, and for comparing MINGFC with two other hierarchical methods of consensus generation.

1. Introduction

The users of cluster analysis are faced with the usual dilemma of choosing among many techniques of classification. Unless an obvious group structure is present in the data, the different clustering algorithms and resemblance coefficients potentially selected may produce quite dissimilar classifications. If interest lies in examining the overall agreement of competing results, a consensus of the alternative classifications provides useful information. Nevertheless, the application of consensus methods is not restricted to such situations. For example, when results based on separate subsets of variables are to be combined into a single classification, consensus methods offer a solution. The overinterpretation of possibly unreliable details of alternative classifications is thus potentially avoided (cf. Lefkovitch 1985).

It has been a common practice in numerical taxonomy to synthesize several dendrograms into a single consensus tree (cf. Adams 1972, Rohlf 1982, and references therein). Whereas the literature abounds with suggestions regarding the construction of consensus trees, it seems that nonhierarchical classifications via partitions are largely neglected from this viewpoint. Although various approaches to the generation of consensus partitions have been available (Régnier 1965, Mirkin 1975, Barthélemy and Monjardet 1981, Lefkovitch 1985), these suggestions apparently escaped the attention of most potential users. This situation is undesirable, because in many applications of cluster analysis (e.g., classification of plant communities, diseases, etc.) simple partitions of objects are equally interesting, if not more interesting than hierarchical, representations of data.

In this paper a new consensus technique for partitions is proposed. Unlike in the approaches referred to above, the starting point employed here is that several consensus partitions can be constructed for a given set of classifications, except in the special case of all classifications being the same. The existence of several meaningful consensus candidates is strongly emphasized in Neumann and Norton's (1986) axiomatic approach. Given a set $T = \{t_1, \ldots, t_i, t_j, \ldots, t_n\}$ of n objects and a profile $\mathbf{P} = (A_1, \ldots, A_c, \ldots, A_p)$ of $p \geqq 2$ partitions of T, those authors suggested that any appropriate consensus partition C (**P**) must lie in the interval $I(\mathbf{P}) = [C_0(\mathbf{P}), C_1(\mathbf{P})]$ in the lattice Π of all partitions of T. The upper bound of this consensus interval is the full join of the p partitions

$$C_1(\mathbf{P}) = A_1 \vee \ldots \vee A_p = \{X_g;\ g=1, \ldots, r\}$$

with $T = X_1 \cup \ldots \cup X_r$. The lower bound is the full meet (or cross-partition)

$$C_0(\mathbf{P}) = A_1 \wedge \ldots \wedge A_p = \{Y_{gh};\ g=1, \ldots, r;$$

$$h=1, \ldots, m(g)\}$$

where m (g) is the number of classes in C_0 (**P**) that are subsets of X_g, so that $Y_{g1} \cup \ldots \cup Y_{gm(g)} = X_g$. m (g) > 1 for at least one g provided that the partitions in the profile **P** are not identical. C_0 (**P**) is termed in systematics as the strict consensus of **P**. Any other consensus candidate C(**P**) may be constructed by amalgamating collections of Y_{gh} for each X_g so that the following inequality is satisfied:

$$C_0 (\mathbf{P}) \leqq C (\mathbf{P}) \leqq C_1 (\mathbf{P}) \qquad (1)$$

which is an obvious consequence of the Pareto axiom on clustering and isolation proposed by Neumann and Norton (1986). Those authors asserted that there is no clear answer as to the selection of a particular consensus partition from I (**P**). The bounds of I (**P**) appear the most straightforward candidates. However, in many practical situations $\sum_g$ m (g) is great and several singleton classes arise. That is, C_0 (**P**) is too fine from a pragmatic viewpoint. On the other hand, a few objects with uncertain class-membership are sufficient to reduce the number of classes in C_1 (**P**) to one, rendering trivial the upper bound of I (**P**). The number of the remaining partitions that satisfy inequality (1) may be enormously large, and Neumann and Norton's approach seems to imply that all are equally good candidates as a consensus partition. This is not necessarily so, however, because further distinction should be made among the partitions within the consensus interval in terms of partial clustering and partial isolation defined in section 2.1. Based on these concepts, I propose a consensus index which may be used to select optimal consensus partitions from I (**P**). This index, denoted by γ (C (**P**)$_k$), measures the appropriateness of C (**P**)$_k$ as a consensus of **P**, where the subscript k refers to the number of classes in the given consensus partition and $|C_0 (\mathbf{P})| \geqq k \geqq \max\{2, r\}$. Since the number of consensus classes (or cells) cannot be fixed according to internal criteria, the problem is to find a partition Z, $|Z| = k$, for each value of k, such that $\gamma (Z_k)$ is the minimum. Z is probably not unique, and there may not exist an efficient algorithm to select the optima. To obtain an approximate solution to this problem, a heuristic method (minimization of global fusion criterion, MINGFC) is proposed. A series of consensus partitions, with the small classes nested in large ones, is generated by an efficient agglomerative hierarchical clustering method in which the consensus index is adopted as the fusion criterion. The graphical illustration of the results is the consensus forest, a set of max {2, r} trees.

Considerable attention is paid to the problem of ties encountered during the classificatory process. Two alternative tie-breaking procedures are suggested, the first based on single linkage fusions and the other utilizing the concept of suboptimal fusions. The application of these procedures leads to a unique result more or less approximating the optimum consensus partition for each value of k. However, if a particular partition at a fixed value of k is sought, rather than the whole consensus forest, uniqueness is subordinate to optimality. In these situations, the result of MINGFC may be improved by an iterative relocation method to get a closer approximation to the optimum.

Artificial and actual data, the latter taken from vegetation science, are used to demonstrate the consensus generation. The performance of MINGFC is compared with that of two other agglomerative methods, the single and complete linkage algorithms. The effect of the two tie-breaking procedures upon the results, and the utility of iterative relocation are also demonstrated using the same data.

2. The Generation of Consensus Partitions

2.1 Partial Clustering and Partial Isolation

The axiomatic approach of Neumann and Norton (1986) utilizes the concepts of clustering and isolation to construct the consensus interval I (**P**) in the lattice Π of all partitions of T. These concepts are fruitful to select all the potential candidates for consensus partitions but are usually insufficient to find optimal solutions to a particular consensus problem. This is because clustering and isolation represent very strict set-theoretical relationships for the consensus classes. However, there are some weaker relationships among the classes which should also be considered in consensus generation. To elaborate this point, suppose that objects t_1 and t_2 are clustered together in all but one of p partitions. Thus, t_1 and t_2 are neither clustered nor isolated in **P**. Yet, it can be said that t_1 and t_2 are partially clustered in **P** to the degree of p-1. At the same time, they are partially isolated in **P** to a degree of 1. A straightforward consequence of this is that if $p > 2$ a consensus partition $R \in I$ (**P**) that clusters t_1 and t_2 together will be more optimal with respect to these objects than another partition $Q \in I$ (**P**) which isolates them. Since partial clustering and isolation of objects are complementary terms, it is sufficient to use only the degree of partial isolation in the sequel. For any objects t_i and t_j of T, the degree of partial isolation d_{ij} is defined as the number of partitions in which t_i and t_j are assigned to different classes in **P**. This measure may be considered as a dissimilarity* of t_i and t_j.

d_{ij} ranges from 0 to p, the extremes indicating (complete) clustering and isolation, respectively. This dissimilarity measure is in fact not new, see e.g., Diday and Simon (1976).

Turning to the problem of partial isolation and clustering of subsets of T in **P**, consider classes Y_{g1} and

* d_{ij} is not a metric since the definiteness property ($d_{ij} = 0$ iff $t_i = t_j$) is not satisfied.

Y_{g2} from X_g. The degree of partial isolation between Y_{g1} and Y_{g2} in **P** is defined as the average of their between-class dissimilarities:

$$\beta\,(Y_{g1}, Y_{g2}) = \sum_{i\epsilon Y_{g1}} \sum_{j\epsilon Y_{g2}} d_{ij}/n_1\, n_2$$

where $n_1 = |Y_{g1}|$ and $n_2 = |Y_{g2}|$. The value of $\beta\,(Y_{g1}, Y_{g2})$ ranges from 0 to p; zero would result only for two nonempty subsets of any Y_{gh} but no such classes occur in the partitions within I (**P**), whereas p is obtained for any pair of classes from C_1 (**P**). That is, 0 corresponds to clustering and p indicates isolation in the sense of Neumann and Norton (1986).

Let $X_{g(1,2)}$ denote the union of Y_{g1} and Y_{g2}. The degree of partial clustering of $X_{g(1,2)}$ in **P** is defined as the average of within-class dissimilarities:

$$\alpha\,(X_{g(1,2)}) = \sum_{i\epsilon X_{g(1,2)}} \sum_{j\epsilon X_{g(1,2)}} d_{ij}/n_{12}\,(n_{12}-1)/2$$

where $i \neq j$ and $n_{12} = |X_{g(1,2)}|$. The value of $\alpha\,(X_{g(1,2)})$ ranges from 0 to p; 0 results for every class of C_0 (**P**) (clustering) whereas p cannot be reached within I (**P**).

2.2 A New Consensus Index for Partitions

There are two fundamental requirements that should be satisfied by an optimum consensus partition in I (**P**), given a fixed value of k. The average of within-class dissimilarities should be minimized for all consensus classes, whereas the average of between-class dissimilarities should be maximized for every pair of consensus classes. This is analogous to the duality of cluster cohesion and separation in a general cluster analysis model (cf. Cormack 1971). $\alpha\,(X_{g(1,2)})$ and $\beta\,(Y_{g1}, Y_{g2})$ cannot be used for optimization because they apply to one class and to one pair of classes, respectively, allowing the optimization of local properties of $C\,(\mathbf{P})_k$ only. A global measure that reflects overall partial clustering and overall partial isolation of classes is necessary. Overall partial clustering of the classes of $C\,(\mathbf{P})_k$ in **P** is defined as the average of all within-class dissimilarities:

$$\alpha\,(C\,(\mathbf{P})_k) = \sum_{c=1}^{k} \sum_{i\epsilon X_c} \sum_{j\epsilon X_c} d_{ij} \Big/ \sum_{c=1}^{k} n_c\,(n_c-1)/2$$

where X_c denotes a consensus class of $C\,(\mathbf{P})_k$, $i \neq j$ and $n_c = |X_c|$. Overall partial isolation of classes of $C\,(\mathbf{P})_k$ in **P** is defined as the average of all between-class dissimilarities:

$$\beta\,(C\,(\mathbf{P})_k) = \sum_{c=1}^{k-1} \sum_{d=c+1}^{k} \sum_{i\epsilon X_c} \sum_{j\epsilon X_c} d_{ij} \Big/ \sum_{c=1}^{k-1} \sum_{d=c+1}^{k} n_c\, n_d\,.$$

An optimum consensus partition of T into k classes could be selected from I (**P**) by minimizing $\alpha\,(C\,(\mathbf{P})_k)$ and maximizing $\beta\,(C\,(\mathbf{P})_k)$. However, the two extremes are not necessarily associated with the same partition. Therefore, I suggest to use the ratio of overall partial clustering and overall partial isolation of classes of $C\,(\mathbf{P})_k$ in **P**,

$$\gamma\,(C\,(\mathbf{P})_k) = \frac{\alpha\,(C\,(\mathbf{P})_k)}{\beta\,(C\,(\mathbf{P})_k)}\,,$$

as a measure of the adequacy of $C\,(\mathbf{P})_k$ as a consensus of the original profile of partitions. The lower bound of the consensus index $\gamma\,(C\,(P)_k)$ is 0, obtained for C_0 (**P**) and of course for any other partition which is finer than C_0 (**P**). The larger this ratio the greater the deviation of the consensus partition from the strict consensus represented by C_0 (**P**), so that $\gamma\,(C\,(\mathbf{P})_k)$ measures dissimilarity between a proposed consensus partition and C_0 (**P**). A sensible upper limit of the consensus index is 1. This would imply equal within- and between-class dissimilarities, but such an extreme situation cannot arise within I (**P**). The actual upper bound of $\gamma\,(C\,(\mathbf{P})_k)$ is $\gamma\,(C_1\,(\mathbf{P}))$ if $r > 1$. For $r = 1$, the consensus index is undefined because the average of between-class dissimilarities is undefined. In this case, the actual upper limit is yielded by the worst two-class partition of T in I (**P**). Note that for randomly generated "consensus" partitions, the value of $\gamma\,(C\,(\mathbf{P})_k)$ may exceed unity.

2.3 A Hierarchical Approach to Consensus Generation

In lieu of algorithms with polynomially-bounded time complexity, the selection of optimum partitions from the consensus interval implies the examination of

$$|I\,(\mathbf{P})| = B\,(m\,(1))\;B\,(m\,(2))\;\ldots\;B\,(m\,(r))$$

partitions, where B (m) is the Bell number that counts all the possible partitions of a set of m elements (cf. Neumann and Norton 1986). For each value of k, that partition is the optimum for which the consensus index takes the minimum value. The prolem is that $|I\,(\mathbf{P})|$ may be very large and that no efficient algorithm is known as yet to select the optimum. Also, the optimum for each k is not necessarily unique. As long as an efficient method is unavailable, it may be practical to use a relatively fast algorithm which provides a unique result that approximates the optimum partitions. For this purpose, an agglomerative hierarchical procedure is suggested. The basic idea is that a hierarchical classification of n objects is generated according to their class-membership relationships in **P**. As in other agglomerative methods, in each cycle of the analysis two objects are fused into a consensus class if the consensus index calculated for the new partition obtained is the minimum. The main steps of the algorithm are as follows.

1. Calculate the dissimilarity matrix $\mathbf{D} \equiv \{d_{ij}\}$ of the n objects. d_{ij} is the number of partitions in which

t_i and t_j are assigned to different classes in **P** (partial isolation of object pairs, see section 2.1). The entries of **D** will be used throughout the analysis. Set the number of classes, k, equal to n.

2. Calculate the secondary matrix $\Gamma \equiv \{\gamma_{ij}\}$ containing the consensus indices for every pair of objects (classes). γ_{ij} measures the goodness of the new consensus partition obtained by amalgamating objects (classes) i and j. The consensus index is used as the fusion criterion, and those two objects are fused for which γ is the minimum. Set k=k—1.

3. Examine whether k is larger than max {2, r}. If so, the analysis proceeds by going back to step 2. Otherwise the analysis stops because:

a. If r=1, the consensus index is undefined for the single class obtained by the fusion of the remaining two classes.

b. If $r \geqq 2$, we have isolated classes whose fusion, although the consensus index is computable, would provide a partition which is not an element of I (**P**).

4. Output the result. Unlike in other hierarchical clustering procedures, the result is not a dendrogram but a collection of max {2, r} trees, in graph theoretical terms: a forest.

In addition to differences in the graph theoretical properties of results, there is another important difference between the commonly used agglomerative procedures and the above algorithm. Instead of local fusion criteria (e.g., the distance between pairs of clusters in the methods compatible with the Lance-Williams (1966) scheme), this strategy utilizes a global fusion criterion: the fusion of two classes is conditioned upon the goodness of the whole partition as a consensus in any step of the analysis. The abbreviation MINGFC (MINimization of Global Fusion Criterion) will be used in the sequel when reference is made to this clustering strategy.

Consensus partitions could be generated from **D** by other clustering algorithms. For example, the method suggested by Diday and Simon (1976) without explicit reference to consensus generation, implies a complete linkage sorting strategy: classes are formed at increasing values of c ($0 \leqq c \leqq p$) such that within-class dissimilarities do not exceed c. Single linkage clustering represents another possibility to recover data structure. The two methods will be compared with MINGFC using both artificial and actual data in section 3.

2.4 Treatment of Ties

The algorithm of MINGFC, in the form presented above, would leave a serious question unanswered: what happens if there is no unique minimum of γ in step 2? Most hierarchical procedures do not offer a solution for this problem, and arbitrary choices among tied pairs are made. Thus, these methods are ill-defined, the result being dependent on the labeling of objects (Jardine and Sibson 1971). The single linkage method is an exception; its result is unique regardless the number of ties occurring in the dissimilarity matrix.

2.4.1 Single Linkage Solution

There are several types of ties that may be found during the analysis. These are best-described using graph theoretical terms. Let G be a graph ("tie-graph") whose vertices represent the objects (or classes) associated with a tie at the dissimilarity level γ_{min}. Two vertices v_q and v_r are connected by an edge e_{qr} in G if $\gamma_{qr} = \gamma_{min}$. According to the connectedness of G four types of ties can be distinguished:

a. G is complete (Fig. 1a);

b. G is disconnected and all isolated subgraphs (i.e., components) are complete (Fig. 1b);

c. G is disconnected and at least one component is not complete (Fig. 1c); and

d. G is not complete but is connected (Fig. 1d).

The resolution of ties is straightforward in the first two situations: all the objects are amalgamated into a single class by a multiple fusion (case a) or several classes are formed simultaneously, each corresponding to a component (case b). In case c, classes represented by complete components may be formed and the others could be excluded from the fusion, thus leaving the decision for the subsequent clustering steps. However, since the results of single linkage clustering are not influenced by ties, the single linkage criterion, as applied to tied pairs, offers a possible remedy of the problem. That is, any two objects are fused into the same class if there exists a path between their corresponding vertices in G. As a result, in case d all the objects associated with the tie will be pooled into the same class. In this way, MINGFC will produce a unique result in which multiple fusions, simultaneous fusions and the single linkage solutions will indicate the type of ties encountered and will depict data structure where the relationships are ambiguous.

The incorporation of this tie-resolving procedure into the algorithm of MINGFC may give rise to another situation in which the analysis must stop. This is when a single class would remain after the fusions. In that case the consensus index is undefined; therefore these fusions are omitted and the clustering process termi-

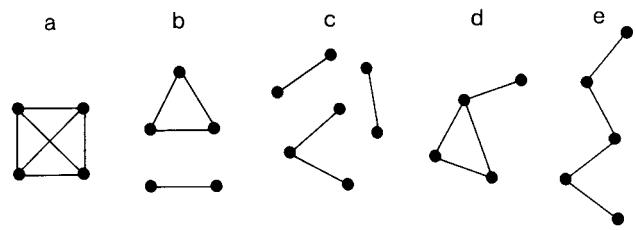

Fig. 1. Graphs illustrating different types of ties (a-d) and chaining, a special case of type d ties (e).

nates with the number of classes present in the previous cycle.

2.4.2 Suboptimal Fusions

A problem with the single linkage solution of ties is that the resulting classes may include two quite dissimilar objects because there is a sequence of objects that connect them ("chaining effect", Fig. 1e). To overcome this problem, I suggest another procedure for eliminating ties in a different way for cases c and d. In cases a and b, the solution is the same as above.

In case c, only the objects associated with isolated complete subgraphs will form classes, and the decision for the other objects is postponed for the next cycle. Therefore, it is sufficient to provide a tie-breaking procedure for case d.

The basic idea is that the fusion of tied objects is ignored and the next lowest value is found in Γ. If there are no ties, or only type a, b or c ties occur at this suboptimal value, simple, multiple, or simultaneous fusions are performed. Then, the analysis proceeds with the next clustering step. Since the optimum value is skipped, these operations may be termed suboptimal fusions. If type d ties occur at the suboptimal level, the search continues as long as the lowest value in Γ with which no type d ties are associated is found.

The use of suboptimal fusions necessitates the application of a new stopping rule. If there is no value in Γ without type d ties, the analysis must terminate with the number of classes present in the previous cycle.

2.5 The Improvement of Consensus Partitions

The series of consensus partitions obtained by MINGFC satisfies the inequalities

$$C(\mathbf{P})_k \leqq C(\mathbf{P})_{k-1},$$

$$|C(\mathbf{P})_k| \rangle |C(\mathbf{P})_{k-1}|, \text{ and}$$

$$|C_0(\mathbf{P})| \geqq k \geqq |C_1(\mathbf{P})|.$$

That is, a consensus partition into k classes is always finer than a consensus partition of objects into k-1 classes; this is a consequence of the agglomerative sorting strategy. Once two objects were classified together at a great value of k, they will remain in the same class for the smaller values of k. As a result, the consensus forest cannot be an equally good approximation to the optimum for every value of k. However, if interest is focused on a consensus partition for a particular value of k, the consensus partition yielded by MINGFC may be further improved by a nonhierarchical clustering procedure. Algorithms for iterative relocation of objects are well-known, for example, from k-means clustering (Hartigan 1975). In the present case, the steps are as follows.

1. Calculate the consensus index for the initial partition.

2. Identify the object whose relocation into a new class yields the minimum value of the consensus index.

3. Examine whether the new value is smaller than the previous one. If so, the object selected is relocated to the new class and the analysis continues with step 2. Otherwise, there is no object whose relocation would improve the partition, and the actual partition is declared as the final consensus.

Of course, even this iterative procedure might not lead to an optimal consensus partition, since the final result of iterative relocation is influenced by the starting configuration. Nevertheless, the procedure does improve the consensus partition and gives a closer approximation to the optimum, as demonstrated in section 3.1.2.

2.6 Computer Programs

Program MINGFC has been written in FORTRAN (IBM System/370 version for mainframe computers and Microsoft V4.0 for IBM XT and AT compatibles) to generate consensus forests. The time requirement of the program is higher than that of other currently used agglomerative methods (cf. Day and Edelsbrunner 1984). The calculation of the consensus index, with a time complexity of $0(n^2)$ for every candidate for consensus partition, is done $0(k^2)$ times in each clustering cycle. Since the tie-resolving procedure requires the scanning of matrix Γ, an additional $0(k^2)$ time is needed to find the minimum in each cycle. The occurrence of ties in fact decreases computation time because the number of clustering steps decreases. The fusions are accelerated at the beginning of the analysis, when classes of $C_0(\mathbf{P})$ are formed. The space requirement of the program is $n^2+14n+s^2+s$, where s is a reasonably large number specified by the user for the maximum number of vertices allowed in a tie-graph. The proportionality to n^2 stems from the simultaneous storage of **D** and Γ in half matrix form; otherwise the tie-breaking procedure would require repeated calculations of γ leading to an increased time complexity of the algorithm.

The input for this program is either a set of p class membership arrays (the ith value in each signifies the class into which object i belongs) or the dissimilarity matrix in half matrix form. The output includes detailed information on each clustering step (objects or classes fused, number of within- and between-class dissimilarities, within- and between-class average dissimilarities). If ties are found, all pairs associated with the tie are listed and the type of tie-resolution is indicated before the list of fusions in that step. The user must decide prior to program execution whether single linkage or suboptimal fusions will resolve ties. The final graphical result, the consensus forest, occurs on the line-

printer.

FORTRAN program PARREL performs iterative relocation of objects to improve a consensus partition for a fixed value of k. In each iteration cycle the consensus index, whose calculation has a time complexity of $0(n^2)$, is calculated $n(k-1)$ times to find the object whose relocation into another class gives the greatest decrease in the consensus index. The input for this program includes the same data used by MINGFC plus an array defining the initial consensus partition.

Programs MINGFC and PARREL have been included in the SYN-TAX III package (Podani 1988).

3. Examples

3.1 Artificial Partitions

Four partitions of 10 objects were constructed for illustrative purposes. These are:

$A_1 = \{1, 2, 3, 4\}\ \{5, 6, 7, 8, 9, 10\}$,

$A_2 = \{1, 2, 3, 4, 5, 6\}\ \{7, 8, 9, 10\}$,

$A_3 = \{1, 2, 3, 4, 5\}\ \{6, 7, 8, 9, 10\}$,

$A_4 = \{1, 2, 3, 7\}\ \{4, 5, 6, 8, 9, 10\}$.

The upper half matrix of dissimilarities is given by

$$\mathbf{D} = \begin{matrix} 0 & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 4 \\ & 0 & 1 & 2 & 3 & 3 & 4 & 4 & 4 \\ & & 1 & 2 & 3 & 3 & 4 & 4 & 4 \\ & & & 1 & 2 & 4 & 3 & 3 & 3 \\ & & & & 1 & 3 & 2 & 2 & 2 \\ & & & & & 2 & 1 & 1 & 1 \\ & & & & & & 1 & 1 & 1 \\ & & & & & & & 0 & 0 \\ & & & & & & & & 0 \end{matrix}$$

The strict consensus, as in apparent from **D**, has six classes:

$C_0(\mathbf{P}) = Z_6 = \{1, 2, 3\}\ \{4\}\ \{5\}\ \{6\}\ \{7\}\ \{8, 9, 10\}$

$\gamma(Z_6) = 0.$

3.1.1 Optimum Consensus Partitions

Since problem size is relatively small, it is easy to find the optimum partitions in $I(\mathbf{P})$ for every $2 \leqq k \leqq 5$. At the five-class level, the optimum is not unique:

$Z_5^{(1)} = \{1, 2, 3\}\ \{4, 5\}\ \{6\}\ \{7\}\ \{8, 9, 10\}$

$Z_5^{(2)} = \{1, 2, 3\}\ \{4\}\ \{5, 6\}\ \{7\}\ \{8, 9, 10\}$

with $\gamma(Z_5) = 0.0565$. Therefore, objects 4, 5 and 6 will be tied in agglomerative clustering. For each smaller value of k, there is a single optimum:

$Z_4 = \{1, 2, 3\}\ \{4, 5\}\ \{6, 7\}\ \{8, 9, 10\}$

$\gamma(Z_4) = 0.1476,$

$Z_3 = \{1, 2, 3, 4\}\ \{5, 6\}\ \{7, 8, 9, 10\}$

$\gamma(Z_3) = 0.1914,$

$Z_2 = \{1, 2, 3, 4, 5\}\ \{6, 7, 8, 9, 10\}$

$\gamma(Z_2) = 0.2848.$

The optimum partitions do not form a nested hierarchical system. For example, the inequality $Z_4 \leqq Z_3$ does not hold. Consequently, agglomerative methods cannot detect the optimum for every k.

3.1.2 Approximate Optima Obtained by Hierarchical Clustering

Two results produced by MINGFC, three alternative complete linkage clustering results, and the single linkage dendrogram will illustrate the complexity of consensus generation by hierarchical clustering. Of course, the methods agree in detecting Z_6 correctly; differences arise at the other levels.

The MINGFC analysis with single linkage resolution of ties amalgamates objects 4, 5 and 6 into the same class (recall that $\gamma_{45} = \gamma_{56} = 0.0565$). Thus, the five-class level is skipped and four classes result directly:

$Q_4 = \{1, 2, 3\}\ \{4, 5, 6\}\ \{7\}\ \{8, 9, 10\}$

$\gamma(Q_4) = 0.172.$

Then, no more ties occur in Γ, and the analysis continues normally:

$Q_3 = \{1, 2, 3\}\ \{4, 5, 6\}\ \{7, 8, 9, 10\}$

$\gamma(Q_3) = 0.213,$

$Q_2 = \{1, 2, 3, 4, 5, 6\}\ \{7, 8, 9, 10\}$

$\gamma(Q_2) = 0.396.$

The results are summarized by the consensus forest of Fig. 2a. Note that except for $k=6$ the partitions are not optimal and the deviation of Q_2 from Z_2 is especially striking.

Tie-breaking with suboptimal fusions in MINGFC provides a completely different fusion sequence. After finding the strict consensus classes, the three lowest

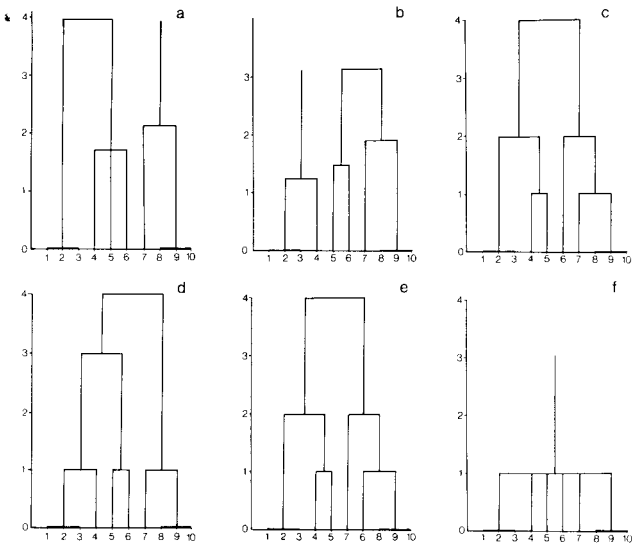

Fig. 2. Consensus hierarchies for four artificial partitions. a: result of MINGFC with single linkage resolution of ties. b: MINGFC with suboptimal fusions. c-e: complete linkage solutions. f: single linkage solution.

values in Γ are associated with ties:

$$\gamma_{45} = \gamma_{56} = 0.0565,$$

$$\gamma_{46} = \gamma_{67} = 0.114,$$

$$\gamma_{6\,(8,\,9,\,10)} = \gamma_{7\,(8,\,9,\,10)} = \gamma_{(1,\,2,\,3)\,4} = 0.127.$$

For the third lowest value, we have a type c tie, resolved by amalgamating class (1, 2, 3) with object 4:

$$R_5 = \{1, 2, 3, 4\}\ \{5\}\ \{6\}\ \{7\}\ \{8, 9, 10\}$$

$$\gamma(R_5) = 0.127.$$

This suboptimal fusion eliminates all ties from the remaining steps:

$$R_4 = \{1, 2, 3, 4\}\ \{5, 6\}\ \{7\}\ \{8, 9, 10\}$$

$$\gamma(R_4) = 0.150,$$

$$R_3 = [1, 2, 3, 4\}\ \{5, 6\}\ \{7, 8, 9, 10\}$$

$$\gamma(R_3) = 0.191,$$

$$R_2 = \{1, 2, 3, 4\}\ \{5, 6, 7, 8, 9, 10\}$$

$$\gamma(R_2) = 0.316.$$

The resulting consensus forest is given in Fig. 2b. As seen, the solution is bad only for $k=5$, R_4 and R_2 are fairly close to the optimum and $R_3 = Z_3$. Thus, in this example suboptimal fusions performed better than single linkage resolution of ties.

Complete linkage clustering is very sensitive to the labeling of objects, because ties are resolved arbitrarily. At a given value of c, there may be several partitions which satisfy the criterion that all within-class dissimilarities are less than or equal to c. Such alternatives may be simply generated by relabeling the objects prior to the analysis. Three variants are shown here (Figs. 2c-e). Although the two-class partitions implied by the dendrograms of Figs. 2c and 2e correspond to the optimum, there is much doubt that complete linkage clustering always performs well. The user is uncertain whether there is only one consensus hierarchy or several. Another point to be made is that the maximum number of partitions possible in a complete linkage dendrogram is $p+1$ (i.e., at levels 0, 1, ..., p). Thus, if $n \gg p$, which is usually the case, the representation of data structure by a complete linkage dendrogram may be too rough.

The single linkage method seems to give the poorest result for the artificial partitions (Fig. 2f). This is obviously the manifestation of the chaining effect. The resulting levels are too few, and it seems unacceptable to treat objects 4, 5, 6 and 7 in the same way.

3.1.3 Iterative Relocation

The results of MINGFC were improved by the iterative relocation technique to see if it is possible to find the optima by the joint application of hierarchical and nonhierarchical consensus seeking methods.

Z_2 was obtained from both Q_2 and R_2 by reallocating a single object in each case. Z_3 was identical to R_3 and was obtained in a single step from Q_3. Also, a single step was necessary to transform Q_4 into Z_4. However, there was no route from R_4 to Z_4 because the relocation of any object in R_4 would have resulted in an increase of the consensus index. This observation shows that the final result of iterative relocation is sensitive to the initial partition. The optimum for $k=5$ can be obtained only from R_5 by relocating object 4, thus yielding $Z_5^{(1)}$. There is no route to find the other optimum, $Z_5^{(2)}$, from the MINGFC results.

3.2 Partitions of Floristic Vegetation Data

Actual data for illustrating the consensus generation are derived from a detailed study of the grassland communities in the Sashegy Nature Reserve, Budapest, Hungary (Podani 1985). The 80 sampling units taken in the field were classified using six different hierarchical clustering methods based on the presence and absence of vascular plant species. The best agreement among these classifications was observed to appear at the three-class level. Thus, it is a natural problem to seek a 3-class consensus partition which synthesizes the-

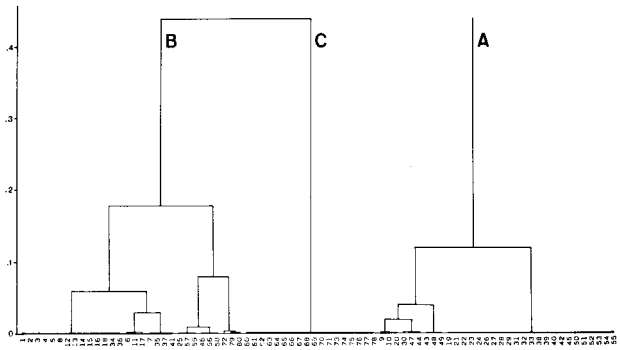

Fig. 3. Consensus forest for six floristic classifications of the Sashegy data. Tie-breaking with suboptimal fusions.

se six classifications into one, and to reveal objects or small classes primarily responsible for differences among the competing results. The hierarchical consensus generation procedures and the iterative relocation method are clearly suitable to these objectives.

MINGFC was run using both tie-breaking procedures, but the corresponding hierarchical levels were much higher when the single linkage solution was applied. This accords well with the study of artificial partitions, suggesting that suboptimal fusions should be preferred to resolve the ties. Therefore, in the sequel reference to MINGFC analysis always implies that suboptimal fusions were incorporated into the clustering algorithm. The MINGFC results were improved at high hierarchical levels by the iterative relocation procedure. In addition, complete linkage and single linkage cluster analyses of the 80 objects were also performed using program NCLAS from the SYN-TAX III package (Podani 1988).

The overall agreement of the six classifications is quite high, since the three biggest classes in the strict consensus contain 23, 14 and 18 objects, respectively. Of course, it is shown by all three consensus seeking procedures used (Figs. 3-5). These classes form the kernel of the classes A, B and C obtained at the 3-class level in the consensus forest. Class C seems the most isolated and most compact, especially in the MINGFC result

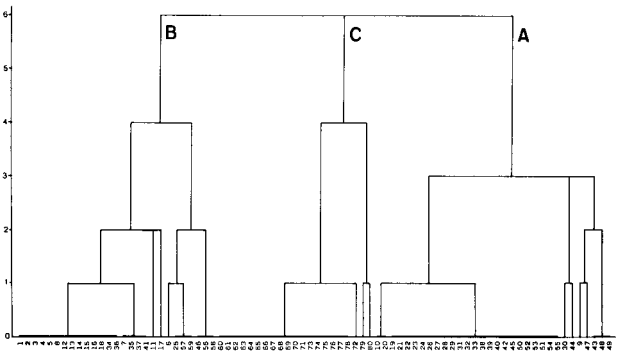

Fig. 4. Consensus dendrogram of six floristic classifications of the Sashegy data. A complete linkage solution.

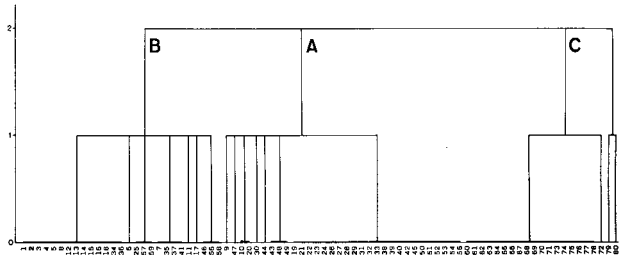

Fig. 5. Consensus dendrogram of six floristic classifications of the Sashegy data. Single linkage analysis.

(Fig. 3). The complete linkage solution (Fig. 4) picks up three objects that are closely related to class C, namely objects 72, 79 and 80. Note, however, that the complete linkage solution illustrated is not necessarily unique, leaving us somewhat uncertain about the relationships above the zero level. Nevertheless, the single linkage consensus (Fig. 5) confirms the above relationship at least for object 72.

The three methods completely agree as to the composition of class A. In case of the single and complete linkage solutions, class B is also invariant, whereas MINGFC added objects 72, 79 and 80 to it. Thus, if we disregard the latter objects, the consensus methods used suggest exactly the same 3-class partition.

The consensus partitions obtained by MINGFC for $k=2$, 3, 4, 5 and 6 were improved by iterative relocation (Table 1). At the two-class level considerable reduction of the consensus index was achieved by relocating objects 72 and 79. For higher numbers of classes up to $k=6$, only 3-4 relocations were made showing that MINGFC provided quite good approximations to the optima.

Table 1. Comparison of the highest five fusion levels of the consensus forest of six floristic classifications after iterative relocation.

Number of consensus classes	γ (C (P) $_k$) before relocation	Number of iterations	γ (C (P) $_k$) after relocation
2	.4479	0	.4479
3	.1840	2	.1597
4	.1231	4	.1077
5	.0825	3	.0764
6	.0602	4	.0521

The phytosociological interpretation of the above results is not difficult. Without entering into details that are irrelevant here, it is worth mentioning that sampling units in class A represent an open, relatively species-poor vegetation type occurring on south-facing slopes, with Festuca cinerea as the dominant species. Class C corresponds to a contrasting vegetation type, a species-rich, closed grassland community on N-NE-facing slo-

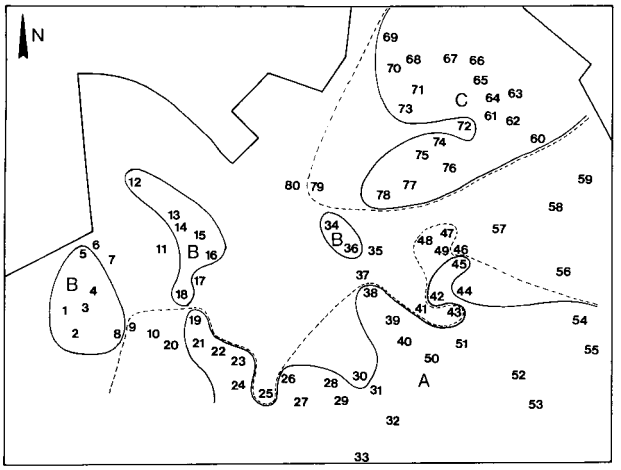

Fig. 6. Allocation of the consensus partition of six floristic classifications onto the schematic map of Sashegy Nature Reserve, Budapest. Solid lines delimit strict consensus classes, dotted lines separate the classes suggested by MINGFC at the 3-class level and improved by iterative relocation. Numbers show the localization of sampling units, letters A, B and C indicate the vegetation types.

pes of the dolomite hills. The sampling units in class B were taken mainly on W-SW-facing slopes and hilltops, and represent a community intermediate between the other two regarding both species richness and plant cover. The spatial separation of these vegetation types is illustrated by the map in Fig. 6.

For the description of the three community types, sampling units forming the largest strict consensus classes (kernels) seem the most appropriate. Using another terminology, the three kernels represent vegetational noda points in an abstract spatial continuum. The sampling units with ambiguous class membership represent transitions between the noda points: the higher the level at which an object is fused with a kernel in the consensus, the stronger the departure of its floristic composition from that of typical sites, i.e., kernel classes. As the MINGFC strategy indicates at the two-class level, the intermediate type is closer to type C than to type A.

4. Concluding Remarks - Future Tasks

A new consensus index defined in terms of partial clustering and partial isolation is suggested to select optimum partitions from the consensus interval. The MINGFC strategy, in which the consensus index is adopted as the fusion criterion, is appropriate to find a unique series of consensus partitions, represented by the consensus forest, in which each partition is an approximation to the optimum. However, hierarchical clustering forces a nested structure on the series of consensus partitions so that an iterative relocation procedure is useful to provide a better solution for a particular level. Because the optimum for each value of k is not necessarily unique, the result of iterative relocation may not be unique as well, showing that uniqueness and optimality are sometimes conflicting requirements. As complementary techniques of consensus generation, complete linkage and single linkage clustering from the dissimilarity matrix of objects may be used. These techniques reveal interesting aspects of the data not apparent from the consensus forest, and the complete linkage method may perform better than MINGFC although its results are affected by chance effects when the objects are labeled.

An advantage of the hierarchical representation of objects is that the consensus partitions may be evaluated at several levels. The number of these levels is usually much more in the consensus forest than for single and complete linkage hierarchies. When the number of classes in the alternative partitions is a fixed value, the evaluation of the consensus forest at that particular level will be the most appealing. This was the case in the examples presented. The more general situation, with unequal number of classes in the input classifications, was not examined in the present paper.

The MINGFC strategy differs essentially from the other agglomerative methods currently used. The fusion criterion calculated for each pair of objects or classes utilizes all the values of the primary dissimilarity matrix. As in multidimensional scaling, all dissimilarities are equally important in affecting the final result. This is not so with the complete and single linkage analyses whose results are influenced only by a minority of entries in the starting matrix, the remaining values can be changed within a relatively broad range without imposing any change on the hierarchy.

If the primary matrix contains integers (as in case of consensus generation), the possibility of finding ties during the clustering process is relatively high. If a unique solution is sought, the resolution of ties is important. Therefore, the problem of ties is examined thoroughly in this paper. Four types of ties have been distinguished, two of them posing no difficulties. For the resolution of the other two situations two alternatives are suggested: the single linkage procedure and tie-breaking by suboptimal fusions. Analyses of artificial and actual data showed that the choice of tie-breaking technique is crucial: the consensus partitions obtained through suboptimal fusions are more optimal than when single linkage resolution of ties is employed by the algorithm. This is due to the chaining effect whose consequence is that the consensus index may be higher for a partition generated using single linkage resolution than for another consensus partition created by breaking ties arbitrarily. Suboptimal fusions perform better, but their application may lead to reversals in the resulting consensus forests (although no reversals occurred in the examples). A future research topic is to investigate the utility of suboptimal fusions in other

agglomerative clustering techniques.

The use of MINGFC is not restricted to the evaluation of partitions. Any kind of distance or dissimilarity matrix can be analyzed by this clustering strategy, so that a more general meaning is attached to the consensus index: this is the ratio of mean within-class and between-class distances. Because this ratio is dimensionless, the actual range and the nature of the distance coefficient in the primary matrix do not matter; the hierarchical levels remain within the range [0, 1]. The question whether alternative hierarchies, obtained by applying different distance measures to the same data set, are directly comparable requires the analysis of the distributional properties of the fusion criterion.

In the general situation, MINGFC may directly inform us about the classifiability of objects. When within-cluster distances are much smaller than between-cluster distances, the fusion levels are low, indicating high cohesion and separation of clusters, that is, high classifiability. The closer the fusion level to unity the more pronounced the tendency to have equal distances within and between clusters, and the classifiability of objects becomes more doubtful. Therefore, contrary to other currently used methods, MINGFC appears to be able to indicate by itself how a group structure is forced upon a set of objects. This suggested property also deserves future investigations using simulated data.

Acknowledgments. I am grateful to two anonymous referees whose suggestions led to substantial improvements of the manuscript. Many thanks are due to P. Juhász-Nagy for his comments. The paper was presented at the 1st Congress of the International Federation of Classification Societies, Aachen, West Germany, in July, 1987. Financial support from the Soros Foundation, New York-Budapest, to cover costs of my participation is gratefully acknowledged.

REFERENCES

ADAMS, E.N. 1972. Consensus Techniques and the Comparison of Taxonomic Trees. Systematic Zoology, 21, 390-397.

BARTHÉLEMY, J.P. and B. MONJARDET. 1981. The Median Procedure in Cluster Analysis and Social Choice Theory. Mathematical Social Sciences, 1, 235-268.

CORMACK, R.M. 1971. A Review of Classification. Journal of the Royal Statistical Society, ser. A. 134, 321-367.

DAY, W.H.E. and H. EDELSBRUNNER. 1984. Efficient Algorithms for Agglomerative Hierarchical Clustering. Journal of Classification, 1, 7-24.

DIDAY, E. and J.C. SIMON. 1976. Clustering Analysis. In Fu K.S. (ed.), Digital Pattern Recognition, New York: Springer. pp. 47-94.

HARTIGAN, J.A. 1975. Clustering Algorithms. New York: Wiley.

JARDINE, J. and R. SIBSON. 1971. Mathematical Taxonomy, London: Wiley.

LANCE, G.N. and W.T. WILLIAMS. 1966. A Generalized Sorting Strategy for Computer Classifications. Nature, 212, 218.

LEFKOVITCH, L.P. 1985. Euclidean Consensus Dendrograms and Other Classification Structures. Mathematical Biosciences, 74, 1-15.

MIRKIN, B.G. 1975. On the Problem of Reconciling Partitions. In H.M. Blalock, A. Aganbegian, F.M. Borodkin, R. Boudon, and V. Capecchi (eds.), Quantitative Sociology, International Perspectives on Mathematical and Statistical Modeling, New York: Academic Press, pp. 441-449.

NEUMANN, D.A. and V.T. NORTON. 1986. Clustering and Isolation in the Consensus Problem for Partitions. Journal of Classification, 3, 281-297.

PODANI, J. 1985. Syntaxonomic Congruence in a Small-Scale Vegetation Survey. Abstracta Botanica, 9, 99-128.

PODANI, J. 1988. SYN-TAX III. Computer Programs for data Analysis in Ecology and Systematics. Coenoses 3, 111-119.

RÉGNIER, S. 1965. Sur quelques aspects mathématiques des problèmes de classification automatique. ICC Bulletin, 4, 175-191.

ROHLF, F.J. 1982. Consensus Indices for Comparing Classifications. Mathematical Biosciences, 59, 131-144.

Manuscript received: April 1988