[7] B.Korte, L. Lovász and R. Schrader, *Greedoids*, Springer (1991).

[8] G.N. Lance, and W.T. Williams (1967) A general theory of classificatory sorting strategies: 1. Hierarchical Systems, *Comp. Journal* **9**, 373-380.

[9] V. Levit (1995) Oracle-defined quasi-convex set functions are exponentially hard to maximize. Personal communication.

[10] J. Mullat, Extremal subsystems of monotone systems: I, II; *Automation and Remote Control* **37**, 758-766, **37**, 1286-1294, (1976).

[11] P.H.A. Sneath and R.R. Sokal (1973) *Numerical Taxonomy*, W.H. Freeman, San Francisco.

[12] Yu. Zaks (Kempner) and I. Muchnik (1989) Incomplete Classifications of a Finite Set of Objects Using Monotone Systems, *Automation and Remote Control, 50*, 553-560.

the corresponding minimum split functions. This allows us to claim that the minimum split functions MS_d for the monotone linkages d present yet another class of greedily maximizable functions, though the greedy series employed are d-series, not F_d-greedy series considered usually (see, for instance, Korte, Lovász and Schrader 1991, Helman, Moret and Shapiro 1993, Dress and W. Terhalle 1995). We have proved also that this class coincides with the class of quasi-convex set functions.

Although the problem of maximizing a quasi-convex set function is exponentially hard when F is considered oracle-defined (V. Levit 1995), it can be resolved greedy-wisely when F is represented through a monotone linkage function by equation (3). Thus, the monotone linkage format may well serve as an easy-to-interpret and easy-to-maximize way for dealing with quasi-convex set functions.

On the other hand, the monotone linkage functions may be used as a sufficiently wide framework for developing clustering techniques based on the entity-to-set linkage functions rather than on usual entity-to-entity dissimilarity measures. What seems important is that the framework allows explicitly treating "unclusterable", 'noisy' entities frequently occurring in the real-world data (see also Zaks and Muchnik 1989).

References

[1] M. Delattre and P. Hansen, Bicriterion cluster analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* **4**, 277-291 (1980).

[2] A.W.M. Dress and W. Terhalle, Well-layered maps - a class of greedily optimizable set functions, *Appl. Math. Lett.* **8**(5), 77-80 (1995).

[3] K. Florek, J. Lukaszewicz, H. Perkal, H. Steinhaus, and S. Zubrzycki (1951) Sur la liason et la division des points d'un ensemble fini, *Colloquium Mathematicum* **2**, 282-285 (1951).

[4] P. Hansen and B. Jaumard, Maximum split single cluster clustering, *Les Cahiers du GERAD*, 1-7 (1993).

[5] P. Helman, B.M.E. Moret, and H.D. Shapiro, An exact characterization of greedy structures, *SIAM J. Disc. Math.***6**, 274-283 (1993).

[6] W.-L. Hsu and G.L. Nemhauser, Easy and hard bottleneck location problems, *Discrete Applied Mathematics* **1**, 209-215 (1979).

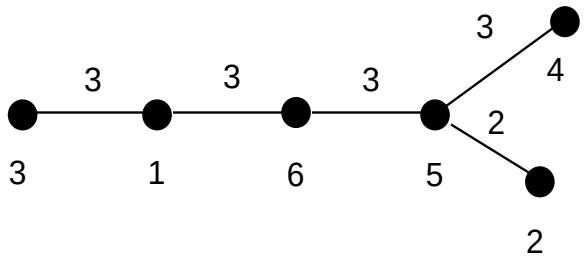

Figure 1: A minimum spanning tree for matrix D.

$\{1\}$, $\{2,5\}$, $\{3\}$, $\{4\}$, $\{6\}$, all corresponding to maximum value $SL(S) = 3$. They form a partition of I since $SL(S) = SL(I - S)$, thus, all the entities must be covered by some of SL maximizers.

The situation is slightly different for the minimum split mix linkage function; its minimal maximizers are $\{1\}$, $\{3\}$, $\{4\}$, and $\{6\}$ while none of the elements 2 and 5 belongs to a maximizer of $MS_{ml}(S)$. Indeed, the six ml-series starting from each of the elements can be taken as follows:
1(3)3(3)2(0)5(1)4(0)6,
2(2)5(2)4(2)6(1)1(0)3,
3(3)1(3)2(0)5(1)4(0)6,
4(3)2(1)5(2)6(1)1(0)3,
5(2)2(2)(2)6(1)1(0)3,
6(3)1(2)3(2)2(0)4(0)5.
The mix linkage values $ml(i_{n+1}, S_n)$ are put in the parentheses between the starting interval S_n seriated and i_{n+1} $(n = 1, ..., 6)$. It can be seen that the maximum value 3 separates each of the four singletons indicated while it never occurs in the series starting with 2 and 5.

6 Conclusion

The monotone linkage functions have been introduced, in clustering framework, by Mullat (1976) who called them "monotone systems" and considered set functions $F(S) = \max_{i \in S} d(i, S)$ as greedily minimizable. In this paper, the concept of minimum split function (Delattre and Hansen 1980) is extended to the case of arbitrary monotone linkage function. We have proved that the monotone linkage clusters are the minimum maximizers of

Proof: Obvious from definition. □.

These two statements imply that the minimal (by inclusion) maximizers of a quasi-convex function $F(S)$ defined by a monotone linkage are not overlapped and they can be found as initial fragments (d-clusters) of some d-series. To do that, d series p_i must be determined starting from each $i \in I$. Then, in each d-series p_i ($i \in I$) the minimal d-cluster is found as a starting fragment S_k with maximum $F(S_k) = d(i_{k+1}, S_k)$ ($k = 1, ..., N-1$). Among the d-clusters, only those maximizing F are left.

The structure of the maximizers of a quasi-convex set function is described as follows.

Statement 7 *Each maximizer of a quasi-convex set function is a union of its minimal maximizers.*

Proof: Indeed, if S^* is a maximizer of $F_d(S)$, than, for each $i \in S^*$, there is a minimal d-cluster containing i, as it follows from the proof of Statement 5.□.

However, the minimal maximizers may not form a partition of I thus leaving some entities unclustered.

5 Example

Let us consider 6-element set $I = \{1, 2, 3, 4, 5, 6\}$ of the rows of a Boolean matrix $X = (x_{ik})$, $k = 1, ..., 7$ presented as follows:

$$X = \left|\begin{array}{c|ccccccc} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 2 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 3 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 4 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 5 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 6 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{array}\right|$$

The matrix of row-to-row Hamming distances (numbers of non coinciding components) is this:

$$D = \begin{pmatrix} 0 & 4 & 3 & 7 & 4 & 3 \\ 4 & 0 & 5 & 3 & 2 & 5 \\ 3 & 5 & 0 & 4 & 5 & 4 \\ 7 & 3 & 4 & 0 & 3 & 4 \\ 4 & 2 & 5 & 3 & 0 & 3 \\ 3 & 5 & 4 & 4 & 3 & 0 \end{pmatrix}$$

A D-based MST is presented in Figure 1. It can be seen from it that the following five subsets are minimal maximizers of the minimum split single linkage function $SL(S)$:

monotone linkage as a tool for defining the quasi-convex set function rather than, reversely, the quasi-convex set function as a tool for representing the monotone linkage.

In relation to this remark, it seems quite obvious that different linkage functions d and d' may produce coinciding minimum split set functions, $F_d = F_{d'}$. In the class of linkage functions producing the same set function $F(S)$ as their minimum split function, the maximum join linkage has a peculiar property: it is the minimum in the class.

Statement 4 *If a set function F is minimum split function for a monotone linkage d, then, for any $S \subset I$ and $i \notin S$, $d_F(i,S) \leq d(i,S)$.*

Proof: For an arbitrary $S \subset I$, let $d_F(i,S) = F(T)$ for some T, $S \subseteq T \subset I - i$. By the condition of the theorem, $F(T) = \min_{j \in I-T} d(j,T)$ which is not larger than $d(i,T)$ since $i \in I-T$. However, $d(i,T) \leq d(i,S)$ since $S \subseteq T$ and d is monotone. Thus, $d_F(i,S) \leq d(i,S)$.□.

4 Maximizing Minimum Split Quasi-Convex Set Function

Let us consider a quasi-convex set function F represented through a monotone linkage function $d(i,S)$ by equation (3). Let us refer to a subset S as a *d-cluster* if S is a maximizer of $F(S)$ with regard to all starting sets for a $d-series$ (that is, $S_k = \{i_1, ..., i_k\}$ for a d-series $(i_1, ..., i_N)$ on I).

Statement 5 *Any maximizer of a quasi-convex function $F(S)$ includes a d-cluster which is a maximizer of F, too.*

Proof: Let S^* be a maximizer of F and p_i is a d-series started from an $i \in S^*$. Then, let S^* is not a starting set; that is, S^* cannot be presented as a continuous segment of the series p_i. In this case, there are some elements between i and the last element of S^* in p_i, which do not belong to S^*; let i^* be the first of them. Let us prove that set T_i of the elements preceding i^* in p_i is a maximizer of F. Indeed, $d(i^*, S^*) \geq F(S^*)$ due to equality (3) applied to $S = S^*$. On the other hand, $d(i^*, T_i) \geq d(i^*, S^*)$ since d is a monotone linkage. Thus, T_i is a maximizer of F. Since T_i is a starting set of p_i, it is a d-cluster, which proves the statement. □.

Another useful statement is this.

Statement 6 *If $S_1, S_2 \subset I$ are overlapping maximizers of a quasi-convex set function $F(S)$, then $S_1 \cap S_2$ is also a maximizer of $F(S)$.*

$d(k, S_2)$. By the definition of F, i does not belong either S_1 or S_2, say, $i \notin S_1$. Then, $d(i, S_1) \geq F(S_1) = d(j, S_1)$ and $F(S_1 \cap S_2) = d(i, S_1 \cap S_2) \geq d(i, S_1)$ since $d(i, S)$ is a monotone linkage, which proves that F is quasi-convex. $\Box$.

Let us define now a linkage function d_F for any set function $F(S)$ with the following equation:

$$d_F(i, S) = \max_{S \subseteq T \subseteq I - i} F(T) \tag{4}$$

The definition means that, for $i \notin S$, $d_F(i, S)$ is either $F(S)$ or $F(T)$ for some $T \subset I$ obtained by adding some elements from $I - S$ (but not i) to S. Function $d_F(i, S)$ will be referred to as *maximum join* linkage function.

Statement 2 *Maximum join linkage function $d_F(i, S)$ is monotone if F is quasi-convex.*

Proof: Obvious, since any increase of the set S makes the set of maximized values in the definition of d_F smaller.$\Box$.

It appears that in the setting defined by conditions of quasi-convexity and monotonicity, the functions d_F and MS_d are dual, that is, for any quasi-convex set function $F : \mathcal{P}(I) \to R$, the minimum split function of its maximum join linkage coinsides with F. Vice versum, for any monotone linkage $d : I \times \mathcal{P}(I) \to R$, the maximum join linkage of its minimum split function coincides with d. It is sufficient to prove only one of these statements; the other obviously follows.

Statement 3 *For any quasi-convex set function $F : \mathcal{P}(I) \to R$, the minimum split function of its maximum join linkage coinsides with F.*

Proof: Let S_i be a maximizer of $F(T)$ with regard to all T satisfying condition $S \subseteq T \subseteq I - i$ for $i \notin S$ so that $d(i, S) = F(S_i)$. Then, the minimum split function for d_F, by definition, is equal to $MS(S) = \min_{i \notin S} F(S_i)$. Thus, $MS(S) \leq F(\cap_{i \notin S} S_i)$, due to quasi-convexity of $F(S)$. But $\cap_{i \notin S} S_i = S$ since $S \subseteq S_i$ and $i \notin S_i$, for every $i \notin S$, which implies $MS(S) \leq F(S)$. On the other hand, $F(S_i) \geq F(S)$, $i \notin S$, since S belongs to the set of the feasible subsets in the definition of S_i as a maximizer of F; this implies that their minimum value, $MS(S) \geq F(S)$, which proves the statement.$\Box$.

The duality proved is asymmetric from the algorithmic point of view: it is quite easy to construct minimum split function MS_d based on a linkage $d(i, S)$ while determining maximum join linkage d_F by $F : \mathcal{P}(I) \to R$ may be a NP-hard problem: the former task involves the elements $i \in I - S$ to enumerate while the latter task requires maximizing a set function $F(T)$ by almost all $T \in \mathcal{P}(I)$. This implies that it would be more appropriate to consider the

Although single and complete linkage functions cannot be presented as the respective increments of $SL(S)$ and $CL(S)$, still it can be proven that seriation with each of these linkages are equivalent to construction of greedy series maximizing $SL(S)$ and minimizing $CL(S)$.

Criterion SL has been considered both for clustering and non-clustering purposes. Non-clustering applications are connected with operation research "bottleneck" problems as, for instance, the problem of location of $|S|$ noxious industrial facilities and $|I| - |S|$ residential facilities among the given sites $i \in I$ in order to maximize the minimum distance between a noxious and a residential facility (Hsu and Nemhauser (1979)). Analysis of the problem in the clustering framework can be found in (Muchnik and Zaks (1989), Delattre and Hansen (1980), Hansen and Jaumard (1993)). The basic observation made by Delattre and Hansen (1980) is that, for every nonempty $S \subset I$, $SL(S)$ is equal to the weight of an edge in every minimum spanning tree (MST). This means, that an optimal cluster is found with cutting MST by an edge of the maximum weight. This provides us with a simple algorithm for optimizing $SL(S)$.

3 Monotone Linkage and Quasi-Convex Set Function

In the sequel, a solution to the problem of connecting monotone linkage functions with set functions will be done.

Let us refer to a dissimilarity linkage function $d(i, S)$, $S \subset I, i \in I - S$ as a *monotone linkage* if $d(i, S) \geq d(i, T)$ whenever $S \subset T$ (for all $i \in I - T$). Examples of the monotone linkage functions are $sl(i, S)$, $tl_\pi(i, S)$ (when all $\pi - d_{ij}$ are non positive), and $ml(i, S)$ in the list above.

A set function $MS_d(S)$ can be defined based on a linkage function $d(i, S)$ as follows:

$$MS_d(S) = \min_{i \in I-S} d(i, S). \tag{3}$$

This set function was considered, among others, by Zaks and Muchnik (1989). Following terminology of Delattre and Hansen (1980), $MS_d(S)$ can be referred to as *minimum split function* for linkage $d(i, S)$. The minimum split function measures the minimum linkage between S, as a whole, and $I - S$ as the set of the "individual" entities.

Statement 1 *Minimum split function for a monotone linkage is quasi-convex.*

Proof: Let $F(S) = \min_{i \in I-S} d(i, S)$ for some monotone linkage $d(i, S)$ and S_1, S_2 be overlapping subsets of I. Let $F(S_1 \cap S_2) = d(i, S_1 \cap S_2)$ while $F(S_1) = d(j, S_1)$ and $F(S_2) =$

Linkage Method	α_S	α_j	β	γ	$const$
Single	1/2	1/2	0	1/2	0
Complete	1/2	1/2	0	-1/2	0
Summary	1	1	0	0	0
Average	$\|S\|/(\|S\|+1)$	$1/(\|S\|+1)$	0	0	0
Threshold	-1	-1	0	0	π
Chain	0	1	0	0	0

Table 1: Coefficients of the modified Lance-Williams formula corresponding to a dissimilarity linkage measure $d(i, S)$.

of greedy optimization of F, and vice versa. The F-greedy series is defined as a series $p = (i_1, i_2, ..., i_N)$ produced by the following rule: i_1 is arbitrary and, for all $k = 1, ..., N-1$, i_{k+1} is any of maximizers of $F(i_1, ..., i_k, i)$ with regard to $i \in I - \{i_1, ...i_k\}$.

For the collection of the six particular linkage functions above, some set functions can be indicated as a kind of solution to the problem. They are, respectively,

1. $SL(S) = \min_{i \in I-S} \min_{j \in S} d_{ij}$;
2. $CL(S) = \max_{i \notin S} \max_{j \in S} d_{ij}$;
3. $SUL(S) = \sum_{i \in S} \sum_{j \in S} d_{ij}$;
4. $AL(S) = \sum_{i \in S} \sum_{j \in S} d_{ij}/\nu(S)$
5. $L(S, \pi) = \sum_{i \in S} \sum_{j \in S} (\pi - d_{ij}) = \pi|S|\nu(S) - \sum_{i \in S} \sum_{j \in S} d_{ij}$;
6. $LL(S) = \sum_k d_{i_k i_{k+1}}$

   where S is considered seriated as $p_S = (i_1, i_2, ..., i_{|S|})$.

In these formulas, $\nu(S)$ equals $|S|$ if the diagonal entries d_{ii}, $i \in N$, are given or $|S| - 1$ if not.

Three of the functions, $SUL(S)$, $L(S, \pi)$, and $LL(S)$, match the problem above perfectly: the increment $F(S) - F(S + i)$ for each of them equals $sul(i, S)$, $l_\pi(i, S)$, and $ll(i, S)$, respectively. Among these functions, $L(S, \pi)$ suggests a stopping rule in the clustering problems: seriation must stop when $l_\pi(i, S)$ becomes negative, which highly depends on the threshold value π.

means that we should be able calculating linkage between i and $S+j$ (where j is the entity added to S at the step considered) for $i \notin S+j$ from the linkages between i and S and the pair-wise linkage between i and j. Six rules considered above admit such a possibility:

1. $sl(i, S+j) = \min(sl(i, S), d_{ij})$;
2. $cl(i, S+j) = \max(sl(i, S), d_{ij})$;
3. $sul(i, S+j) = sl(i, S) + d_{ij})$;
4. $al(i, S+j) = (|S|al(i, S) + d_{ij})/(|S|+1)$;
5. $l_\pi(i, S+j) = l_\pi(i, S) + \pi - a_{ij}$;
6. $ll(i, S+j) = d_{ij}$.

These formulas can be expressed in a unified way with the formula proposed by Lance and Williams (1967) in a somewhat different context of agglomerative clustering. Adapted to the seriation framework, Lance-Williams formula is as follows:

$$d(i, S+j) = \alpha_S d(i, S) + \alpha_j d(i, j) + \beta d(j, S) + \gamma |d(i, S) - d(i, j)| + const \qquad (2)$$

where α_S, α_j, β, γ, $const$ are coefficients to be specified according to the particular linkage method used. Although the $const$ when S is fixed does not change the order of the values and, thus, does not affect the seriation process, it does change the values of the linkage measure (across different S_ks) and is included to take into account the threshold linkage formula.

Table 1 contains the values of the coefficients corresponding to the six linkage measures, 1 to 6, above.

The mix linkage function cannot be expressed this way since adding a point to S may make changes in a 'deeper' level of $ml(i, S)$, within the terms for each of the coordinates $k \in K$.

The major problem we would like to deal with concerns relationship between the linkage and set functions. A most standard way of relating them involves equation $F(S+i) - F(S) = -d(i, S)$ where $F(S)$ is a set function, $d(i, S)$ is a corresponding linkage function, and $S+i$ denotes set S with $i \in I$ added to. The equation can be considered an analog of the equation between a numeric function and its derivative in calculus, and it has a similar meaning leading to the fact that F-greedy series are simultaneously d-series, that is, every d-series is a result

5. *Threshold linkage*

$$l_\pi(i,S) = \sum_{j\in S}(\pi - d_{ij}) = \pi|S| - \sum_{j\in S} d_{ij};$$

where π is a fixed threshold value.

In the situation when the elements of S are ordered in a series, $p_S = (i_1, ..., i_{|S|})$, yet another measure might be useful,

6. *Chain linkage* or *Last neighbor*:

$$ll(i,S) = d_{ii_{|S|}}.$$

Linkage measures can be defined also for different kinds of data. For example, if a data table is of form $X = (x_{ik})$ where x_{ik} is value of variable $k \in K$ for entity $i \in I$, a linkage measure can be defined as

7. *Mix linkage*

$$ml(i,S) = \sum_{k\in K} \min_{j\in S} |x_{ik} - x_{jk}|.$$

The mix linkage is an example of a "holistic" linkage measure which cannot be reduced to a function of the pair-wise dissimilarities between all i and $j \in S$.

One might think also of a situation when such a linkage measure arises just as a kind of primary data, which seems possible in technical applications connected to VLSI or image processing. However, the present authors have nothing to say on that.

For any linkage measure $d(i,S)$, a one-by-one adding procedure for d-seriation of the entities in I can be defined as follows.

The procedure starts with $S = \{i_1\}$ where $i_1 \in I$ is an arbitrary entity.

The general step: when S seriated is given, find $i^* \in I - S$ minimizing dissimilarity linkage measure $d(i,S)$ with regard to all $i \in I - S$ and join i^* to S as the last element in the series. This is repeated unless $S = I$.

The output of the process is a d-series along with the sequence of the linkage measure values between every starting set and the next to it entity. Obviously, the number of the basic comparison operations here is $O(N^2)$ since at every step S is compared with all the remaining elements (average number of them is $N/2$), and the number of steps is $O(N)$.

To simplify the computations, it would be nice to calculate the linkage $d(i,S)$ at every step not just from the scratch but updating the linkage value from the previous step. That

minimum split functions for monotone linkages coincides with the set of quasi-convex set functions $F : \mathcal{P}(I) \to R$ satisfying the condition that

$$F(S_1 \cap S_2) \geq \min(F(S_1), F(S_2)), \tag{1}$$

for any overlapping $S_1, S_2 \in \mathcal{P}(I)$ (Zaks and Muchnik, 1989).

This provides both a simple algorithm for maximizing quasi-convex set functions presented as minimum split functions for some monotone linkage and a natural mechanism for generation of the quasi-convex set functions.

The remainder consists of the following. In Section 2, the concept of linkage function is introduced, cluster-analysis-based examples are given, and an interpretation of the linkage-function clustering as the greedy-like optimization procedure is discussed. In Section 3, the major concept of the minimum split function is introduced and its duality toward the quasi-convex set function is stated. In Section 4, it is shown that the starting sets of minimum linkage d-series are minimal maximizers of the corresponding minimum split functions. An example is given in Section 5. The results are discussed in Section 6.

2 Linkage Clustering

Let, for every subset $S \subset I$ and every $i \in \bar{S} = I - S$, a dissimilarity measure, $d(i, S)$, between i and S, be known. Such a measure could be called a dissimilarity *linkage* between i and S. A set of the linkages can be defined based on an entity-to-entity dissimilarity matrix given, $D = (d_{ij})$, $i, j \in I$, as follows.

1. *Single linkage* or *Nearest neighbor*

$$sl(i, S) = \min_{j \in S} d_{ij};$$

2. *Complete linkage* or *Farthest neighbor*

$$cl(i, S) = \max_{j \in S} d_{ij};$$

3. *Summary linkage*

$$sul(i, S) = \sum_{j \in S} d_{ij};$$

4. *Average linkage* or *Average neighbor*

$$al(i, S) = \sum_{j \in S} d_{ij}/|S|;$$

1 Introduction

The subject of the paper originated in cluster analysis framework. It can be outlined as a generalization of a well-known method called single-linkage clustering (Sneath and Sokal 1973). Let $d = (d_{ij})$ be a symmetric $N \times N$ matrix of the dissimilarities d_{ij} between all $i, j \in I$. For a subset $S \subset I$ and an entity $i \in I - S$, let us define $sl(i, S) = \min_{j \in S} d_{ij}$, which can be referred to as single linkage dissimilarity between i and S. Series $s = (i_1, i_2, ..., i_N)$ of the elements of I will be referred to as a single linkage series if, for every $k = 1, ..., N-1$, i_{k+1} is a minimizer of $sl(i, \{i_1, ..., i_k\})$ with regard to $i \in I - S$. Every set $S_k = \{i_1, ..., i_k\}$ corresponding to a starting interval of the series s will be referred to as a starting set of series s. A single linkage cluster is defined as a starting set S_k of a single linkage series having $sl(i_{k+1}, S_k)$ maximum by all $k = 1, ..., N-1$.

Let us consider a complete weighted graph G with I as its vertex set and a dissimilarity matrix $D = (d_{ij})$ as its weight function. The length of a spanning tree $H = (I, E(H))$ is defined as the sum of the weights of its edges: $l(H) = \sum_{(i,j) \in E(H)} d_{ij}$. A spanning tree is called *minimum spanning tree (MST)* if its length is minimum over the set of all spanning trees in G. The MST concept was originally invented, in the beginning of this century (Wrozlaw, Poland, see K. Florek et al. 1951), as a tool in biological taxonomy and it has become an important concept in the mathematical theory of a clustering approach. It turns out rather simple; finding MST is nothing but finding a sl-series along with all the pairs (i_{k+1}, j), where j minimizes $sl(i_{k+1}, S_k)$ by $j \in S_k$, giving the edge set $E(MST)$. This means that any single linkage cluster can be found with cutting MST by an edge $(i, j) \in H$ of maximum weight. As it is known, the greedy algorithm finds the global solutions since the MSTs are the bases in the matroid of all the acyclical edge subsets (see Korte, Lovász and Schrader 1991, Helman, Moret, and Shapiro 1993).

The single linkage dissimilarity function $sl(i, S)$ satisfies the monotonicity property (coined by Mullat 1976), that is, its value can only decrease when some elements (not coinciding with i) are added to S. This allows us extending all the clustering concepts above to an arbitrary monotone linkage function, $d(i, S)$, whether it is defined based on a dissimilarity matrix or not. It can be easily seen that a set function, $SL(S) = \min_{i \in I-S} sl(i, S)$, called minimum split function, is maximized by a single linkage cluster (Delattre and Hansen 1980). This paper is aimed at proving that, for any monotone linkage series, corresponding minimum split function, $D(S) = \min_{i \in I-S} d(i, S)$, has all its minimal maximizers among the monotone linkage clusters (starting sets of d-series). Moreover, it appears that the entire stock of the

ABSTRACT

Greedy selection of the subsets with adding the elements one by one is implicitly employed in many heuristical clustering procedures. Such a seriation procedure can be described generally in terms of a linkage function measuring entity-to-set dissimilarities.

A quite known clustering technique, single linkage clustering, can be considered an example of the seriation procedure (actually, based on the minimum spanning tree construction) leading to the global maximum of a corresponding "minimum split" set function. The purpose of this work is to have the property extended to a much wider class of so called monotone linkage functions. It is proved that the minimum split functions of monotone linkage functions can be maximized greedy-wisely, with the linkage function seriating. Moreover, it appears that this class of set functions coincides with the class of so called quasi-convex set functions. Thus, a class of set functions is characterized 'analytically', and a greedily maximizable format for it is indicated.

DIMACS Technical Report 95-44
September 1995

Monotone Linkage Clustering and Quasi-Convex Set Functions

by

Yulia Kempner
School of Mathematical Sciences
Tel-Aviv University
Ramat-Aviv 69978 Israel

Boris Mirkin [1,2]
DIMACS, Rutgers University
P.O.Box 1179 Piscataway NJ 08855-1179
and CEMI of Russian Academy of the Sciences, Moscow, Russia

Ilya Muchnik [3]
RUTCOR, Rutgers University
New-Brunswick NJ 08903

[1] Research Associate.

[2] The author thanks the Office of Naval Research for its support under grant number N00014-93-1-0222 to Rutgers University.

[3] The author gratefully acknowledges the support of DIMACS (grant NSF-CCR-91-1999) and RUTCOR (grant F49620-93-1-0041).

DIMACS is a cooperative project of Rutgers University, Princeton University, AT&T Bell Laboratories and Bellcore.

DIMACS is an NSF Science and Technology Center, funded under contract STC–91–19999; and also receives support from the New Jersey Commission on Science and Technology.