Monotone Linkage Clustering and Quasi-Convex Set Functions

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

Boris Mirkin*†
DIMACS, Rutgers University
P.O. Box 1179, Piscataway NJ 08855-1179, USA
and CEMI of Russian Academy of the Sciences, Moscow, Russia

Ilya Muchnik ‡
RUTCOR, Rutgers University
New-Brunswick NJ 08903, USA

Abstract

Greedily seriating objects one by one is implicitly employed in many heuristic clustering procedures, which can be described in terms of a linkage function measuring entity-to-set dissimilarities.

A well-known clustering technique, single linkage clustering, can be considered as an example of the seriation procedures (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 extend this property to a wide class of the so-called monotone linkages. It is shown that the minimum split functions of the monotone linkages can be greedily maximized. Moreover, this class of set functions is proven to coincide with the class of so-called quasi-convex set functions.

Key words: clustering, monotone linkage, quasi-convexity, greedy optimization.

1 Introduction

The subject of this paper originated in cluster analysis as a generalization of a well-known method called single linkage clustering [1]. Let $D = (d_{ij})$ be a symmetric $N \times N$ matrix of the dissimilarities d_{ij} between elements, i, j, of an N-element set I. For a subset $S \subset I$ and an element $i \in I - S$, let us define $l(i, S) = \min_{j \in S} d_{ij}$, single linkage dissimilarity between i and S. A sequence $s = (i_1, i_2, ..., i_N)$ consisting of all elements of I will be referred to as a *series*, and sets $S_k = \{i_1, ..., i_k\}$ consisting of the initial fragments of s, as its *starting sets* ($k = 1, 2, ..., N-1$). A series $s =$

*The person to communicate with.

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

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

$(i_1, i_2, ..., i_N)$ is a *single linkage series* if, for every $k = 1, ..., N-1$, the element i_{k+1} is a minimizer of $l(i, S_k)$ with regard to $i \in I - S_k$. A starting set S_k of a series s can be referred to as a *single linkage cluster* if it is maximally separated from the other elements along the series, that is, if $l(i_{k+1}, S_k)$ is maximum over all $k = 1, ..., N-1$. Basically, a single linkage series defines a minimum spanning tree (MST) of the graph whose vertex set is I and edge weight function is $d = (d_{ij})$, in the framework of the well-known Dijkstra-Prim algorithm for finding a MST. By cutting any MST at any of the links (i_k, i_{k+1}) whose value $d_{i_k i_{k+1}}$ is maximum over $k = 1, ..., N-1$, the set I is partitioned into two parts, one of which is a single linkage cluster. It can be shown that a set function, $L(S) := \min_{i \in I-S} l(i, S)$, called the *minimum split function*, is maximized by every single linkage cluster [2].

The single linkage $l(i, S)$ satisfies a monotonicity property [3]: its value can only decrease when some elements (not coinciding with i) are added to S. All the clustering concepts above can be extended to an arbitrary monotone linkage function, $d(i, S)$, whether it is defined in terms of a dissimilarity matrix or not. This paper is aimed at proving that, for any monotone linkage function, d, its minimum split function, $M_d(S) := \min_{i \in I-S} d(i, S)$, has all its minimal maximizers among the monotone linkage clusters defined over the monotone linkage series. Moreover, it appears that the entire stock of the minimum split functions for the monotone linkages coincides with the set of quasi-convex set functions $F : \mathcal{P}(I) \to R$ defined by condition that

$$F(S_1 \cap S_2) \geq \min(F(S_1), F(S_2)),$$

for any overlapping $S_1, S_2 \in \mathcal{P}(I)$ [4].

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

The remainder consists of the following. In Section 2, the monotone linkage and corresponding minimum split function concepts are discussed, and their duality is proven as related to the quasi-convex set functions. In Section 3, it is shown that the minimal maximizers of a minimum split function are starting sets of the corresponding linkage series, while every non-minimal maximizer is just a union of some of the minimal ones. An example is given in Section 4. The results are discussed in Section 5.

2 Monotone Linkage and Quasi-Convex Set Function

Let, for every $S \subset I$ and $i \in I - S$, a dissimilarity measure, $d(i, S)$, be given. Such a measure, referred to as a *linkage* between i and S, can be defined in terms of different data formats. For example, for a data table $X = (x_{ik})$ where x_{ik} is the value of a variable $k \in K$ for any entity $i \in I$,

a linkage measure can be defined as

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

The *ml* linkage is an example of a 'holistic' measure which cannot be reduced to a function of pair-wise dissimilarities. One might think also of a situation when a linkage measure arises just as a primary data, e.g., in applications connected to VLSI or image processing.

Let us refer to a 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 $i \in I - T$). Both of the specific linkage functions considered, $l(i,S)$ and $ml(i,S)$, are monotone.

Based on a linkage function d, a set function M_d can be defined, as follows:

$$M_d(S) := \min_{i\in I-S} d(i,S). \tag{1}$$

This set function was considered, among others, in [4]. Following the terminology introduced in [2], M_d can be referred to as the *minimum split function* for linkage d. The minimum split function measures the minimum linkage between S, as a whole, and $I - S$ as the set of the "individual" entities. A set function $F : \mathcal{P}(I) \to R$ is called *quasi-convex* [4] if

$$F(S_1 \cap S_2) \geq \min(F(S_1), F(S_2)), \tag{2}$$

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

Statement 1 *The minimum split function of any monotone linkage is quasi-convex.*

Proof: Let $F(S) = M_d(S) := \min_{i\in I-S} d(i,S)$ for some monotone linkage d, and let S_1, S_2 be overlapping subsets of I. Assume $F(S_1 \cap S_2) = d(i, S_1 \cap S_2)$, $F(S_1) = d(j, S_1)$ and $F(S_2) = d(k, S_2)$. By the definition of F, i does not belong either to S_1 or to S_2, say, $i \notin S_1$. Then, $d(i, S_1) \geq F(S_1) = d(j, S_1)$ and $F(S_1 \cap S_2) = d(i, S_1 \cap S_2) \geq d(i, S_1)$ due to monotonicity of d, which proves that F is quasi-convex. □

Let us define now the *maximum join* linkage function d_F for any set function F by:

$$d_F(i,S) := \max_{S\subseteq T\subseteq I-i} F(T) \tag{3}$$

for any $S \subset I$ and $i \in I - S$.

Statement 2 *The maximum join linkage d_F is monotone.*

Proof: Obvious since any increase of S makes the set of maximized values in (3) smaller. □

Next, we show that in the setting defined by conditions of quasi-convexity and monotonicity, the functions d_F and M_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 coincides with F. This immediately implies that, for any monotone linkage d, the maximum join linkage of its minimum split function coincides with d.

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

Proof: For $S \subset I$ and $i \in I - S$, let S_i be a maximizer of $F(T)$ over all T satisfying the condition $S \subseteq T \subseteq I - i$, so that $d_F(i, S) = F(S_i)$. The minimum split function for d_F, by definition, is equal to $M(S) = \min_{i \notin S} F(S_i)$. Thus, $M(S) \leq F(\cap_{i \notin S} S_i)$, 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 $M(S) \leq F(S)$. 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 $M(S) \geq F(S)$, which proves the statement. □

The duality proven is asymmetric from the algorithmic point of view: it is quite easy to construct the minimum split function M_d associated with a linkage d while determining the maximum join linkage d_F by $F : \mathcal{P}(I) \to R$ may be an exponentially hard problem: the former task involves the elements $i \in I - S$ to enumerate while the latter task requires maximizing a set function $F(T)$. This implies that it would be more appropriate to consider the monotone linkage as a means for defining the quasi-convex set function rather than, reversely, the quasi-convex set function as a tool for representing the monotone linkage.

Different linkage functions d and d' may produce coinciding minimum split functions, $M_d = M_{d'}$. The maximum join linkage is peculiar: it is the minimum in its class.

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

Proof: For an arbitrary $S \subset I$, assume $d_F(i, S) = F(T)$ for some T with $S \subseteq T \subset I - i$. By definition, $F(T) = \min_{j \in I - T} d(j, T) \leq d(i, T)$ since $i \in I - T$. However, $d(i, T) \leq d(i, S)$ since $S \subseteq T$ and d is monotone. Thus, $d_F(i, S) \leq d(i, S)$. □

3 Maximizing Minimum Split Quasi-Convex Set Function

Let us consider a quasi-convex set function F such that $F = M_d$ in (1) for a monotone linkage function d. Let us refer to a series $(i_1, ..., i_N)$ as a *d-series* if $d(i_{k+1}, S_k) = \min_{i \in I - S_k} d(i, S_k)$ for any starting set $S_k = \{i_1, ..., i_k\}$, $k = 1, ..., N - 1$. This definition can be considered as a description of a greedy procedure for construction of a d-series starting with any $i_1 \in I$: having S_k defined, take any i minimizing $d(i, S_k)$ over all $i \in I - S_k$ as i_{k+1}, $k = 1, ..., N - 1$. A subset $S \subset I$ will be referred to as a *d-cluster* if there exists a d-series, $s = (i_1, ..., i_N)$, such that S is a maximizer of $F(S)$ over all starting sets S_k of s. Greedily found d-clusters play important part in maximizing of the quasi-convex set functions.

Statement 5 *Any maximizer of F includes a d-cluster which is a maximizer of F, also.*

Proof: Let S^* be a maximizer of F and p_i be a d-series starting from an $i \in S^*$. Then, let S^* be not a starting set of p_i, which means that 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, that 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, $F(T_i) = d(i^*, T_i)$ by definition. Then, $d(i^*, S^*) \geq F(S^*)$ due to (1) applied to $S := S^*$. On the other hand, $d(i^*, T_i) \geq d(i^*, S^*)$ since d is a monotone linkage. Thus, $F(T_i) \geq F(S^*)$, and 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. □

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 $F(S)$.*

Proof: Obviously follows from (2). □

This implies that the minimal (by inclusion) maximizers of a quasi-convex set function are not overlapping and, thus, the number of them is not larger than N. Moreover, every non-minimal maximizer can be partitioned into a set of the minimal ones:

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

Proof: Indeed, if S^* is a maximizer of $F = M_d$, then, for each $i \in S^*$, there exists a minimal d-cluster containing i, as it follows from the proof of Statement 5. □

Finding all the minimal maximizers of a quasi-convex set function $F = M_d$ for a monotone linkage d is not a difficult task. It can be solved with the following three-step *extended greedy procedure* (EGP):

(A) For each $i \in I$, greedily define a d-series p_i starting from i as its first element.

(B) For each of the d-series found, p_i, find T_i, a minimal d-cluster as its smallest starting fragment S_k having maximum $F(S_k) = d(i_{k+1}, S_k)$ over all $k = 1, ..., N-1$.

(C) Among the non-coinciding minimal d-clusters T_i, $i \in I$, choose those maximizing F.

Performing EGP takes $O(N^2 g)$ time where g is the average time required to calculate the values $d(i, S)$, which is determined by the step (A) where N series are constructed, each taking $O(Ng)$ time.

Statement 8 *EGP finds all the minimal maximizers.*

Proof: Assume that, for an $i \in I$, there exists a series q_i starting with i, whose minimal d-cluster Q_i does not belong to the set of clusters found with EGP. Then, $T_i \cap Q_i$ contains i and, thus, is a maximizer of F, strictly included in T_i, which contradicts the minimality of T_i along p_i. □

4 Example

Let us consider set $I = \{1, 2, 3, 4, 5, 6\}$ of the rows of a 6×7 Boolean matrix X:

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

The row-to-row Hamming distances (numbers of non-coinciding components) form the following matrix:

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

A D-based MST, presented in Fig. 1, shows that the following five subsets are the minimal maximizers of the minimum split single linkage function L: $\{1\}$, $\{2,5\}$, $\{3\}$, $\{4\}$, $\{6\}$, all corresponding to the maximum value $L(S) = 3$. They form a partition of I since $L(S) = L(I - S)$, implying that all the entities must be covered by the maximizers of L.

The situation is slightly different for the minimum split of ml; its minimal maximizers are $\{1\}$, $\{3\}$, $\{4\}$, and $\{6\}$ while none of the elements 2 and 5 belongs to a maximizer of M_{ml}. Indeed, let us take a look at six ml-series starting from each of the elements of I:

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 value $ml(i_{k+1}, S_k)$ is put in parentheses between every starting interval S_k seriated and i_{k+1} ($k = 1, ..., 5$). 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 or 5.

5 Conclusion

The monotone linkage functions have been introduced, in clustering framework, by Mullat [3] who called them 'monotone systems' and considered set functions $G(S) := \max_{i \in S} d(i, S)$ as greedily minimizable. In this paper, the concept of minimum split function [2] is extended to the case of the monotone linkage functions. We have proven that the minimal maximizers of a minimum split function are monotone linkage clusters that can be found with the extended greedy procedure EGP. This allows us to claim that the minimum split functions M_d for the monotone linkages d present yet another class of greedily maximizable functions, though the greedy series employed are d-series

rather than M_d-greedy series considered usually (see, for instance, [5]). We have proven also that this class coincides with the class of quasi-convex set functions.

Although the problem of maximizing the quasi-convex set functions is exponentially hard when they are oracle-defined [6], it can be resolved with the extended greedy procedure, when they are associated with the monotone linkages via (1). Thus, the monotone linkage format may well serve as an easy-to-interpret and easy-to-maximize means for dealing with the quasi-convex set functions.

On the other hand, the monotone linkage concept may be used as a framework for developing clustering techniques based on the entity-to-set linkage functions rather than on the conventional entity-to-entity dissimilarity measures. The 'unclusterable', 'noisy' entities frequently occurring in the real-world data can be explicitly treated in this framework.

References

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

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

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

[4] Yu. Zaks (Kempner) and I. Muchnik, Incomplete classifications of a finite set of objects using monotone systems, *Automation and Remote Control* **50**, 553-560 (1989).

[5] 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).

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

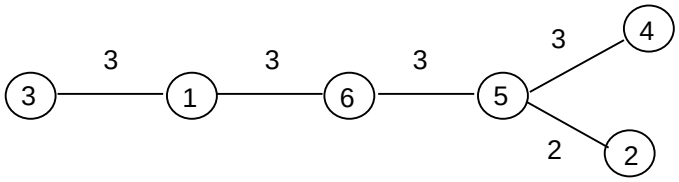

Figure 1: A minimum spanning tree for matrix D.