Layered Clusters of Tightness Set Functions*

Boris Mirkin† Ilya Muchnik‡

March 22, 2000

Abstract

A method for structural clustering is proposed involving data on subset-to-entity linkages that can be calculated with structural data such as graphs or sequences or images. The method is based on the layered structure of the problem of maximization of a set function defined as the minimum value of linkages between a set and its elements and referred to as the tightness function. When the linkage function is monotone, the layered cluster can be easily found with a greedy type algorithm.

Key words: layered cluster, monotone linkage, greedy optimization.

1 Introduction

The problem of separation of a dense core in a given set of interrelated objects has attracted attention of researchers in such disciplines as Operations Research (knapsack and location problems), Social Choice (selection of representatives), Data Mining (finding a pattern in data), Clustering (single cluster clustering), etc. The problem is traditionally formalized in terms of a set function $F(H)$ defined on the set of subsets H of a finite set I. The function is assumed to score the subsets according to their "density" so that an H^* maximizing $F(H)$ can be referred to as a maximum density core in I. Typically, finding a solution to the maximization problem may involve enumeration of an exponential number of subsets with regard to the cardinality of I (NP-hard problems), but when $F(H)$ is relatively simple, this can be done with a polynomial-time algorithm (see, for instance, [1]– [3]).

In particular, Mullat [1] introduced a specific framework for defining a set function (called here the tightness function) as an integral characteristic of a function scoring linkages between subsets and their elements. For any non-empty subset $H \subseteq I$, the tightness function, $F_\pi(H)$, is the minimum of a monotone linkage function $\pi(i, H)$ over $i \in H$. The tightness functions can be maximized with greedy-type algorithms. This framework has been effectively applied for finding maximally dense cores in such applications as ecology and organization design [1], [4].

In this paper, we introduce and explore the concept of layered cluster which involves not only the maximally dense core of I but also a nested chain of its "shells" whose densities monotonely decrease with the growth of the shells. This concept can be considered an abstract implementation of the idea of multiresolutional view at the structure of a system of interrelated elements.

We define a set function pattern as a subset which is strongly separated from the rest: its score, according to the set function, decreases if any supplementary elements are added, even if some of its elements are removed. The set of patterns, proven to be nested, is referred to as the layered cluster if the smallest pattern is a global maximizer of the function. The existence of the layered cluster is proven for the tightness functions of monotone set-to-element linkage functions, and a greedy type "serial partitioning" algorithm for finding the layered cluster is proposed.

*The work was done in the Center for Discrete Mathematics and Theoretical Computer Science (DIMACS) at Rutgers University. The authors thank DIMACS' Director Dr. F. Roberts for his support. The authors are indebted to T. Fenner and G. Loizou for thorough discussions of the material and help in shaping it.

†School of Computer Science and Information Systems, Birkbeck College, London WC1E 7HX, UK, e-mail: mirkin@dcs.bbk.ac.uk

‡DIMACS, Rutgers University, Piscataway NJ, 08854-8018, USA, e-mail: muchnik@dimacs.rutgers.edu

2 Layered patterns of a set function

When a set function $F(H)$, $H \subseteq I$, reflects the density of interrelations within sets H in such a way that the greater $F(H)$ the greater the density of H, one may wish to investigate relatively dense subsets H.

Let us refer to a subset $H \subseteq I$ as to a pattern with regard to $F(H)$ if H is separated from the rest in such a way that $F(H)$ is greater than $F(H')$ for any H' which is not part of H, that is, $F(H) > F(H')$ for any $H' \subseteq I$ such that $H' \cap (I - H) \neq \emptyset$.

Thus defined, patterns must be chain-nested.

Assertion 1 *The set of all patterns, P, is nonempty and chain-nested, that is, $H_1 \subseteq H_2$ or $H_2 \subseteq H_1$ for any $H_1, H_2 \in P$.*

Proof: Indeed, if H_1, H_2 are patterns and H_1 is not a part of H_2, then $F(H_2) > F(H_1)$. If, moreover, H_2 is not a part of H_1, then $F(H_1) > F(H_2)$. The contradiction proves the nesticity. Besides, $H = I$ makes the definition of a pattern true because of the false premise, which proves that I is always a pattern. □

In general, more dense subparts may occur within the smallest pattern S: nothing prevents one or more $H \subset S$ with $F(H) > F(S)$ to exist. When this is not the case, that is, when the smallest pattern is a global maximizer of the function $F(H)$, the set of patterns can be considered as a complete representation of the density structure in I. The set of patterns will be referred to as the layered cluster of I according to function $F(H)$ if the minimum pattern is a global maximizer of $F(H)$. Obviously, the layered cluster is unique.

The patterns of a layered cluster can be considered as levels of resolution of the overall similarity modelled by the set function.

3 Monotone linkage and tightness functions

To catch the similarity structure in a system such as a digitallized image or folded protein, the concept of an element-to-set linkage function can be utilized. A linkage function $\pi(i, H)$ measures proximity between subsets $H \subseteq I$ and their elements $i \in H$. This measure can be defined in terms of pair-wise distances as, for instance, $\pi(i, H) = \sum_{j \in H} d_{ij}$, or similarities as, for instance, $\pi(i, H) = \max_{j \in H} s_{ij}$. (Some other ways for defining linkage functions are considered in [5].) In these examples, an important feature is that the linkage functions are monotone. A linkage function π is referred to as a monotone one if for any $H, G \subset I$ and any $i \in H$, $\pi(i, H) \leq \pi(i, H \cup G)$.

As an example, let us consider the set I whose similarity structure is presented by the edge-weighted graph in Figure 1. The linkage function $\pi(i, H)$ in this example is defined as the sum of the weights of edges connecting i with $j \in H$. For instance, in the set $H = \{c, d, f, j\}$, $\pi(c, H) = 13$, $\pi(d, H) = 13+8+8 = 29$, $\pi(f, H) = 8$ and $\pi(j, H) = 8$. Obviously, thus defined $\pi(i, H)$ is monotone.

A linkage function, $\pi(i, H)$, can be used to estimate the overall density of a subset $H \subseteq I$ by "integrating" its values $\pi(i, H)$ over $i \in H$. In particular, an integral function defined as

$$F_\pi(H) = \min_{i \in H} \pi(i, H) \tag{1}$$

will be referred to as the tightness function when π is monotone.

In the example above, $F_\pi(H) = 8$ for $H = \{c, d, f, j\}$.

A property of the tightness function (easily following from the monotonicity of π) is that it satisfies the so-called quasi-convexity condition: for any $H, G \subseteq I$,

$$F(H \cup G) \geq \min(F(H), F(G)). \tag{2}$$

Actually, inequality (2) is a characteristic of the tightness functions [6], [3], but this result will not be used in this paper.

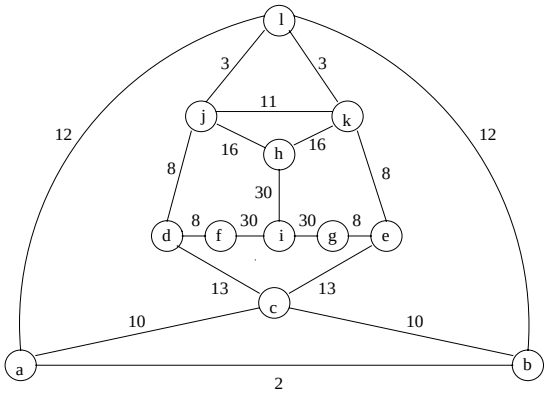

Figure 1: Weighted graph generating the summary linkage function π.

4 Minimum pattern of a tightness function

The following statement shows that the set of patterns of a tightness function $F(H)$ is always a layered cluster.

Assertion 2 *If $F(H)$ is a tightness function, then its minimum pattern is the largest global maximizer of $F(H)$ over all $H \subseteq I$ (with regard to the set-theoretic inclusion).*

Proof: Let S be the minimum pattern in the chain nested set of patterns of $F(H)$. If S is not a global maximizer of F, then $F(H) > F(S)$ for some $H \subseteq I$. In fact, all such H must fall within S, because S is a pattern. Let us take a maximal subset $H \subset S$ in the set of all H such that $F(H) > F(S)$ and prove that H is a pattern as well. Indeed, let us take any $S' \subseteq S$ such that $S' \cap (I - H) \neq \emptyset$; the existence of such S' follows from the fact that H does not coincide with S. Then $F(H) > F(H \cup S')$ because of the assumed maximality of H within S. But $F(H \cup S') \geq \min(F(S'), F(H))$ according to (2), that is, $F(H) > F(S')$. Let us consider now an S' which is not contained in S and still satisfies the condition $S' \cap (I - H) \neq \emptyset$. (This may only happen when S is not equal to I.) By the definition of S', $F(S) > F(S')$ because S is a pattern. Therefore, $F(H) > F(S')$. This implies that H is a pattern, which contradicts the assumption of minimality of S. Thus, no $H \subset S$ exists with $F(H) > F(S)$, which implies that S is the maximum global maximizer of $F(H)$. □

5 Finding layered clusters with serial partitioning

Let us denote by $m(H)$ the set of elements $i \in H$ at which the value of $F(H)$ is reached:

$$m(H) = \{i : \pi(i, H) = \min_{j \in H} \pi(j, H)\}.$$

Obviously, $m(H)$ is not empty if H is not empty. Iteratively applying the operation m to I, $I - m(I)$, etc., one can build what will be referred to as the serial partition of I.

Algorithm "Serial Partitioning".

Input: Monotone linkage function $\pi(i, H)$ defined for all pairs i, H such that $i \in H \subseteq I$.

Output: Serial partition $M = (M_0, M_1, ..., M_n)$ of I along with class values $F = \{F_0, F_1, ..., F_n\}$ defined as follows.

Step 0. Initial setting: Put $t = 0$ and define $I_0 = I$.

Step 1. Find class $M_t = m(I_t)$ and define $I_{t+1} = I_t - M_t$. Define class value $F_t = F_\pi(I_t) = \pi(i, I_t)$ for $i \in M_t$.

Step 2. If $I_{t+1} = \emptyset$, end. Otherwise, add 1 to t and go to Step 1.

This algorithm extends the traditional greedy procedure [7] to the situations in which:

(1) entities are selected according to a function of similarity between sets and their elements, the linkage function, rather than to the set function optimized;

(2) at each step, all the minimizing elements are selected rather than just one of them.

The layered cluster of $F_\pi(H)$ can be easily extracted from the serial partition M thus produced.

From the sequence $F = \{F_0, F_1, ..., F_n\}$, pick up the smallest index t^* among those maximising F_t, $t = 0, 1, ..., n$. Then apply the same selection rule to the starting part of the sequence F, $F^{t^*} = (F_0, ..., F_{t^*-1})$ obtained by removing F_{t^*} and all the consequent elements. Reiterating this pick-and-removal process until all elements of F are removed, we obtain set T^* of all the picked up indices.

The sets I_{t^*}, $t^* \in T^*$, form the layered cluster of F_π.

Assertion 3 *Subset H is a pattern if and only if $H = I_{t^*}$ for some $t^* \in T^*$.*

Proof: First, let us prove that I_{t^*} is a pattern for any $t^* \in T^*$. Indeed, for any $H \subset I$ containing elements outside of I_{t^*} there exists the minimum I_t such that $H \subseteq I_t$ and $t < t^*$. The minimality of I_t implies that $m(I_t) \cap H \neq \emptyset$. Thus, $F(H) \leq \pi(i, H) \leq \pi(i, I_t) = F(I_t) = F_t$ where $i \in m(I_t) \cap H$. But $F_t < F_{t^*}$ by the definition of T^*; therefore, $F(H) < F_{t^*}$, which proves this part of the statement.

Now, let $H \subseteq I$ be a pattern that doesn't coincide with I_{t^*} for any $t^* \in T^*$. Let I_{t*} be the smallest of the sets I_t with $t \in T^*$ containing H. There can be either $m(I_{t^*}) \cap H = \emptyset$ or not. In the latter case, for an $i \in m(I_{t^*}) \cap H$, $F(H) \leq \pi(i, H) \leq \pi(i, I_{t^*}) = F(I_{t^*})$, which contradicts the assumption that H is a pattern. In the former case, H must be part of an I_t with $t > t^*$. Let I_t be the smallest of these sets so that $m(I_t) \cap H \neq \emptyset$. Then, for an $i \in m(I_t) \cap H$, $F(H) \leq \pi(i, H) \leq \pi(i, I_t) = F(I_t) \leq F(I_{t^*})$ which contradicts the assumption that H is a pattern. □

Let us apply the serial partitioning algorithm to the example of summary linkage function in Figure 1. We can see that $m(I) = \{a, b\}$ because $\pi(a, I) = \pi(b, I) = 24$ is the minimum of $\pi(i, I)$ over all $i \in I$. With a, b removed from I, the minimum of $\pi(i, I - \{a, b\})$ is reached at l with $\pi(l, I - \{a, b\}) = 6$. The next entity to be removed is c, with $\pi(c, I - \{a, b, l\}) = 26$. In the remaining set $I_3 = I - \{a, b, l, c\}$, obvious leaders are d and e with minimum $\pi(d, I - \{a, b, l, c\}) = \pi(e, I - \{a, b, l, c\}) = 16$. This yields $I_4 = \{f, i, g, h, j, k\}$ with the minimum link, 27, reached at $m(I_4) = \{j, k\}$. At what remains, $I_5 = \{f, i, g, h\}$, $m(I_5) = \{f, g, h\}$ with the link equal to 30. This leaves $I_6 = \{i\}$ and no nonzero links within I_6 so that $F_\pi(I_6) = 0$. The results can be represented as a labelled sequence,

$$(ab)^{24}(l)^6(c)^{26}(de)^{16}(jk)^{27}(fgh)^{30}(i)^0,$$

where the parentheses contain sets $M_t = m(I_t)$ removed at each step of the algorithm, their order reflecting the order of removals, and the labels corresponding to the values of the linkage function $F_\pi(I_t)$ for $t = 0, 1, ...6$. The maxima are 30, 27, 26, and 24; the corresponding patterns, $H_3 = \{f, g, h, i\}$, $H_2 = H_3 \cup \{j, k\}$, $H_1 = H_2 \cup \{c, d, e\}$, and $H_0 = I$, form the layered cluster.

6 Universality of the monotone linkage functions

Let us consider a chain-nested set $P = \{H_0, H_1, ..., H_p\}$ where $H_0 = I$ and $H_t \subset H_{t-1}$ for all $t = 1, ..., p$. The question is if there exists a monotone linkage function $\pi(i, H)$ such that P is the layered cluster of F_π. The answer is yes.

Let us define $G_t = H_t - H_{t+1}$ for each $t = 0, 1, ..., p-1$, and $G_p = H_p$. Obviously, the set $G = \{G_0, G_1, ..., G_p\}$ forms a partition of I.

Let us define now a linkage function, $\chi(i, H)$, by the condition: $\chi(i, H) = t$ if $i \in G_t$ and $H_t \subseteq H$ (t=1,..., p); otherwise, $\chi(i, H) = 0$. This linkage function is monotone by its definition. Its tightness function, $F_\chi(H)$, is equal to 0 for all $H \subseteq I$ except for the cases when $H = H_t$ and $F_\chi(H_t) = t$ $(t = 1, ..., p)$. Set function F_χ can be referred to as the characteristic function of the chain nested set P. When $p = 1$, F_χ is the conventional characteristic function of H_1: $F_\chi(H) = 0$ for all H except

for the $H = H_1$; $F_\chi(H_1) = 1$. Since P is obviously the pattern set of F_χ, the following statement is proved.

Assertion 4 *The characteristic function F_χ of a chain-nested set P is a tightness function whose layered cluster is P.*

References

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

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

[3] Y. Kempner, B. Mirkin, and I. Muchnik, Monotone linkage clustering and quasi-concave set functions, *Applied Mathematics Letters*, 10, **4**, 19-24 (1997).

[4] E.N. Kuznetsov and I.B. Muchnik, Analysis of the distribution of functions in an organization, *Automation and Remote Control*, **43** (10), 1325-1331 (1982).

[5] B. Mirkin, and I. Muchnik, Combinatorial optimization and clustering, In: D.-Z. Du and P.Pardalos (Eds.) *Handbook of Combinatorial Optimization, 2*, Kluwer Academic Publishers, Boston, Ma., 261-329 (1998).

[6] A. Malishevski, Properties of ordinal set functions, A paper at the All-Union Conference on Optimization, Dushanbe, 1986. In: A. Malishewski, *Qualitative Models in the Theory of Complex Systems*, Nauka, Moscow, 169-173 (1998, in Russian).

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