Clustering on antimatroids and convex geometries

YULIA KEMPNER[1] , ILYA MUCHNIK[2]
[1]Department of Computer Science
Holon Academic Institute of Technology
52 Golomb Str., P.O. Box 305, Holon 58102
ISRAEL
[2]Department of Computer Science
Rutgers University, NJ, DIMACS
96 Frelinghuysen Road , Piscataway, NJ 08854-8018
US

Abstract: - The clustering problem as a problem of set function optimization with constraints is considered. The behavior of quasi-concave functions on antimatroids and on convex geometries is investigated. The duality of these two set function optimizations is proved. The greedy type Chain algorithm, which allows to find an optimal cluster, both as the "most distant" group on antimatroids and as a dense cluster on convex geometries, is described.

Key-Words: - Quasi-concave function, antimatroid, convex geometry, cluster, greedy algorithm

1 Introduction

In this paper we consider the problem of clustering on antimatroids and on convex geometries. There are two approaches to clustering. In the first one we try to find extreme dense clusters (either as a partition of the considered set of objects into homogeneous groups, or as a single tight set of objects). The second approach finds a cluster which is the "most distant" from its complementary subset in the whole set (as in the single linkage clustering method) without a control how close are objects within a cluster. The problem of density (in the first approach) or isolation (in the second approach) of a subset may be formalized in terms of a set function - the function that its values score the subsets according to their tightness or to distant relations with their complementary subsets.

Usually, in cluster analysis each subset may be referred as a cluster. However, in last years the problem to find clusters as optimization problem with constraints became actual, particularly in scientific database mining, where objects have complex structure. In those cases their similarity measurement requires to consider the problem from multi-criteria perspectives. For instance, if one considers transportation problems in a large city and wants to apply clustering technique, she or he has to take into account at least two criteria: a population flow density along transportation lines and a capacity of all types of transports along the lines. In this paper we also consider restricted situations for clustering. We found that such situations can be characterized as constructively defined families of subsets, such that only elements from these families can be considered as feasible clusters. For example, let us consider a boolean matrix of data in which variables (columns of matrix) are divided into several groups. Variables, belonging to one group, are ordered. In this case every group of variables can be interpreted as a complex variable which represent an ordinal scale for some property: value 1 of the first (according to the order) boolean variable means that an object represents the corresponding property on very general level, 1 for the second variable means that the object represents the property on more specific level, etc. Thus the value 1 of a boolean variable with rank "k" in its group means that all previous variables of the group also have this value 1. In other words, it is an ordinal data matrix. The problem is to find an extreme in this ordinal data which represents a group of points, that are smaller (according to the ordinal vector comparison) than the choosen extreme. It is obviously that this case can be generalized on a common ordinal data. If one interprets ordinal scales as criteria for a particular multicriteria evaluation then he can interpret the above cluster as the best choice solution.

Thus the cluster problem can be formulated as follows: for a given *set system* over E, i.e. for a pair (E,S) where $S \subseteq 2^E$ is a family of feasible subsets of finite E, and for a given function $F : S \to \mathbf{R}$, find

elements of S for which the value of the function F is maximal. In general, this optimization problem is NP-hard, but for some specific function and for some specific set systems polynomial algorithms are known. The well-known examples are modular cost functions that can be optimized over matroids by using polynomial greedy algorithm [2] and bottleneck function that can be maximized over greedoids [5]. The other example is set function defined as the minimum value of monotone linkage functions. Such set function can be maximized by a greedy type algorithm over a family of all subsets of E [4,7] and over antimatroids [3]. In this paper we refine our results obtained for antimatroids and extend them to convex geometries.

In Section 2 some basic information about convex geometries and antimatroids are listed. Section 3 gives a brief introduction to the optimization of quasi-concave functions and investigates the optimization problem on antimatroids. In Section 4 the optimization problem is considered on convex geometries and duality of two problems are established.

2 Preliminaries

Let E be a finite set. A *set system* over E is a pair (E,S) where $S \subseteq 2^E$ is a family of feasible subsets of E, called *feasible sets*. We will use $X \cup x$ for $X \cup \{x\}$ and $X - x$ for $X - \{x\}$.

Definition 2.1 *A closure operator, $\tau : 2^E \to 2^E$, is a map satisfying the closure axioms:*

C1: $X \subseteq \tau(X)$
C2: $X \subseteq Y \Rightarrow \tau(X) \subseteq \tau(Y)$
C3: $\tau(\tau(X)) = \tau(X)$

A set X is said to be *closed* if $\tau(X) = X$. The pair (E,τ) is called a *closure space*.

To obtain an equivalent definition consider a set system (E,S) such that

(i) $E \in S$
(ii) $X,Y \in S \Rightarrow X \cap Y \in S$

Then the operator

(1) $\tau(A) = \cap \{X : A \subseteq X \text{ and } X \in S\}$

is a closure operator.

A set system (E,S) satisfying (*i*) and (*ii*) with closure operator τ defined in (1) is an equivalent axiomatization of closure space [5].

Definition 2.2 *A closure space is called a convex geometry if the following anti-exchange property is satisfied:*

if $y,z \notin \tau(X)$ *then* $z \in \tau(X \cup y)$ *implies* $y \notin \tau(X \cup z)$

We call an element $x \in A$ an *extreme point* of A if $x \notin \tau(A - x)$. For a *convex set* X, i.e. for a set from a convex geometry (E,S), this is equivalent to $X - x \in S$. The set of extreme points of A is denoted $ex(A)$.

Consider another set system.

Definition 2.3 *A nonempty set system (E,H) is an antimatroid if*

(A1) for each nonempty $X \in H$ there is an $x \in X$ such that $X - x \in H$

(A2) for all $X,Y \in H$, and $X \not\subset Y$, there exists an $x \in X - Y$ such that $Y \cup x \in H$

Any set system satisfying (*A1*) is called *accessible*.

Another antimatroid definition is based on the following property:

Definition 2.4 *A set system (E,H) has the interval property without upper bounds if for all $X,Y \in H$ with $X \subseteq Y$ and for all $x \in E - Y$, $X \cup x \in H$ implies $Y \cup x \in H$.*

Theorem 2.1 [1,5] *For an accessible set system (E,H) the following statements are equivalent:*

(i) (E,H) *is an antimatroid*
(ii) H *is closed under union*
(iii) (E,H) *satisfies the interval property without upper bounds*

A maximal feasible subset of set $X \subseteq E$ is called a *basis* of X, and is denoted by B_X. Clearly, (by (*ii*)), there is only one basis for each set.

Now consider the Theorem that establishes duality between two structures.

Theorem 2.2 [5] *A set system (E,H) is an antimatroid if and only if (E,S) is a convex geometry, where $S = \{E - X : X \in H\}$.*

As an immediate consequence we have

(2) $\tau(E - X) = E - B_X$

For a set $X \in H$, let

$$\Gamma(X) = \{x \in E - X : X \cup x \in H\}$$

be the set of *feasible continuations* of X, then

(3) $\Gamma(X) = ex(E - X)$

3 Quasi-concave functions and antimatroids

In this section we consider set functions defined as the minimum value of monotone linkage functions. Such set functions can be maximized by a greedy type algorithm over a family of all subsets of E [10] and over antimatroids [3]. Here we detail the behavior of these functions defined on antimatroids in order to use them for clustering.

We consider the problem to find a cluster which is "most distant" from its complementary subset in the whole considered set. To define a distance between the subset and rest objects, the concept of an element-to-set linkage function can be used. A linkage function $\pi(x,X)$ measures a distance between subset X and element x. The well-known example is the single linkage function $\pi(x,X) = \min_{y \in X} d_{xy}$ which is defined in terms of pair-wise distances d_{xy}. As another example consider a data table $A = (x_{ik})$ where x_{ik} is a value of variable $k \in K$ for entity $i \in E$. A linkage function can be defined as

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

An important feature of these examples is the monotonicity of linkage functions. The monotone linkage functions were introduced by Mullat [9].

Definition 3.1 *A function* $\pi : E \times 2^E \to \mathbf{R}$ *is a monotone linkage function if for all* $X,Y \subseteq E$ *and* $x \in E$

(4) $X \subseteq Y$ *implies* $\pi(x,X) \geq \pi(x,Y)$

Consider $F : 2^E \to \mathbf{R}$ defined for each $X \subset E$

$$F(X) = \min_{x \in E - X} \pi(x,X)$$

This function may be considered as the function that measures "isolation" of the subset X, i.e. it measures how this subset is distant from rest objects. Hence a subset maximizing this set function can be referred as a cluster.

It was shown [6], that the function F satisfies the next condition:

for each $X,Y \subset E$, $F(X \cap Y) \geq \min \{F(X), F(Y)\}$

Such functions are called *quasi-concave* set functions. These functions were studied in [10,4]. In particular, the simple polynomial algorithm which finds the minimal set $X \subset E$ such that

$$F(X) = \max \{F(Y) : Y \subset E\}$$

was developed and used for incomplete clustering constructor [10].

In this section we extend our results to antimatroids. For this purpose we define a new function specified for antimatroids:

$$F_H(X) = \min_{x \in \Gamma(X)} \pi(x,X)$$

and if $\Gamma(X) = \varnothing$, then $F_H(X) = -\infty$.

It should be mentioned, that an antimatroid (E,H) has one and only one maximal feasible set, $E_H = \cup_{X \in H} X$. Thus, for each feasible set $X \in H - E_H$ the continuation $\Gamma(X)$ is not-empty, i.e., F_H is well defined on $H - E_H$ for any antimatroid (E,H).

Additional notation is that the feasible sets H of an antimatroid (E,H), ordered by inclusion, form a lattice L_H, with lattice operations:

$$X \vee Y = X \cup Y, X \wedge Y = B_{X \cap Y}$$

Theorem 3.1. *If a set system* (E,H) *is an antimatroid, then the function* F_H *is a quasi-concave function on* L_H, *i.e.,*

for each $X,Y \in H$, $F_H(X \wedge Y) \geq \min \{F_H(X), F_H(Y)\}$

Proof. The case $X \wedge Y = E_H$ is trivial, then assume that $X \wedge Y \in H - E_H$, i.e.,

$$F_H(X \wedge Y) = \min_{x \in \Gamma(X \wedge Y)} \pi(x, X \wedge Y).$$

Let $F_H(X \wedge Y) = \pi(x^0, X \wedge Y)$, where $x^0 \in \Gamma(X \wedge Y)$. Hence, $x^0 \in E - (X \cap Y)$, since $\Gamma(B_X) \subseteq E - X$. Indeed, suppose that $x \in \Gamma(B_X)$ and $x \in X$, then $B_X \cup x \subseteq X$ and $B_X \cup x \in H$. Thus $B_X \cup x$ is also a basis, a contradiction.

Assume, without loss of generality, that $x^0 \in E - X$. Hence, we have $x^0 \in E - X$, and $X \wedge Y \subseteq X$, and $x^0 \in \Gamma(X \wedge Y)$, so from the interval property without upper bounds follows that $x^0 \in \Gamma(X)$. Then

$$F_H(X \wedge Y) = \pi(x^0, X \wedge Y) \geq \pi(x^0, X) \geq$$

$$\min_{x \in \Gamma(X)} \pi(x,X) = F_H(X) \geq \min \{F_H(X), F_H(Y)\}$$

■

Consider the following optimization problem - given a monotone linkage function π and a set system (E,H), find the feasible set $X \in H$ such that $F_H(X) = \max \{F_H(Y) : Y \in H\}$. From Theorem 3.1 it follows that the set of the optimization problem solutions is a meet-semilattice with unique minimal element. To find this minimal element we use the following algorithm [3]:

The Chain Algorithm (*E, H,* π)

1. $X := \varnothing$
2. $X^0 := \varnothing$
3. *While* $\Gamma(X) \neq \varnothing$ *do*
   - 3.1 *if* $F_H(X) > F_H(X^0)$, $X^0 := X$
   - 3.2 *choose* $x \in \Gamma(X)$ *such that* $\pi(x,X) \leq \pi(y,X)$ *for all* $y \in \Gamma(X)$
   - 3.3 $X := X \cup x$
4. *Return* X^0

Thus, the algorithm generates a chain of sets $\varnothing = X_0 \subset X_1 \subset \ldots \subset X_k$, where $X_i = X_{i-1} \cup x_i$ and $x_i \in \Gamma(X_{i-1})$, and returns the minimal set X^0 of the chain on which the value $F_H(X^0)$ is maximal.

Theorem 3.2 *Let a set system* (E,H) *be an antimatroid, then for all monotone linkage function* π *the Chain Algorithm finds a minimal feasible set that maximizes the function* F_H.

Proof. Let X^0 be a set obtained by the Chain Algorithm. To prove that X^0 is a minimal feasible set that maximizes F_H, we have to prove two statements:

(*i*) $F_H(X) < F_H(X^0)$ for each $X \in H$ and $X^0 \not\subset X$

Let $\varnothing = X_0 \subset X_1 \subset \ldots \subset X_k$ be the chain generated by the Chain Algorithm. Let j be the least integer for which $X_j \not\subset X$. Then $X_{j-1} \subseteq X$, $x_j \notin X$ and $X_{j-1} \cup x_j \in H$, that implies (from the interval property without upper bounds) $x_j \in \Gamma(X)$. Hence

(5) $F_H(X) \le \pi(x_j, X) \le \pi(x_j, X_{j-1}) = F_H(X_{j-1})$

Since $X^0 \not\subset X$ we have $X_{j-1} \subset X^0$, so $F_H(X_{j-1}) < F_H(X^0)$

(*ii*) $F_H(X) \le F_H(X^0)$ for each $X \in H$ and $X^0 \subseteq X$

If $X = E_H$ or $X = X_k$ the statement is obviously, otherwise, by analogy with previous case, we obtain (5), but now $F_H(X_{j-1}) \le F_H(X^0)$, because it is possible that $X^0 \subseteq X_{j-1}$.

■

The Chain Algorithm is a greedy type algorithm since it based on the best choice principle: it chooses on each step the extreme element (in sense of linkage function) and thus approaches the optimal solution. Let P is the maximum complexity of π computation, then the Chain Algorithm finds the minimal set that maximizes the function F_H in $O(P|E|^2)$ time.

Notice, that a quasi-concave function determines on an antimatroid (E,H) the special structure. It has been already noted that the family of feasible sets maximizing the function F_H is a meet-semilattice with unique minimal element. Denote this family by T^0 and let a^0 will be the value of the function F_H on the sets from T^0. The family of sets, that maximize the function F_H over $H - T^0$, we will denote by T^1 and by a^1 will be denoted the value of the function F_H on these sets. Continuing this process we have $H = \bigcup_{i=0}^{t} T^i$. It is easy to see that $L_j = \bigcup_{i=0}^{j} T^i$ is a subsemilattice of H, where $L_j = \{X \in H : F_H(X) \ge a^j\}$. We will call these subsemilattices by *level semilattices*.

Denote by K^j the minimal element (null) of level semilattice L_j. Since $L_j \subseteq L_{j+1}$, we obtain $K^0 \supseteq \ldots \supseteq K^t$.

Theorem 3.3. *Let* $\varnothing = A_0 \subseteq \ldots \subseteq A_p$ *be a chain of sets that were stored as local optimal subsets* X^0 *in the Chain Algorithm running on an antimatroid* (E,H) *, then this chain coincides with the chain of level semilattice nulls* $K^0 \supseteq \ldots \supseteq K^t$.

Proof. From the algorithm construction follows that if $\varnothing = X_0 \subset X_1 \subset \ldots \subset X_k$ is a sequence of sets generated by the Chain Algorithm, then for each X_i such that $A_{l-1} \subseteq X_i \subset A_l$, we have $F_H(X_i) \le F_H(A_{l-1}) < F_H(A_l)$. Hence,

(6) $X_i \subset A_l \Rightarrow F_H(X_i) \le F_H(A_{l-1})$

To prove the Theorem, we first prove the following claim:

Claim. *For each* $l = 0,1,\ldots,p$ *and for each* $X \in H$, *if* $A_l \not\subset X$, *then* $F_H(X) \le F_H(A_{l-1}) < F_H(A_l)$

By analogy with the proof of Theorem 3.2 (case (i)) we obtain that there exist the set X_i belonging to the chain generated by the Chain Algorithm, such that $X_i \subset A_l$ and $F_H(X) \le F_H(X_i)$, then the Claim follows from (6).

Now prove that for each $l = 0,1,\ldots,p$, A_l is a null of some level semilattice L_j, i.e., $A_l = K^j$. Let $a^j = F_H(A_l)$. From the Claim immediately follows that A_l is a null of L_j.

It remains to show that for each null K^j exists $A_l = K^j$. Since $F_H(A_p)$ is a maximal value of the function F_H, for each null K^j exists $1 \le l \le p$, such that $F_H(A_{l-1}) < F_H(K^j) \le F_H(A_l)$ or $F_H(K^j) \le F_H(A_l)$ where $A_l = \varnothing$. Show that $K^j \supseteq A_l$. In the second case ($A_l = \varnothing$) it is obviously. Suppose, that in the first case $A_l \not\subset K^j$, then from the Claim we have $F_H(K^j) \le F_H(A_{l-1})$, a contradiction. On the other hand, $A_l \in L_j$, i.e., $K^j \subseteq A_l$ because K^j is a null of L_j. Then $A_l = K^j$.

■

The obtained chain-nested structure, referred in [7,8] as a layered cluster, can be considered an abstract implementation of the idea of multiresolutional view on a system of interrelated objects.

4 Quasi-concave functions and convex geometry

In this section we consider convex geometries. As set functions, defined on convex geometries, we investigate functions which are dual to quasi-concave functions introduced in the previous sections.

Let (E,S) be a convex geometry and let $F_S(X) = F_H(E - X)$ for each $X \in S$. From Theorem 2.2 follows that the set system $H = \{E - X : X \in S\}$ is an antimatroid, then the definition is correct. From the definition follows that $F_S(X) = \min_{x \in \Gamma(E-X)} \pi(x, E - X)$. Then, by using (3) we obtain

$$F_S(X) = \min_{x \in ex(X)} \pi(x, E - X) = \min_{x \in ex(X)} \pi^*(x, X)$$

It is easy to see that the function π^* is a *monotone increasing linkage* function where

if $X, Y \in S$ *and* $X \subseteq Y$ *then* $\pi^*(x,X) \le \pi^*(x,Y)$.

It should be mention, that we not assume that the original given function is just a monotone linkage function π. If a given function is a monotone increasing linkage function π^*, then we straightforwardly obtain the set function F_S defined on the convex geometry. Consider, for example, $\pi^*(x, X) = \max_{y \in X} s_{xy}$, defined in terms of pair-wise similarity s_{xy}. Obviously, thus defined π^* is a monotone increasing function. Then the function F_S may be referred as the tightness function [8], and a subset maximizing this function can be considered as a dense cluster.

Notice that the feasible sets S of a convex geometry ordered by inclusion form a lattice L_S, with lattice operations:

$$X \vee Y = \tau(X \cup Y), X \wedge Y = X \cap Y$$

then we obtain:

Lemma 4.1 *If a set system* (E,S) *is a convex geometry, then for two lattices* L_S *and* L_H *we have*

$$F_S(X \vee Y) = F_H(\overline{X} \wedge \overline{Y})$$

The proof is based on the Theorem 2.2 and on (2).

Theorem 4.1 *If a set system* (E,S) *is a convex geometry, then the function* F_S *is a quasi-concave function on* L_S*, i.e.,*
for each $X, Y \in H$, $F_S(X \vee Y) \ge \min\{F_S(X), F_S(Y)\}$

Proof. From the Lemma 4.1 and from the Theorem 3.1 we have

$$F_S(X \vee Y) = F_H(\overline{X} \wedge \overline{Y}) \ge \min\{F_H(\overline{X}), F_H(\overline{Y})\} = \\ = \min\{F_S(X), F_S(Y)\}$$

■

Consider the following optimization problem - given a monotone linkage function π^*, and a convex geometry (E,S), find the feasible set $X \in S$ such that $F_S(X) = \max\{F_S(Y) : Y \in S\}$. From Theorem 4.1 it follows that the set of the optimization problem solutions is a join-semilattice with unique maximal element. To find this maximal element we use the algorithm that is a mirror version of the Chain Algorithm:

The Mirror Chain Algorithm (*E, S,* π^*)

1. $X := E$
2. $X^0 := E$
3. *While* $ex(X) \neq \varnothing$ *do*
   - 3.1 *if* $F_S(X) > F_S(X^0)$, $X^0 := X$
   - 3.2 *choose* $x \in ex(X)$ *such that* $\pi^*(x, X) \le \pi^*(y, X)$ *for all* $y \in ex(X)$
   - 3.3 $X := X - x$
4. *Return* X^0

Thus, the algorithm generates a chain of sets $E = X_0 \supset X_1 \supset \ldots \supset X_k$, where $X_i = X_{i-1} - x_i$ and $x_i \in ex(X_{i-1})$, and returns the maximal set X^0 of the chain on which the value $F_S(X^0)$ is maximal.

Theorem 3.2 *Let a set system* (E,S) *be a convex geometry, then for all monotone linkage function* π^* *the Mirror Chain Algorithm finds a maximal feasible set that maximizes the function* F_S.

The proof of this Theorem immediately follows from Theorem 3.2, since each step $X \to X - x$ of the Mirror Chain algorithm correspondences to the step $\overline{X} \to \overline{X} \cup x$ of the original Chain Algorithm, and the result of the Mirror Chain Algorithm is a complement of the minimal set obtained by the Chain Algorithm.

By analogy with a quasi-concave function F_H a quasi-concave function F_S determines on a convex geometry (E,S) the chain-nested structure of join-semilattices $L_0 \subseteq \ldots \subseteq L_t$, where $L_j = \{X \in S : F_S(X) \ge a^j\}$ and $a^0 > \ldots > a^t$ are the all values of function F_S on a convex geometry (E,S). We also call these subsemilattices by *level semilattices*. Denote by I^j the maximal element (one) of level semilattice L_j. Since $L_i \subseteq L_{i+1}$, we obtain $I^0 \subseteq I^1 \subseteq \ldots \subseteq I^t$.

Theorem 4.3 *Let* $E = A_0 \supset A_1 \supset \ldots \supset A_p$ *be a chain of sets that were stored as a local maximal subsets* X^0 *in the Mirror Chain Algorithm running on a convex geometry* (E,S), *then this chain coincides with the chain of level semilattice ones* $I^0 \subseteq I^1 \subseteq \ldots \subseteq I^t$, *such that the maximal optimal subset* $A_p=I^0$ *and* $E=A_0=I^t$.

The proof of the Theorem 4.3 is based on the duality of antimatroids and convex geometries, and on Theorem 3.3

5 Conclusion

In this paper we have investigated the problem of clustering with constraints. The clustering problem was considered as optimization problem on antimatroids and on convex geometries with quasi-concave goal function. The greedy type Chain algorithm, which allows to find an optimal cluster, also as the "most distant" group on antimatroids and as dense cluster on convex geometries, was described.

References:

[1] A. Björner and G.M. Ziegler, Introduction to greedoids , in "*Matroid applications*", ed. N. White, Cambridge Univ.Press, Cambridge, UK,1992

[2] J.Edmonds, Matroid and the greedy algorithm, *Mathematical Programming 1*, (1971),127-136

[3] Y. Kempner, and V.Levit, Algorithmic characterizations of antimatroids, *Third Haifa Workshop on Interdisciplinary Applications of Graph Theory, Combinatorics and Algorithms,* Haifa, Israel,2003

[4] Y.Kempner, B.Mirkin, and I.Muchnik, Monotone linkage clustering and quasi-concave functions, *Appl.Math.Lett.* 10 ,No.4 (1997) 19-24

[5] B.Korte, L.Lovász, and R.Schrader, "*Greedoids*", Springer-Verlag, New York/Berlin, 1991

[6] A.Malishevski, Properties of ordinal set functions, in A.Malishevski ,*"Qualitative Models in the Theory of Complex Systems"*, Nauka, Moscow,1998 (in Russian)

[7] B.Mirkin, and I.Muchnik, Layered clusters of tightness set functions, *Appl.Math.Lett*,15 (2002),147-151

[8] B. Mirkin, and I.Muchnik, Induced layered clusters, hereditary mappings and convex geometries, *Appl.Math.Lett*,15 (2002),293-298

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

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