Aplikace matematiky

Mirko Křivánek

A note on the computational complexity of hierarchical overlapping clustering

Aplikace matematiky, Vol. 30 (1985), No. 6, 453–460

Persistent URL: http://dml.cz/dmlcz/104174

A NOTE ON THE COMPUTATIONAL COMPLEXITY OF HIERARCHICAL OVERLAPPING CLUSTERING

MIRKO KŘIVÁNEK

(Received October 15, 1984)

Summary. In this paper the computational complexity of the problem of the approximation of a given dissimilarity measure on a finite set X by a k-ultrametric on X and by a Robinson dissimilarity measure on X is investigated. It is shown that the underlying decision problems are NP-complete.

I. INTRODUCTION

In the past a large variety of clustering definitions and methods have been developed and used. To introduce the topic of hierarchical overlapping clustering let $X = \{x_1, \ldots, x_n\}$ denote n objects which are to be clustered and d a *dissimilarity measure* on X, i.e. d: $X \times X \to R_0^+$ (nonnegative rational numbers), $d(x, y) = 0$ iff $x = y$ and $d(x, y) = d'(y, x)$ for $x, y \in X$.

A *clustering* is any partition of X into k non-empty sets, i.e. clusters. Informally speaking the problem of *hierarchical clustering* is to find a sequence of nested clustering (with respect to the partition refinement) which must induce an ultrametric on X. The optimization problem of hierarchical clustering is formulated as the approximation of a given dissimilarity measure on X by an ultrametric on X. Recently this problem has been shown to be NP-hard [6].

Some authors [1, 2, 4] proposed a more general problem of hierarchical clustering in which the aim is to construct a certain sequence of coverings of X which starts with the partition of X into singletons and ends with the trivial partition $\{\{X\}\}$. As the clusters may overlap this latter problem is often referred to as the problem of *hierarchical overlapping clustering.*

In this note we study the NP-completeness of the computational problems of hierarchical overlapping clustering. We show that two underlying decision problems are NP-complete. The first is the problem of the approximation of a given dissimilarity measure on X by a k-ultrametric on X [3, 4] and the second is the problem of the approximation of a given dissimilarity measure by a Robinson dissimilarity measure on X [1].

Finally we state one open problem using graph-theoretical concepts.

Our NP-completeness terminology using graphs is that of [2].

II. BACKGROUND

Throughout this paper let $X = \{x_1, \ldots, x_n\}$ be a set of objects and d a dissimilarity measure on X.

The dissimilarity measure d on X is said to be a *k-ultrametric* on X if

$$(1) \qquad \forall S \subset X\,, \quad |S| = k\,, \quad \forall x, y \in X$$
$$d(x, y) \leqq \max \{d(v, w) \mid v \in S \cup \{x, y\},\ w \in S\}\,.$$

The 1-ultrametric on X is simply called an *ultrametric* on X.

The dissimilarity measure d on X is said to be *Robinson* if there is a permuation θ of the set $\{1, \ldots, n\}$ such that

(i) $d(x_{\theta(i)}, x_{\theta(i)+1}) \leqq d(x_{\theta(i)}, x_{\theta(i)+2}) \leqq \ldots \leqq d(x_{\theta(i)}, x_n)$
(ii) $d(x_{\theta(i)}, x_{\theta(i)+1}) \leqq d(x_{\theta(i)-1}, x_{\theta(i)+1}) \leqq \ldots \leqq d(x_1, x_{\theta(i)+1})$
for all $i = 1, \ldots, n$.

Every set-function pair (P, f) satisfying the following conditions (i)–(vi) is called a pyramid on X:

(i) $P \in \mathscr{P}(\mathscr{P}(X))$,
(ii) $X \in P$,
(iii) $\emptyset \notin P$,
(iv) $(\forall x \in X)\ \{x\} \in P$,
(v) $f: P \to Z_0^+$ (nonnegative integers) and
$(\forall h, h' \in P)\ f(h) = 0 \Leftrightarrow |h| = 1$,
$f(h) < f(h') \Leftrightarrow h \subset h'$ and $h \neq h'$,
(vi) the function $r_P: X \times X \to Z_0^+$ defined by
$r_P(x, y) = \min \{f(h) \mid \{x, y\} \subset h\}$
is a Robinson dissimilarity measure on X.

Remark. If r is an ultrametric on X then the pyramid (P, f) on X is called the *hierarchy* on X.

It can be easily observed [1] that the set of all hierarchies on X is strictly included in the set of all pyramids on X.

The *height* ϱ of a pyramid (P, f) on X is defined as follows:

$$\varrho(P) = |\text{Range} f| - 1\,.$$

Obviously $1 \leqq \varrho(P) \leqq n - 1$ for every pyramid (P, f) on X.

Now we introduce the decision problems of hierarchical overlapping clustering whose NP-completeness we shall be interested in.

Problem μ. *Instance*: Dissimilarity measure d on X, positive integer k;
Question: Is d a k-ultrametric on X?

Problem π. *Instance*: Dissimilarity measure d on X, positive integer k;
Question: Is there a pyramid (P, f) on X such that
$\sum_{x,y \in X} |d(x, y) - r_P(x, y)| \leqq k$?

III. RESULTS

Theorem 1. *The problem* μ *is* NP-*complete.*

Proof. As is customary with such proofs, we omit the trivial verification that μ belongs to NP.

Let d be a dissimilarity measure on X such that $d(x, y) \in \{1, 2\}$ $(x \neq y \in X)$. Let us define the graph $G = (X, E)$, where

$$\{x, y\} \in E \Leftrightarrow d(x, y) = 1 .$$

There is a very simple condition which is equivalent to the k-ultrametric inequality (1). The condition is that

(2) d is k-ultrametric on X iff G contains no subgraph isomorphic to the graph $K_{k+2} - e$ (i.e. complete graph on $(k + 2)$ vertices without precisely one edge).

In what follows we give a polynomial transformation from the problem **3-satisfiability** [2], page 259, to μ. The problem **3-satisfiability** is defined as follows:

Instance: Set U of variables, collection C of clauses over U such that each clause has $|c| = 3$;

Question: Is there a satisfying truth assignment for C?

So let U, $C = \{c_1, \ldots, c_k\}$ be an arbitrary instance of **3-satisfiability**. Let $G = (V, E)$ be the graph such that

$$V = \{\langle \sigma, i \rangle \mid \sigma \in c_i\} , \quad E = \{\{\langle \sigma, i \rangle, \langle \delta, j \rangle\} \mid i \neq j \text{ and } \sigma \neq \bar{\delta}\} .$$

R. Karp has shown [5] that this graph G contains a complete graph on k vertices as its subgraph iff **3-satisfiability** has "yes"-solution.

Further let us consider the graph $G^` = (V^`, E^`)$ where

$$V^` = V \cup \{v', v''\} , \quad v' \neq v'' \notin V \text{ are "new" vertices joined to } V ;$$

$$E^` = E \cup \{v, v'\} \cup \{v, v''\} \quad (v \in V) .$$

Clearly the construction of the graph G' can be carried out in polynomial time. Now we shall prove that

(3) G contains a subgraph K_k iff G' contains a subgraph $K_{k+2} - e$.

Let the set $\{v_1, \ldots, v_k\}$ induce in G the complete graph K_k. Then the set $\{v_1, \ldots, v_k, v', v''\}$ induces in G' the graph isomorphic to $K_{k+2} - e$.

Conversely, let $\{v_1, \ldots, v_{k+2}\}$ be the subset of V' which induces a subgraph $K_{k+2} - e$ in G'. As the graph $K_{k+2} - e$ contains two subgraphs K_{k+1} we have $|V \cap \{v_1, \ldots, v_{k+2}\}| = k$ and the set $\{v_1, \ldots, v_{k+2}\} - \{v', v''\}$ induces in G the complete graph K_k. Let us set

$$X = X',$$

$$d(x, y) = 0 \quad \text{if} \quad x = y,$$

$$= 1 \quad \text{if} \quad \{x, y\} \in E',$$

$$= 2, \quad \text{otherwise}.$$

Using (2) and (3) we obtain that the dissimilarity measure d on X is not a k-ultrametric if and only if **3-satisfiability** has "yes"-solution. This concludes the proof. □

Now we turn our attention to the problem π. First we prove one auxiliary lemma.

Lemma 1. *Let d be a dissimilarity measure on X such that* Range $d = \{0, 1, 2\}$ *and let (P, f) be the optimal solution of π with respect to this instance. Then* Range $f = \{0, 1, 2\}$.

Proof. Let d be a dissimilarity measure on X such that Range $d = \{0, 1, 2\}$ and let (P, f) be the optimal solution of π with respect to this instance. Let us suppose that Range $f \neq$ Range d. Let $x, y \in X$ be two objects such that $d(x, y) = 1$. Let us consider the pyramid (P', f') defined as follows:

1) If Range $f \cap \{1\} \neq \emptyset$ then

$$P' =^{\text{df}} \bigcup_{i=1}^{n} \{x_i\} \cup X \cup \{h \mid h \in P \text{ and } f(h) = 1\},$$

$f'(h) =^{\text{df}} 1$ for all $h \in P$ with the property $f(h) = 1$,

$f'(\{x_i\}) =^{\text{df}} 0$ $(i = 1, \ldots, n)$ and $f'(X) =^{\text{df}} 2$.

2) If Range $f \cap \{1\} = \emptyset$ then

$$P' =^{\text{df}} \bigcup_{i=1}^{n} \{x_i\} \cup X \cup \{x, y\}, \; f'(\{x, y\}) =^{\text{df}} 1,$$

$f'(\{x_i\}) =^{\text{df}} 0$ $(i = 1, \ldots, n)$ and $f'(X) =^{\text{df}} 2$.

Now one can easily observe that

$$\sum_{x,y \in X} |d(x, y) - r_{P'}(x, y)| < \sum_{x,y \in X} |d(x, y) - r_{P}(x, y)|. \quad □$$

Theorem 2. *The problem π is* NP-*complete.*

Proof. The problem π is obviously in NP. To prove the NP-hardness of π we use the problem **Hamiltonian path** (cf. [2], page 199), defined as follows.

Instance: Planar cubic graph $G = (V, E)$ which has no face with less than 5 edges.

Question: Does G contain a Hamiltonian path?

Let $G = (V, E)$, $|V| = n$, be an arbitrary instance of **Hamiltonian path**. The instance of π will be constructed as follows:

$$X = V(G),$$
$$d(x, y) = 0 \quad \text{if} \quad x = y,$$
$$1 \quad \text{if} \quad \{x, y\} \in E(G),$$
$$2, \quad \text{otherwise}.$$

Let (P, f) be the solution of π with respect to this instance. It follows from Lemma 1 that $\text{Range} f = \{0, 1, 2\}$. We complete the proof by proving the following equivalence:

The graph G contains a Hamiltonian path iff

$$\sum_{x,y \in X} |d(x, y) - r_P(x, y)| \leqq \frac{n}{2} + 1 .$$

Let G contain a Hamiltonian path $H = \{\{x_1, x_2\}, \{x_2, x_3\}, \ldots, \{x_{n-1}, x_n\}\}$. Then for the pyramid (P, f) on X where

$$P = \bigcup_{i=1}^{n} \{x_i\} \cup H \cup X, \quad f(\{x_i\}) = 0 \quad i = 1, \ldots, n,$$
$$f(h) = 1 \quad \text{for} \quad h \in H \quad \text{and} \quad f(X) = 2,$$

we get

$$\sum_{x,y \in X} |d(x, y) - r_P(x, y)| = \tfrac{1}{2}n + 1 .$$

Conversely, let us suppose that there exists a pyramid (P, f) on X such that $\text{Range} f = \{0, 1, 2\}$ and that $\sum_{x,y \in X} |d(x, y) - r_P(x, y)| \leqq \frac{1}{2}n + 1$. Further, let G contain no Hamiltonian path. We examine two cases:

a) $|P| = 2n$.

Then the pyramid (P, f) on X has exactly $(n - 1)$ subsets $h_1, \ldots, h_{n-1}$ such that $|h_i| = 2$. As $\sum_{x,y \in X} |d(x, y) - r_P(x, y)| \leqq \frac{1}{2}n + 1$ the set $H = \bigcup_{i=1}^{n} h_i$ is a Hamiltonian path in G, a contradiction.

b) $|P| < 2n$.

Then there exist m, $1 \leqq m \leqq n - 2$, elements $h_1, \ldots, h_m$ of P such $2 \leqq |h_i| \leqq n - 1$, $i = 1, \ldots, m$. We transform the case b) to the case a) in such a way that using the pyramid (P, f) on X we construct a sequence of pyramids (P_i, f_i) on X. Each member, say (P_{i+1}, f_{i+1}), is constructed from the precedent member (P_i, f_i) by the following recursive rule: "Replace a set $h \in P_i$, $h = \{x_{i_1}, x_{i_2}, \ldots, x_{i_l}\}$, $l \geqq 3$, by $(l - 1)$ sets $\{x_{i_1}, x_{i_2}\}, \{x_{i_2}, x_{i_3}\}, \ldots, \{x_{i_{l-1}}, x_{i_l}\}$ and put $f_{i+1}(\{x, y\}) = 1$ if $x, y \in h$, $f_{i+1}(h) = f_i(h)$, $h \in P_i$, otherwise."

Further we shall use the equality

(4) $$(\forall h \in P_i - P_{i+1},\ |h| \geqq 3)$$
$$\sum_{x,y \in X} |d(x, y) - r_{P_i}(x, y)| - \sum_{x,y \in X} |d(x, y) - r_{P_{i+1}}(x, y)| = \varphi(h) - \psi(h),$$

where

$\varphi(h) = |\{\{x, y\} \mid x, y \in h \text{ and } d(x, y) = 2\}|$,
$\psi(h) =$ the minimum number of edges in a graph G', on the set h, such that $(G' - G(h)) \cup (G(h) - G')$ is a Hamiltonian path, where $G(h)$ is the subgraph of G induced by the set of vertices h.

We claim that

(5) $$\varphi(h) > \psi(h), \quad |h| \geqq 3.$$

For $h \in V(G)$, $3 \leqq |h| \leqq 5$, this inequality can be checked e.g. by exhaustive search, utilizing the fact that an induced subgraph of G contains neither a circuit C_3 nor C_4. For greater cardinalities of h this follows directly from the selfevident inequality

$$\binom{i}{2} > 2i + 1 \quad \text{for} \quad i \geqq 6,$$

since each subgraph of G has the maximum degree 3. Thus in virtue of (4) and (5) we have

$$\sum_{x,y \in X} |d(x, y) - r_{P_i}(x, y)| > \sum_{x,y \in X} |d(x, y) - r_{P_{i+1}}(x, y)|.$$

So starting from the pyramid $(P, f) = (P_1, f_1)$ and taking into account the pyramid (P_*, f_*) (from the constructed sequence of pyramids) such that $|P_*| = 2n$ we obtain

$$\sum_{x,y \in X} |d(x, y) - r_P(x, y)| > \sum_{x,y \in X} |d(x, y) - r_{P_*}(x, y)|.$$

The proof is complete. □

In the rest of this section we shall deal with special variants of the problem π. Let us denote by π_i the decision computational problem defined in precisely the same way as the problem π with the exception that the aim is to find a pyramid on X with the height i.

Lemma 2. *We have*

$$\pi_i \propto \pi_{i+1}\,, \quad i = 1, 2, \ldots .$$

Proof. Let (d, k) be an instance of the problem π_i. The corresponding instance (d', k') of the problem π_{i+1} will be constructed as follows:

1) d' is a dissimilarity measure on $X' = X \cup \{z\}$, where $z \notin X$ is a "new" object joined to X and
   $d'(x, y) = d(x, y)$ if $x, y \in X$,
   $d'(x, z) = n^2 \max \{d(x, y) \mid x, y \in X\}$,
   $d'(z, z) = 0$,
2) $k' = k$.

To conclude the proof it is sufficient to verify the obvious equivalence

$$\sum_{x,y \in X} |d(x, y) - r_P(x, y)| \leqq k \Leftrightarrow \sum_{x,y \in X} |d(x, y) - r_{P'}(x, y)| \leqq k'\,,$$

where

$$P' = P \cup \{z\} \cup X'\,, \quad (\forall h \in P)\; f'(h) = f(h) \quad \text{and}$$

$$f'(\{z\}) = 0\,, \quad f'(X') = n^2 \max \{d(x, y) \mid x, y \in X\}\,. \quad \square$$

Using the transitivity of $\propto$, Lemma 1, Lemma 2 and Theorem 2 we obtain the following assertition:

Theorem 3. *The problems* π_i, $i \geqq 2$, *are* NP-*complete.* $\square$

IV. CONCLUDING REMARKS

Let π^i denote the computational problem of hierarchical overlapping clustering defined in precisely the same way as the problem π where we subject the pyramid (P, f) on X to the additional condition $|P| - n - 1 = i$. Similarly as in Lemma 2 we have $\pi^i \propto \pi^{i+1}$, $i = 1, 2, \ldots$. It is of particular interest even from the point of view of hierarchical custering to decide the NP-completeness of the problem π^2. Note that the problem π^1 (as the problem π_1) has the trivial solution in polynomial time and that its solution is a hierarchy on X. On the other hand the solution of π^2 is a hierarchy on X as well. The special variant of the problem π^2 can be equivalently restated in the graph-theoretical framework as follows:

"Given a graph, find the minimum number of edge-changes (i.e. additions or deletions of an edge) which results in a graph which is exactly the union of one complete and one discrete graph."

We conjecture that even this variant of π^2 is NP-complete.

Acknowledgment. The author wishes to thank Dr. J. Morávek for his critical reading of the manuscript.

References

[1] *E. Diday:* Une représentation visuelle des classes empiétantes -- les pyramides. Rap. de Recherche No 291, INRIA, Rocquencourt, 1984.

[2] *M. R. Garey* and *D. S. Johnson:* Computers and Intractability: a Guide to the Theory of NP-completeness. W. H. Freeman, San Francisco, 1979.

[3] *L. Hubert:* A set-theoretical approach to the problem of hierarchical clustering. Journal of Mathematical Psychology, 15 (1977), 1 pp. 70—88.

[4] *N. Jardine* and *R. Sibson:* Mathematical Taxonomy. Wiley, London, 1971.

[5] *R. M. Karp:* Reducibility among combinatorial problems. Complexity of computer computations, Proceedings, Plenum Press 1972. Editors: R. E. Miller, J. W. Thatcher.

[6] *M. Křivánek* and *J. Morávek:* On NP-hardness in hierarchical clustering. COMPSTAT 1984, Proceedings, Physica-Verlag 1984. Editors: T. Havránek, Z. Šidák, M. Novák.

Souhrn

POZNÁMKA O VÝPOČETNÍ SLOŽITOSTI HIERARCHICKÉHO POKRÝVÁNÍ

MIRKO KŘIVÁNEK

V tomto článku se zkoumá výpočetní složitost problému aproximace dané míry nepodobnosti na konečné množině X pomocí k-ultrametriky na X a Robinsonovy míry nepodobnosti na X. V obou případech je ukázáno, že se jedná o NP-úplné problémy.

Author's address: RNDr. *Mirko Křivánek*, Výzkumný ústav matematických strojů, Loretánské nám. 3, 118 55 Praha 1.