On the m-clustering Problem on the Line

Anna A. Kurochkina[2] and Alexander A. Kurochkin[1] *

[1] Sobolev Institute of Mathematics, Koptyuga 4, 630090, Novjsibirsk, Russia
[2] SibGUTI, Kirova str., 86, 630102, Novosibirsk, Russia
a.potapova@ngs.ru, alkurochkin@ngs.ru

Abstract. We consider a special case of the Euclidean m-clustering problem, namely the problem of clustering on the real axis with given capacities of separate clusters. We prove that the problem considered is NP-hard. An example of non-optimality of the solution obtained for the problem with connected clusters is presented. A special case of the problem with connected clusters is solved by exact algorithm with running time $O(mn2^m)$ that is, depending linearly on n if m is fixed.

Keywords: Clusterization problem, real axes, NP-hard, algorithm, dynamic programming

1 Introduction

Clustering is the classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines [1]–[4], [6]–[7].

However, clustering is a difficult combinatorial problem.

We consider a special case of Euclidean clustering problem, namely the problem of clustering on the real axis.

Consider a set $X = \{x_i\}$ of n elements (points) on the real axis. Each point $x \in X$ has a weight $w(x)$.

The set of points is divided into m disjoint subsets (clusters) $C_1, C_2, \ldots, C_m$.

For each cluster C_k, one of the points $\bar{x}(C_k)$ on the axis is selected as the center of this cluster, depending on the subset of points included in the cluster and on the cluster cost function $\rho(x, \bar{x}(C_k))$. Further we will consider the cost function in the form

$$\rho(x, y) = \alpha + (1 - \alpha)||x - y||,$$

where $||x - y||$ is the distance between the points $x, y \in X$ on the line, α is some non-negative constant, that is less or equal 1.

* This research was supported by the Russian Foundation for Basic Research (grant 150100976)

In: A. Kononov et al. (eds.): DOOR 2016, Vladivostok, Russia, published at http://ceur-ws.org

As an example of selecting the cluster center may be the following point

$$\bar{x}(C_k) = \arg\min_{y \in C_k} \sum_{x \in C_k} w(x)\rho(x, y). \tag{1}$$

Another example of selecting a cluster center is the so-called centroid of the set C_k:

$$\bar{x}(C_k) = \frac{1}{|C_k|} \sum_{x \in C_k} x. \tag{2}$$

Each cluster C_k has the weight capacity W, where $W \geq \frac{1}{m} \sum_{x \in X} w(x)$.

Denote the costs associated with the cluster C_k by

$$f(C_k) = \sum_{x \in C_k} w(x)\rho(x, \bar{x}(C_k)).$$

m-CLUSTERING Problem Formulation:
Given:

- a set $X = \{x_i\}$ of n elements (points) on the real axis,
- a natural number $m \leq n$,
- a weight function $w : X \to \mathbb{R}$,
- the cluster weight capacity W,
- a cost function $\rho : E \to \mathbb{R}$,

Find: a partition of the set $X = \{x_i\}$ into disjoint subsets (clusters) $C_1, C_2, \ldots, C_m$ such that

$$\sum_{k=1}^{m} f(C_k) \to \min_{\{C_k\}}, \tag{3}$$

subject to

$$C_k \cap C_{k'} = \varnothing, \quad k \neq k', \ \cup_{k=1}^{m} C_k = X, \tag{4}$$

$$\sum_{x \in C_k} w(x) \leq W, \ k = 1, \ldots, m. \tag{5}$$

2 Algorithmic analysis of the m-clustering problem (6)–(8)

2.1 NP-hardness of the problem

Statement *The m-clustering problem (6)–(8) is NP-hard.*

Proof. Formulate the following Bin-Packing Problem in the verification form (BPP-ver).
Given: A finite set U of items with nonnegative numbers $w_1, \ldots, w_n$, and a positive integer bin capacity B.
Question: Is there a partition of U into disjoint sets $B_1, \ldots, B_m$ such that the sum of the items in each B_k is B or less?

It is known that the Bin-Packing Problem is NP-hard [5].

Consider the particular case of the m-clustering problem ((6)–(8) for $\alpha = 1$, and $B = W$:

$$\sum_{k=1}^{m} \sum_{x \in C_k} w(x) \to \min_{\{C_k\}}, \tag{6}$$

subject to

$$C_k \cap C_{k'} = \varnothing, \quad k \neq k', \ \cup_{k=1}^{m} C_k = X), \tag{7}$$

$$\sum_{x \in C_k} w(x) \leq W, \ k = 1, \ldots, m. \tag{8}$$

It easy to see that the solution of the NP-hard problem BPP-ver polynomially reduces to the problem (6)–(8).

2.2 An example of an non-optimal solution to problem (6)–(8

Given: A set of elements $X = \{0, 10, 10 + \varepsilon, 20\}$
with the weights of elements $\{5, 9, 1, 4\}$, $m = 2$, $B = 13$, $k = 1, 2$.

Find: A partition of the set of elements into $m = 2$ disjoint subsets C_1 and C_2 solving the minimization problem

$$\min_{\{C_k\}} \sum_{k=1}^{m} \sum_{x_i \in C_k} w(x_i) \rho(x_i, \bar{x}(C_k)) \tag{9}$$

subject to the capacity constraints (5) in two special cases:

CASE 1: selecting the cluster center by formula (2). It is easy to see that the optimal solution of problem (3)–(5) is is given by the disjoint clusters clusters $C_1 = \{0, 10 + \varepsilon\}$, $C_2 = \{10, 20\}$.

With connectivity requirement, we obtain
$C_1 = \{0\}$, $C_2 = \{10,\ 10 + \varepsilon,\ 20\}$ with the sum of $w_2 + w_3 + w_4 = 14 > B = 13$, which does not satisfy constraint (5).

CASE 2: selecting the cluster center by formula (1). In this case, we have the same solution as in the previous case.

3 Dynamic Programming for the special m-clustering problem

Next will also address the problem (3)–(5) under the additional conditions:
(a) connectivity of clusters on the real line;
(b) $\sum_{x_i \in C_k} w_i \leq B_k$, $k = 1, \ldots, m$, where $\sum_{k=1}^{m} B_k \geq \sum_{i=1}^{n} w_i$;
(c) $|C_k| = L_k$, $k = 1, \ldots, m$, where $L_1 + \ldots + L_m = n$;
(d) $\alpha = 1$.

The connectivity conditions mean that if any two points belong to the same cluster, then all points between them must belong to this clust too :

$$x_q \le x_{i_1}, \ldots, x_{i_r} \le x_p, and x_q, x_p \in C_k \Rightarrow x_{i_1}, \ldots, x_{i_r} \in C_k. \tag{10}$$

Problem (3)–(5) can be solved with an exhausting search among all $m!$ different location of the clusters on a real line from point x_1 to point x_n. The following means of dynamic programming is done in time $O(mn2^m)$, which is considerably less than $m!$.

By Y^μ denote the set of m-vectors $Y^\mu = (y_1, y_2, \ldots, y_m),\ \ y_k \in \{0, 1\}$, such that

$$\sum_{k=1}^{m} y_k = \mu.$$

The initial problem in the set of points $\{x_1, x_2, \ldots, x_n\}$ is denoted by $\left\langle n, m, Y^m \right\rangle$.

Along with this, also consider the family of problems

$$\{\left\langle i, \mu, y^\mu \right\rangle, 1 \le i \le n,\ y^\mu \in Y^\mu,\ 1 \le \mu \le m\}$$

for the subsets $\{x_1, x_2, \ldots, x_n\}$ of the real axis, $\ 1 \le i \le n$.

The problem $\left\langle n, m, Y^m \right\rangle$ is solved by Dynamic Programming using the recurrence relations:

for $\mu = 1$:

$$F_i(1, Y^1) = \begin{cases} \min\limits_{1 \le k \le n} \{f_k(i - L_k, i) \mid i = L_k, \sum\limits_{x_i \in C_k} w_i \le B_k\}; \\ \infty, \text{ if there is no k satisfying the specified conditions.} \end{cases}$$

$1 \le k \le m,\ 1 \le i \le n$;

when $\mu > 1$:

$$F_i(\mu, Y^\mu) = \min_{1 \le k \le m} \{F_{i - L_k}(\mu - 1, Y^\mu - \boldsymbol{e}_k) + f_k(i - L_k, i)\},$$

$$1 \le i \le n,\ 1 < \mu \le m.$$

Here $f_k(i - L_k, i)$ is the associated cost for the cluster $C_k = \{x_j | n - L_k < j \le i\}$. Whenever the condition $\sum\limits_{x_i \in C_k} w_i \le B_k$ fails, we set $f_k(i - L_k, i)$ equal to ∞.

The time complexity of the algorithm can be estimated as follows:

$$mn \sum_{\mu=1}^{m} |Y^\mu| = mn \sum_{\mu=1}^{m} C_m^\mu = O(mn2^m).$$

If the number of clusters is fixed there is no decision under the given constraints.

Remark. In the case of final response $F_n(m, Y^m) = \infty$, decision under the given conditions is missing.

4 Conclusions

1) We prove that the considered problem is NP-hard.

2) We present an example of non-optimality of the solution obtained for the problem with connected clusters.

3) To solve a special case of the problem with connected clusters an exact algorithm is constructed with the computational complexity $O(mn2^m)$ that is, depending linearly on n if m is fixed.

References

1. Anil K., Jain K.: Data Clustering: 50 Years Beyond k-Means. Pattern Recognition Letters 31. 651–666 (2010).
2. Bishop M.C.: Pattern Recognition and Machine Learning. New York: Springer Science+Business Media, LLC, 738 p (2006).
3. Brucker P.: On the Complexity of Clustering Problems. Lecture Notes in Economics and Mathematical Systems 157, 45-54 (1978).
4. Flach P.: Machine Learning: The Art and Science of Algorithms that Make Sense of Data. New York: Cambridge University Press, 396 p (2012).
5. Garey M. R. and Johnson D. S.: Computers and Intractability, Freeman, San Francisco (1979).
6. James G., Witten D., Hastie T., Tibshirani R: An Introduction to Statistical Learning. New York: Springer Science+Business Media, LLC, 426 p (2013).
7. Rao M. R.: Cluster Analysis and Mathematical Programing. Journal of the American Statistical Assosiation 66(335), 622-626 (1971).