Annales Univ. Sci. Budapest., Sect. Comp. 56 (2024) 83–86

LLOYD'S CLUSTERING METHOD IS NOT 1-SEPARABILITY DETECTING

Katalin Bene and László Szabó

(Budapest, Hungary)

Communicated by László Szili

(Received June 15, 2023; accepted July 30, 2023)

Abstract. In this note we construct a data set in the plane with a 1-separable k-clustering for any $k \geqslant 2$ such that Lloyd's method doesn't terminate with this clustering regardless of the initialization method.

1. Introduction

Clustering is a fundamental tool for data analysis. Its goal is natural: to identify groups of similar items within data. While the goal of clustering is simple, formalizing this task is much more challenging. It is well known that most of the common clustering objectives are NP-hard to optimize. In practice, however, clustering is being routinely carried out. One approach for providing theoretical understanding of this seeming discrepancy is to introduce notions of clusterability that distinguish realistically interesting input data from worst-case data sets. Here we focus on one such notion.

We consider a space (X, d) where X is a set of data elements and d is a distance function on X. It is assumed that d is symmetric and non-negative,

Key words and phrases: Clustering, point set, k-means clustering, separability detection.
2010 Mathematics Subject Classification: 62H30.
The work was supported by the project "Integrált kutatói utánpótlás-képzési program az informatika és számítástudomány diszciplináris területein (Integrated program for training new generation of researchers in the disciplinary fields of computer science)", No. EFOP-3.6.3-VEKOP-16-2017-00002. The project has been supported by the European Union and co-funded by the European Social Fund.

and $d(x,x) = 0$ for all $x \in X$. For an integer $k \geqslant 1$, a k-clustering of X is a partition $\mathcal{C} = \{C_1, \ldots, C_k\}$ of X into k disjoint non-empty sets. For a k-clustering $\mathcal{C}$ of X and data elements $x_1, x_2 \in X$, we write $x_1 \sim_{\mathcal{C}} x_2$ if x_1 and x_2 belong to the same cluster in $\mathcal{C}$, and $x_1 \not\sim_{\mathcal{C}} x_2$ otherwise. A k-clustering $\mathcal{C}$ of X is called 1-separable if for any $x_1, x_2, x_3, x_4 \in X$ such that $x_1 \sim_{\mathcal{C}} x_2$ and $x_3 \not\sim_{\mathcal{C}} x_4$ the inequality $d(x_1, x_2) < d(x_3, x_4)$ holds. Note that the 1-separable k-clustering of any given data set is unique, if it exists.

One of the most widely used algorithm for clustering is Lloyd's method [4]. For a given data set X and initial center set S in the n-dimensional Euclidean space Lloyd's method performs the following steps until two consecutive iterations return the same clustering: (1) assign each point in X to its closest element of S producing a clustering of X, (2) replace S with the set of the centers of gravity of data elements assigned to each cluster.

A common initialization for Lloyd's method is to select k random centers from the input data set [2]. Another well-known initialization method is the so-called furthest-centroid initialization [3]. Using this method, given a set X, the initial centers $c_1, c_2, \ldots, c_k$ in S are chosen as follows: center c_1 is an arbitrary point in X, then, for each $i = 2, 3, \ldots, k$, center c_i is set to be the point in X that maximizes the distance from the set of the other centers that were already chosen.

Margareta Ackerman, Shai Ben-David, David Loker and Sivan Sabato stated in [1], as Lemma 6.4, that Lloyd's clustering method with furthest centroid initialization is 1-separability detecting, i.e., it always terminates with a 1-separable clustering if there is such a clustering of the data set.

In this short note we show that this is not true by constructing a data set in the plane with a 1-separable k-clustering for any $k \geqslant 2$ such that Lloyd's method doesn't terminate with this clustering regardless of the initialization method.

2. The counterexample

Consider first the data set X consisting of the 12 points

$$x_i = (-5 + i, 45) \text{ for } i = 1, 2, \ldots, 9,$$
$$x_{10} = (0, 0),\ x_{11} = (-22, -40),\ x_{12} = (22, -40)$$

in the plane (see Figure 1).

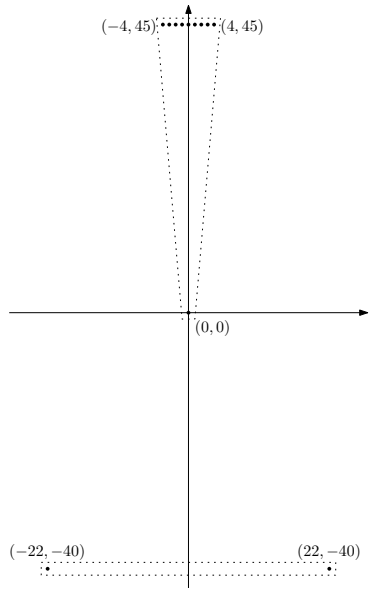

Figure 1

Now

$$\mathcal{C} = \{\{x_1, x_2, \ldots, x_9, x_{10}\}, \{x_{11}, x_{12}\}\}$$

is a 1-separable 2-clustering of X since the within-cluster distances are

$$d(x_i, x_j) \leqslant d(x_1, x_9) = 8$$

for $1 \leqslant i < j \leqslant 9$,

$$d(x_i, x_{10}) \leqslant d(x_1, x_{10}) = d(x_9, x_{10}) = \sqrt{2041} = 45.17\ldots$$

for $i = 1, 2, \ldots, 9$, and

$$d(x_{11}, x_{12}) = 44,$$

while the between-cluster distances are

$$d(x_i, x_k) \geqslant d(x_1, x_{11}) = d(x_9, x_{12}) = \sqrt{7549} = 86.88\ldots$$

for $i = 1, 2, \ldots, 9$ and $k = 11, 12$, and

$$d(x_{10}, x_{11}) = d(x_{10}, x_{12}) = \sqrt{2084} = 45.65\ldots.$$

Run Lloyd's method on X and suppose that it returns $\mathcal{C}$ in some iteration. Now the algorithm calculates the centers of gravity of the clusters which are

$(0, 40.5)$ and $(0, -40)$ for $\{x_1, x_2, \ldots, x_9, x_{10}\}$ and $\{x_{11}, x_{12}\}$, respectively, and assigns each point in X to its closest center. In this way we obtain the clustering

$$\mathcal{C}' = \{\{x_1, x_2, \ldots, x_9\}, \{x_{10}, x_{11}, x_{12}\}\}$$

which is different from $\mathcal{C}$. The algorithm calculates again the centers of gravity of the clusters which are $(0, 45)$ and $(0, -80/3)$ for $\{x_1, x_2, \ldots, x_9\}$ and $\{x_{10}, x_{11}, x_{12}\}$, respectively, and assigns each point in X to its closest center. Now we obtain the the same clustering $\mathcal{C}'$ as before, thus Lloyd's method terminates with $\mathcal{C}'$.

This implies that Lloyd's method running on X never terminates with $\mathcal{C}$ regardless of the initialization method.

This counterexample can easily be generalized to any $k \geqslant 3$. By adding the points

$$x_{12+j} = (0, 100j) \text{ for } j = 1, 2, \ldots, k-2$$

to X we obtain a data set with a 1-separable k-clustering

$$\{\{x_1, x_2, \ldots, x_9, x_{10}\}, \{x_{11}, x_{12}\}, \{x_{13}\}, \ldots, \{x_{13+k-3}\}\}$$

such that Lloyd's method running on this data set never terminates with this clustering regardless of the initialization method.

Acknowledgement. The authors would like to thank the anonymous referee for the insightful comments and suggestions.

References

[1] **Ackerman, M., S. Ben-David, D. Loker and S. Sabato,** Clustering oligarchies, in: C. M. Carvalho, P. Ravikumar (Eds.) *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, PMLR* **31** (2013), 66–74.

[2] **Forgy, E.W.,** Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics,* **21** (1965), 768–769.

[3] **Katsavounidis, I., C.C. Jay Kuo and Z. Zhang,** A new initialization technique for generalized Lloyd iteration, *IEEE Signal Processing Letters,* **1** (1994), 144–146.

[4] **Lloyd, S.P.,** Least squares quantization in PCM, *IEEE Transactions on Information Theory,* **28** (1982), 129–137.

K. Bene and L. Szabó
Department of Algorithms and their Applications
Eötvös Loránd University
Budapest
Hungary
`koecwg@inf.elte.hu` and `szabolaszlo@inf.elte.hu`