Size constrained clustering problems in fixed dimension

Jianyi Lin

Dipartimento di Informatica, Università degli Studi di Milano, Italy
jianyi.lin@unimi.it

Extended Abstract

Clustering or cluster analysis [1] is a classical method in unsupervised learning and one of the most used techniques in statistical data analysis. Clustering has a wide range of applications in many areas like pattern recognition, medical diagnostics, data mining, biology, market research and image analysis among others. A cluster is a set of data points that in some sense are similar to each other, and clustering is a process of partitioning a data set into disjoint clusters. In *distance clustering*, the similarity among data points is obtained by means of a *distance* function.

Fixed a norm $\| \ \|_p$ ($p \geq 1$), the *clustering problem* consists in finding for a finite point set $X \subset \mathbb{Q}^d$ and an integer k, a k-partition $\{A_1, ..., A_k\}$ of X that minimizes the cost function

$$W(A_1, ..., A_k) = \sum_{i=1}^{k} \sum_{x \in A_i} \|x - C_{A_i}\|_p^p \tag{1}$$

where C_{A_i} is the *p-centroid* of A_i, i.e.

$$C_{A_i} = \arg\min_{\mu} \sum_{x \in A_i} \|x - \mu\|_p^p$$

Distance clustering is a difficult problem. For an arbitrary dimension d, assuming the Euclidean norm ($p = 2$), the problem is NP-hard even if the number k of clusters equals 2 [2]; the same occurs if $d = 2$ and k is arbitrary [3,4]. For the Euclidean distance, a well-known heuristic is Lloyd's algorithm [5,6], also known as the k-Means Algorithm; however there is no guarantee that the solution yielded by this procedure approximates the global optimum. This algorithm is usually very fast, but it can require exponential time in the worst case [7].

In real-world problems, often people have some information on the clusters: incorporating this information into traditional clustering algorithms can increase the clustering performance. Problems that include background information are called *constrained clustering* problems and are divided in two classes.

On the one hand, clustering problems with instance-based constraints typically comprise a set of must-link constraints or cannot-link constraints [8], defining pairs of elements that must be included, respectively, in the same cluster or in different clusters.

On the other hand, clustering problems with cluster-based constraints [9,10] incorporate constraints concerning the size of the possible clusters. Recently, in [11] cluster size constraints are used for improving clustering accuracy; this approach, for instance, allows one to avoid extremely small or large clusters in standard cluster analysis.

Here we study a constrained clustering problem where the size of clusters is included in the instance. This problem, called *Size Constrained Clustering Problem* (SCC), is formally defined as follows: given a set $X \subset \mathbb{Q}^d$ of n points and k many positive integers $m_1, ..., m_k$ such that $\sum_1^k m_i = n$, find a k-partition $\{A_1, ..., A_k\}$ of X that minimizes the cost function $W(A_1, ..., A_k)$ such that $|A_i| = m_i$ for each $i = 1, ..., k$. This problem was studied in [12,13] and it is known to be a difficult problem. More precisely, the following results hold [13]:

1) For every norm $\| \ \|_p$ with $p > 1$, SCC with fixed clustering size k is NP-hard, even in the case $k = 2$ and $m_1 = m_2 = \frac{n}{2}$.
2) For every norm $\| \ \|_p$ with $p \geq 1$, SCC with fixed dimension d is NP-hard, even in the case $d = 1$.

As a consequence, we can't expect to obtain a polynomial-time algorithm for solving the general SCC problem.

In this paper we investigate SCC in the plane ($d = 2$) with a fixed clustering size $k = 2$. In particular, we consider the following two problems:
• 2-SCC in the Plane:
Given a point set $X = \{x_1, ..., x_n\} \subset \mathbb{Q}^2$ and a positive integer $m \leq \frac{n}{2}$, find a 2-partition $\{A, \bar{A}\}$ of X with $|A| = m, |\bar{A}| = n - m$, that minimizes

$$W(A, \bar{A}) = \sum_{x \in A} \|x - C_A\|_2^2 + \sum_{x \in \bar{A}} \|x - C_{\bar{A}}\|_2^2$$

where C_A and $C_{\bar{A}}$ are the centroid of A and $\bar{A}$ respectively.
• Full 2-SCC in the Plane:
Given a point set $X = \{x_1, ..., x_n\} \subset \mathbb{Q}^2$, for all integers m, $1 \leq m \leq \frac{n}{2}$, find the optimal 2-partition $\pi_m = \{A_m, \bar{A}_m\}$, with $|A_m| = m$.

The main results we obtain are the following:

1) There is an algorithm for solving Full 2-SCC in the Plane in time $O(n^2 \cdot \log n)$.
2) There is an algorithm for solving 2-SCC in the Plane in time $O(n \sqrt[3]{m} \cdot \log^2 n)$.

It should be observed that, the algorithm for solving 2-SCC in the plane requires the application of methods for enumerating the k-sets of a collection of points in the plane, which is a challenging problem [14] in combinatorial geometry.

Here we also study the problem 2-SCC in fixed dimension d. First, we use a separation result [13] stating that if $\{A, \bar{A}\}$ is an optimal solution of an instance of the 2-SCC problem, then A and $\bar{A}$ are separated by an hypersurface of the form

$$\|x - \alpha\|_p^p - \|x - \beta\|_p^p = c$$

for some constant parameters $\alpha, \beta \in \mathbb{R}^d$, $c \in \mathbb{R}$. By applying a suitable method for decomposing the parameter space $\mathbb{R}^{2d+1}$, one can compute a set of optimal 2-partitions $\pi_m = \{A_m, \bar{A}_m\}$ such that $|A_m| = m$, for $m = 1, ..., \lfloor \frac{n}{2} \rfloor$. This allows us to design an algorithm for the Full 2-SCC problem in fixed dimension d that works in polynomial time both in n and p. To obtain this result we make use of concepts and methods of real algebraic geometry, and in particular we apply the cylindrical algebraic decomposition [15].

In this work we also study another variant of the clustering problem, called *Relaxed Constraints Clustering Problem* (RCC), which is defined as follows: given a point set $X = \{x_1, ..., x_n\} \subset \mathbb{Q}^d$, an integer $k > 1$ and a finite set $\mathcal{M}$ of positive integers, find a k-partition $\{A_1, ..., A_k\}$ of X with

$$|A_i| \in \mathcal{M} \quad \text{for all } i = 1, ..., k$$

that minimizes the cost function

$$W(A_1, ..., A_k) = \sum_{i=1}^{k} \sum_{x \in A_i} \|x - C_{A_i}\|_p^p.$$

We prove that for the euclidean norm $\| \ \|_2$, the decision version of RCC in dimension $d = 2$ is NP-complete even in the case $\mathcal{M} = \{2, 3\}$. On the contrary, RCC in dimension 1 is known to be solvable in polynomial time by a dynamic programming technique [12].

Acknowledgements

The author wishes to acknowledge helpful discussions with Alberto Bertoni and Massimiliano Goldwurm.

References

1. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd edn. Springer-Verlag (2009)
2. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of Euclidean sum-of-squares clustering. Machine Learning **75** (2009) 245–249
3. Mahajan, M., Nimbhorkar, P., Varadarajan, K.: The Planar k-Means Problem is NP-Hard. In Das, S., Uehara, R., eds.: WALCOM: Algorithms and Computation. Volume 5431 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg (2009) 274–285
4. Vattani, A.: The hardness of k-means clustering in the plane. manuscript (2009)

5. Lloyd, S.: Least squares quantization in PCM. IEEE Transactions on Information Theory **28**(2) (1982) 129–137
6. MacQueen, J.B.: Some method for the classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Structures. (1967) 281–297
7. Vattani, A.: K-means requires exponentially many iterations even in the plane. In: Proceedings of the 25th Symposium on Computational Geometry (SoCG). (2009)
8. Wagstaff, K., Cardie, C.: Clustering with instance-level constraints. In: Proc. of the 17th Intl. Conf. on Machine Learning. (2000) 1103–1110
9. Bradley, P.S., Bennett, K.P., Demiriz, A.: Constrained K-Means Clustering. Technical Report MSR-TR-2000-65, Miscrosoft Research Publication (May 2000)
10. Tung, A., Han, J., Lakshmanan, L., Ng, R.: Constraint-Based Clustering in Large Databases. In Van den Bussche, J., Vianu, V., eds.: Database Theory ICDT 2001. Volume 1973 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg (2001) 405–419
11. Zhu, S., Wang, D., Li., T.: Data clustering with size constraints. Knowledge-Based Systems **23**(8) (2010) 883–889
12. Saccà, F.: Problemi di Clustering con Vincoli: Algorithmi e Complessità. PhD thesis, University of Milan, Milan (2010)
13. Bertoni, A., Goldwurm, M., Lin, J., Saccà, F.: Size Constrained Distance Clustering: Separation Properties and Some Complexity Results. Fundamenta Informaticae **115**(1) (2012) 125–139
14. Erdős, P., Lovász, L., Simmons, A., Straus, E.G.: Dissection graphs of planar point sets. In: A survey of combinatorial theory (Proc. Internat. Sympos., Colorado State Univ., Fort Collins, Colo., 1971). North-Holland, Amsterdam (1973) 139–149
15. Collins, G.E.: Quantifier Elimination for Real Closed Fields by Cylindrical Algebraic Decomposition. In Barkhage, E., ed.: Proc. 2nd GI Conf. on Automata Theory and Formal Lang. Volume 33 of LNCS., Berlin, Springer (1975) 134–183