アルゴリズム 33−12
（1993. 5. 29）

点集合を分散の総和が最小となるように k個のクラスターに分割するアルゴリズム

長谷川 進*, 今井 浩*, 稲葉 真理*,
加藤 直樹†, 中野 淳‡

* 東京大学理学部情報科学科　† 神戸商科大学管理科学科　‡ 日本IBM 東京基礎研究所

d次元空間内のn点をk個のクラスターに分割し、クラスターの分散の総和を最小化するアルゴリズムについて考察する。この問題はNP困難であるがkやdの値が十分小さい場合に対しては効率良く解けることを示す（$k=2$のときは$O(n^d)$時間; $k=3$, $d=2$のときには$O(n^5 \log n)$時間、等）。さらに次元によらずに$O(n^{k+1})$時間で終了する近似比2のアルゴリズムを示す。

Efficient Algorithms for Variance-Based k-Clustering

Susumu Hasegawa*, Hiroshi Imai*, Mary Inaba*,
Naoki Katoh†, and Jun Nakano‡

* Department of Information Science, University of Tokyo, Tokyo 113, Japan
† Department of Management Science, Kobe University of Commerce, Kobe 651-21, Japan
‡ IBM Tokyo Research Laboratory, Kanagawa 242, Japan

We are given n points in the d-dimensional space. The k-clustering problem that we consider is to find a partition of these points into k disjoint sets that minimizes the sum of variances of clusters. We show that although this problem is NP-hard, we can solve it efficiently for some special cases where k and/or d are small: in $O(n^d)$ time for $k=2$; in $O(n^5 \log n)$ time for $k=3$, $d=2$, etc. We also give an $O(n^{k+1})$-time approximation algorithm with the performance ratio of 2 for any d.

Abstract

In this paper we consider the k-clustering problem for n points in the d-dimensional space, motivated from the problem of computing a color lookup table for frame buffer display, and that of compressing two-dimensional image data. Using the technique of computational geometry, this clustering problem is investigated in a unified manner.

A k-clustering is a partition of the given set of n points into k disjoint sets, called clusters. Representing each cluster by its centroid, the sum-of-squared error of the cluster is defined to be the sum of squared distance of the centroid and each point in the cluster. The all-pairs sum-of-squared error of the cluster is defined to be the sum of squared distance of every pair of points in the cluster. The min-sum k-clustering problem with some specified error for clusters is to find a k-clustering such that the sum of the error of each cluster is minimum. These criteria are suited to the color image quantization problem. The all-pairs sum-of-squared error criterion makes the sizes of clusters more balanced compared with the sum-of-squared error criterion.

This paper shows the following.

- The min-sum 2-clustering problem with sum-of-squared error can be solved in $O(n^d)$ time.
- The min-sum 3-clustering problem with sum-of-squared error in the planar case ($d = 2$) can be solved in $O(n^5 \log n)$ time and $O(n)$ space (or, in $O(n^5)$ time and $O(n^2)$ space). This result can be generalized to the $(d+1)$-clustering problem in the d-dimensional space.
- If k is considered to be a constant, the min-sum k-clustering problem with sum-of-squared error can be solved in polynomial time in n.
- For any d, there is an $O(n^{k+1})$-time approximation algorithm for the min-sum k-clustering with sum-of-squared error such that the sum-of-squared error of its approximate clustering is larger by at most twice than that of the optimum clustering.

Our results for small k in the two- and three-dimensional cases can be used in top-down heuristics for the clustering problem with large k, and are quite powerful since the proposed algorithms for small k are fast enough.

1 Introduction

Clustering problems are very fundamental in various fields in computer science. For example, the color image quantization in computer graphics is a clustering problem in the three-dimensional space. This problem arises in displaying a color image represented by part of 256^3 colors on a display with frame buffer of limited size, say 256. To produce a good approximate image in this situation, points in the RGB space are clustered into 256 clusters, and points in each cluster is represented by a representative point of the cluster. The color quantization is also useful in compressing the amount of color image data [9].

In this paper we consider the k-clustering problem for a set S of n points in the d-dimensional space, motivated by the above-mentioned color quantization problem. In general, the problem of finding an optimum clustering is NP-hard [6, 13], and therefore is intractable. However, if a specified clustering criterion has nice properties, and k and/or d

are small, the problem may be solved rigorously in polynomial time in n, which is the theme of this paper. In the above color quantization problem, $k = 256$ and $d = 3$, and hence k is not small at all. For large k, the problem is intractable as stated above, and heuristic algorithms are used in practice. Many heuristics employ a top-down clustering approach, where as subproblems the clustering problems with small k are repeatedly solved. Efficient algorithms for small k presented in this paper are very useful in such stages.

We now define the k-clustering problem, and describe our results together with related problems. A k-clustering is a partition of the given set S of n points p_i $(i = 1, \ldots, n)$ in the d-dimensional space into k disjoint sets $S_1, \ldots, S_k$, called clusters. Representing each cluster S_i by its centroid, the sum-of-squared error $\mathrm{Error}_{\mathrm{ss}}(S_i)$ of the cluster S_i is defined to be the sum of squared distance of the centroid and each point in S_i. The all-pairs sum-of-squared error $\mathrm{Error}_{\mathrm{apss}}(S_i)$ of the cluster S_i is defined to be the sum of squared distance of every pair of points in S_i. The variance $\mathrm{Var}(S_i)$ of the cluster is $\mathrm{Error}_{\mathrm{ss}}(S_i)$ divided by $|S_i|$. Concerning these three, we have the following.

$$\mathrm{Var}(S_i) = \frac{1}{|S_i|}\mathrm{Error}_{\mathrm{ss}}(S_i) = \frac{1}{|S_i|^2}\mathrm{Error}_{\mathrm{apss}}(S_i) \tag{1}$$

As is seen from (1), the all-pairs sum-of-squared error criterion makes the sizes of clusters more balanced compared with the sum-of-squared error criterion. The min-sum k-clustering problem with some specified error for clusters finds a k-clustering such that the sum of the error of each cluster is minimum. The minimax k-clustering with some specified error finds a k-clustering such that the maximum of the error of each cluster is minimum.

There are different criteria for the error of a cluster. The diameter $\mathrm{Diam}(S_i)$ of a cluster S_i of points is the maximum distance over all pairs of points in the cluster. For the min-sum and minimax k-clustering problem with diameters as errors, efficient algorithms have been proposed for the case of $k = 2$ and $d = 2$ in computational geometry [3, 11]. These algorithms solve the min-diameter 2-clustering problems in the plane in $O(n \log n)$ time. However, concerning the color quantization, the diameter is not a good criterion to represent clusters, because the diameter is determined only by two points, and the other points do not affect the diameter much. The radius of a cluster of points is defined to be the radius of the minimum enclosing circle of the points. The minimax k-clustering based on the radius is equivalent to the so-called k-center problem. Using the radius as a criterion is not good for the color quantization by a similar reason. The k-clustering problem based on the variance is thus more suited to the color quantization, and, for this application, the three-dimensional problem should be solved.

In this paper, we first consider the min-sum k-clustering problem with sum-of-squared error. The case of $k = 2$ is the most fundamental case, and, we show that, in an optimum 2-clustering, two clusters are linearly separable. With this property combined with computational-geometric algorithms and a wise way of updating the sum-of-squared error of a cluster of points, we present an $O(n^d)$-time algorithm. Next, the case of $k = 3$ and $d = 2$ is investigated, and an $O(n^5 \log n)$-time $O(n)$-space algorithm is developed. This result can be generalized to the $(d + 1)$-clustering problem in the d-dimensional space. If k is consid-

ered to be a constant, the k-clustering problem with sum-of squared error can be solved in polynomial time in n. Here, we utilize an idea of Voronoi partition of n points generated by k generators (Hasegawa [8]). The algorithm for general d takes $O(n^d)$ time, which is very costly for large d. For the case of large d and small k, we present an approximation algorithm for the k-clustering with minimum variance such that the sum of variances of its approximate clustering is larger by at most twice than that of the optimum clustering. The time complexity of the approximation algorithm is $O(n^{k+1})$.

2 Min-sum k-clustering with sum-of-squared error

We first provide some more definitions and fundamental relations. Let S be a set of n points p_i $(i = 1, \ldots, n)$ in the d-dimensional space with the $(x_1, x_2, \ldots, x_d)$-coordinate. For a subset S' of S, we denote by $x_j(S')$ the sum of the x_j-coordinate values of points p_i in S'. Similarly, we denote by $x_j^2(S')$ the sum of the squared values of the x_j-coordinate of points p_i in S'. The centroid of S is $\bar{p} = \frac{1}{|S|}(x_1(S), x_2(S), \ldots, x_d(S))$. As is well known, the sum of squared distances of the centroid and points in S is expressed as follows.

$$\mathrm{Error}_{ss}(S) = \sum_{p_i \in S} d(\bar{p}, p_i)^2 = \sum_{j=1}^{d} (x_j^2(S) - x_j(S)^2/|S|) \tag{2}$$

where $d(\bar{p}, p_i)$ is the Euclidean distance between $\bar{p}$ and p_i and $x_j(S)^2 = (x_j(S))^2$.

2.1 Voronoi partition and k-clustering with minimum variance

In this subsection, we will first define the Voronoi partition of n points generated by k points and characterize a solution of the clustering problem using it. This characterization leads to efficient algorithms.

Consider k points q_l $(l = 1, \ldots, k)$ in the d-dimensional space. Define the Voronoi region $\mathrm{Vor}(q_l)$ of q_l by

$$\mathrm{Vor}(q_l) = \bigcap_{h=1}^{k} \{ p \mid d(p, q_l) \leq d(p, q_h) \}$$

For any point in $\mathrm{Vor}(q_l)$, q_l is the closest point among q_h $(h = 1, \ldots, k)$. $\mathrm{Vor}(q_l)$ $(l = 1, \ldots, k)$ partitions the space, which is called the Voronoi diagram generated by these k points q_l.

By the Voronoi diagram generated by these k points, n points in the given set S are naturally partitioned into k clusters. We call this partition a Voronoi partition of n points in S by k generators. Apparently, not all k-clusterings are Voronoi partitions. In fact, we can characterize optimal solutions of the min-sum k-clustering problem with sum-of-squared error as follows.

Theorem 2.1 ([4, 15]) *An optimum min-sum k-clustering with sum-of-squared error is a Voronoi partition.*

By this theorem, this k-clustering problem can be solved by enumerating all the Voronoi partitions of n points generated by k points, and finding a partition with minimum variance. Concerning the number of distinct Voronoi partitions, Hasegawa [8] shows the following.

Lemma 2.1 ([8]) *The number of Voronoi partitions of n points generated by k points in the d-dimensional space is $O(n^{dk})$.*

Hasegawa proves this lemma by extending the VC-dimension, known in computational geometry and computational learning theory, to the k-label space. A similar extension was given by Natarajan [12]. It is conjectured that the approach in [8] is more suitable to counting the number of Voronoi partitions. For this general theorem, we refer to Hasegawa [8], but, for its special case, we will give a tight bound in the next subsection.

There has been known no algorithm which enumerates all the Voronoi partitions in $O(n^{dk})$ time. We only have an $O(n^{dk^2})$-time algorithm to do so. This complexity is polynomial in n, and in general we obtain the following theorem.

Theorem 2.2 *The min-sum k-clustering problem for n points with sum-of-squared error in the d-dimensional space can be solved in polynomial time in n when d and k are regarded as constants.*

2.2 Cases of small k

We now turn to the case of small k. We first consider the case of $k = 2$. Since a Voronoi partition in this case is a partition by a hyperplane, an optimal min-sum 2-clustering with sum-of-squared consists of two clusters which are linearly separable.

To list up linearly separable two clusters, the hyperplane arrangement can be used via the duality transform, which is well know in computational geometry. By using the topological sweep algorithm [5], all linearly separable 2-clusterings can be arranged in $O(n^d)$ time and $O(n)$ space in such a way that most of consecutive pairs of 2-clusterings are very similar. Combining this with a fact that the sum-of-squared error of a point set can be maintained in a constant time for addition or deletion of a point by using the relation (2).

Theorem 2.3 *The min-sum 2-clustering problem for n points with sum-of-squared error in the d-dimensional space can be solved in $O(n^d)$ time and $O(n)$ space.*

Next, we consider the case of $k = 3$ and $d = 2$. By the characterization theorem, an optimum 3-clustering is a Voronoi partition generated by three points. As in the definition of Voronoi partitions, define a partition of n points in S into six clusters by three lines meeting at a point in the plane. Apparently, for each Voronoi partition of S generated by three points, we can assign a unique partition by three lines meeting at a point (just extend Voronoi edges (half lines) to lines). Concerning the latter partition, we have the following.

Lemma 2.2 *The number of partitions of n points in S induced by three lines meeting at a point is $O(n^5)$.*

Proof. Consider a partition of n points by three lines meeting at a point. Then, moving these three lines so that they continue to meet at a point, we can find another set of three lines such that this set induces the same partition of S with that for the original set of three lines and that two lines among them pass two points in S and the remaining line passes a point in S.

Such partition can be enumerated by choosing two points from S for the first line, choosing two points from S for the second line, and then choosing a point from S and letting the third line pass the chosen point and the point of intersection of the first and second lines. Hence, the number of partitions of S by three lines meeting at a point is $O(n^5)$. □

In enumerating all the partitions based on this proof, if the first and second lines are fixed, the third line may be regarded as a circular sweeping line passing the point of intersection of the first and second lines. By this sweeping technique combined again with maintaining the variance of point set efficiently, we have the following.

Theorem 2.4 *The min-sum 3-clustering problem with sum-of-squared error in the plane can be solved in $O(n^5 \log n)$ time and $O(n)$ space. If we allow $O(n^2)$ space, an $O(n^5)$ time algorithm can be designed using dual transformation of the set of points.*

We can generalize this result to 4-clustering in the three-dimensional space, and in this case the complexity becomes $O(n^{11})$. Although this complexity is high, we may make use of this bound through randomization in future work.

2.3 An efficient approximation algorithm with worst case ratio at most 2

By the algorithms described so far, even when k is small, the algorithms are not efficient if d is large. In such situations, we may utilize an approximation algorithm in practice. In this subsection, we consider the constrained version of the k-clustering problem, and use it to obtain such an approximate algorithm.

In the k-clustering problem with sum-of-squared error, the representative point of each cluster is its centroid. The centroid is not contained in the original set of points in most cases. We may restrict the representative point to be in the original set of points. Define the constrained centroid to be a point p in the cluster such that the sum of squares of distances of p and points in the cluster is minimum. Further, define the constrained sum-of-squared error of the cluster to be the sum of squares of distances of the constrained centroid and points in the cluster. This problem is called the constrained k-clustering problem. Due to the constraint, it is easy to solve the constrained problem.

Lemma 2.3 *The min-sum constrained k-clustering problem with constrained sum-of-squared error for n points in any fixed dimensions d can be solved in $O(n^{k+1})$ time and $O(n)$ space.*

Furthermore, between the unconstrained and constrained errors of a cluster of points, we have the following.

Lemma 2.4 *The sum-of-squared error of a constrained optimum solution is at most twice the sum-of-squared error of an unconstrained optimum solution.*

Proof. Let S be an arbitrary set of n points. Let UOPT $= \{(p_1, S_1), \ldots, (p_k, S_k)\}$ be an optimum solution of the unconstrained clustering problem for S, where p_i is the representative point of the cluster S_i. We denote the sum-of-squared error of this solution by $C_{\mathrm{UOPT}}(S)$. Then we define a solution CUOPT of the constrained problem for S by CUOPT $= \{(q_1, S_1), \ldots, (q_k, S_k)\}$, where $q_i \in S_i$ is defined by

$$\sum_{r \in S_i} d(q_i, r)^2 = \min_{r' \in S_i} \sum_{r \in S_i} d(r', r)^2.$$

We denote the sum-of-squared error of this solution by $C_{\mathrm{CUOPT}}(S)$. Since

$$\sum_{r' \in S_i} \sum_{r \in S_i} d(r', r)^2 = 2|S_i| \sum_{r \in S_i} d(p_i, r)^2,$$

we have from the definition of q_i

$$\sum_{r \in S_i} d(q_i, r)^2 \leq 2 \sum_{r \in S_i} d(p_i, r)^2.$$

Thus we obtain

$$\begin{aligned} C_{\mathrm{CUOPT}}(S) &= \sum_{i=1}^{k} \sum_{r \in S_i} d(q_i, r)^2 \\ &\leq 2 \sum_{i=1}^{k} \sum_{r \in S_i} d(p_i, r)^2 \\ &= 2C_{\mathrm{UOPT}}(S). \quad \square \end{aligned}$$

Combining these two lemmas, we have the following theorem for large d.

Theorem 2.5 *For n points in any fixed dimensions d, there is an approximate algorithm for the min-sum k-clustering problem with sum-of-squared error such that its time complexity is $O(n^{k+1})$ and the sum-of-squared error of an approximate k-clustering is at most twice as much as that of an optimum k-clustering.*

3 Concluding Remarks

As discussed in the introduction, as a criterion for the color image quantization, the sum-of-squared error and the all-pairs sum-of-squared error are more suitable than the diameter and the radius of point set. Using these criteria, two types of clustering problems, min-sum and minimax, have been considered. As far as the time complexity is concerned, the min-sum k-clustering with sum-of-squared error is easier to solve than the others. It is left open which of the criteria and types is most suited to the color quantization. Some computational experiments will be executed for comparing them.

Acknowledgment

This work was partially supported in part by the Grant-in-Aid by the Ministry of Education, Science and Culture of Japan.

References

[1] A. Aggarwal, H. Imai, N. Katoh and S. Suri, Finding k points with minimum diameter and related problems, *Proceedings of the 5th Annual ACM Symposium on Computational Geometry*, 1989, pp.283–291.

[2] H. C. Andrews, *Introduction to Mathematical Techniques in Pattern Recognition*, Wiley-Interscience, New York, 1972.

[3] T. Asano, B. Bhattacharya, M. Keil and F. Yao, Clustering algorithms based on minimum and maximum spanning trees, *Proceedings of the 4th Annual Symposium on Computational Geometry*, Urbana, 1988, pp.252-257.

[4] E. Boros and P.L. Hammer, On clustering problems with connected optima in Euclidean spaces, Discrete Mathematics 75 (1989), pp.81–88.

[5] H. Edelsbrunner and L. J. Guibas, Topologically sweeping an arrangement, *Proceedings of the 18th ACM Annual Symposium on Theory of Computing*, 1986, pp.389–403.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, CA, 1979.

[7] P. Hansen and B. Jaumard, Minimum sum of diameters clustering, *Journal of Classification*, (1987), 215-226.

[8] S. Hasegawa, A study on ϵ-net and ϵ-approximation, Master's Thesis, Department of Information Science, University of Tokyo, 1993.

[9] P. Heckbert, Color image quantization frame buffer display, *ACM Transactions on Computer Graphics*, Vol.16, No.3, (1982), pp.297–304.

[10] E.L. Lawler, Fast approximation algorithms for knapsack problems, *Mathematics of Operations Research*, 4 (1979), pp.339–356.

[11] C. Monma and S. Suri, Partitioning points and graphs to minimize the maximum or the sum of diameters, *Proceedings of the 6th International Conference on Theory and Applications of Graphs*, 1988.

[12] B. K. Natarajan, *Machine Learning*, Morgan Kaufmann, 1991.

[13] C.H. Papadimitriou and K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, New Jersey, 1982.

[14] S. Sahni, Approximate algorithms for the 0/1 knapsack problem, *Journal of the Association for Computing Machinery*, 22 (1975), pp.115–124.

[15] S. J. Wan, S. K. M. Wong, and P. Prusinkiewicz, An algorithm for multidimensional data clustering, *ACM Transactions on Mathematical Software*, 14, 2, (1988), pp.153–162.