数理解析研究所講究録
第871巻 1994年 124-129

Randomized Algorithms for Variance-Based k-Clustering

Mary Inaba*, Naoki Katoh† and Hiroshi Imai*

稲葉真理、加藤直樹、今井浩

Abstract

In this paper we consider the k-clustering problem for a set S of n points in the d-dimensional space with minsum sum of squared errors as clustering criteria, which is motivated from a problem, called color quantization problem, of computing a color lookup table for frame buffer display. Using the technique of computational geometry and random sampling, we present an efficient randomized algorithm which, roughly speaking, finds an ϵ-approximate 2-clustering in $O(n(1/\epsilon)^d)$ time.

1 Introduction

Clustering is the grouping of similar objects and a clustering of a set is a partition of its elements that is chosen to minimize some measure of dissimilarity. It is very fundamental and used in various fields in computer science such as pattern recognition, learning theory, image processing and computer graphics. There are various kinds of measure of dissimilarity, called criteria, in compliance with the problem. In this paper, we investigate the clustering problem suited for the color quantization problem.

Definition of the k-clustering problem: The general k-clustering problem can be defined as follows. A k-clustering is a partition of the given set S of n points $p_i = (\boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ in the d-dimensional space into k disjoint nonempty subsets $S_1, \ldots, S_k$, called clusters. A k-clustering is measured by the following two criteria.

(Intra-cluster criterion) For each cluster S_j, the measure (or error) $\mathrm{Intra}(S_j)$ of S_j, representing how good the cluster S_j is, is defined appropriately by applications. Typical intra-cluster criteria are the diameter, radius, variance, and sum of squared errors, namely variance multiplied by $|S_j|$ and sometimes called variance-based, of point set S_j.

(Inter-cluster criterion) The inter-cluster criterion defines the total cost of the k-clustering, which is a function of $\mathrm{Intra}(S_j)$ $(j = 1, \ldots, k)$ and is denoted by $\mathrm{Inter}(y_1, y_2, \ldots, y_k)$ where $y_j = \mathrm{Intra}(S_j)$. Typical function forms are $\max\{y_j \mid j = 1, \ldots, k\}$ and $\sum_{i=1}^{k} y_k$.

Then, the k-clustering problem is to find a k-clustering which minimizes the inter-cluster criterion:

$$\min\{\,\mathrm{Inter}(\mathrm{Intra}(S_1), \ldots, \mathrm{Intra}(S_k)) \mid k\text{-clustering } (S_1, \ldots, S_k) \text{ of } S\,\}$$

Previous results concerning diameter and radius: In computational geometry, many results have been obtained for the clustering problem. The diameter and radius problems are

*Department of Information Science, University of Tokyo, Tokyo 113, Japan
†Department of Management Science, Kobe University of Commerce, Kobe 651-21, Japan

rather well studied. They include an $O(n \log n)$-time algorithm for finding a 2-clustering of n points in the plane which minimizes the maximum diameter (Asano, Bhattacharya, Keil and Yao [1]), an $O(n^2 \log^2 n)$-time algorithm for finding a 3-clustering of planar point set which minimizes the mxhaximum diameter (Hagauer and Rote [3]), and an $O(n \log^2 n / \log \log n)$-time algorithm for finding a 2-clustering which minimizes the sum of the two diameters (Hershberger [6]). When k is regarded as a variable, most k-clustering problems become NP-hard (e.g., see Megiddo and Supowit [7]). For fixed k, the k-clustering problem using the diameter and radius as the intra-cluster criterion and a monotone function, including taking the maximum and the summation, as the inter-cluster criterion can be solved in a polynomial time (Capoyleas, Rote and Woeginger [2]).

Motivation for the variance-based clustering: In this paper, we consider the k-clustering problem with variance-based measures as an intra-cluster criterion. This is motivated from the color quantization problem of computing a color lookup table for frame buffer display. Typical color quantization problems cluster hundreds of thousands of points in the RGB three-dimensional space into $k = 256$ clusters. Since k is large, a top-down approach to recursively divide the point set into 2 clusters is mostly employed. In this problem, the diameter and radius are not suited as an intra-cluster criterion, and sum of squared errors criterion, sometimes called variance-based, (Wan, Wong and Prusinkiewicz [8]) and L_1-based (median cut; Heckbert [5]) criterion are often used. In [8], [5], the top-down approach is used and in solving the 2-clustering problem both only treat separating planes orthogonal to some coordinate axis. These algorithms are implemented in `rlequant` of Utah Raster Toolkit, and `ppmquant` of X11R5 or `tiffmedian` of Tiff Soft. Although these implementations run rather fast in practice, roughly speaking in $O(n \log n)$time, there is no theoretical guarantee about how good their solution k-clusterings are.

Rigorous definition of the variance-based clustering: Therefore, it is required to develop a fast 2-clustering algorithm and to determine the complexity of the k-clustering problem for the variance-based case. Before describing the existing computational-geometric results concerning variance-based case, let us define the variance-based intra-clutser criterion in a rigorous way. The variance $\mathrm{Var}(S)$ of S of points $p_i = (\boldsymbol{x}_i)$ in S is defined by

$$\mathrm{Var}(S) = \frac{1}{|S|} \sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S)\|^2$$

where

$$\bar{\boldsymbol{x}}(S) = \frac{1}{|S|} \sum_{p_i \in S} \boldsymbol{x}_i.$$

The sum of squared errors $\mathrm{Error}(S)$ with respect to the centroid of S is defined by

$$\mathrm{Error}(S) = \sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S)\|^2.$$

Previous results on the variance-based clustering: For the variance-based criteria, unlike the diameter and radius, the k-clustering problem adopting the maximum function as the inter-cluster criterion becomes hard to solve (Hasegawa, Imai, Inaba, Katoh and Nakano [4]). Also, in applications such as the color quantization problem, the summation function is adopted as an inter-cluster criterion [8]. In this paper, we consider only the summation case, that is, the k-clustering problem to minimize the summation of variance-based intra-cluster costs among clusters.

For the variance-based clustering problem with the summation function as an inter-cluster metric, it is known that an optimum 2-clustering is linearly separable and that an optimum k-clustering is induced by the Voronoi diagram generated by k points (e.g., see [4, 8]). Using this

characterization together with standard computational-geometric techniques, the 2-clustering problem can be solved in $O(n^2)$ time and $O(n)$ space, and the k-clustering problem is solvable in a polynomial time when k is fixed [4].

Our results: To develop a practically useful 2-clustering algorithm with the most typical intra-cluster criterion of the sum of squared errors, we present an efficient randomized algorithm which, roughly speaking, finds an ϵ-approximate 2-clustering in $O(n(1/\epsilon)^d)$ time, which is quite practical and can be used to real large-scale problems such as the color quantization problem. This randomized algorithm can be easily generalized to the k-clustering problem.

2 Randomized algorithms for the case of the sum of squared errors

It has been shown that the k-clustering problem for fixed k can be solved in $O(n^{dk})$time, which is polynomial in n. But its degree is large even for moderate values of d and k, and even for $k = 3, 4, 5$, its polynomial degree is quite high, which makes it less interesting to implement the algorithms for practical problems such as the color quantization problem. The k-clustering problem is NP-complete in general when k is regarded as a variable, and in this respect the results are best possible we may expect to have.

To develop a practically useful algorithm, utilizing randomization may be a good candidate, since the intra-cluster metric we are using has its intrinsic statistical meanings. In this section, we develop randomized algorithms for the k-clustering problem.

Here we mainly consider the 2-clustering problem, but most of the following discussions carry over to the k-clustering problem. First, let us consider how to estimate Error(S) for the set S of n points $p_i = (\boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ by random sampling. Let T be a set of m points obtained by m independent draws at random from S. If the original point set S are uniformly located, $(n/(m-1))$Error(T) may be a good estimate for Error(S). However, this is not necessarily the case. For example, suppose that a point p_i in S is far from the other $n-1$ points in S, and the other $n-1$ points are very close to one another. Then, Error(S) is nearly equal to the squared distance between p_i and a point in $S - \{p_i\}$, while with high probability Error(T) is almost zero. This indicates that Error(T) cannot necessarily provide a good estimate for Error(S).

On the other hand, the centroid $\bar{\boldsymbol{x}}(T)$ of T is close to the centroid $\bar{\boldsymbol{x}}(S)$ of S with high probability by the law of large numbers, and we obtain the following lemma.

Lemma 1 *With probability* $1 - \delta$,

$$\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2 < \frac{1}{\delta m}\mathrm{Var}(S).$$

Proof: First, observe that

$$\mathrm{E}(\bar{\boldsymbol{x}}(T)) = \bar{\boldsymbol{x}}(S), \qquad \mathrm{E}(\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2) = \frac{1}{m}\mathrm{Var}(S)$$

and then apply the Markov inequality to obtain the following.

$$\Pr\Big(\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2 > \frac{1}{\delta m}\mathrm{Var}(S)\Big) < \delta. \quad \square$$

Lemma 2 *With probability* $1 - \delta$,

$$\sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(T)\|^2 < (1 + \frac{1}{\delta m})\mathrm{Error}(S).$$

Proof: Immediate from Lemma 1 and the following.

$$\sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(T)\|^2 = \mathrm{Error}(S) + |S| \cdot \|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2. \quad \square$$

Thus, we can estimate Error(S) by random sampling. For the 2-clustering problem, we have to estimate Error(S_1) and Error(S_2) for a 2-clustering (S_1, S_2) by estimating the centroids of S_1 and S_2. Now, consider the following algorithm.

A randomized algorithm for the 2-clustering:

1. Sample a subset T of m points from S by m independent draws at random;
2. For every linearly separable 2-clustering (T_1, T_2) of T, execute the following:

   Compute the centroids t_1 and t_2 of T_1 and T_2, respectively;

   Find a 2-clustering (S_1, S_2) of S by dividing S by the perpendicular bisector of line segment connecting t_1 and t_2;

   Compute the value of Error(S_1)+Error(S_2) and maintain the minimum among these values;

3. Output the 2-clustering of S with minimum value above.

The idea of this randomized algorithm is to use all pairs of centroids of linearly separable 2-clusterings for the sampled point set T. Let (S_1^*, S_2^*) be an optimum 2-clustering of S, and let s_1^* and s_2^* be the centroids of S_1^* and S_2^*, respectively. By considering all linearly separable 2-clusterings for T, the algorithm handles the 2-clustering (T_1', T_2') obtained by dividing T by the perpendicular bisector of line segment connecting s_1^* and s_2^*. Then, from the centroids of T_1' and T_2', we obtain a 2-clustering (S_1', S_2') in the algorithm.

Since T is obtained from m independent draws,

$$\mathrm{E}(|T_j'|) = \frac{m}{n}|S_j^*| \quad (j = 1, 2).$$

From Lemma 2, Error(S_j^*) can be estimated by using $|T_j'|$. The sizes $|T_j'|$ $(j = 1, 2)$ are determined by independent Bernoulli trials, and is dependent on the ratio of $|S_1^*|$ and $|S_2^*|$. For the sampling number m, we say that S is $f(m)$-*balanced* if there exists an optimum 2-clustering (S_1^*, S_2^*) with

$$\frac{m}{n} \min\{\, |S_1^*|, |S_2^*| \,\} \geq f(m),$$

and the optimum 2-clustering is called an $f(m)$-balanced optimum 2-clustering. We then have the following.

Lemma 3 *Suppose there exists a* $(\log_e m)$*-balanced optimum 2-clustering* (S_1^*, S_2^*)*. Then, with probability* $1 - \dfrac{2}{m^{\beta^2/2}}$

$$\min\{\, |T_1^*|, |T_2^*| \,\} > (1 - \beta)\frac{m}{n} \min\{\, |S_1^*|, |S_2^*| \,\} \geq (1 - \beta) \log m.$$

Proof: Set $\mu' = \dfrac{m}{n} \min\{\, |S_1^*|, |S_2^*| \,\}$. For m independent Bernoulli trials $X_1, X_2, \ldots, X_m$ with $\Pr(X_i = 1) = \mu'/m \leq \Pr(X_i = 0) = 1 - \mu'/m$, the Chernoff bound implies, for $X = X_1 + \cdots + X_m$,

$$\Pr(X < (1 - \beta)\mu') < \exp(-\mu'\beta^2/2).$$

From the assumption,

$$\exp(-\mu'\beta^2/2) \leq \exp(-(\log m)\beta^2/2) = \frac{1}{m^{\beta^2/2}}. \quad \square$$

Theorem 1 *Suppose that the point set S is $f(m)$-balanced with $f(m) \geq \log m$. Then, the randomized algorithm finds a 2-clustering whose total value is within a factor of* $1 + \frac{1}{\delta(1-\beta)f(m)}$ *to the optimum value with probability* $1 - \delta - \frac{2}{m^{\beta^2/2}}$ *in* $O(nm^d)$ *time.*

Proof: From Lemmas 2 and 3, with probability $1 - \delta - \frac{2}{m^{\beta^2/2}}$,

$$\sum_{j=1}^{2} \sum_{p_i \in S'_j} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(T'_j)\|^2 \leq (1 + \frac{1}{\delta(1-\beta)f(m)}) \sum_{j=1}^{2} \mathrm{Error}(S^*_j)$$

holds. Furthermore, the left hand side is bounded from below by $\sum_{j=1}^{2} \mathrm{Error}(S'_j)$, whose value is computed in the algorithm. Hence, the minimum value found in the algorithm is within the factor.

Concerning the time complexity, all linearly separable 2-clusterings for T can be enumerated in $O(m^d)$ time. For each 2-clustering (T_1, T_2) of T, finding a pair of centroids and a 2-clustering of S generated by the pair together with its objective function value can be done in $O(n)$ time. Thus the theorem follows. $\square$

We have developed a randomized algorithm only for the 2-clustering problem so far, but this can be directly generalized to the k-clustering problem. If there exists a balanced optimum k-clustering, similar bounds can be obtained. It may be noted that the technique employed here has some connection with the technique used to obtain a deterministic approximate algorithm with worst-case ratio bounded by 2 for the k-clustering problem in [4].

The above theorem assumes some balancing condition. In some applications, a very small cluster is useless even if its intra-cluster is small. In such a case, the randomized algorithm naturally ignores such small-size cluster. Also, for the case of finding a good and balanced 2-clustering, such as the VLSI layout partition problem, we have only to apply the randomized algorithm directly. Restating the theorem for such cases, we have the following.

Theorem 2 *For the problem of finding an optimum 2-clustering among (γm)-balanced 2-clusterings for a constant γ, the randomized algorithm finds a 2-clustering which is almost at least (γm)-balanced and whose value is within a factor of* $1 + O(1/(\delta m))$ *to the optimum value of this problem with probability* $1 - \delta$ *for not so small δ.* $\square$

In the proof of this theorem, we use results concerning the ϵ-net and ϵ-approximations. On the other hand, if very small clusters with small intra-cluster metric should be found, we may enumerate such small clusters deterministically or in a randomized manner, since the number of such small clusters is relatively small. For the 2-clustering problem in the two-dimensional case, the number of linearly separable 2-clustering such that one cluster consists of at most k' points is $O(k'n)$ and can be enumerated efficiently. By enumerating k'-sets for an appropriate value of k', we obtain the following theorem.

Theorem 3 *The 2-clustering problem for n points in the plane with* Error *as the intra-cluster metric can be solved in* $O(n^{5/3}(\log n)^3)$ *time with the approximation ratio within a factor of* $1 + O(1/\log m)$ *with probability* $1 - O(1/\log m)$.

Proof: We set $m = n^{1/3} \log n$, and by the randomized algorithm find a good $(\log m)^2$-balanced 2-clustering and by the deterministic algorithm enumerating $(\leq n^{2/3} \log n)$-sets find a best unbalanced 2-clustering. Setting $\delta = 1/\log m$ and β to a constant, the time complexity of the randomized algorithm is $O(n(n^{1/3} \log n)^2) = O(n^{5/3}(\log n)^2)$ and the approximation ratio is bounded by $1 + O(1/\log m)$ with probability $1 - O(1/\log m)$. The deterministic algorithm runs in $O(n(n^{2/3} \log n)(\log n)^2) = O(n^{5/3}(\log n)^3)$ time. □

It should be noted that the time complexity in this theorem is subquadratic, compared with the deterministic quadratic exact algorithm.

References

[1] 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.

[2] V. Capoyleas, G. Rote and G. Woeginger: Geometric clustering. *Journal of Algorithms*, Vol.12 (1991), pp.341–356.

[3] J. Hagauer and G. Rote: Three-clustering of points in the plane. *Proceedings of the 1st Annual European Symposium on Algorithms (ESA '93)*, Lecture Notes in Computer Science, Vol.726, 1993, pp.192–199.

[4] S. Hasegawa, H. Imai, M. Inaba, N. Katoh and J. Nakano: Efficient algorithms for variance-based k-clustering. *Proceedings of the First Pacific Conference on Computer Graphics and Applications*, World Scientific, 1993, pp.75–89.

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

[6] J. Hershberger: Minimizing the sum of diameters efficiently. *Computational Geometry: Theory and Applications,* Vol.2 (1992), pp.111-118.

[7] N. Megiddo and K. J. Supowit: On the complexity of some common geometric location problems. *SIAM Journal on Computing*, Vol.13 (1984), pp.182–196.

[8] S. J. Wan, S. K. M. Wong and P. Prusinkiewicz: An algorithm for multidimensional data clustering. *ACM Transactions on Mathematical Software*, Vol.14, No.2 (1988), pp.153–162.