PAPER

Variance-Based k-Clustering Algorithms by Voronoi Diagrams and Randomization

Mary INABA[†], ***Nonmember***, **Naoki KATOH**[††], ***and*** **Hiroshi IMAI**[†], ***Regular Members***

SUMMARY In this paper we consider the k-clustering problem for a set S of n points $p_i = (\boldsymbol{x}_i)$ in the d-dimensional space with variance-based errors as clustering criteria, motivated from the color quantization problem of computing a color lookup table for frame buffer display. As the inter-cluster criterion to minimize, the sum of intra-cluster errors over every cluster is used, and as the intra-cluster criterion of a cluster S_j,

$$|S_j|^{\alpha-1} \sum_{p_i \in S_j} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S_j)\|^2$$

is considered, where $\|\cdot\|$ is the L_2 norm and $\bar{\boldsymbol{x}}(S_j)$ is the centroid of points in S_j, i.e., $(1/|S_j|)\sum_{p_i \in S_j} \boldsymbol{x}_i$. The cases of $\alpha = 1, 2$ correspond to the sum of squared errors and the all-pairs sum of squared errors, respectively. The k-clustering problem under the criterion with $\alpha = 1, 2$ are treated in a unified manner by characterizing the optimum solution to the k-clustering problem by the ordinary Euclidean Voronoi diagram and the weighted Voronoi diagram with both multiplicative and additive weights. With this framework, the problem is related to the generalized primary shatter function for the Voronoi diagrams. The primary shatter function is shown to be $O(n^{O(kd)})$, which implies that, for fixed k, this clustering problem can be solved in a polynomial time. For the problem with the most typical intra-cluster criterion of the sum of squared errors, we also 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 may be used to real large-scale problems such as the color quantization problem.

key words: *geometric clustering, Voronoi diagram, randomization*

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. Hence, this introduction first defines general clustering problems, summarizes existing results for the various types of clustering problems, and then proceeds to describing the variance-based clustering problem with its rigorous definitions and statements of our results in the following subsections.

Manuscript received August 4, 1999.
Manuscript revised January 24, 2000.

[†]The authors are with the Department of Information Science, University of Tokyo, Tokyo, 113–0033 Japan.
[††]The author is with the Department of Architecture and Architectural Systems, Kyoto University, Kyoto-shi, 606–8501 Japan.

1.1 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) 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, variance multiplied by $|S_j|$ (sum of squared errors) and variance multiplied by $|S_j|^2$ (all-pairs sum of squared errors) 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 Intra(S_j) $(j = 1, \ldots, k)$ and is denoted by Inter$(y_1, y_2, \ldots, y_k)$ where y_j = 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:

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

1.2 Previous Results Concerning Diameter and Radius

In computational geometry, many results have been obtained for the clustering problem. The diameter and radius problems are 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 maximum diameter (Hagauer and Rote [7]), 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 [12]). When k is regarded as a variable, most k-clustering problems become NP-hard (e.g.,

see Megiddo and Supowit [20], Feder and Greene [5]). 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 [4]).

There are also proposed approximate algorithms for the diameter and radius whose approximation ratio is theoretically guaranteed. Feder and Greene [5] gave optimal approximate algorithms whose running time is $O(n \log k)$ for n points in the d-dimensional space for fixed d, and whose worst-case ratio is 2. It should be noted here that this constant worst-case ratio might be seen as a more powerful method for clustering, but in the case the sum of squared errors has statistical meanings the diameter clustering does not necessarily guarantee producing a good clustering since the objective function to minimize is completely different from the statistical viewpoint.

1.3 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 the variance-based (Wan, Wong and Prusinkiewicz [22]) and L_1-based (median cut; Heckbert [11]) criteria are often used. In [11], [22], 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.

1.4 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-cluster criterion in a rigorous way. The variance $\mathrm{Var}(S)$ 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 $\bar{\boldsymbol{x}}(S)$ is the centroid of S:

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

For a parameter α, define $\mathrm{Var}^\alpha(S)$ by

$$\mathrm{Var}^\alpha(S) = |S|^\alpha \mathrm{Var}(S).$$

Var^0 is exactly the variance itself. Var^1 is represented as

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

and hence is the sum of squared errors with respect to the centroid of S. Var^2 is represented as

$$\begin{aligned}\mathrm{Var}^2(S) &= |S| \sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S)\|^2 \\ &= \sum_{p_i, p_l \in S,\ i<l} \|\boldsymbol{x}_i - \boldsymbol{x}_l\|^2\end{aligned}$$

and hence is the all-pairs sum of squared errors in S. Adopting Var^α as the intra-cluster metric, as α becomes larger, the sizes of clusters in an optimum k-clustering becomes more balanced.

1.5 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. For this inter-cluster criterion with the all-pairs sum of squared errors, only a pseudo-polynomial approximation scheme is known (Hasegawa, Imai, Inaba, Katoh and Nakano [9]). Also, in applications such as the color quantization problem, the summation function is adopted as an inter-cluster criterion [22]. 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 over clusters.

For the variance-based clustering problem with the summation function as an inter-cluster metric, the following are known. Concerning Var^1, the sum of squared errors, it is well 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 [3], [9], [22]). Using this characterization together with standard computational-geometric techniques, the 2-clustering problem with Var^1 as the intra-cluster metric 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 [9]. Concerning Var^2, the all-pairs sum of squared errors, an optimum 2-clustering is circularly separable (Boros and Hammer [3]), and a finer characterization by using the higher-order Voronoi diagram is given in [9]. Using this characterization, the 2-clustering problem with Var^2 as the intra-cluster metric can be solved in $O(n^{d+1})$ time, and also it is seen that the k-clustering problem for this case can be solved in a polynomial time $O(n^{(d+1)k(k-1)/2})$ when k is fixed [8].

There is also proposed an approximate algorithm for the k-clustering problem with Var^1 as an intra-cluster metric. Hasegawa, Imai, Inaba, Katoh and Nakano [9] gave an $O(n^{k+1})$-time algorithm for fixed d whose worst-case ratio is 2. This algorithm solves the k-clustering problem with constraining the representative point of each cluster to be one of points in the cluster.

For the k-clustering problem with Var^1 as an intra-cluster metric, the iterative improvement algorithm, known as the k-means algorithm [6], [21], is widely used. The approximation algorithms mentioned so far can be used to produce an initial good k-clustering, to which the k-means algorithm is applied, as was checked in [22].

1.6 Results of This Paper

In this paper, theoretical analyses on the k-clustering problem from the viewpoint of algorithmic complexity and approximation ratio are presented.

First, The k-clustering problem under the intra-cluster criterion Var^α with $\alpha = 1, 2$ is treated in a unified way by characterizing the optimum solution to the k-clustering problem by the ordinary Voronoi diagram and the weighted Voronoi diagrams with both multiplicative and additive weights.

With this framework, the problem is related to the generalized primary shatter function for the Voronoi diagrams, which is roughly the number of partitions of n points in the d-dimensional space induced by the Voronoi diagram generated by k generator points. The primary shatter function of the Euclidean Voronoi diagram is shown to be $O(n^{dk})$, and that for the Voronoi diagram with additive and multiplicative weights $O(n^{(d+2)k})$. Based on these, the k-clustering problem for n points in the d-dimensional space with a variance-based criterion can be solved in $O(n^{O(dk)})$ time. This greatly improves the previous bound $O(n^{O(dk^2)})$. We have thus given a polynomial-time algorithm for the case of fixed k, but its degree is large even for moderate values of d and k.

To cope with the problem with large k, it is often used to apply a 2-clustering algorithm recursively in a top-down fashion. To solve such 2-clustering problem for general d, we develop a practically useful approximation algorithm having some theoretical guarantee bounds. For the problem 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 may be used to real large-scale problems such as the color quantization problem. In the analysis, a fact that this intra-cluster cost has its statistical meanings by definition is used. This randomized algorithm can be easily generalized to the k-clustering problem. Some preliminary computational results are given in Inaba, Imai and Katoh [13], where results of applying the k-means algorithm for a computed k-clustering obtained by recursive application of this randomized 2-clustering algorithm are also reported. The connection of such an approach with a continuous clustering problem is mentioned in Inaba and Imai [15].

2. A Unified Approach to the Variance-Based k-Clustering by Weighted Voronoi Diagrams

The variance-based k-clustering problem is described as follows:

$$\min\left\{\sum_{j=1}^{k}\mathrm{Var}^\alpha(S_j) \mid k\text{-clustering}(S_1,\ldots,S_k) \text{ of } S\right\}$$

In [9], a parametric characterization was given for the case of $\alpha = 2$ (all-pairs case) by using a general parametric technique for minimizing quasiconcave functions developed by Katoh and Ibaraki [18], which enabled us to characterize an optimal 2-clustering for $\alpha = 2$ by means of higher-order Voronoi diagram, and to obtain a pseudo polynomial approximation scheme for the 2-clustering problem for Var^2 and the maximum function as the inter-cluster metric.

In this paper, we concentrate on the case where the summation function is adopted for the inter-cluster criterion, and give a more direct characterization for the problem with $\alpha = 1, 2$.

We may make use of partitions of n points induced by weighted Voronoi diagrams. Consider k points $q_j = (\boldsymbol{\mu}_j)$ in the d-dimensional space with multiplicative weight ν_j and additive weight σ_j $(j = 1,\ldots,k)$. Define the Voronoi region $\mathrm{Vor}(q_j)$ of q_j by

$$\mathrm{Vor}(q_j) = \bigcap_{l=1}^{k}\{p = (\boldsymbol{x}) \mid \nu_j\|\boldsymbol{x} - \boldsymbol{\mu}_j\|^2 + \sigma_j \le \nu_l\|\boldsymbol{x} - \boldsymbol{\mu}_l\|^2 + \sigma_l\}$$

For any point in $\mathrm{Vor}(q_j)$, q_j is the closest point among q_l $(l = 1,\ldots,k)$ with respect to the weighted distance. $\mathrm{Vor}(q_j)$ $(j = 1,\ldots,k)$ partitions the space, which is called the *weighted Voronoi diagram* generated by these k points q_j. When $\sigma_j = 0$ and $\nu_j = 1$ $(j = 1,\ldots,k)$,

this weighted Voronoi diagram reduces to the ordinary Euclidean Voronoi diagram.

By the Voronoi diagram generated by these k weighted points, n points in the given set S are naturally partitioned into k clusters (we here ignore the case in this definition where a point in S is equidistant from two points among these k weighted points). We call this partition a *Voronoi partition* of n points in S by k weighted generators. Apparently, not all k-clusterings are Voronoi partitions. In fact, we can characterize optimal k-clusterings by the Voronoi partition. The case of $\alpha = 1$ is well known, and its characterization is stated as follows (see [3], [9], [22]).

Theorem 1: Suppose that $(S_1^*, \ldots, S_k^*)$ is an optimum k-clustering for the k-clustering problem with Var^1 ($\alpha = 1$) as the intra-cluster metric. Then, an optimum k-clustering is a Voronoi partition by the ordinary Euclidean Voronoi diagram for k points $q_j = (\boldsymbol{\mu}_j^*)$ $(\boldsymbol{\mu}_j^* = \bar{\boldsymbol{x}}(S_j^*))$. □

Now, we prove the following theorem for the case of Var^2, i.e., all-pairs sum of squared errors.

Theorem 2: Suppose that $(S_1^*, \ldots, S_k^*)$ is an optimum k-clustering for the k-clustering problem with Var^2 ($\alpha = 2$) as the intra-cluster metric. Then, an optimum k-clustering is a Voronoi partition by the weighted Voronoi diagram for k points $q_j = (\boldsymbol{\mu}_j^*)$ $(\boldsymbol{\mu}_j^* = \bar{\boldsymbol{x}}(S_j^*))$ with multiplicative weight $\nu_j^* = |S_j^*|$ and additive weight σ_j defined by $\sigma_j = \sum_{p_i \in S_j^*} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S_j^*)\|^2$.

Proof: First, observe the following relation.

$$\begin{aligned}
&\sum_{p_i \in S_j} \|\boldsymbol{x}_i - \boldsymbol{x}\|^2 \\
&\quad = \sum_{p_i \in S_j} \|(\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S_j)) + (\bar{\boldsymbol{x}}(S_j) - \boldsymbol{x})\|^2 \\
&\quad = |S_j| \cdot \|\boldsymbol{x} - \bar{\boldsymbol{x}}(S_j)\|^2 + \sum_{p_i \in S_j} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S_j)\|^2
\end{aligned}$$

where it should be noted that $\sum_{p_i \in S_j} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}(S_j)) = 0$.

Now, suppose that, in the weighted Voronoi diagram above, a point $p_i \in S_j^*$ is not contained in $\mathrm{Vor}(q_j)$ for $q_j = (\bar{\boldsymbol{x}}(S_j^*))$, and is in $\mathrm{Vor}(q_{j'})$. Then, moving p_i from S_j^* to $S_{j'}^*$, the total cost is strictly reduced from the above formula (note that Var^2 is the all-pairs sum of squared errors), which contradicts the optimality of $(S_1^*, \ldots, S_k^*)$. Hence, each $p_j \in S_j^*$ is contained in $\mathrm{Vor}(q_j)$ for $j = 1, \ldots, k$, and the theorem follows. □

By Theorem 1 and Theorem 2, the variance-based k-clustering problem with $\alpha = 1, 2$ can be solved by enumerating all the Voronoi partitions of n points generated by k weighted points, and finding a partition with minimum one.

The number of distinct Voronoi partitions has strong connection with the generalized primary shatter function for k-label space introduced by Hasegawa [8], [10] which has applications in computational learning theory. For the rigorous definition of the generalized primary shatter function, we refer to Hasegawa, Imai and Ishiguro [10]. In this case, the corresponding generalized primary shatter function is the number of Voronoi partitions multiplied by $k!$, and hence, regarding k as a constant, these two are of the same order. Hasegawa [8] shows that this generalized primary shatter function is $O(n^{dk(k-1)/2})$. We here improve the bound on this number by showing that Voronoi partitions are duals of arrangements of algebraic surfaces in the $O(dk)$-dimensional space.

Theorem 3: The number of Voronoi partitions of n points by the Euclidean Voronoi diagram generated by k points in the d-dimensional space is $O(n^{dk})$.

Proof: In the ordinary Voronoi diagram, all multiplicative weights are one and all additive weights are zero, i.e., $\nu_j = 1$, $\sigma_j = 0$ $(j = 1, \ldots, k)$. Parameters are $\boldsymbol{\mu}_j$ $(j = 1, \ldots, k)$. Consider the (dk)-dimensional vector space consisting of $\boldsymbol{\mu}$ with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k)$. In this (dk)-dimensional space, we can define an equivalence relation among points such that two points are in the equivalence relation if their corresponding Voronoi partitions are identical. The equivalence relation produces a subdivision of this space into equivalence classes.

For each pair of distinct $\boldsymbol{\mu}_{j_1}$ and $\boldsymbol{\mu}_{j_2}$ among $\boldsymbol{\mu}_j$ $(j = 1, \ldots, k)$ and each point $p_i = (\boldsymbol{x}_i)$ among p_i $(i = 1, \ldots, n)$, consider an algebraic surface in this dk-dimensional space defined by

$$\|\boldsymbol{x}_i - \boldsymbol{\mu}_{j_1}\|^2 - \|\boldsymbol{x}_i - \boldsymbol{\mu}_{j_2}\|^2 = 0$$

where $\boldsymbol{x}_i$ is regarded as a constant vector. The number of such surfaces is $nk(k-1)/2$. The arrangement of these $nk(k-1)/2$ algebraic surfaces coincides with the subdivision defined by the equivalence relation from Voronoi partitions. The number of Voronoi partitions is bounded by the combinatorial complexity of the arrangement of $nk(k-1)/2$ constant-degree algebraic surfaces, which is bounded by $O(n^{dk})$ (e.g. see Warren [23]). Hence, the theorem follows. □

A further detailed analysis about the primary shatter function of the k-Voronoi space is done by Ishiguro [17] and further to the case of generalized Voronoi diagrams by divergence in Inaba and Imai [16]. In the papers, the linearization technique is applied in the analysis, and an algorithm using hyperplane arrangements is given based on it. Using algorithms to construct the hyperplane arrangement, the Voronoi partition can be enumerated in $O(n^{dk+k-d-2})$ time [16].

A similar analysis yields the following theorem, whose proof is omitted.

Theorem 4: The number of Voronoi partitions of n points by the weighted Voronoi diagram generated by k

weighted points with a multiplicative weight and an additive weight in the d-dimensional space is $O(n^{(d+2)k})$.

The weighted case can also be algorithmically solved via the hyperplane arrangement algorithm in $O(n^{(d+2)k-1})$ time [16].

It should be noted that, from the linear separability or circular separability, of an optimum 2-clustering for Var^1 and Var^2, respectively, as shown by [3], [9], the k-clustering problem for Var^1 and Var^2 is readily seen to be solvable in $O(n^{dk(k-1)/2})$ and $O(n^{(d+1)k(k-1)/2})$ time, respectively, using similar arguments in [4] for the diameter and radius. Only with the linear/circular separability for the 2-clustering, an algorithm of order $n^{O(dk^2)}$ may be best possible for the k-clustering problem. Our algorithms run in $O(n^{O(dk)})$ time, and improve the $O(n^{O(dk^2)})$ bound greatly. This becomes possible by the fine characterization of optimum k-clusterings by the weighted Voronoi diagram, and by evaluating the primary shatter function of the weighted Voronoi partitions in a tighter manner.

3. Randomized Algorithms for the Case of the Sum of Squared Errors

The results in the previous section are interesting from the theoretical viewpoint, and the time complexity is polynomial when k is considered as a constant. However, 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 with Var^1, the sum of squared error, as the intra-cluster metric.

In this extended abstract, we mainly consider the 2-clustering problem with Var^1, but most of the following discussions carry over to the k-clustering problem. First, let us consider how to estimate $\mathrm{Var}^1(S)$ 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, $(n/(m-1))\mathrm{Var}^1(T)$ may be a good estimate for $\mathrm{Var}^1(S)$. 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, $\mathrm{Var}^1(S)$ is nearly equal to the squared distance between p_i and a point in $S - \{p_i\}$, while with high probability $\mathrm{Var}^1(T)$ is almost zero. This indicates that $\mathrm{Var}^1(T)$ cannot necessarily provide a good estimate for $\mathrm{Var}^1(S)$.

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 $1-\delta$,

$$\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2 < \frac{1}{\delta m}\mathrm{Var}^0(S).$$

Proof: First, observe that

$$\mathrm{E}(\bar{\boldsymbol{x}}(T)) = \bar{\boldsymbol{x}}(S),$$
$$\mathrm{E}(\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2) = \frac{1}{m}\mathrm{Var}^0(S)$$

and then apply the Markov inequality to obtain the following.

$$\Pr\left(\|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2 > \frac{1}{\delta m}\mathrm{Var}^0(S)\right) < \delta. \qquad \square$$

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

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

Proof: Immediate from Lemma 1 and the following.

$$\sum_{p_i \in S} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(T)\|^2 = \mathrm{Var}^1(S) + |S| \cdot \|\bar{\boldsymbol{x}}(T) - \bar{\boldsymbol{x}}(S)\|^2. \qquad \square$$

Thus, we can estimate $\mathrm{Var}^1(S)$ by random sampling. For the 2-clustering problem, we have to estimate $\mathrm{Var}^1(S_1)$ and $\mathrm{Var}^1(S_2)$ 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 $\mathrm{Var}^1(S_1)+\mathrm{Var}^1(S_2)$ 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 for Var^1, 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, $\mathrm{Var}^1(S_j^*)$ 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 $(\log_{\mathrm{e}} m)$-balanced optimum 2-clustering (S_1^*, S_2^*). Then, with probability $1 - \dfrac{2}{m^{\beta^2/2}}$ for a constant β $(0 < \beta < 1)$, the following holds.

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

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 5: 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 $1 + \dfrac{1}{\delta(1-\beta)f(m)}$ to the optimum value with probability $1 - \delta - \dfrac{2}{m^{\beta^2/2}}$ in $O(nm^d)$ time.

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

$$\begin{aligned}&\sum_{j=1}^{2}\sum_{p_i \in S_j'} \left\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}(T_j')\right\|^2\\ &\quad\leq \left(1 + \frac{1}{\delta(1-\beta)f(m)}\right)\sum_{j=1}^{2}\mathrm{Var}^1(S_j^*)\end{aligned}$$

holds. Furthermore, the left hand side is bounded from below by $\sum_{j=1}^{2}\mathrm{Var}^1(S_j')$, 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 [9].

The above theorem assumes some balancing condition. In some applications, a very small cluster is useless even if its intra-cluster measure is small. For example, when we apply a 2-clustering algorithm recursively in a top-down fashion to solve the k-clustering problem, a balancing condition on 2-clusterings may be imposed to 2-clustering subproblems so that the sizes of subproblems may become small quickly and the total clustering may have nicer properties. 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, we have only to apply a slightly modified version of the randomized algorithm directly. This is typical for the clustering problem in VLSI layout design. See, for example, Kernighan and Lin [19]. Generalizing Theorem 5 for such cases is partially discussed in [14].

4. Concluding Remarks

We have demonstrated that optimum solutions to the variance-based k-clustering can be characterized by the (weighted) Voronoi diagram generated by k points, and have evaluated the primary shatter function of the k-Voronoi space. This primary shatter function can be used in computational learning theory in learning k-Voronoi spaces.

We have then presented a simple randomized algorithm for the k-clustering problem with Var^1 as an intra-cluster metric. This algorithm is practically useful when k is small and balanced k-clusterings are preferable. For example, for the problem of finding an optimum 2-clustering for n planar points among almost completely balanced 2-clusterings, an approximate 2-clustering which is approximately balanced and whose cost is within a factor of $1 + 1/3 = 4/3$ on the average to the optimum cost can be found by sampling 10 points

from n points and spending $O(10^2 n) = O(n)$ time with probability $\sum_{i=3}^{7}\binom{10}{i}/2^{10} \approx 0.89$, or by sampling 20 points and spending $O(20^2 n) = O(n)$ time with probability $\sum_{i=3}^{17}\binom{20}{i}/2^{20} \approx 0.9996$.

The randomized algorithm, however, is not so suitable to find a good unbalanced 2-clustering. Also, although the randomized algorithm itself is valid for large k, the running time becomes inherently large since the primary shatter function for m sampled points is $O(m^{O(dk)})$. To solve the variance-based k-clustering for large k practically, say for $k = 256$ of the typical color quantization problem, we may apply the randomized 2-clustering algorithm proposed in this paper recursively in a top-down manner with sampling only a small number of points at each stage as mentioned above.

Preliminary computational reports in the two-dimensional case are given in [13], [15]. It is observed that a small set of sample points provide rather good solutions, and also that the recursive application of the 2-clustering algorithm to obtain a good k-clustering performs well. Further experiments on the higher-dimensional spaces and other cases are required.

Acknowledgment

The author would like to thank referees for their helpful comments. This paper is a refined version of the paper [14]. The work of the authors was supported in part by the Grant-in-Aid of Ministry of Education, Science, Sports and Culture of Japan.

References

[1] T. Asano, B. Bhattacharya, M. Keil, and F. Yao, "Clustering algorithms based on minimum and maximum spanning trees," Proc. 4th Annual Symp. on Computational Geometry, pp.252–257, Urbana, 1988.
[2] P. Auer, R.C. Holte, and W. Maass, "Theory and applications of agnostic PAC-learning with small decision trees," Proc. 12th Int. Conf. on Machine Learning, pp.21–29, Morgan Kaufmann, 1995.
[3] E. Boros and P.L. Hammer, "On clustering problems with connected optima in Euclidean spaces," Discrete Mathematics, vol.75, pp.81–88, 1989.
[4] V. Capoyleas, G. Rote, and G. Woeginger, "Geometric clustering," J. Algorithms, vol.12, pp.341–356, 1991.
[5] T. Feder and D.H. Greene, "Optimal algorithms for approximate clustering," Proc. 20th Annual ACM Symp. on Theory of Computing, pp.434–444, 1988.
[6] A. Gersho and R.M. Gray, Vector Quantization and Signal Compression, Kluwer Academic, 1992.
[7] J. Hagauer and G. Rote, "Three-clustering of points in the plane," Computational Geometry: Theory and Applications, vol.8, no.2, pp.87–95, 1997.
[8] S. Hasegawa, A Study on ϵ-Net and ϵ-Approximation, Master's Thesis, Department of Information Science, University of Tokyo, 1993.
[9] S. Hasegawa, H. Imai, M. Inaba, N. Katoh, and J. Nakano, "Efficient algorithms for variance-based k-clustering," Proc. 1st Pacific Conf. on Computer Graphics and Applications, World Scientific, pp.75–89, 1993.
[10] S. Hasegawa, H. Imai, and M. Ishiguro, "ϵ-approximations of k-label spaces," Theoretical Computer Science, vol.137, pp.145–175, 1995.
[11] P. Heckbert, "Color image quantization frame buffer display," ACM Trans. on Computer Graphics, vol.16, no.3, pp.297–304, 1982.
[12] J. Hershberger, "Minimizing the sum of diameters efficiently," Computational Geometry: Theory and Applications, vol.2, pp.111–118, 1992.
[13] M. Inaba, H. Imai, and N. Katoh, "Experimental results of randomized clustering algorithm," Proc. 12th Annual ACM Symp. on Computational Geometry, pp.C1–C2, 1996.
[14] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," Proc. 10th ACM Symp. on Computational Geometry, pp.332–339, 1994.
[15] H. Imai and M. Inaba, "Geometric clustering with applications," Zeitschrift für Angewandte Mathematik und Mechanik (ZAMM), vol.76, Suppl., pp.183–186, 1996.
[16] M. Inaba and H. Imai, "The number of partitions of n points induced by the Voronoi diagram via the conjugacy generated by k points," "Proc. 1st Japanese-Hungarian Symp. on Discrete Mathematics and Its Applications, pp.83–90, Kyoto, March 1999.
[17] M. Ishiguro, Evaluation of Combinatorial Complexity for Hypothesis Spaces in Learning Theory with Applications, Master's Thesis, Department of Information Science, University of Tokyo, 1994.
[18] N. Katoh and T. Ibaraki, "A parametric characterization and an ϵ-approximation scheme for the minimization of a quasiconcave program," Discrete Applied Mathematics, vol.17, pp.39–66, 1987.
[19] B.W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," The Bell System Technical J., vol.49, no.2, pp.291–307, 1970.
[20] N. Megiddo and K.J. Supowit, "On the complexity of some common geometric location problems," SIAM J. Computing, vol.13, pp.182–196, 1984.
[21] B. Mirkin, "Mathematical classification and clustering," Nonconvex Optimization and Its Applications Series, vol.11, Kluwer Academic Publishers, Boston, 1996.
[22] S.J. Wan, S.K.M. Wong, and P. Prusinkiewicz, "An algorithm for multidimensional data clustering," ACM Trans. Mathematical Software, vol.14, no.2, pp.153–162, 1988.
[23] H.E. Warren, "Lower bounds for approximation by nonlinear manifolds," Amer. Math. Soc., vol.133, pp.167–178, 1968.

Mary Inaba obtained B. Eng. in Architecture, and M.Sc. and D.Sc. in information Science, University of Tokyo in 1984, 1995 and 1999, respectively. She was a research associate in 1996–1999, and has been a lecturer since 1999 at Faculty of Science, University of Tokyo. Her research interests include computational geometry, data mining, and networking. She is a member of IPSJ.

Naoki Katoh received the B.Eng, M.Eng. and Dr. Eng. degrees in Applied Mathematics and Physics from Kyoto University in 1973, 1975 and 1981, respectively. He was an Assistant Professor during 1981–1982, an Associate Professor during 1982–1990, and a Professor during 1990–1997, respectively, at Department of Management Science of Kobe University of Commerce. In 1997 he joined Kyoto University where he is currently a Professor at Department of Architecture and Architectural Systems. His research interests include the design and analysis of combinatorial and geometric algorithms, data mining and architectural information systems. He is a member of IPSJ, OR Soc. Japan, Japan SIAM, and ACM.

Hiroshi Imai obtained B.Eng. in Mathematical Engineering, and M.Eng. and D.Eng. in Information Engineering, University of Tokyo in 1981, 1983 and 1986, respectively. In 1986–1990, he was an associate professor of Department of Computer Science and Communication Engineering, Kyushu University. Since 1990, he has been an associate professor at Department of Information Science, University of Tokyo. His research interests include algorithms, computational geometry, and optimization. He is a member of IPSJ, OR Soc. Japan, JSIAM, ACM and IEEE.