INVITED PAPER *Special Issue on Surveys on Discovery Science*

Divergence-Based Geometric Clustering and Its Underlying Discrete Proximity Structures

Hiroshi IMAI[†], ***Member and*** **Mary INABA**[†], ***Nonmember***

SUMMARY This paper surveys recent progress in the investigation of the underlying discrete proximity structures of geometric clustering with respect to the divergence in information geometry. Geometric clustering with respect to the divergence provides powerful unsupervised learning algorithms, and can be applied to classifying and obtaining generalizations of complex objects represented in the feature space. The proximity relation, defined by the Voronoi diagram by the divergence, plays an important role in the design and analysis of such algorithms.

key words: *unsupervised learning, geometric clustering, Voronoi diagram, computational geometry, information geometry*

1. Introduction

Clustering is a powerful tool in unsupervised learning. Clustering is the grouping of similar objects, from which generalizations of each cluster formed by such similar objects are obtained. When a learned clustering of observed objects is at hand and a new object is given as a query, the problem of answering which class the given query object belongs to arises. This problem may be regarded as a kind of the nearest neighbor query against the clusters.

For clustering, the definition of similarity among objects is crucial. In many applications, especially those concerned with multimedia objects, objects have multiple attributes, say d attributes, whose values are integers or real numbers. Then each of such objects can be modeled as a point in the d-dimensional space in a direct manner. This space is called a feature space. This enables us to treat these objects in a geometric setting, to which many fertile properties of geometry together with efficient geometric algorithms can be applied. In this space, the similarity of objects is represented by their dissimilarity, which is a kind of 'distance' of points in the space. Throughout this paper we often use the term 'distance' as a measure of dissimilarity in an informal way and use the term 'metric' to denote the distance satisfying the distance axiom in the mathematical sense when necessary. Again, the definition of such a distance of points in the feature space becomes crucial.

To represent the distance of points in the space, we often use the Euclidean distance as the most simple form for the general distance. The Euclidean distance is a metric. However, the Euclidean distance is never a unique choice. In fact, in some cases, it may not be invariant with respect to natural transformations which the target objects admit. Especially, for objects arising from stochastic phenomena, statistical and information-theoretic measures are meaningful.

Information geometry has been proposed as a theoretically sound model representing objects having stochastic and statistical properties by Amari [1], Amari and Nagaoka [2] (see also [10]). In their work, differential-geometric properties of information geometry has been clarified. The divergence is naturally introduced as a measure representing the distance of points in the space, thus information geometry can be a basis for learning problems mentioned above. The divergence in information geometry is a generalization of the squared Euclidean distance in the Euclidean space and the Kullback-Leibler divergence for the exponential family of probability distributions. Information geometry has been applied to learning problems with stochastic nature, but its combinatorial structures have not yet been understood well compared with differential-geometric structures of the space.

Recently, Onishi, Imai [11]–[13], Inaba, Imai and Sadakane [7], [9] shed light on the the most fundamental discrete proximity relations, represented by the Voronoi diagrams, in an information-geometric space. They develop a combinatorial and algorithmic approach to the space of information geometry. It has been shown that the proximity structures induced by the divergence in information geometry is combinatorially quite similar to those in the Euclidean space. This indicates that the space of information geometry has almost the same combinatorial complexity, and hence can be computationally handled in a similar way. Furthermore, geometric clustering becomes more natural in information geometry, and, using the combinatorial proximity properties, the computational complexity of geometric clustering in information geometry can be discussed.

From the viewpoint of clustering algorithms, these proximity structures can be used to identify the intrinsic computational complexity of such unsupervised-learning algorithms. Specifically, divergence-based clustering has strong connection with learning the mixture model. In fact, for the exponential family, which is

Manuscript received September 12, 1999.
Manuscript revised September 20, 1999.

[†]The authors are with the Department of Information Science, the University of Tokyo, Tokyo, 113–0033 Japan.

the most fundamental class of probability distributions including normal distributions, multinomial distributions, etc., the divergence-based clustering corresponds to the classification likelihood method by Celeux and Govaert [3]. In learning the mixture model, there have been proposed many algorithms such as EM algorithms, but their computational complexity has not yet been analyzed well, and the approach in [7], [9], [12], [13] clarifies the combinatorial complexity of underlying discrete structures.

This paper surveys such geometric-clustering approaches in the space of information geometry. First, examples of the feature space are described for texts and images. We then describe a well-known example to reveal how the definition of the distance is important. Furthermore, an additional problem of statistical clustering is described to show the discrete structure of the clustering. Then, the Voronoi diagram by the divergence in information geometry is introduced. Results on this Voronoi diagram are explained, with emphasis on its connection with geometric clustering.

2. Examples of the Feature Space

Before going into details of the geometric structures of the feature space, it is advisable to have an intuitive understanding of the feature space. We here select the feature spaces of texts and images as examples, and try to describe the importance of geometric structures.

In the full text databases, a geometric approach, called the vector-space method, has been developed for advanced information retrieval of full text databases (Salton et al. [16], [17]). Roughly speaking, this method first fixes a set of d terms, and maps each text to a point in the d-dimensional space such that the value in the i-th coordinate is the frequency of term i. In image databases, it is also natural to adopt geometric approaches in querying images by their content, as discussed in [4]. By counting the frequencies of d colors, each image is mapped to a point in the k-dimensional space, called the feature space of images.

These are now described in more detail with commonly used distances in such application fields.

2.1 Vector Space Model for Texts

In the vector-space model, all information items of stored texts are represented by vectors, or points, of terms, or keywords, in the space whose dimension is the number of terms. A term is typically a word. In automatic processing of various texts, the terms are derived directly from the texts under consideration.

Since all the terms do not equally represent the contents of texts, it is important to use a term-weighting system which assigns high weights to terms deemed important and lower weights to the less important terms. There are many term-weighting systems, and a typical one described in [16] is given by the equation $f_t \times 1/f_c$ (term frequency times inverse collection frequency), which favors terms with a high frequency (f_t) in particular documents but with a low frequency overall in the collection (f_c). General nouns appear frequently everywhere, and hence their weights are low, while technical nouns insensitively appear in some specific places, and may have relatively higher weights.

Then, all texts and text queries are represented by weighted term vectors $\boldsymbol{t}_i$ in the d-dimensional space where d is the number of terms. Of course, the l-th element in $\boldsymbol{t}_i$ is the weight assigned to the l-th term in the document i. The similarity $\mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ between two vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ of two given documents is defined by

$$\mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \frac{\boldsymbol{t}_i^{\mathrm{T}} \boldsymbol{t}_j}{\|\boldsymbol{t}_i\| \|\boldsymbol{t}_j\|} = \cos\theta$$

where $\| \, \|$ denotes the L_2 norm, and θ is the angle between two vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$. The similarity value ranges from 0 (low similarity) to 1 (high similarity).

Here, instead of similarity, we define dissimilarity $\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j)$ between two vectors $\boldsymbol{t}_i$ and $\boldsymbol{t}_j$ to be

$$\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j) = 1 - \mathrm{sim}(\boldsymbol{t}_i, \boldsymbol{t}_j).$$

Geometrically, text vector $\boldsymbol{t}_i$ is mapped to a normalized vector $\tilde{\boldsymbol{t}}_i$ by the projection on the sphere with radius 1, then, the dissimilarity between two texts $\boldsymbol{t}_i, \boldsymbol{t}_j$ can be considered as the squared Euclidean distance between these normalized vector $\tilde{t}_i$, $\tilde{t}_j$. That is,

$$\mathrm{dis}(\boldsymbol{t}_i, \boldsymbol{t}_j) = \frac{1}{2} \|\tilde{\boldsymbol{t}}_i - \tilde{\boldsymbol{t}}_j\|^2.$$

The squared Euclidean distance is a special case of the divergence in information geometry, and, we may impose information-geometric structure on a manifold consisting of normalized vectors.

We may adopt another normalization such that each frequency vector is divided by its total frequency. Then, for the normalized vector, the sum of coordinate values becomes one. This corresponds to using the hyperplane $x_1 + x_2 + \cdots + x_d = 1$, instead of sphere $x_1^2 + x_2^2 + \cdots + x_d^2 = 1$ as above, in the normalization process.

With this normalization, we may regard this space as that of parameters of multinomial (or, discrete) distributions. That is, the l-th term with the normalized vector $(x_1, \ldots, x_d)$ is considered to appear with probability x_l. This would be the simplest probabilistic model for texts, and would be too naive to capture complex structures of texts. Yet, it provides an approximation based on the theoretical model. In this case, information geometry tells us that the distance from a point to another should be measured by the Kullback-Leibler divergence.

2.2 Quadratic Form Model for Images

In the Query-by-Image-Content (QBIC) system [4] the

following measure is used to estimate the distance of two images by their color frequency vectors. First, find $\tilde{d}$ representative colors for the original color images, and quantize the images by using the $\tilde{d}$ colors. Let $B = (b_{ij})$ be a matrix of order $\tilde{d}$ such that b_{ij} is the negative of some dissimilarity between color i and color j. Let $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ be the normalized frequency vectors of two images with respect to the $\tilde{d}$ colors, where normalization is done by dividing the frequency vector by the total count so that the sum becomes 1. Then, the distance-like function between two images is defined to be $\mathrm{dis} = \Delta\boldsymbol{\xi}^{\mathrm{T}} B \Delta\boldsymbol{\xi}$ with $\Delta\boldsymbol{\xi} = \boldsymbol{\xi}_1 - \boldsymbol{\xi}_2$.

When b_{ij} is defined to be the negative of squared Euclidean distance between two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the d-dimensional space, i.e., $b_{ij} = -\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$, which is the case in the original framework of the QBIC system based on the *Luv* color system with $\tilde{d} = 64$ or 256 and $d = 3$, then the distance dis is expressed as

$$\mathrm{dis} = 2\|A\Delta\boldsymbol{\xi}\|^2$$

where A is the matrix whose i-th column is $\boldsymbol{x}_i$ (Inaba [7]). Hence, simply applying the singular value decomposition of A, we can reduce the clustering problem by the distance dis to the geometric clustering problem with our objective function in the d-dimensional space. Thus, the distances used in this case are all represented as the squared Euclidean distances, which can be modeled as the divergence in information geometry.

3. The Definition of the Distance Matters

3.1 Case of Normal Distributions

The following example of normal distributions is a well-known to understand the Euclidean distance is not necessarily a unique choice to measure the distance. The probability density function of one-dimensional normal distribution is given by

$$p(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

where μ is the mean and σ is the standard deviation. A normal distribution can be identified with a point (μ,σ) in the (μ,σ)-plane, more specifically, in the (μ,σ) upper half plane since $\sigma > 0$ (concerning the proximity relations in the upper half plane, see [14]). Then, consider four normal distributions $p(x;1,1)$, $p(x;1,3)$, $p(x;-1,1)$, and $p(x;-1,3)$. In the upper half plane, these correspond to $(1,1)$, $(1,3)$, $(-1,1)$, and $(-1,3)$, respectively. See Fig. 1.

In the upper half plane, these four points form a square, and the Euclidean distance between $(1,1)$ and $(-1,1)$ and that between $(1,3)$ and $(-1,3)$ are the same. However, as is seen from the density functions in the figure, $p(x;1,1)$ and $p(x;-1,1)$ are much easier to distinguish from each other than $p(x;1,3)$ and $(x;-1,3)$.

In fact, we can take another pair of coordinates, such as (μ,σ^2), $(\mu,\mu^2+\sigma^2)$, rather freely instead of the pair (μ,σ), and the Euclidean distance between two points in such a coordinate changes. We should adopt a measure which is statistically meaningful and invariant under such transformations, which has been studied from the viewpoint of differential geometry and extended in the framework of information geometry.

3.2 Finding a Cluster in the Mixture Case

Consider a probability distribution $p(x;\xi)$ with a probability variable (vector) x parameterized by a parameter (vector) ξ. Suppose that there are k distributions $p(x;\xi_j)$ $(j = 1,\ldots,k)$, and an observation $\tilde{x}$ is drawn by first choosing $p(x;\xi_j)$ with probability q_j among the k distributions ($\sum_{j=1}^k q_j = 1$, $q_j \geq 0$), and then from the chosen distribution. An observation x is drawn by the following probability density function:

$$\sum_{j=1}^{k} q_j p(x;\xi_j).$$

This is called the mixture distribution. We will return to the so-called mixture clustering later in this paper, and here we consider the problem of, for an observation $\tilde{x}$ drawn from the mixture, finding a distribution among the k given distributions from which the observation is originally drawn. The posterior probability that $\tilde{x}$ belongs to a distribution $p(x;\xi_{\tilde{j}})$ is given by

$$\frac{q_{\tilde{j}}p(\tilde{x};\xi_{\tilde{j}})}{\sum_{j=1}^k q_j p(\tilde{x};\xi_j)}.$$

This may be regarded as a fuzzy membership function. When it is required to identify one most likely distribution for $\tilde{x}$, a distribution $p(x;\xi_{\tilde{j}})$ attaining the following maximum,

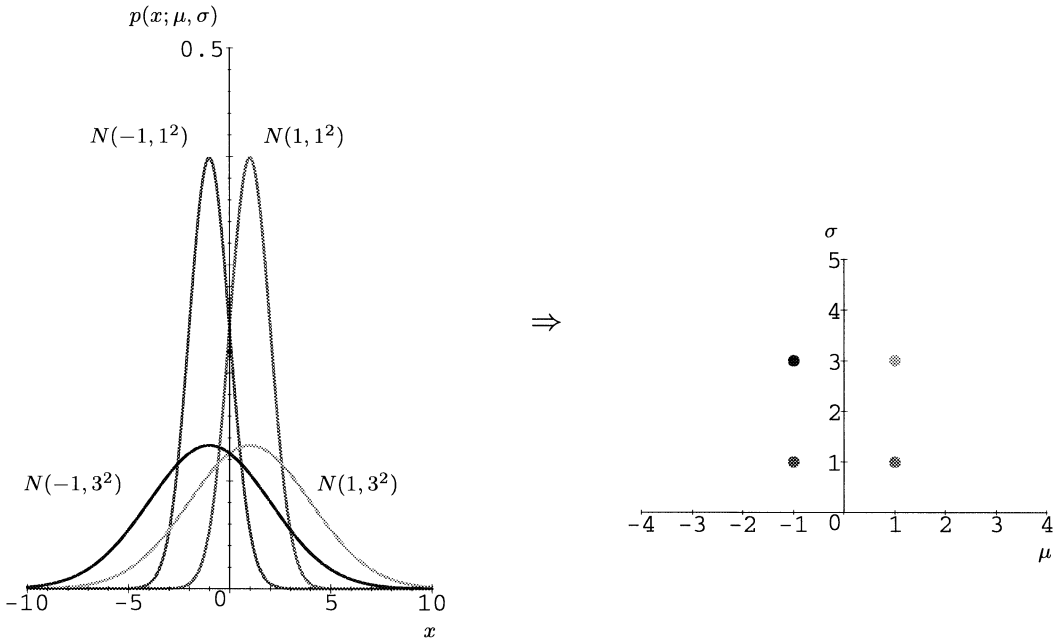
$$\max_{j=1}^{k} q_j p(\tilde{x};\xi_j)$$

is selected in most studies. Then, the domain of probability variable vector x is partitioned into the territory V_j of each $p(x;\xi_j)$ $(j = 1,\ldots,k)$:

$$V_j = \prod_{\tilde{j}\neq j}\{x \mid q_j p(\tilde{x};\xi_j) > q_{\tilde{j}}p(\tilde{x};\xi_{\tilde{j}})\}$$

The partition of the domain by V_j $(j = 1,\ldots,k)$ is nothing but a generalized Voronoi diagram based on the distance function $p(x;\xi_j)$ (for the definition of ordinary and some generalized Voronoi diagrams, see [5]). This diagram is directly obtained from the likelihood function.

In the sequel, we describe the Voronoi diagram in the space of parameters of distributions with respect to



(a) (b) (μ,σ)-plane

Fig. 1 Four normal distributions.

the divergence. For the exponential family, the Voronoi diagram in the manifold of observed data, discussed in this subsection, and the Voronoi diagram in the space of distribution parameters are both the projections of the same upper envelope of hyperplanes to respective spaces. This relation will be discussed elsewhere.

4. Statistical Manifolds of Probability Distributions

A set of parameterized probability distributions form a Riemannian manifold $\mathcal{M}$ by their d parameters. For example, a class of one-dimensional normal distribution with mean μ and standard deviation σ form a manifold $\mathcal{M} = \{[\mu,\sigma] \mid \sigma > 0\}$, the upper half plane. This section describes fundamental properties of this manifold for a wide and well-behaved class of probability distributions, called the exponential family. Since we will use two dual coordinates, θ-coordinate and η-coordinate, which generalizes the polarity with respect to a paraboloid, we will use the tensor notation.

4.1 Exponential Family

A probability distribution parameterized by $\theta = [\theta^i]$ belongs to the exponential family if its probability density function $f(x;\theta)$ with probability variable (vector) x is expressed as

$$f(x;\theta) = \exp[C(x) + \sum_i \theta^i F_i(x) - \psi(\theta)].$$

Since $\int f(x;\theta)\mathrm{d}x = 1$, ψ is given by

$$\psi(\theta) = \log \int \exp[C(x) + \sum_i \theta^i F_i(x)]\mathrm{d}x$$

For this $\theta = [\theta^i]$, we define $\eta = [\eta_i]$ by

$$\eta_i = \int F_i(x) f(x;\theta)\mathrm{d}x.$$

θ and η are two coordinate systems on the manifold $\mathcal{M}$ of parameters of the distributions in the exponential family. η is also given by

$$\eta_i = \frac{\partial\psi(\theta)}{\partial\theta^i}$$

In the case of the exponential family, the dual potential function $\varphi(\eta)$ is defined in the η-coordinate system by

$$\varphi(\eta) = \int f(x;\theta)(\log f(x;\theta) - C(x))\mathrm{d}x$$

where θ in the right-hand side is that corresponding to η in the left-hand side. Note that when $C(x) \equiv 0$, this potential function φ becomes the minus of entropy of distribution,

$$\varphi(\theta) = \int f(x;\theta)\log f(x;\theta)dx = -H(f_\theta).$$

θ is then given by

$$\theta^i = \frac{\partial\varphi}{\eta_i}.$$

In fact, $\theta = \theta(p)$ and $\eta = \eta(p)$ give two coordinate systems on the manifold $\mathcal{M}$ of points p.

The exponential family includes many fundamental probability distributions, such as the normal distribution, Poisson distribution, exponential distribution and finite discrete distribution.

4.2 Properties of the Divergence

Statistical manifold $\mathcal{M}$ of the exponential family has very good properties as a dually flat space [2]. We consider the θ-coordinate and the η-coordinate of the manifold $\mathcal{M}$ for the exponential family. $\theta(p)$ and $\eta(p)$ denote the θ- and η-coordinate values for a point p on $\mathcal{M}$, that is, $\theta(p) = [\theta^1(p), \ldots, \theta^d(p)]$, and $\eta(p) = [\eta_1(p), \ldots, \eta_d(p)]$. Then, the *divergence* between two points p and q on $\mathcal{M}$ is defined as follows.

Definition 1 (Divergence): Consider the two potential functions $\psi, \varphi : \mathcal{M} \to \mathbf{R}$ for the exponential family. For two points $p, q \in \mathcal{M}$, define the divergence $D(p\|q)$ by

$$D(p\|q) = \psi(p) + \varphi(q) - \sum_i \theta^i(p)\eta_i(q)$$

The pair of potential functions are connected via the Legendre transformation, that is,

$$\theta^i = \frac{\partial \varphi}{\partial \eta_i}, \qquad \eta_i = \frac{\partial \psi}{\partial \theta^i}$$

ψ, φ are strictly convex, and

$$\varphi(q) = \max_{p \in S}\{\sum_i \theta^i(p)\eta_i(q) - \psi(p)\}$$

$$\psi(p) = \max_{q \in S}\{\sum_i \theta^i(p)\eta_i(q) - \varphi(q)\}$$

Hence, $D(p\|q) \geq 0$, and $D(p\|q) = 0$ iff $p = q$.

$$D(p\|p) = \psi(p) + \varphi(p) - \sum_i \theta^i(p)\eta_i(p) = 0$$

But, unlike the metric, $D(p\|q) \neq D(q\|p)$, in general.

Next, we consider the relation of $D(p\|q)$ with the potential function φ and a tangent hyperplane. Add a new coordinate z, corresponding to the height, to the η-coordinate system, and consider the graph $z = \varphi$ in the $[\eta, z]$-space. For $p \in \mathcal{M}$, lift it up to the graph $(\eta_1(p), \eta_2(p), \ldots, \eta_d(p), \varphi(p))$, and consider the tangent hyperplane

$$\begin{aligned} z - \varphi(p) &= \textstyle\sum_i \frac{\partial \varphi}{\partial \eta_i}(p)(\eta_i - \eta_i(p)) \\ &= \textstyle\sum_i \theta^i(p)(\eta_i - \eta_i(p)) \end{aligned}$$

Then, for a point $q \in \mathcal{M}$, the height difference of a point lifted to the graph $z = \varphi(\eta)$

$$(\eta_1(q), \eta_2(q), \ldots, \eta_d(q), \varphi(q))$$

to a point lifted to the above tangent hyperplane

$$(\eta_1(q), \ldots, \eta_d(q), \sum_i \theta^i(p)(\eta_i(q) - \eta_i(p)) + \varphi(p))$$

is given by

$$\begin{aligned} &\varphi(q) - \sum_i \theta^i(p)\eta_i(q) + \sum_i \theta^i(p)\eta_i(p) - \varphi(p) \\ &\quad = \psi(p) + \varphi(q) - \sum_i \theta^i(p)\eta_i(q) = D(p\|q) \end{aligned}$$

By the duality of the definition of divergence, this linearization technique can be also applied in the θ-coordinate system; namely, the divergence $D(p\|q)$ is also the difference of the height at the point p between the potential function ψ and tangent hyperplane on ψ on the point q in the θ-coordinate system.

The divergence has such a nice and natural meaning, which was used to analyze the ∇^*-Voronoi diagram as will be stated in Theorem 1.

Example 1 (Euclidean case): This corresponds to a self-dual case: $\psi = \varphi = \sum_{i=1}^d x_i^2/2$ and $\theta^i = \eta_i = x_i$. The divergence is a half of the square of the Euclidean distance.

Example 2 (Exponential family): For this family, the divergence coincides with the Kullback-Leibler divergence $D_K(q\|p)$, also known as the relative entropy, as follows:

$$D(p\|q) = D_K(q\|p)$$

In the case of the finite discrete distributions p and q such that $(\xi_1(p), \ldots, \xi_d(p))$ and $(\xi_1(q), \ldots, \xi_d(q))$ are the parameters for p and q,

$$D_K(q\|p) = \sum_{i=0}^d \xi_i(q) \log \frac{\xi_i(q)}{\xi_i(p)}$$

where $\xi_0(p) = 1 - \sum_{i=1}^d \xi_i(p)$ and $\xi_0(q) = 1 - \sum_{i=1}^d \xi_i(q)$. Thus, this dually flat structure is an extension of the ordinary Euclidean case, and the divergence is an extension of the squared Euclidean distance.

4.3 Maximum Likelihood Method, Minimizing the Sum of Divergences and Centroid in η-Coordinate

For a parameterized probability distribution $f(x; \theta)$, suppose we are given a set S_x of n observations $\{x^{(1)}, \ldots, x^{(n)}\}$. For these data, the likelihood function is defined as

$$L(\theta) = \prod_{l=1}^n f(x^{(l)}; \theta)$$

and the maximum likelihood method finds θ that maximizes $L(\theta)$.

For the exponential family, we can consider the log likelihood. Let $l(x^{(l)}; \theta) = \log f(x^{(l)}; \theta)$, and then $L(\theta)$

is maximized when

$$\hat{L}(\theta) = \sum_{l=1}^{n} l(x^{(l)};\theta) = \sum_{l=1}^{n} C(x^{(l)}) + \sum_{i} \theta^i F_i(x^{(l)}) - \psi(\theta).$$

By partial differentiation by θ^i

$$\sum_{l=1}^{n} F_i(x^{(l)}) - \eta_i(\theta)$$

$L(\theta)$ is maximized when $\eta_i(\theta) = \frac{1}{n}\sum_{l=1}^{n} F_i(x^{(l)})$. Recall the definition $\eta_i \equiv \int F_i(x) f(x;\theta)\mathrm{d}x$, the maximum likelihood estimator is nothing but the centroid of the manifold $\mathcal{M}$ in the η-coordinate system. Consequently, given a set S_p of n probability distribution $\{p^{(1)},\ldots,p^{(n)}\}$, the centroid of the set S_p in the η-coordinate system also becomes a maximum likelihood estimator of the whole distribution.

On the manifold $\mathcal{M}$, the distance between two distributions are measured by the divergence. Suppose, given a set S_p of n probability distribution in the exponential family, $\{p^{(1)},\ldots,p^{(n)}\}$, the centroid of the set S_p in the η- coordinate system is $\eta_i(\overline{p}) = \frac{1}{n}\sum_{j=1}^{n}\eta_i(p^{(j)})$. The sum of divergences is expressed as

$$\sum_{l=1}^{n} D(p\|p^{(l)}) = nD(p\|\overline{p}) + \sum_{l=1}^{n} D(\overline{p}\|p^{(l)})$$

Since the divergence of two identical points is 0 and the divergence of two distinct points is positive, it is seen that the sum of divergences is achieved only at the centroid of points corresponding to $p^{(1)},\ldots,p^{(n)}$ in the η-coordinate, thus having strong connection with the maximum likelihood estimator.

4.4 Information Theoretic Interpretation

The divergence is not a metric, and it does not satisfy even the symmetric property, i.e., in general $D(p\|q) \neq D(q\|p)$. This also implies that the θ-coordinate and the η-coordinate have different properties.

In the above discussion, we describe that the centroid of observations in the η-coordinate is the maximum likelihood estimator, and hence is meaningful. In this subsection, we briefly describe this point from the standpoint of information theory.

The Kullback-Leibler divergence $D_K(p^{(l)}\|p)$ represents the average redundancy when the probability distribution $p^{(l)}$ is expressed by p. This is explained as follows. The Kullback-Leibler divergence is expanded as

$$D_K(p^{(l)}\|p) = \sum \eta_i(p^{(l)}) \log \frac{\eta_i(p^{(l)})}{\eta_i(p)} = \sum \eta_i(p^{(l)}) \log \frac{1}{\eta_i(p)} - \sum \eta_i(p^{(l)}) \log \frac{1}{\eta_i(p^{(l)})}$$

The first term corresponds to the code length of encoding, by the approximate probability p, codes originally generated by $p^{(l)}$, and the second term corresponds to the optimal code length of encoding, by the original probability $p^{(l)}$, codes generated by $p^{(l)}$.

Hence, minimizing $\sum D_K(\eta^{(l)}\|\theta)$ corresponds to minimizing the sum of redundancies when expressing n probability distributions by a distribution among the same parameterized family of distributions.

5. Voronoi Diagrams by Divergence

5.1 ∇^*-Voronoi Diagrams by Divergence

The Voronoi diagram by the divergence is investigated in [12], [13], which is defined bas follows.

Definition 2 (∇^*-Voronoi diagram): For k generator points $r^{(j)}$ $(j = 1,\ldots,k)$, the ∇^*-Voronoi diagram consists of Voronoi regions $V(r^{(j)})$ defined as follows in [13].

$$V(r^{(j)}) = \bigcap_{j'\neq j} \{p \mid D(p^{(j)}\|p) < D(p^{(j')}\|p)\}$$

See Fig. 2 for the case of normal distributions.

For the ∇^*-Voronoi diagram, the following holds.

Theorem 1 (Onishi, Imai [13]): The ∇^*-Voronoi diagram can be obtained as the projection to the manifold $\mathcal{M}$ of the upper envelope of hyperplanes which are tangent hyperplanes in the $[\eta, z]$-coordinate of the graph $z = \varphi(p)$ at $[\eta(p), \varphi(p)]$.

By this theorem, the combinatorial complexity of the ∇^*-Voronoi diagram can be bounded by the upper bound theorem for convex polytopes.

6. Clustering by Divergence

For a given set S of n points $p^{(l)}$ $(l = 1,\ldots,n)$ on the manifold $\mathcal{M}$, a k-clustering is a partition of S into nonempty k disjoint subsets $S_1,\ldots,S_k$ whose union is S.

Problem 1 (Divergence-sum clustering):

$$\min_{r^{(j)},\ S_j\ (j=1,\ldots,k)} \sum_{j=1}^{k} \sum_{p^{(l)}\in S_j} D(r^{(j)}\|p^{(l)})$$

Here, $r^{(j)}$ is a representative point for S_j, and, since the sum of divergence is minimized at the centroid, $r^{(j)}$ is simply set to the centroid of S_j in the η-coordinate.

This clustering criterion corresponds to maximizing the Classification Maximum Likelihood (CLM) for the exponential family [3]. The following theorem establishes connection between optimal clustering and the underlying discrete proximity structures.

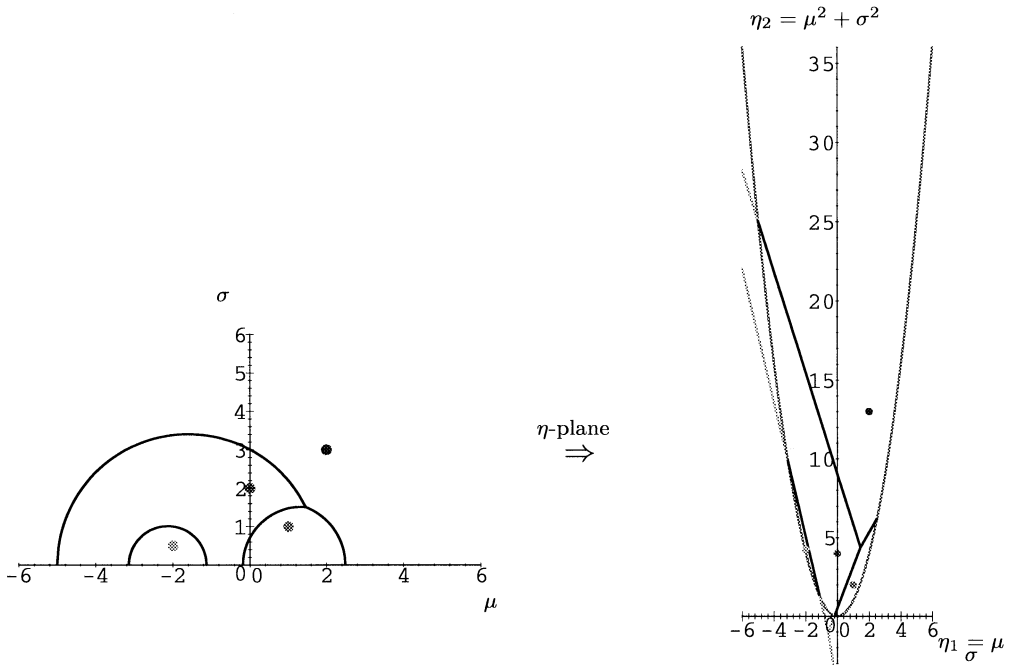


Fig. 2 Four normal distributions and their Voronoi diagram by the divergence in the (μ, σ) upper half plane (left) and the η-plane (right), where, in the η-plane, $\eta_1 = \mu$ and $\eta_2 = \mu^2 + \eta^2$ and the manifold is $\{(\eta_1, \eta_2) \mid \eta_2 > \eta_1^2\}$.

Theorem 2 (Inaba, Imai, Sadakane [9]): An optimal clustering for the divergence-sum clustering problem is identical with a partition by the ∇^*-Voronoi diagram generated by the centroids of clusters.

This kind of property was known only for the case of the sum of squared Euclidean distances. This theorem generalizes it to the divergence-sum case, and hence to the classification likelihood method.

This theorem considers clustering of points on the statistical manifold formed by parameters. For clustering observed data points, as described in Sect. 3.2, the Voronoi diagram with respect to the likelihood was defined. This diagram is the projection of the upper envelope in Theorem 1 onto the manifold formed by data points, and a similar theorem as above can be obtained.

7. Complexity of the Voronoi Partitions

By Theorem 2, k-clustering problem by divergence can be solved by enumerating all the partitions of n points induced by the corresponding Voronoi diagram generated by k points, and finding a partition with the minimum one. We call a partition of n points induced by such a Voronoi diagram a Voronoi partition.

The number of all possible Voronoi partitions by k generators corresponds to evaluation of the generalized primary shatter function for a label space induced by the Voronoi diagrams [6]. which has connection with the VC dimension. For the primary shatter function as well as the VC dimension, refer to [6]. That is, k generators are numbered from 0 to $k-1$, and, each of n points is labeled by the label of a generator whose Voronoi region contains the point. The generalized primary shatter function of this label space is the number of all possible partitions.

In this section, utilizing the dual structure between the η- and θ-coordinate system, we evaluate the number of all possible partitions $\pi_S(m)$ for the label space $S = (X, \mathcal{L})$ defined for the ∇^*-Voronoi diagram, where X is a set of infinite points on the d-dimensional statistical manifold, and $\mathcal{L}$ is a set of functions from X to $\{0, \ldots, k-1\}$. The function for the weighted ∇^*-Voronoi diagram can be evaluated in a similar way. We evaluate $\pi_S(m)$ by counting the number of cells of an arrangement of hyperplanes in the $(d+1)k$-dimensional representative space.

In the d-dimensional statistical manifold with a dually flat structure, and given k representative points for k-clustering, each of k points can be considered to move independently. Denote by R a set of k generator points $\{\boldsymbol{r}^{(1)}, \boldsymbol{r}^{(2)}, \ldots, \boldsymbol{r}^{(k)}\}$, and, denote by X a set of n observed points which are partitioned. We will consider two spaces, one is the dk-dimensional space of

$$(\theta^1(\boldsymbol{r}^{(1)}), \theta^2(\boldsymbol{r}^{(1)}), \ldots, \theta^d(\boldsymbol{r}^{(1)}), \ldots, \\ \theta^1(\boldsymbol{r}^{(k)}), \theta^2(\boldsymbol{r}^{(k)}), \ldots, \theta^d(\boldsymbol{r}^{(k)})),$$

which we call *representative space*, and the other is the $k(d+1)$-dimensional space,

$$(\theta^1(\boldsymbol{r}^{(1)}), \ldots, \theta^d(\boldsymbol{r}^{(1)}), \theta^{d+1}(\boldsymbol{r}^{(1)}) = \psi(\boldsymbol{r}^{(1)}), \\ \ldots, \\ \theta^1(\boldsymbol{r}^{(k)}), \ldots, \theta^d(\boldsymbol{r}^{(k)}), \theta^{d+1}(\boldsymbol{r}^{(k)}) = \psi(\boldsymbol{r}^{(k)})).$$

Definition 3: (Equivalence relationship with respect to partitioning) Suppose we are given a set X of n points, and sets R and R' of k generator points. If the partitioning $\{X_1, \ldots, X_k\}$ and $\{X'_1, \ldots, X'_k\}$ induced by R and R', respectively are identical, R and R' are in *equivalence relationship* concerning to partitioning.

By the definition of the ∇^*-Voronoi Diagram, this equivalence relationship changes only when $\exists \boldsymbol{x}, \boldsymbol{r}^{(j)}, \boldsymbol{r}^{(l)}$, the sign of $D(\boldsymbol{r}^{(j)}\|\boldsymbol{x}) - D(\boldsymbol{r}^{(l)}\|\boldsymbol{x})$ changes. Hence, consider a hypersurface in the dk-dimensional representative space:

$$D(\boldsymbol{r}^{(j)}\|\boldsymbol{x}) - D(\boldsymbol{r}^{(l)}\|\boldsymbol{x}) = \psi(\boldsymbol{r}^{(j)}) - \psi(\boldsymbol{r}^{(l)}) - \textstyle\sum_i (\theta^i(\boldsymbol{r}^{(j)}) - \theta^i(\boldsymbol{r}^{(l)})\eta_i(\boldsymbol{x})) = 0$$

This can be regarded as a hyperplane in the above-mentioned $(d+1)k$-dimensional space, and the total number of the hyperplanes is $n\binom{k}{2} = O(nk^2)$. Then, the problem is reduced to evaluating the number of cells of this hypersurface arrangements. Using the convexity of potential functions, and the linearization technique, we obtain the following.

Theorem 3 (Inaba, Imai, Sadakane [9]): The number of distinct partitions of n points induced by the ∇^*-Voronoi diagram generated by k points on $\mathcal{M}$ is bounded by $O(n^{(d+1)k})$.

Constructing this $(d+1)k$-dimensional hyperplane arrangement and its section as above, all the Voronoi partitions can be enumerated. Regarding d and k as constants, this yields a polynomial-time algorithm to solve our divergence-sum clustering problem.

8. Random Sampling Algorithm for 2-Clustering

If we regard k and d to be constant, the complexity of exact algorithm runs in polynomial time, but, even for small k and d, it becomes quite large. We extend an approximate algorithm for 2-clustering using random sampling technique to the divergence-sum problem, based on the algorithm in the Euclidean case [8].

The random sampling technique surely captures some outline of the point distribution, but it is not powerful enough to make the divergence sum relatively small with respect to the minimum value. That is, sampled data by themselves might not reflect the divergence sum of the whole data.

The algorithms in [8], [9] proceeds as follows. Sampled data can reflect the centroid with high probability, so, try all possible partitions on sampled data, compute the centroid using sampled data, then, compute cost function using the whole data and get the minimum one.

In the Euclidean case [8], the divergence is directly connected with the variance, the cost function, while in general cases it is not. Hence, as for analysis of approximation ratio we restrict ourselves to the case of finite discrete distribution.

[Randomized 2-clustering algorithm with divergence]

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 in the η-coordinate system, execute the following:

   Compute the centroids t_1 and t_2 of T_1 and T_2 in the η-coordinate system, respectively;
   Find a 2-clustering (S_1, S_2) of S by dividing S by the hyperplane with the same divergence between t_1 and t_2 in the η-coordinate system,
   Compute the value of $\mathrm{Cost}(S_1)+\mathrm{Cost}(S_2)$ and maintain the minimum among these values;

This randomized algorithm is an approximation algorithm, and its approximation ratio may be evaluated as follows. First, we consider the error of cost function for one cluster. Consider a set S of n points and its subset T randomly sampled from S. The absolute error for one cluster depends on how the estimated centroid $\overline{q}(T)$ is deviated from the centroid $\overline{q}(S)$ of S. Recall the following.

$$\begin{aligned}\sum_{l=1}^{n} D(\overline{q}(T)\|p^{(l)}) - \sum_{l=1}^{n} D(\overline{q}(S)\|p^{(l)}) &= nD(\overline{q}(T)\|\overline{q}(S)) \\ &= n\sum_{i=1}^{d} \overline{q}(S)_i \log \frac{\overline{q}(S)_i}{\overline{q}(T)_i}\end{aligned}$$

This can be bounded by using the Hoeffding inequality. We obtain the following theorem.

Theorem 4 (Inaba, Imai, Sadakane [9]): Suppose there is an optimal 2-clustering such that the sizes of clusters are within some constant factor to each other. Let D be the minimum among the averages of $D(\overline{q}(S_j)\|p^{(l)})$ for each cluster in the optimal clustering. Then, for some constant α' with $\alpha > \alpha' > 0$, the randomized algorithm finds a 2-clustering in $O(nm^d)$ time, whose sum of divergences is within a factor of $1+c$ with probability at least $1 - 4d\exp(-2\alpha'\left(1-\exp(-\frac{cD}{n}\right)^2 m)$.

When the divergence is the squared Euclidean distance, a tighter analysis can be done and the following holds.

Theorem 5 (Inaba, Katoh, Imai [8]): For the problem of finding an optimum 2-clustering, which is assumed to be moderately balanced in size, the randomized algorithm finds a 2-clustering whose value is within a factor of $1+O(1/(\delta m))$ to the optimum value of this problem with probability $1-\delta$ for arbitrary small δ.

Acknowledgment

The authors would like to thank anonymous referees for their helpful comments, which improve the presentation of this paper very much. Part of this work was supported by the Grant-in-Aid on Priority Areas, "Discovery Science," of the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] S. Amari, "Differential Geometrical Methods in Statistics," Lecture Notes in Statistics, vol.28, Springer-Verlag, New York, 1985.
[2] S. Amari and H. Nagaoka, "Method of Information Geometry," Iwanami Shoten, Tokyo, 1993.
[3] G. Celeux and G. Govaert, "A classification EM algorithm for clustering and two stochastic versions," Computational Statistics & Data Analysis, vol.14 (1992), pp.315–332.
[4] C. Faloutsos, R. Barber, M. Flickner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," Journal of Intelligent Information Systems, vol.3 (1994), pp.231–262.
[5] H. Edelsbrunner, Algorithms in Combinatorial Geometry, Springer-Verlag, 1987.
[6] S. Hasegawa, H. Imai, and M. Ishiguro, "ϵ-approximations of k-label spaces," Theoretical Computer Science, vol.137, pp.145–175, 1995.
[7] M. Inaba, Geometric Clustering on Feature Manifold, Doctoral Dissertation, The University of Tokyo, 1999.
[8] M. Inaba, N. Katoh, and H. Imai, "Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering," Proc. 10th ACM Symposium on Computational Geometry, pp.332–339, 1994.
[9] M. Inaba, H. Imai, and K. Sadakane, Geometric Clustering on the Statistical Manifold, Preprint, 1998.
[10] M.K. Murray and J.W. Rice, Differential Geometry and Statistics, Chapman & Hall, 1993.
[11] K. Onishi, Riemannian Computational Geometry — Convex Hull, Voronoi Diagram and Delaunay-type Triangulation, Doctoral Dissertation, The University of Tokyo, 1998.
[12] K. Onishi and H. Imai, "Voronoi diagrams for an exponential family of probability distributions in information geometry," Japan-Korea Joint Workshop on Algorithms and Computation, pp.1–8, Fukuoka, 1997.
[13] K. Onishi and H. Imai, Riemannian Computational Geometry — Voronoi Diagram and Delaunay-type Triangulation in Dually Flat Space, submitted for publication.
[14] K. Onishi and N. Takayama, "Construction of Voronoi diagram on the upper half-plane," IEICE Trans. Fundamentals, vol.E79-A, no.4, pp.533–539, 1996.
[15] K. Sadakane, H. Imai, K. Onishi, M. Inaba, F. Takeuchi, and K. Imai, "Voronoi Diagrams by Divergences with Additive Weights," Proc. 14th Annual ACM Symposium on Computational Geometry, pp.403–404, 1998.
[16] G. Salton, J. Allan, C. Buckley, and A. Singhal, "Automatic analysis, theme generation, and summarization of machine-readable texts," Science, vol.264, pp.1421–1426, 1994.
[17] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol.18, no.11, pp.613–620, 1975.

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. 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 IEICE, IPSJ, OR Soc. Japan, JSIAM, ACM and IEEE.

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. Since 1996, she has been a assistant professor at Faculty of Science, University of Tokyo. She is a member of IPSJ.