数理解析研究所講究録
1093 巻 1999 年 21-26

Computational Geometry on Statistical Manifolds for Clustering

(Extended Abstract)

Mary Inaba*, Hiroshi Imai* and Kunihiko Sadakane*

Abstract

This paper investigates computational-geometric aspects of clustering problems on the statistical manifolds in information geometry. This extends the traditional Euclidean computational geometry to a new one in the space of information geometry. Computational-geometric concepts such as Voronoi diagrams in the space of information geometry are described, and then some results on the geometric clustering problem based on the concepts are outlined.

1 Introduction

In information geometry [1, 2, 7], differential geometric properties of probabilistic distributions and other stochastic systems have been studied. A set of parametrized probability distributions form a Riemannian manifold by their parameters, and, the exponential family of probabilistic distribution is the most typical and well-behaved family. This family contains normal distribution, Poisson, finite discrete, and exponential distribution as special cases. Furthermore, this has a nice differential-geometric property, called "dually flat," and a divergence in general form, which is a distance-like function between two probability distributions and contains both Kullback-Leibler divergence and squared Euclidean distance as its special cases [1, 2]. The Voronoi diagram in the dually flat space is studied in [8, 9, 10].

In this paper, we extend our results for the Euclidean clustering problem [6] to clustering problems by divergence on statistical manifolds. Clustering problem is to group similar objects under some criteria, and, in general it is NP-hard. Geometric k-clustering problem is to find a good partition, called a k-clustering, 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 $S_1, \ldots, S_k$. We first introduce the weighted Voronoi diagram by divergence, and analyze its complexity. This generalized Voronoi diagram share nice properties with the Euclidean diagrams. Our Voronoi diagram by divergence is then applied to the general mixture clustering case. With this unified approach via the divergence, we could characterize an optimal clustering for the pure variance criterion in the Euclidean case which was left open in [6]. We propose a randomized 2-clustering approximation algorithm using random sampling technique, together with analyzing its approximation performance theoretically.

2 Statistical manifolds of probability distributions

For the statistical estimation, in a traditional form, first we assume an underlying distribution such as normal distribution or Poisson distribution, then, from a set of observed data, we estimate the parameters of the distribution, such as mean or deviation in the normal distribution case. In this sense, once distribution has been assumed, statistical estimation can be regarded as the estimation of parameter of the distribution. Here, we regard a statistical distribution characterized by d parameters, as a point in the d-dimensional parametric space, geometrically structures by the properties of the distributions.

*Department of Information Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

A set of parametrized probability distributions form a Riemannian manifold $\mathcal{M}$ by their 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.

2.1 Exponential family

A probability distribution parametrized 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. 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.

Examples of the exponential family are normal distributions, finite discrete distribution, exponential distribution, etc., see [1, 2].

2.2 Properties of the divergence

In the sequel, we adopt the Einstein's notation.

$$\theta^i \eta_i \equiv \sum_i \theta^i \eta_i$$

The statistical manifold of the exponential family has nice differential-geometric structure, called dually flat. In order to investigate the discrete proximity structure of this manifold, in this paper, it suffices to consider the θ-coordinate and the η-coordinate of the manifold $\mathcal{M}$ for the exponential family as two dual affine coordinate systems. $\theta(p)$ 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 $\eta(p) = [\eta_1(p), \ldots, \eta_d(p)]$. In the dually flat space, we can define a distance-like function *divergence* between two points p and q on $\mathcal{M}$.

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*

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

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

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

Hence, $D(p\|q) \geq 0$, and $D(p\|q) = 0$ iff $p = q$. But, unlike the distance, $D(p\|q) \neq D(q\|p)$ 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. Then, for a point $q \in \mathcal{M}$, the height difference of a point lifted to the graph $z = \varphi(\eta)$ is given by

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

By the symmetric duality, this holds also in θ-coordinate system, i.e., 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.

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

In the Euclidean case, which 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. For the exponential family, the divergence coincides with the Kullback-Leibler divergence $D_K(q\|p)$, also known as the relative entropy. Thus, this dually flat structure is an extension of the ordinary Euclidean case, and the divergence is an extension of the squared Euclidean distance.

Furthermore, the maximum likelihood method in statistical inference can be interpreted in a natural way in the η-coordinate system, say, taking the average or the orthogonal projection to obtain maximum likelihood estimators. For details, see [1, 2].

3 Weighted Voronoi diagrams by divergence

The Voronoi diagram by the divergence is investigated in [8, 9]. In extending the all-pair sum of squared Euclidean distances to the divergence case, multiplicatively and additively weighted Voronoi diagrams are useful. Hence, this section investigates such weighted diagrams. As will be seen, the weighted diagram has similar structures as the weighted Euclidean diagram [3], and this result may be viewed as an extension of [3].

We begin with a non-weighted case.

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 [9].*

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

$V(r^{(j)})$ $(j = 1, \ldots, k)$ partition the manifold, which is called the ∇^*-Voronoi diagram.

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

Theorem 1 (Onishi, Imai [9]) *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.

The weighted Voronoi diagram by the divergence is defined as follows.

Definition 3 (Weighted ∇^*-Voronoi diagram) *Suppose that, for each k points $r^{(j)}$, a multiplicative weight $w^{(j)}$ and an additive weight $\tilde{w}^{(j)}$ are given. The Voronoi region of point $r^{(j)}$ is given by*

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

The collection of $V(r^{(j)})$ $(j = 1, \ldots, k)$ is called the weighted ∇^*-Voronoi diagram, or simply the weighted Voronoi diagram.

Concerning the the combinatorial complexity of this weighted diagram, we can show that each Voronoi region has complexity $O(n^{\lfloor \frac{d+1}{2} \rfloor})$, and hence obtain the following.

Theorem 2 *The combinatorial complexity of the weighted divergence Voronoi diagram is $O(n^{\lfloor \frac{d+3}{2} \rfloor})$.*

4 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 [4].

Next, generalizing the weighted Euclidean case, we consider the following.

Problem 2 (Multiplicatively weighted divergence-sum clustering)

$$\min \sum_{j=1}^{k} w(|S_j|) \sum_{p^{(l)} \in S_j} D(r^{(j)} \| p^{(l)})$$

where $w(s) = s^\alpha$, and the following cases have their own meanings:

- $w(s) = 1$ (this case corresponds to the simple sum case)
- $w(s) = s$ (this case corresponds to the all-pair sum of squared Euclidean distances in the Euclidean case)
- $w(s) = 1/s$ (this case corresponds to the variance)

As in the first problem, $r^{(j)}$ for cluster S_j is simply set to the centroid of S_j in the η-coordinate.

We further extend the problem to a mixture case.

Problem 3 (Mixed divergence clustering) *In this clustering model, each point $p^{(l)}$ belongs to every cluster in some sense. Specifically, $\zeta(j,l)(\geq 0)$ denotes how much point p^l belongs to cluster S_j, where*

$$\sum_{j=1}^{k} \zeta(j,l) = 1 \text{ for each } l.$$

Then, the clustering problem is to find an optimal k-clustering for

$$\min \sum_{j=1}^{k} w\left(\sum_{l=1}^{n} \zeta(j,l)\right) \sum_{l=1}^{n} \zeta(j,l) D(\tilde{r}^{(j)} \| p^{(l)})$$

In this problem, $\tilde{r}^{(j)}$ is the weighted centroid defined by

$$\eta_i(\tilde{r}^{(j)}) = \frac{1}{\sum_{l=1}^{n} \zeta(j,l)} \sum_{l=1}^{n} \zeta(j,l)\eta_i(p^{(l)})$$

Now, we describe our results for these problems. First, we have the following.

Theorem 3 *An optimal clustering for the divergence-sum clustering problem is identical with a partition by the ∇^*-Voronoi diagram generated by the centroids of clusters.*

The number of partitions of n points induced by the *nabla**-Voronoi diagram is shown to be $O(n^{\min\{dk, dk-d+k-2\}})$ and furthermore such partitions can be enumerated in time proportional to the bound in [5]. This number is polynomial when d and k are regarded as constants, thus revealing the geometry helps in solving the clustering problem for moderate d and k.

Concerning the Problem 2, at this point, it seems hard to generalize a result for the Euclidean case in [6] to this general case, but its mixture version of Problem 3 can be again characterized by the weighted ∇^*-Voronoi diagram as follows. Furthermore, this characterization solves unsolved pure variance problem in the Euclidean case with $w(s) = 1/s$.

To analyze optimal clusterings in the mixed divergence clustering problem, we adopt infinitesimal analysis.

Now, define W_j by

$$W_j = w\left(\sum_{l=1}^{n} \zeta(j,l)\right) \sum_{l=1}^{n} \zeta(j,l) D(\tilde{r}^{(j)} \| p^{(l)})$$

where we recall $\tilde{r}^{(j)}$ also depends on $\zeta(j,l)$.

We now have the following lemma.

Lemma 1 $$\frac{\partial W_j}{\partial \zeta(j,l')} = w\left(\sum_{l=1}^{n} \zeta(j,l)\right) D(\tilde{r}^{(j)} \| p^{(l')}) + \frac{\partial w(\sum_{l=1}^{n} \zeta(j,l))}{\partial \zeta(j,l')} \sum_{l=1}^{n} \zeta(j,l) D(\tilde{r}^{(j)} \| p^{(l)})$$

Theorem 4 *Suppose $w(s) = s^\alpha$. For an optimal solution of the mixed divergence clustering problem, consider the weighted ∇^*-Voronoi diagram for weighted centroids $\tilde{r}^{(j)}$ with multiplicative weight $w(\sum_{l=1}^{n} \zeta(j,l))$ and additive weight*

$$\alpha\left(\sum_{l=1}^{n} \zeta(j,l)\right)^{\alpha-1} \sum_{l=1}^{n} \zeta(j,l) D(\tilde{r}^{(j)} \| p^{(l)}).$$

Then, for point $p^{(l)}$ in the Voronoi region of $\tilde{r}^{(j)}$, $\zeta(j,l) = 1$, and, for points $p^{(l)}$ with $0 < \zeta(j,l) < 1$, it is on some boundary face on the Voronoi region of $\tilde{r}^{(j)}$.

As mentioned above, this general characterization solves an unsolved pure variance-clustering problem in the Euclidean case with $w(s) = 1/s$ to which no characterization by the Voronoi diagram has not yet been known.

5 Random sampling algorithm

We extend approximate algorithm for 2-clustering using random sampling technique to divergence-sum problem, based on the algorithm in the Euclidean case [6]. The idea of this algorithm is that, sampled data might not reflect the cost function of whole data, but, sampled data can reflect the centroid with high probability, so, try all possible partitioning on sampled data, and, compute the centroid using sampled data, then, compute cost function using whole data and get minimum one.

However, in the Euclidean case, the divergence is directly connected with the variance, the clustering 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. Using the Kullback-Leibler divergence, we apply the tail distribution analysis using this divergence.

[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, and we have the following.

Theorem 5 *Suppose that there is an optimal 2-clustering such that the sizes of each cluster 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 $1 - 4d\exp\left(-2\alpha'\left(1-\exp(-\frac{cD}{n})\right)^2 m\right)$.*

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* (in Japanese). Iwanami Shoten, Tokyo, 1993.

[3] F. Aurenhammer and H. Edelsbrunner: An Optimal Algorithm for Constructing the Weighted Voronoi Diagram in the Plane. *Pattern Recognition*, Vol.17 (1984), 251–257.

[4] 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.

[5] M. Inaba and H. Imai: The Number of Partitions of n Points Induced by the Voronoi Diagram via the Conjugacy Generated by k Points. *Japanese-Hungarian Symposium on Discrete Mathematics and Its Applications*, March 1999.

[6] M. Inaba, N. Katoh and H. Imai: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering. *Proceedings of the 10th ACM Symposium on Computational Geometry*, 1994, pp.332–339.

[7] M. K. Murray and J. W. Rice: *Differential Geometry and Statistics.* Chapman & Hall, 1993.

[8] 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*, Fukuoka, 1997, pp.1–8.

[9] K. Onishi and H. Imai: Riemannian Computational Geometry — Voronoi Diagram and Delaunay-type Triangulation in Dually Flat Space, submitted for publication.

[10] K. Sadakane, H. Imai, K. Onishi, M. Inaba, F. Takeuchi and K. Imai: Voronoi Diagrams by Divergences with Additive Weights. *Proceedings of the 14th Annual ACM Symposium on Computational Geometry*, 1998, pp.403–404.