Preliminary Work on Multiple Clustering

Yingyu Liang LYYCS42@GMAIL.COM
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China

Abstract

In this paper, a new approach using multiple clustering results to quantify data distribution is proposed. Theoretical analysis is provided to prove that the proposed approach can achieve guaranteed accuracy with much fewer cluster centers compared to traditional approach using one optimal cluster result. This suggests its potential usage in high-dimensional data analysis.

1. Introduction

Clustering has long been an important issue in data analysis. It is regarded as quantization problem in the signal processing field. As a mathematical topic it concerns the best approximation of a d-dimension random vector X with probability distribution P by a random vector Y with at most k values in its image(Graf & Luschgy, 2000). From its beginning quantization or clustering occurs in various scientific fields, such as information theory(Shannon, 1959; Gersho & Gray, 1992), pattern recognition, numerical integration (Pages, 1997) and mathematical models in economics(Bollobas, 1973). The widely used K-means clustering algorithm or Lloyd's algorithm goes back at least to (Steinhaus, 1956; Lloyd, 1957) and (Forgy, 1965; Jancey, 1966; MacQueen, 1967).

In the field of statistical machine learning, cluster analysis lies in the center of unsupervised learning. It also has close connection to dimension reduction or data coarse representation. A recent positive development has been the realization that a lot of data that superficially lie in a high-dimensional space $\mathbb{R}^d$, actually have low intrinsic dimension, in the sense of lying close to a manifold of dimension $d << D$. There has thus been a huge interest in algorithms that learn this manifold, or sparse representation from data, with the intention that future data can then be transformed into this low-dimensional space, in which the usual nonparametric (and other) methods will work well (Tenenbaum et al., 2000; Roweis & Saul, 2000; Belkin & Niyogi, 2003; Lafon & Lee, 2006). The recent works on compressed sensing (Donoho, 2006; Candes & Tao, 2006; Dasgupta & Freund, 2009) can be viewed as an extend of vector quantization and the K-SVD(Aharon et al., 2006) is the extended version of K-means algorithm.

The vector quantization is susceptible to the same curse of dimensionality that has been the bane of other nonparametric statistical methods (Dasgupta & Freund, 2009). The resulting quantization error has been shown to be roughly $k^{-2/d}$ under a variety of different assumptions on P (Graf & Luschgy, 2000), which is discouraging in the high-dimensional scene. We propose to use multiple clustering results to quantify data distribution, in order to achieve satisfying accuracy with acceptable size of representing centers. Theoretical analysis is provided to show the error bound of the proposed approach compared to that of the optimal set of clustering centers under assumptions on P, which suggests its potential usage in high-dimensional data analysis.

Preliminary work. Use latex template modified from ICML2010 style. Do not distribute.

2. Clustering and Quantization

In this section, we briefly review the definitions and notations, which are mainly from (Graf & Luschgy, 2000). Let X denote a $\mathbb{R}^d$-valued random variable with distribution P. Let $1 \leq r < \infty$ and assume $E\|X\|^r < \infty$. The quantization and cluster center problems are defined as follows.

Definition For $n \in \mathbb{N}$, let $\mathcal{F}_n$ be the set of all Borel measurable maps $f : \mathbb{R}^d \to \mathbb{R}^d$ with $|f((R)^d| \leq n$. The elements of $\mathcal{F}_n$ are called **n-quantizers**.

Definition The **n-th quantization error for P of order r** is defined by

$$V_{n,r}(P) = \inf_{f \in \mathcal{F}_n} E\|X - f(X)\|^r.$$

We will also write $V_{n,r}(X)$ instead of $V_{n,r}(P)$.

Definition A quantizer $f_0 \in \mathcal{F}_n$ is called **n-optimal quantizer for P of order r** if

$$V_{n,r}(P) = E\|X - f_0(X)\|^r.$$

Definition An **n-centers** is a set $\alpha \subset \mathbb{R}^d, |\alpha| \le n$. The **n-th clustering error for P of order r** is defined by

$$U_{n,r}(P) = \inf_{\substack{\alpha \subset \mathbb{R}^d, \\ |\alpha| \le n}} E \min_{a \in \alpha} ||X - a||^r.$$

Definition A set of cluster centers $\alpha_0 \subset \mathbb{R}^d, |\alpha_0| \le n$ is called **n-optimal set of centers for P of order r** if

$$U_{n,r}(P) = E \min_{a \in \alpha_0} ||X - a||^r.$$

Lemma A.1 shows that if f is an n-optimal quantizer, then $f(\mathbb{R}^d)$ is an n-optimal set of centers. Conversely, if $\alpha \subset \mathbb{R}^d$ is an n-optimal set of centers and $A_a : a \in \alpha$ is a Voronoi partition of $\mathbb{R}^d$ with respect to α, then $f = \sum_{a \in \alpha} a 1_{A_a}$ is an n-optimal quantizer. Let $C_{n,r}(P)$ denote the set of all n-optimal sets of centers for P of order r. We also write $C_{n,r}(X)$ instead of $C_{n,r}(P)$.

3. Multiple Clustering

Previous works on quantization and clustering provide important tools for analyzing data. However, high-dimensional data analysis requires much larger set of clustering centers since the clustering error increases rapidly with the dimension. Lemma A.2 shows that the clustering error of the grid centers for uniform distribution on a cube grows linearly with the dimension, even when the number of centers grows exponentially. We propose to use multiple sets of centers, which are much smaller and can achieve similar clustering error compared to one single large set of centers. First the notations about multiple clustering are defined and then the main result concerning the error bound is presented.

Definition An **m fold n-centers** is a set $\mathscr{A} = \{\alpha_i : \alpha_i \subset \mathbb{R}^d, |\alpha_i| \le n, 1 \le i \le m\}$. The **m fold n-th clustering error for P of order r** is defined by

$$U^m_{n,r}(P) = \inf_{\mathscr{A}} E\Big[\frac{1}{2}||X - Y||^r | h(X,\alpha) = h(Y,\alpha), \alpha \in \mathscr{A}\Big]$$

where

$$h(X,\alpha) = \arg\min_{a \in \alpha} ||X - a||^r.$$

Definition An **m fold n-centers** $\mathscr{A}_0$ is called **m fold n-optimal set of centers for P of order r** if

$$U^m_{n,r}(P) = E\Big[||X - Y||^r | h(X,\alpha) = h(Y,\alpha), \alpha \in \mathscr{A}\Big].$$

Theorem 3.1 *Let* $P = \sum_{i=1}^k \frac{1}{k} U(C_i)$ *where*

$$C_i = x_i + [\frac{-1}{2}, \frac{1}{2}]^d, 1 \le i \le k$$

and

$$\inf_{\substack{x \in C_i, \\ y \in C_j, i \ne j}} ||x - y|| > \max_{1 \le i \le k} \sup_{x,y \in C_i} ||x - y||.$$

Then with sufficient large d,

$$\frac{U^d_{2k,2}(P)}{U_{2^d k,2}(P)} \le \frac{2\pi e}{3}.$$

Proof Suppose α is the $2^d k$-optimal set of centers for P (the existence of α is guaranteed (Pollard, 1982)). Let

$$B(a,r) = \{x : ||x - a|| < r\}$$

$$I(X) = \begin{cases} 1 & \text{if } X \in \cup_{a \in \alpha} B(a,r) \\ 0 & \text{if } X \notin \cup_{a \in \alpha} B(a,r) \end{cases}$$

and

$$\beta_i = (x_i + \{-\frac{1}{4}, \frac{1}{4}\}^d), \beta = \cup_{i=1}^k \beta_i$$

$$L = \sup_x \min_{b \in \beta} ||x - b|| = \frac{1}{4}\sqrt{d}.$$

Then we have

$$\begin{aligned} U_{2^d k,2}(P) &= E \min_{a \in \alpha} ||X - a||^2 \\ &\ge Pr[I(X) = 0] E\Big[\min_{a \in \alpha} ||X - a||^2 | I(X) = 0\Big] \\ &\ge Pr[I(X) = 0] E\Big[\min_{b \in \beta} \Big(\frac{r}{L}\Big)^2 ||X - b||^2 | I(X) = 0\Big] \\ &\ge \Big(\frac{r}{L}\Big)^2 E \min_{b \in \beta} ||X - b||^2 \\ &\quad - Pr[I(X) = 1] E\Big[\min_{b \in \beta} \Big(\frac{r}{L}\Big)^2 ||X - b||^2 | I(X) = 1\Big] \\ &\ge \Big(\frac{r}{L}\Big)^2 E \min_{b \in \beta} ||X - b||^2 - Pr[I(X) = 1] r^2. \end{aligned}$$

The condition

$$\inf_{\substack{x \in C_i, \\ y \in C_j, i \ne j}} ||x - y|| > \max_{1 \le i \le k} \sup_{x,y \in C_i} ||x - y||$$

leads to if $x \in C_i$, $\arg\min_{b \in \beta} ||x - b|| \in \beta_i$, so

$$\begin{aligned} E \min_{b \in \beta} ||X - b||^2 &= E\Big[E[\min_{b \in \beta_i} ||X - b||^2 | X \in C_i]\Big] \\ &= E\Big[E[\min_{b \in \beta_i} ||X - b||^2 | X \in C_i]\Big] \\ &= E[\min_{b \in \beta_1} ||X - b||^2 | X \in C_1] \\ &= \frac{d}{48}. \end{aligned}$$

Also the volume of unit ball is $\frac{\pi^{d/2}}{\Gamma(d/2+1)}$, so

$$Pr[I(X) = 1] \le 2^d \frac{\pi^{d/2}}{\Gamma(d/2+1)} r^d.$$

Let $r = \sqrt{\frac{d}{8\pi e}}$, then

$$\lim_{d \to \infty} Pr[I(X) = 1] = 0.$$

So with sufficient large d,

$$\begin{aligned} U_{2^d k,2}(P) &\geq \left(\frac{r}{L}\right)^2 \frac{d}{48} - Pr[I(X)=1]r^2 \\ &\geq \frac{d}{32\pi \mathrm{e}}. \end{aligned}$$

Let

$$\mathscr{A} = \{\alpha_i, 1 \leq i \leq d\}, \alpha_i = \{x_j \pm \frac{1}{4}\epsilon_i, 1 \leq j \leq k\}$$

where $\{\epsilon_i, 1 \leq i \leq d\}$ is the natural base of $\mathbb{R}^d$. Since

$$\inf_{\substack{x \in C_i, \\ y \in C_j, i \neq j}} \|x-y\| > \max_{1 \leq i \leq k} \sup_{x,y \in C_i} \|x-y\|,$$

if $h(X,\alpha_i) = h(Y,\alpha_i)$ and $X \in C_i$ then $Y \in C_i$. Also, if $h(X,\alpha_i) = h(Y,\alpha_i), 1 \leq i \leq d$ then $\{h(X,\alpha_i), 1 \leq i \leq d\} \subset \{x_j \pm \frac{1}{4}\epsilon_i, 1 \leq i \leq d\}$. So

$$E\Big[\frac{1}{2}\|X-Y\|^r | h(X,\alpha) = h(Y,\alpha), \alpha \in \mathscr{A}\Big] = E \min_{b \in \beta} \|X-b\|^2.$$

Finally,

$$\begin{aligned} \frac{U^d_{2k,2}(P)}{U_{2^d k,2}(P)} &\leq \frac{E\Big[\frac{1}{2}\|X-Y\|^r | h(X,\alpha) = h(Y,\alpha), \alpha \in \mathscr{A}\Big]}{U_{2^d k,2}(P)} \\ &\leq \frac{E \min_{b \in \beta} \|X-b\|^2}{U_{2^d k,2}(P)} \\ &\leq \frac{2\pi \mathrm{e}}{3}. \end{aligned}$$

4. Conclusion

A new approach using multiple clustering results to quantify data distribution is proposed. Theoretical analysis is provided to prove that the proposed approach can achieve guaranteed accuracy with much fewer cluster centers compared to traditional approach using one optimal cluster result.

References

Aharon, M., Elad, M., and Bruckstein, A.M. The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representation. In *IEEE Trans. On Signal Processing*, 2006.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. In *Neural-Computation, 15(6)*, pp. 1373–1396, 2003.

Bollobas. The optimal arrangeme of producers. In *J. London Math. Soc. 6*, pp. 605–613, 1973.

Candes, E. and Tao, T. Near optimal signal recovery from random projections: universal encoding strategies? In *IEEE Transactions on Information Theory, 52(12)*, pp. 5406–5425, 2006.

Dasgupta, Sanjoy and Freund, Yoav. Random projection trees for vector quantization. In *IEEE Transactions on Information Theory July 2009.Vol 55, Issue 7*, 2009.

Donoho, D. Compressed sensing. In *IEEE Transactions on Information Theory, 52(4)*, pp. 1289–1306, 2006.

Forgy, E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications(abstract). In *Biometrics 21*, pp. 768–769, 1965.

Gersho, A. and Gray, R. M. (eds.). *Vector Quantization and Signal Compression*. Kluwer, Boston, 1992.

Graf, Siegfried and Luschgy, Harald (eds.). *Foundations of Quantization for Probability Distributions*. Springer, 2000.

Jancey, R. Multidimensional group ananlysis. In *Austral. J. Botany*, pp. 127–130, 1966.

Lafon, S. and Lee, A. B. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. In *IEEE Trans. on Pattern Analysis and Machine Intelligence 28(9)*, pp. 1393–1403, 2006.

Lloyd, S. Least squares quantization in pcm. Technical report, Bell Laboratories, Published in 1982 in IEEE Trans. Inf. Theory 28 128-137, 1957.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

Pages. A space quantization method for numerical integration. In *J. Comput. Appl. Math. 89*, pp. 1–38, 1997.

Pollard, D. Quantization and the method of k-means. In *IEEE Trans. Inform. Theory 28*, pp. 199–205, 1982.

Roweis, S. T. and Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. In *Science(290)*, pp. 2323–2326, 2000.

Shannon, C. E. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record, Part 4*, pp. 142–163, 1959.

Steinhaus, H. Sur la division des corps materiels en parties. In *Bull. Acad. Polon. Sci. 4*, pp. 801–804, 1956.

Tenenbaum, J., de Silva, V., and Langford, J. A global geometric framework for nonlinear dimensionality reduction. In *Science, 290(5500)*, pp. 2319–2323, 2000.

A. Appendix A

Lemma A.1 *(Graf & Luschgy, 2000)The n-optimal quantization problem is equivalent to the n-centers problem, which means*

$$\inf_{f\in\mathcal{F}_n} E\|X-f(X)\|^r = \inf_{\substack{\alpha\subset\mathbb{R}^d,\\ |\alpha|\le n}} E\min_{a\in\alpha}\|X-a\|^r.$$

Proof For $f\in\mathcal{F}_n$, let $\alpha = f(\mathbb{R}^d)$ and $A_a = f = a, a\in\alpha$. Then

$$\begin{aligned} E\|X-f(X)\|^r &= \sum_{a\in\alpha}\int_{A_a}\|x-a\|^r dP(x)\\ &= \sum_{a\in\alpha}\int_{A_a}\min_{b\in\alpha}\|x-b\|^r dP(x)\\ &= E\min_{b\in\alpha}\|X-b\|^r. \end{aligned}$$

Conversely, for $\alpha\subset\mathbb{R}^d$ with $|\alpha|\le n$, let $\{A_a : a\in\alpha\}$ be a Voronoi partition of $\mathbb{R}^d$ with respect to α and let $f = \sum_{a\in\alpha} a1_{A_a}$. Then $f\in\mathcal{F}_n$ and

$$\begin{aligned} E\min_{a\in\alpha}\|X-a\|^r &= \sum_{a\in\alpha}\int_{A_a}\|x-a\|^r dP(x)\\ &= E\|X-f(X)\|^r. \end{aligned}$$

Lemma A.2 *Let $P = U([-\frac{1}{2},\frac{1}{2}]^d)$. Let $\alpha = \{a_1,...,a_n\} = \{-\frac{1}{4},\frac{1}{4}\}^d$. Then*

$$E\min_{a\in\alpha}\|X-a\|^2 = \frac{d}{48}.$$

Proof Consider a tesselation of $[\frac{-1}{2},\frac{1}{2}]^d$ consisting of $n = 2^d$ cubes $S = \{C_1,...,C_n\} = \{C : C = \{(x_1,...,x_d) : b_i \le x_i \le b_i+\frac{1}{2}, 1\le i\le d\}, b_i = -\frac{1}{2} \text{ or } b_i = 0\}$. It can be seen that α is the set of centers for cubes in S. Without loss of generality, suppose a_i is the center of $C_i\in S$. We have

$$a_i = \arg\min_{a\in\alpha}\|x-a\|, \forall x\in C_i.$$

Let $V(C_i)$ denote the volume of C_i.Then

$$\begin{aligned} E\min_{a\in\alpha}\|X-a\|^2 &= E\Big[E[\min_{a\in\alpha}\|X-a\|^2|X\in C_i]\Big]\\ &= \sum_{i=1}^n Pr[X\in C_i]\int_{C_i}\frac{1}{V(C_i)}\|x-a_i\|^2dx\\ &= \sum_{i=1}^n Pr[X\in C_i]\int_{[-\frac{1}{2},\frac{1}{2}]^d}\|x\|^2dx\\ &= \Big(\sum_{i=1}^n Pr[X\in C_i]\Big)\Big(\int_{[-\frac{1}{2},\frac{1}{2}]^d}\sum_{j=1}^d x_j^2dx\Big)\\ &= \frac{d}{48}. \end{aligned}$$