On the choice of initial seed values for the Lloyd's k-means algorithm

Kalev Pärna
University of Tartu, Tartu, Estonia – kalev.parna@ut.ee

This paper is about efficient ways of converting continuous data into a compact discrete form. In information theory such a conversion is called quantization and it is used for transmission of data through a discrete channel which is capable of admitting a finite number (k) of values, only. In statistics and data mining the same method is called k-means clustering and its aim is to partition a data set into k non-overlapping clusters by minimizing the within-sum of squares of deviations from their respective cluster centers (k-means). Efficient calculation of k-means, especially in multivariate setting, is still a problem which needs further research. Well-known Lloyd's iterative method for calculation of k-means is sensitive with respect to initial values and, therefore, much research has been focused on the choice of initial seed values of k-means. In this paper we propose to use certain theoretical results about asymptotic density of points of k-means in the process where the number k grows infinitely. Namely, it is known that for large values of k the optimal k points are distributed in accordance with density $f^*(x)$ which is a power function of the initial data density $f(x)$ (for example, in one-dimensional case $f^*(x)$ is proportional to $f(x)^{1/3}$). Our main idea is to use this asymptotic distribution for placement of initial seeds of k-means. In order to benefit from the asymptotic theory, we propose a 3-steps method for calculation of k-means, consisting of 1) estimation of $f^*(x)$ from the data, 2) generation of k points from $f^*(x)$, and, 3) using these points as initial values in the Lloyd's iterative algorithm.

Keywords: quantization; asymptotic distribution of k-means, Lloyd's algorithm.

1. Introduction

Cluster analysis aims at partitioning of a given data set into groups (clusters) so that the data points within a group are more similar to each other than the points in different groups. Numerous methods have been proposed to solve grouping problem. One of the most popular method is k-means clustering algorithm developed by Mac Queen in 1967 [1]. The simplicity of k-means clustering algorithm made this algorithm popular in various fields like image compression, distribution of resources, cellular biology, statistics, the territorial behavior of animals, etc. The k-means clustering algorithm is a partitioning clustering method that separates data into k groups. It is scalable and can handle even massive data [1], [2], [3], [4], [5].

In practice, k-means clustering is based on the use of Lloyd's 2-phase iterative algorithm [6]. The algorithm starts with some initial set of k points in R^d and finds its corresponding Voronoi partition i.e. minimum distance partition. As a result, one obtains k regions, each containing data points that are closest to a certain point among k initial points. Next, for each cluster its mass center (mean value) is calculated, resulting in a new k-set – an improved version of the previous k-mean. These two steps are repeated until the sequence of k-sets and their respective clusters converge to a limit. However, the difficulty with Lloyd's method is that, generally speaking, the described process does not end at the global optima – depending on the choice of the initial k-set the process can easily converge to a local optima. The usual remedy is to try with many different initial values and to choose the best result. It is understood, therefore, that much research has been focused on the good choice of initial values of k-means. [2], [5], [7], [8]. Initial seeds of k -means also have an influence on the number of iterations required for convergence of the original k-means algoritm.

Various ideas have been proposed in the literature to increase the efficiency of the k-means clustering algorithm. Papers like [2], [3], [7] try to take into account, at least in some extent, the distribution of the initial data, by allocating more centres to the regions with higher density, and, undoubtedly, there is good rational behind that. However, there is clear theoretical indication that this is not the data distribution itself that should be followed when placing initial k points, but its certain transform. Namely, it is known that for large k the optimal k points are distributed in accordance with certain density $f^*(x)$ which can be computed from intial data density $f(x)$.

In this paper, our basic idea is to make use of asymptotic theory of optimal points density in order to place initial values of k-means for the Lloyd's algorithm. While this approach is well studied and used in quantization theory, especially when high-resolution (large k) is required, it seems to be neglected in data mining community. In fact, this can easily be understood because in cluster analysis very often k is not a big number.

The paper is organized as follows. In Section 2, we give a brief overview of asymptotic results about the points density of optimal k-means. In Section 3, we show how the asymptotic distribution of k-means, $f^*(x)$, can be estimated from the initial data, and how to place k points in accordance with the distribution $f^*(x)$.

2. Overview of some asymptotic results in quantization theory.

Let us recall some asymptotic results about k optimal quantizing points in the situation where k increases unboundedly. The following result has been proved by several authors, including Bennett [9].

> Let $A = \{a_1, a_2, \dots, a_k\} \subset \mathbb{R}$ be a set of k optimal quantizing points for the density $f(x)$ on the real line $\mathbb{R}$, i.e. A minimises the distortion error $E[\min_i \|X - a_i\|^2]$ over all possible choices of A. If $k \to \infty$, then for any given subset $D \subset \mathbb{R}$ the fraction of points in A which belong to the subset D has the limit
>
> $$\frac{\#\{a_i : a_i \in D\}}{k} \to \int_D f^*(x)dx,$$
>
> where $f^*(x) = C \cdot f^{1/3}(x)$ is points density function with normalizing constant, $C = \left[\int_{\mathbb{R}} f^{1/3}(x)dx\right]^{-1}$.

As a corollary, it is seen that if D is a subset of f^*-probability $\frac{1}{k}$, then for k large enough, D must contain just one element of A. In practice, this simple rule works quite well already for moderate values of k (k=5,6,…). Furthermore, if we divide the real line into k non-overlapping intervals of equal probability, then each segment should contain exactly one element of the k-mean A. Thus, for e.g. k=10, each decile interval is supposed to contain one element of a k-mean. A natural idea is to place that element at the median point of respective interval. Thus, the quantization points will be the α −quantiles of density $f^*(x)$, with $\alpha = \frac{1}{2k}, \frac{3}{2k}, \dots, \frac{2k-1}{2k}$. The quantizer obtained by this method is called *compander*.

A similar result can be formulated in higher-dimensional spaces. Namely, in [10], [11] it was shown that for the distortion error $E[\min_i \|X - a_i\|^2]$the density of optimal points can be expressed as $f^*(x) = C \cdot f^{d/d+2}(x)$ with normalizing constant $C = \left[\int_{\mathbb{R}} f^{d/d+2}(x)dx\right]^{-1}$. It is seen that in case of high dimension the fraction $\frac{d}{d+2}$ is close to 1 and, hence, the density function of optimal quantizing

points does not differ from the initial data density too much. However, in case of low-dimensional data the density of quantizing points is more uniform than the initial density.

3. Initial seed generation for *k*-means

It is easy to display *k* points on the real line to conform with a known density function $f^*(x)$. For doing that, one should simply use α −quantiles of the density $f^*(x)$ by taking $\alpha = \frac{1}{2k}, \frac{3}{2k}, \ldots, \frac{2k-1}{2k}$. If $f^*(x)$ is not known explicitly, one can use empirical α −quantiles from ordered sample. This is a well established and efficient method in scalar quantization. In *k*-means clustering with moderate size of *k* we can use these quantiles as initial seeds for *k*-means in Lloyd's algorithm.

Unfortunately, the method described above for scalar data, stops working *d*-dimensional case, $d > 1$, because multivariate quantiles are not defined, at least in a straightforward way. Therefore, as a remedy, we propose to use some density estimation method to get an estimate $\hat{f}(x)$ of the initial density *f(x)*, and then apply the transform $\hat{f}^*(x) = C \cdot \hat{f}^{d/d+2}(x)$ to obtain the estimate of the optimal points distribution. Thus, we propose the following procedures:

1) Using a kernel density estimation method, calculate $\hat{f}(x_i)$ at each data point x_i.
2) Reweight all data points by using weights $\hat{f}^*(x_i) = C \cdot \hat{f}^{d/d+2}(x_i)$.
3) Select *k* random points among the reweighted data points (using the weights as probabilities).
4) Use the *k* points chosen at previous step as initial seeds for *k*-means in the Lloyd's algorithm.
5) Apply Lloyd's iterative algorithm to improve initial seeds until the process converges.

4. Conclusions

Our method of calculation of initial values of *k*-means makes use of deep results about asymptotic point density of optimal quantizers in quantization theory. These asymptotic methods, although suboptimal for small and moderate values of *k*, can be used for efficient placement of initial seeds of *k*-means in Lloyd's iterative method. Our examples demonstrate that the use of asymptotic theory is justified used already for small values of *k*.

References

[1] J. Mc Queen, "Some methods for classification and analysis of multivariate observations," Proc. 5th Berkeley Symp. Math. Statist. Prob., (1): 281–297, 1967.
[2] A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, "An Efficient enhanced k-means clustering algorithm," Journal of Zhejiang University, 10(7): 16261633, 2006.
[3] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the accuracy and efficiency of the k-means clustering algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009),Vol 1, July 2009, London, UK.
[4] F. Yuan, Z. H. Meng, H. X. Zhangz, C. R. Dong, " A New Algorithm to Get the Initial Centroids," proceedings of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.
[5] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: an algorithm for Centroids initialization for k-means," department of information science and Electrical Engineering. Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.
[6] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory,Vol. IT-28, No. 2, March,1982.

[7] Chen Zhang and Shixiong Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.
[8] M.Yedla, Srinivasa Rao Pathakota, T. M. Srinivasa. Enhancing K-means Clustering Algorithm with Improved Initial Center. International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, 121-125.
[9] W. R. Bennett, "Spectra of quantized signals," Bell Syst. Tech. J., vol. 27, pp. M-472, July, 1948.
[10] P. Zador, "Asymptotic quantization of continuous random variables," unpublished memorandum, Bell Laboratories, 1966.
[11] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," IEEE Trans. Znform. Theory, vol. IT-14, pp. 676-683, Sept. 1968.