IQ-Means

Evangelos Anagnostopoulos

National and Kapodistrian University of Athens

March 31, 2017

Problem Definition

Definition (Clustering)

Given a set of objects partition them into disjoint sets such that objects within a group are more "similar" compared to those in other groups.

Definition (k-means Clustering)

Given a pointset $X \subset \mathbb{R}^d$ of n points and a parameter k, find k point centers $C^* = \{c_1, c_2, \ldots, c_k\} \subset \mathbb{R}^d$ such that the sum of squared distances of each point in X to its nearest center is minimized.

Objective function:

$$\min \sum_{x \in X} ||x - c(x)||^2,$$

where $c(x) \in C^*$ is the center closest to x.

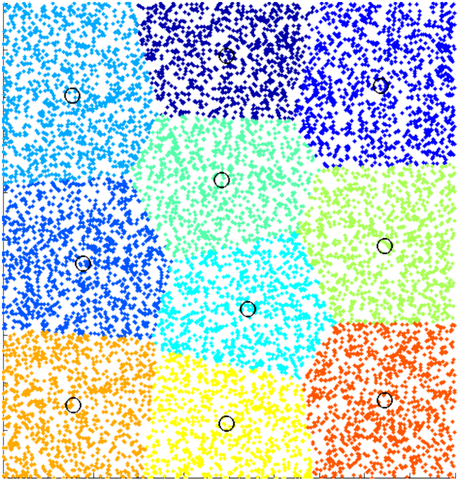

Figure: *A k-means clustering example. Notice how the cluster regions correspond to a a Voronoi diagram of the centroids.*

Base Algorithm (Lloyd's)

Input: X s.t. $|X| = n, k, j$ optional
Output: C^*

- Initialize C^* to k points selected uniformly at random from X.
- Until convergence (or for j iterations)
  - **Assignment step:**
    Assign each point to its nearest center
  - **Update step:**
    Compute the mean μ_i of each cluster i, and assign that as the new center c_i

Complexity: $O(nkd(j))$

Related work

- **Approximate k-means**
  [Philbin et al. '07]
  Replace assignment step with approximate nearest neighbor (ANN) from points to centers.
- **Binary k-means**
  [Gong et. al '15]
  Binarize points and centers, followed by ANN in Hamming space
- **Ranked Retrieval**
  [Broder et al. '14]
  ANN queries from centroids to points

Related work cont.

- **Dimensionality-Recursive Vector Quantization (DRVQ)**
  [Avrithis '13]
  Centroids to point queries on a two-dimensional grid
- **Expanding Gaussian Mixtures (EGM)**
  [Avrithis et al. '12]
  On the fly estimation of the number of clusters by a statistical approach.

IQ-Means

Goal: Web scale clustering (i.e. hundreds of millions of points into millions of clusters)
IQ-Means combined with powerful deep learned representations, achieves clustering of a 100 million image collection on a single machine in less than one hour.[Avrithis '15].
Compare to distributed k-means on 300 machines which takes 2.2 hours per iteration on average, i.e. one order of magnitude slower.

IQ-Means idea

- Adopt subspace quantization from DRVQ.
- Modify search algorithm to imitate Ranked Retrieval's approach.
- Estimate k dynamically by purging clusters, as in EGM.

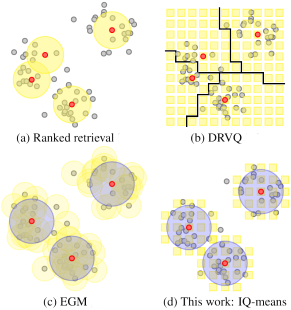

Figure: *Different k-means variants.*

Vector Quantization

Given a pointset $X \subset \mathbb{R}^d$, where $|X| = n$ and assuming that d is even:

Dimension Decomposition

$\mathbb{R}^d$ is expressed as the Cartesian product of two orthogonal subspaces S^1, S^2. In the simplest form $S^1 = S^2 = \mathbb{R}^{d/2}$, i.e.

$$x = (x^1, x^2), \text{where } x^1 \in S^1 = \mathbb{R}^{d/2}, x^2 \in S^2 = \mathbb{R}^{d/2}$$

This can continue recursively until we reach $\mathbb{R}$ and then we can perform a one-dimensional clustering.

Vector Quantization cont.

$X \subset \mathbb{R}^d$, $|X| = n$; $\mathbb{R}^d = S^1 \times S^2$

Representation of Points

Assume two clusters U^1, U^2 trained indepedently on the projection of P onto S^1 and S^2, where each cluster contains s centroids.
Then, $U = U^1 \times U^2$ contains $s \times s$ centroids and partitions $\mathbb{R}^d$ into $s \times s$ cells.
We view U as a two-dimensional grid and map each $p \in P$ to cell $q(x) = (q^1(x^1), q^2(x^2))$, where $q^i(x^i)$ is the closest centroid to x^i in U^i.

Quantization

For each cell $u_\alpha, \alpha \in I = [s] \times [s]$, compute:

- Empirical frequency: $p_\alpha = |X_\alpha|/n$, where $X_\alpha = \{x \in X \mid q(x) = u_\alpha\}$.
- Mean: $\mu_\alpha = \frac{1}{|X_\alpha|} \sum_{x \in X_\alpha} x$

We can now discard X.

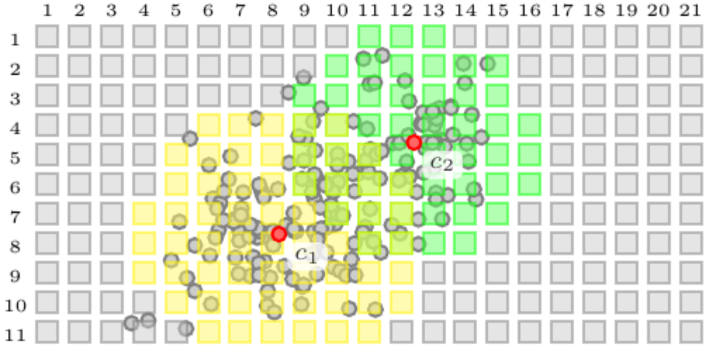

Figure: *Example of a two-dimensional grid U, composed of the Cartesian product of two sub-codebooks U^1, U^2. The points can now be mapped onto this grid and be discarded.*

IQ-Means Algorithm - I

Start with an arbirtrary set C of k centroids.

Update Step

For all centroids $c_m \in C$:

$$c_m \leftarrow \frac{1}{P_m} \sum_{\alpha \in A_m} p_\alpha \mu_\alpha,$$

$A_m = \{\alpha \in I \mid \hat{q}(u_\alpha) = m\}$ and $\hat{q}(u) = \arg\min_{c_m \in C} ||u - c_m||$ and
$P_m = \sum_{\alpha \in A_m} p_\alpha$

IQ-Means Algorithm - II

Assignment Step

For each centroid c_i the w nearest sub-centroids are found in U^1, U^2 and ordered by ascending distance to c_i.
The $w \times w$ cells are then visited in order via a priority queue.
Upon visiting a cell a function f is called. In this case, it updates the curent assignment α and lowest distance *dist* found for each cell u_α. It also terminates upon visiting a specified target T of points.

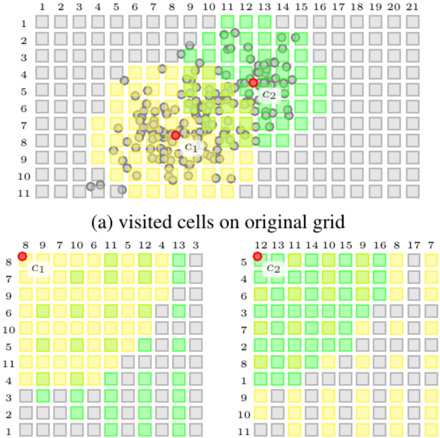

Figure: *Example assignment step. For the centroids c_1, c_2 we have computed the $w \times w$ nearest cells and re-arranged them such that nearest cells appear in the top left corner.*

Small Scale Experiments I

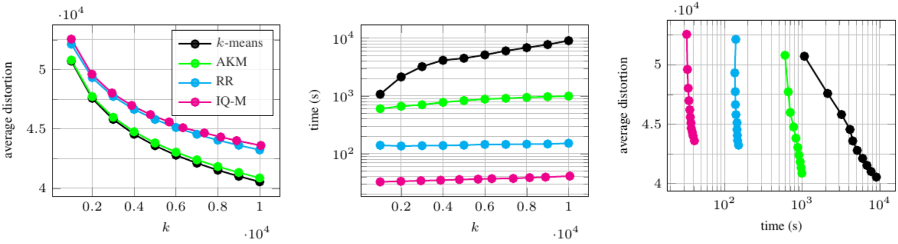

Figure: *Average distortion and total time for* 20 *iterations on SIFT1M for varying number of clusters* k*. Time for IQ-means includes encoding of data points that is constant in* k*, but not codebook learning, which is performed on a different dataset.*

Small Scale Experiments II

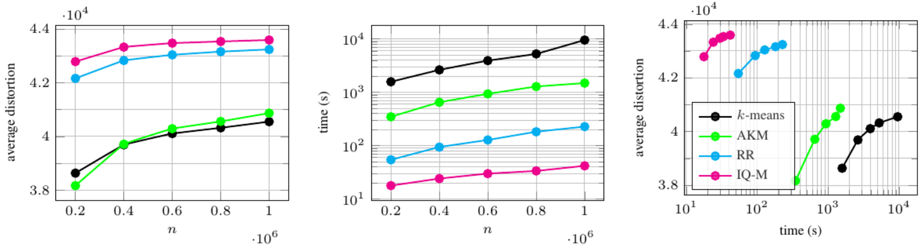

Figure: *Average distortion and total time for 20 iterations on SIFT1M for* $k = 10^4$ *and varying number of data points n. Time for IQ-means includes encoding of data points that is linear in n, but not codebook learning.*

Large Scale Experiments

Figure: *Mining example: subsets of similar clusters for (a) Paris and (b) Paris+YFCC100M. Images in red outline are from the Paris ground truth.*

Dynamic IQ-Means

No-cost purging

Quantize centroids by assigning each centroid c_i to cell u_α by using the nearest sub-centroids returned in the assignment step above.
Maintain a list for each centroid keeping the other centroids encountered in search.
Model the distribution of points assigned to a centroid c_m by an isotropic normal density $\mathcal{N}(x|c_m, \sigma_m)$, where

$$\sigma_m^2 \leftarrow \frac{1}{P_m} \sum_{\alpha in A_m} p_\alpha ||\mu_\alpha - c_m||^2$$

Iterate over all centroids in descending order of population and purge clusters that overlap too much with previous ones.

Dynamic IQ-Means Experiments

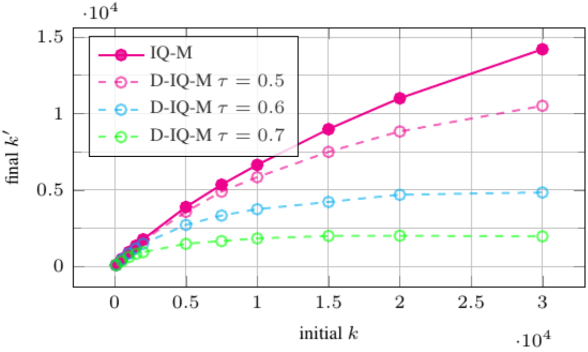

Figure: *Final k' versus initial k number of centroids on SIFT1M for varying overlap threshold τ .*

Thank you!

References I

- Y. Avrithis, Y. Kalantidis, E. Anagnostopoulos, I. Z. Emiris
  Web-scale Image Clustering revisited
  ICCV 2015
- Y. Avrithis.
  Quantize and Conquer: A dimensionality-recursive solution to clustering, vector quantization, and image retrieval
  ICCV 2013
- A. Broder, L. Garcia-Pueyo, V. Josifovski, S. Vassilvitskii, and S. Venkatesan
  Scalable k-means by ranked retrieval
  Web Search and Data Mining 2014.

References II

Y. Gong, M. Pawlowski, F. Yang, L. Brandy, L. Boundev, and R. Fergus
Web scale photo hash clustering on a single machine
CVPR 2015

J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisser- man.
Object retrieval with large vocabularies and fast spatial matching
CVPR 2017

Y. Avrithis, Y. Kalantidis
Approximate Gaussian Mixtures for Large Scale Vocabularies
ECCV 2012.