Tree-Structured Vector Quantization for Similarity Queries

Hanwei Wu, Qiwen Wang, and Markus Flierl
School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm
{hanwei, qiwenw, mflierl}@kth.se

Summary

The problem of compression for similarity queries is characterized by the tradeoff between the rate of the compressed data and the reliability of the answers to a given query [1]. In this work, we propose tree-structured vector quantizers that hierarchically cluster the data into K sphere-shaped quantization cells $C_k, k = 1, \ldots, K$, for efficient similarity queries. In particular, we consider two clustering methods to obtain partitions S for the data $\mathbf{x}$:

k-means: $\min_S \sum_{k=1}^{K} \sum_{i:\mathbf{x}_i \in C_k} (\mathbf{x}_i - \mu_k)^T (\mathbf{x}_i - \mu_k)$

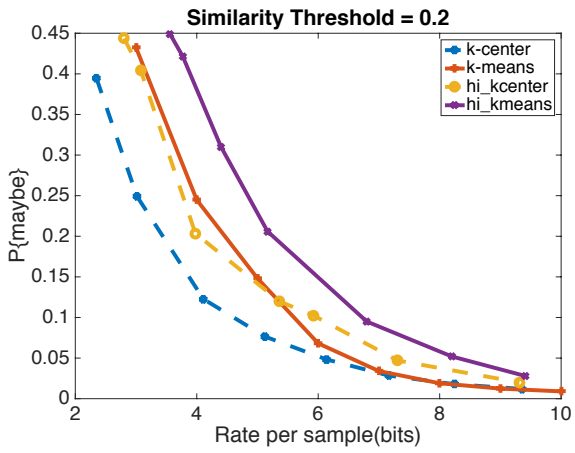
k-center: $\min_S \max_{k=1,\ldots,K} \max_{i:\mathbf{x}_i \in C_k} (\mathbf{x}_i - \mu_k)^T (\mathbf{x}_i - \mu_k)$

We consider D-admissible systems [1] that do not allow false negatives in the retrieval process. Hence, the cells will be characterized by the maximum distance between the cell centroid μ_k and any data point within the cell $\mathbf{x}_i \in C_k$. Compared to k-means, k-center clustering directly minimizes the size of the cells that is proportional to the probability of the output 'maybe', i.e., $P\{g(Q(\mathbf{X}), \mathbf{Y}) = \text{maybe}\}$ with the query function g, the signature assignment $\mathrm{Q}(\mathbf{x})$ of the data $\mathbf{x}$, and the query $\mathbf{y}$.

For tree structures, the query function is guided by decision rules derived from triangle inequalities.

$$g(Q(\mathbf{x}), \mathbf{y}) = \begin{cases} \text{maybe} & |d(\mathbf{y}, \mathbf{P}) - d(\mathbf{P}, \hat{\mathbf{x}})| \leq d^*(\mathbf{x}, \hat{\mathbf{x}}) + D; \\ \text{no} & \text{otherwise} \end{cases}$$

where $\mathbf{y}$ is the query, D is the similarity threshold, $d(\mathbf{P}, \hat{\mathbf{x}})$ is the parent-child distance, and $d^*(\mathbf{x}, \hat{\mathbf{x}})$ is the cell similarity.



Experimental results show that k-center clustering has a better performance than k-means clustering in terms of $P\{\text{maybe}\}$ for a given rate. At high rates, tree-structured quantizers require less computations in the retrieval process when compared to unstructured quantizers in order to maintain a similar probability $P\{\text{maybe}\}$.

[1] A. Ingber, T. Courtade, and T. Weissman, "Compression for quadratic similarity queries," *IEEE Trans. on Information Theory*, vol. 61, no. 5, pp. 2729 –2747, May 2015.