technische universität dortmund

sisap

Accelerating k-Means Clustering with Cover Trees

Andreas Lang and Erich Schubert

TU Dortmund, Data Mining, 44221 Dortmund

{firstname.lastname}@tu-dortmund.de

SISAP 2023

k-Means

- The sum of squared deviations is minimized.
- Current State of the Art algorithms use the triangle inequality to omit unnecessary distance computations.

Cover Tree

- Tree based index structure.
- Representation uses routing objects and radii.

Cover Tree properties:

1. (nesting) $N_i \subset N_{i-1}$,
2. (cover) $\forall q \in N_{i-1} \exists p \in N_i : d(p,q) \leq 2^i$ and exactly one p is the parent of q,
3. (separation) $\forall p,q \in N_i : d(p,q) \geq 2^i$.

Cover Tree and k-Means

Input: Node x with routing object p_x and radius r_x, candidate cluster centers $c_i \in C$.

1. calculate $\forall c_i \in C : d_i = d(p_x, c_i)$
2. prune all $c_i : d_i - 2r_x \geq min(d_i)$
3. $\forall y \subset x$ prune c_i if $d_i - 2(d(p_x,p_y) - r_y) \geq min(d_i)$
4. continue with *Step 1* for each y with the reduced candidate set

Assign node x to a cluster if there is only one remaining for x.
When switching strategies to Hamerly's or derived algorithms, set:

$$u_{q \in y} = d(p_x, c_1) + d(p_x, p_y) + r_y,$$
$$l_{q \in y} = d(p_x, c_2) - d(p_x, p_y) - r_y.$$

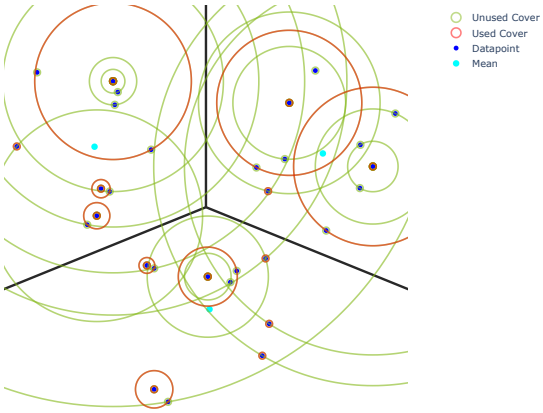
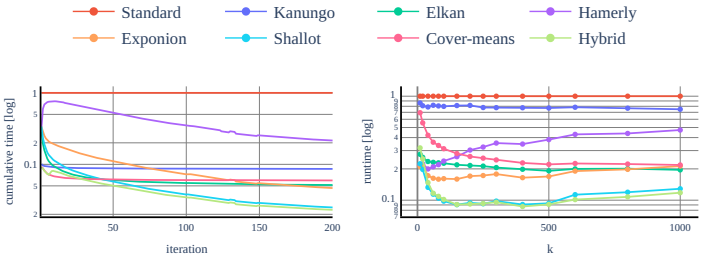


Figure 1: Exemplary k-Means clustering using a Cover Tree. Necessary rounting objects and radii are highlighted.

Evaluation

- Room for improvement in the first iterations.
- Improvements mainly for medium to high k.
- Tree construction overhead less impactful when doing multiple runs.



(a) Runtime over iterations. **(b)** Scaling with k.

Andreas Lang

Corresponding Author
PhD Student at TU Dortmund
Interested in clustering and data mining.

* Supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center Sonderforschungsbereich 876 "Providing Information by Resource-Constrained Data Analysis", DFG project number 124020371, project A2.