# Multi-Step k-Nearest Neighbor Search Using Intrinsic Dimension


Michael E. HOULE[1] Xiguo MA[2] Michael NETT[1,3] Vincent ORIA[2]

[1] National Institute of Informatics, Japan [2] New Jersey Institute of Technology, USA [3] The University of Tokyo, Japan


## Motivation

Most existing solutions for similarity search fail in handling queries with respect to high-dimensional or adaptable distance functions. For such situations, researchers have proposed multi-step search approaches consisting of two stages: filtering and refinement. The filtering stage of the state-of-the-art multi-step search algorithm of Seidl and Kriegel is known to examine the minimum number of candidates necessary to *guarantee* a correct query result; however, the number of examined candidates may be unacceptably large for some applications.
We propose a multi-step search heuristic (MAET) that utilizes a measure of intrinsic dimension, the *generalized expansion dimension*, as the basis of an early termination condition.

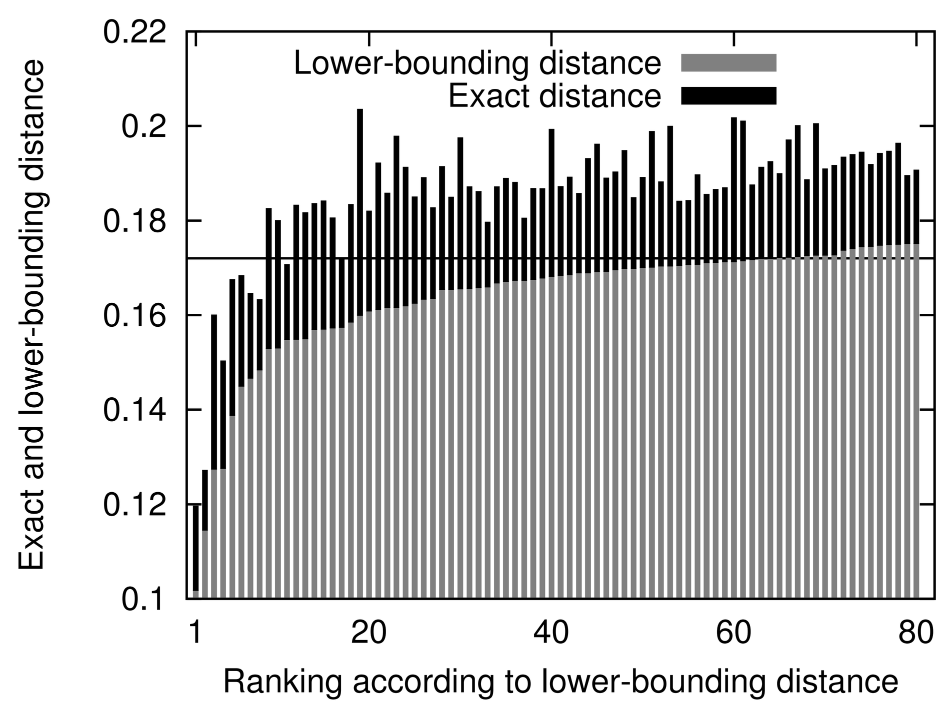
## Seidl and Kriegel's Algorithm

Find $k$-NN of query $q$ with respect to a target distance function $d$, given a lower-bounding distance function $d'$:

- Sequentially scan the neighborhoods of $q$ with respect to $d'$ to retrieve the candidates.
- From the candidate set, the $k$-NN of $q$ with respect to $d$ are stored as tentative query result, and the $k$-th smallest target distance ($d_{\max}$) is maintained. If the size of the candidate set is smaller than $k$, keep the value of $d_{\max}$ being infinity.
- The algorithm terminates when the value of $d_{\max}$ is no greater than the largest lower-bounding distance value encountered so far.
- At termination, the value of $d_{\max}$ will have been decreased to the exact $k$-th smallest distance from $q$ to all the objects in $S$, ensuring the optimality of the algorithm.

## Justification of Heuristic Approach

Although Seidl and Kriegel's Algorithm (SK) examines the minimal number of candidates required in order to guarantee a correct query result, the total number of examined objects can be unacceptably large. The following figure shows distance values encountered while processing a typical 10-NN query on a real data set. In this example, the SK algorithm examines 64 candidates, although the correct neighbor set is available once the 17th candidate has been examined.

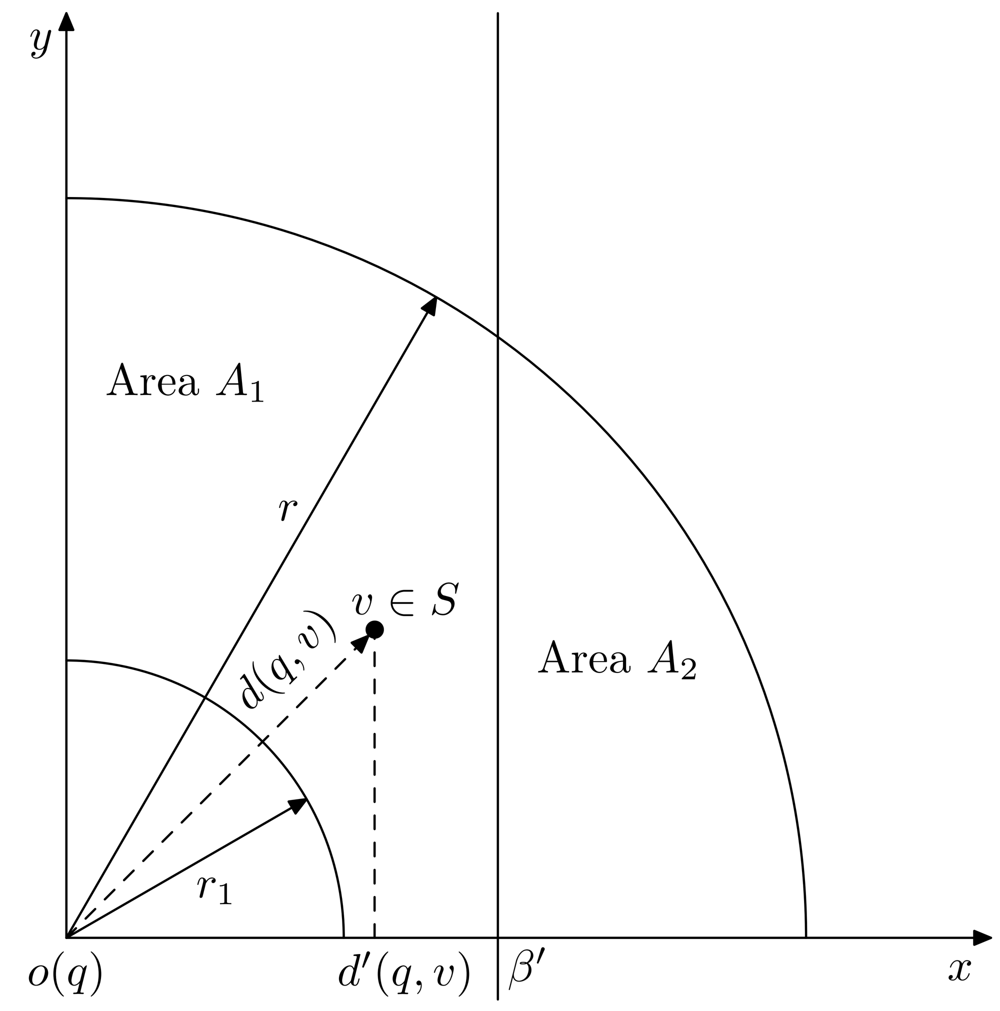



## Generalized Expansion Dimension

- Let $B(q,r_1)$ and $B(q,r_2)$ be two co-centric balls with radii $0 < r_1 < r_2$, each containing $0 < k_1 < k_2$ points from $S$. The *generalized expansion dimension* with respect to those balls is defined as
$$\mathrm{GED}(B(q,r_1),B(q,r_2)) = \frac{\log k_2 - \log k_1}{\log r_2 - \log r_1}.$$
- We define the *inner ball set* relative to a point $q \in S$ and a neighborhood size $k \geqslant 2$ as
$$\mathcal{B}(q,k) = \{B(q,\delta_j) \mid j \leqslant k-1, \delta_j > 0\} \setminus \{B(q,\delta_k)\}.$$
Here $\delta_k$ denotes the distance from $q$ to its $k$-nearest neighbor.
- The *maximum generalized expansion dimension* relative to a point $q$ and a neighborhood size $k \geqslant 2$ is defined as follows:
$$\mathrm{MGED}(q,k) = \max\{\mathrm{GED}(B, B(q,\delta_k)) \mid B \in \mathcal{B}(q,k)\}.$$
- Allows for the estimation of intrinsic dimension in the vicinity of $q$.
- Can be utilized to dynamically guide decisions made in search algorithms.

## Algorithm MAET

- Sequentially scan the neighborhoods of $q$ with respect to $d'$ to retrieve the candidates.
- From the candidate set, the $k$-NN of $q$ with respect to $d$ are stored in $P$ as tentative query result, and the largest lower-bounding distance is maintained as $\beta'$.
- From the set $P$, find the number of points ($k_1$) that have target distances to $q$ less than $\lambda\beta'$, where $\lambda \geqslant 1$ is the lower-bounding ratio.
- From the set $P$, compute $r = \delta_k$ and $r_1 = \delta_{k_1}$.
- The algorithm terminates when $k_1 = k$ or $k_1(r/r_1)^t < k+1$, where $t$ is a termination parameter controlling the performance trade-off between accuracy and computational cost.



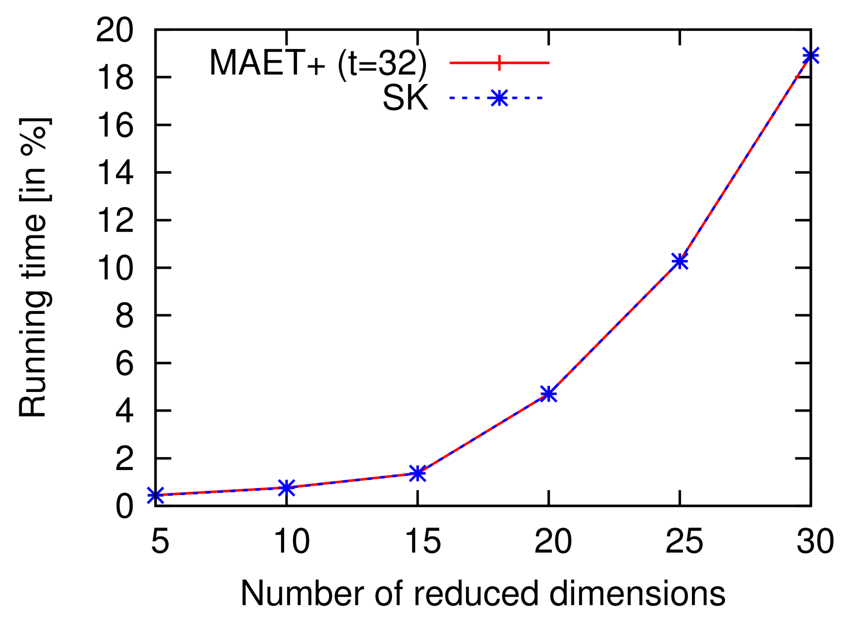

- For a particular query $q$, the correctness of the algorithm can be guaranteed whenever $t \geqslant \mathrm{MGED}(q,k+1)$.
- Termination parameter $t$ can be chosen through sampling of potential queries for Algorithm MAET, in order to correctly answer a desired proportion of potential queries with high probability.

## Algorithm MAET+

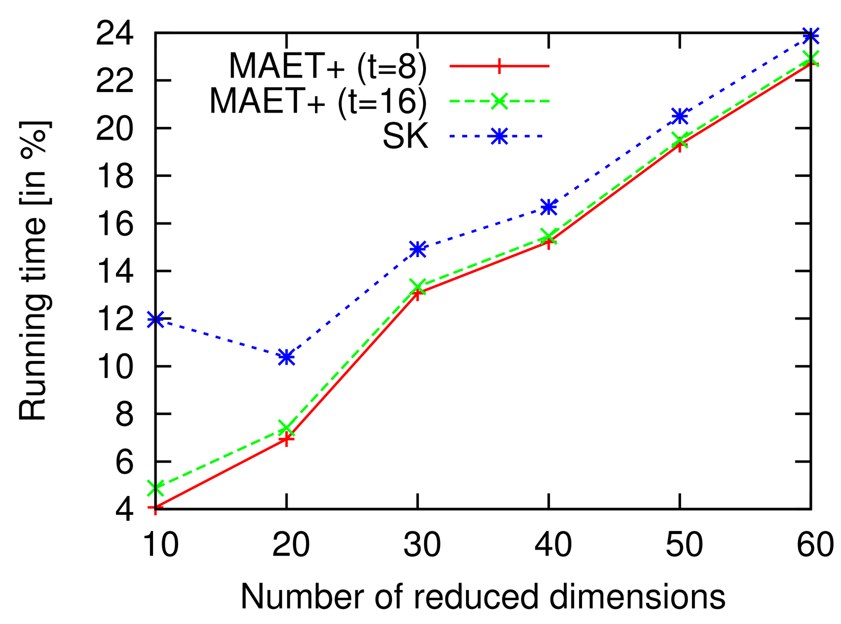
- Compared to MAET, the only change made in MAET+ is that observed distance values are used to dynamically estimate the lower-bounding ratio $\lambda$.
- After each candidate $v$ is retrieved, the ratio of $d(q,v)$ over $d'(q,v)$ is computed, and the smallest ratio encountered so far is used as the estimate of $\lambda$.

## Experiments

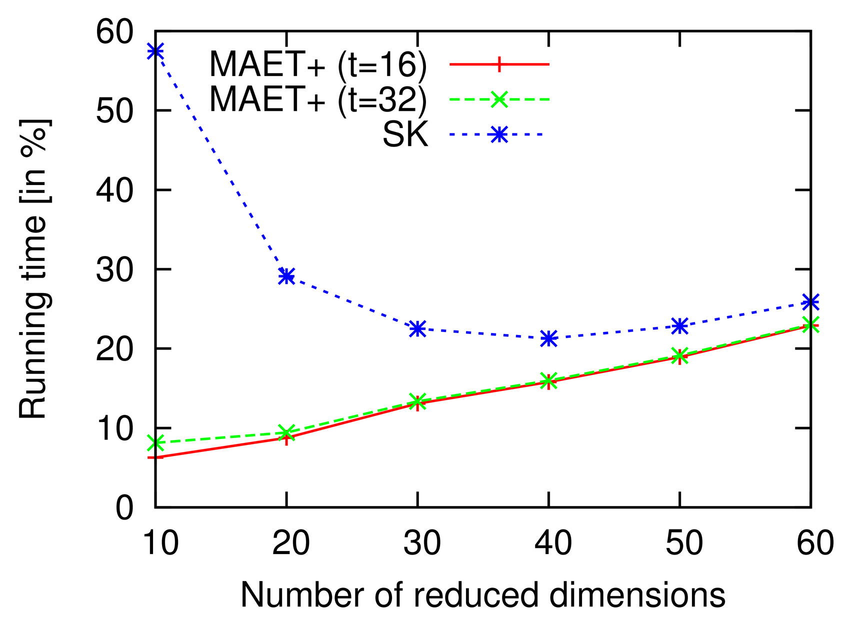
- Three real data sets are used for the experiments: Forest Cover Type (FCT), Amsterdam Library of Object Images (ALOI) and MNIST.
- Lower-bounding distance functions are generated by projecting the original feature space of a data set to new feature spaces with reduced dimensions using the Karhunen-Loève Transform.
- MAET+ is consistently competitive with MAET.
- For the FCT data set, the performances of MAET+ and SK are similar. However, for the other two data sets, MAET+ shows much improvement over SK while sacrificing little in accuracy (less than 5%).
- The performance variance over three data sets can be explained through the differences in their MGED values.




(a) FCT




(b) ALOI




(c) MNIST


NII 連絡先: Michael E. HOULE (フール マイケル) / 国立情報学研究所 客員教授
☎: 03-4212-2538 FAX: 03-3556-1916 ✉: meh@nii.ac.jp