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 \( d_{\text{max}} \) is maintained. If the size of the candidate set is smaller than \( k \), keep the value of \( d_{\text{max}} \) being infinity.
- The algorithm terminates when the value of \( d_{\text{max}} \) is no greater than the largest lower-bounding distance value encountered so far.
- At termination, the value of \( d_{\text{max}} \) 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 k-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:

\[
\text{GED}(B(q, r_1), B(q, r_2)) = \frac{\log k_2 - \log k_1}{\log r_2 - \log r_1}.
\]

We define the inner ball set relative to a point \( q \in S \) and a neighborhood size \( k \geq 2 \) as:

\[
B(q, k) = \{ B(q, \delta_j) | j \leq k-1, \delta_j > 0 \} \setminus B(q, \delta_k).
\]

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 \( k \geq 2 \) is defined as follows:

\[
\text{MGED}(q, k) = \max\{ \text{GED}(B(q, \delta_k)) | B \in B(q, k) \}.
\]

- 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 \( \lambda > 1 \) is the lower-bounding ratio.
- From the set \( P \), compute \( r = \delta_1 \) and \( r_1 = \delta_1 \).
- The algorithm terminates when \( k_1 = k \) or \( p_{(r/r')} > E + 1 \), 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 \( t \geq \text{MGED}(q, k+1) \).

- 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.