Random Projections and Sampling Algorithms for Clustering of High-Dimensional Polygonal Curves

Stefan Meintrup, Alexander Munteanu, Dennis Rohde

Univariate Time-Series
- plethora of analysis techniques exists
- most of them hardly applicable when sampling rates or lengths differ too much

Fallback-option [Driemel et al (SODA '16)]
1. drop concrete instants of time
2. treat remaining data as one-dimensional
3. comparison based on continuous dissimilarity measure for curves

Frechet Distance
- a.k.a. dog-meat distance
  \[ \max_i \left| \max_j \left( f_i(t) - f_j(t) \right) \right| \]
- maximum distance a point on the first curve has to travel when the first curve is continuously and monotonously deformed into the second curve under optimal speed adjustment

Algorithm
- Alt-Godau Algorithm
  - input two polygonal curves
  - output exact Frechet distance between input curves
  - running-time \( O(n^2 \log n) \)

Discrete 1-Median Clustering
1. uniformly at random, draw a candidate sample
2. uniformly at random, draw a witness sample
3. evaluate each candidate using the witnesses
4. return a candidate that evaluates best

\( \sigma \left( \frac{1}{2 - \gamma(T)} \right) \) candidates sufficient
\( \sigma \left( \frac{\ln(1/2 - \gamma(T))}{2} \right) \) witnesses sufficient
to obtain \( (1 + \epsilon) \)-Median with positive constant probability

\( \gamma(T) : \) fraction of outliers in \( T \)
can serve as plugin for k-median \( (1 + \epsilon) \) approximation
[Ackermann et al (TALG '10)]

What about Multivariate Time-Series?
- similar possibilities and problems
- our approach
  1. drop concrete instants of time
  2. treat remaining data as high-dimensional
  3. employ dimension reduction technique
  4. comparison based on continuous dissimilarity measure for curves

Johnson-Lindenstrauss Embedding for Polygonal Curves
1. obtain a Johnson-Lindenstrauss projection for all vertices of all curves
2. apply projection to all vertices
3. re-connect vertices with edges, in order

with positive constant probability the following bounds hold:
- \( d_F(f_1(t), f_2(t)) \leq \sqrt{(1 + \epsilon) \beta_{2n}^2 \| \sigma \| + 2 \alpha(\epsilon, \sigma)^2} \)
- \( \left( 1 - \epsilon \right) \beta_{2n}^2 \| \sigma \| - 2 \alpha(\epsilon, \sigma)^2 \leq d_F(f_1(t), f_2(t)) \)

\( \alpha(\epsilon, \sigma) : \) length of longest edge of both curves

Reducing the Curve's Complexity?
- let \( E, S \) be data oblivious randomized functions such that
  \( d_F(\sigma(\epsilon), \sigma(\tau)) \leq E(S(\sigma), \tau) \leq c \cdot d_F(\sigma, \tau) \)
holds with positive constant probability
- For \( c \in [1, \sqrt{2}] \) the sketch \( S(\sigma) \)
  must use \( \Omega(M) \) bits

one can not obtain a \( (1 + \epsilon) \) approximation
by reducing the dependency on \( M \)
- parallelized Frechet distance computation

Quality

Running-Time
- measured simultaneously
- known to correlate

Experiments


comparison based on continuous dissimilarity measure for curves

Technical University of Dortmund
SFB 876 Providing Information by Resource-Constrained Data Analysis
Dortmund Data Science Center
www.dennisrohde.work/rp4frechet-paper