Sequences and Time Series

Time Series Motifs

K. Selçuk Candan, Professor of Computer Science and Engineering

CASCADE
ASU Center for Assured and SCAlable Data Engineering

Ira A. Fulton Schools of Engineering
Arizona State University

Strings, sequences, time series

| A *string* or *sequence*, $S = (c_1, c_2, \ldots, c_N)$, *is a finite sequence of symbols.*

abcbbbaabbaabcbbbaaabbc

| A *time series*, $T = (d_1, d_2, \ldots, d_N)$, *is a finite sequence of data values.*

Interest in the term “big data”

https://trends.google.com/trends/explore?date=2008-12-19%202018-01-19&q=big%20data

Motifs

"Big Data"
"Machine Learning"
"Deep Learning"
100
75
50
25
Feb 24, 2013
Nov 23, 2014
Aug 21, 2016

Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with a given length)
  - Apply a clustering algorithm to identify groups of similar subsequences

Motif Quality

Motifs

(a) motif A has a higher support

(b) motif A has smaller distortion/higher purity

(c) motif A has higher dynamicity

Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with a given length)
  - Apply a clustering algorithm to identify groups of similar subsequences
  - Eliminate clusters with
    - too few subsequences (not enough support)
    - too imperfect matches (not well defined motifs)
    - too low dynamicity (patterns that do not carry much useful information or are not interesting)

Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with a given length)
    - all subsequences
    - subsequences that are sufficiently different from their immediate neighborhood
    - subsequences that have high dynamicity
    - subsequences that have at least one other similar subsequence in the series
  - Apply a clustering algorithm to identify groups of similar subsequences
  - Eliminate clusters with
    - too few subsequences (not enough support)
    - too imperfect matches (not well defined motifs)
    - too low dynamicity (patterns that do not carry much useful information or are not interesting)

Other questions

- Do all instances of the motif need to be of the same (average) amplitude?

Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with a given length)
    - all subsequences
    - subsequences that are sufficiently different from their immediate neighborhood
    - subsequences that have high dynamicity
    - subsequences that have at least one other similar subsequence in the series
  - Apply a clustering algorithm to identify groups of similar subsequences
    - allow for amplitude differences
  - Eliminate clusters with
    - too few subsequences (not enough support)
    - too imperfect matches (not well defined motifs)
    - too low dynamicity (patterns that do not carry much useful information or are not interesting)

Other questions

- What is the motif length?
- Do all instances of a motif need to have the same length?

"Olympics"

Motif Search Algorithm

- General approach
  - Enumerate subsequences of the time series (with different lengths)
    - all subsequences
    - subsequences that are sufficiently different from their immediate neighborhood
    - subsequences that have high dynamicity
    - subsequences that have at least one other similar subsequence in the series
  - Apply a clustering algorithm to identify groups of similar subsequences
    - allow for amplitude differences
    - account for length differences
  - Eliminate clusters with
    - too few subsequences (not enough support)
    - too imperfect matches (not well defined motifs)
    - too low dynamicity (patterns that do not carry much useful information or are not interesting)