Cluster analysis of time series data

Tomáš Bartoň

Supervisor: Ing. Pavel Kordík, Ph.D.

Department of Theoretical Computer Science
Faculty of Information Technology
Czech Technical University in Prague

January 5, 2012
The Problem

- find suitable method for identification of patterns
- assign samples into (unknown) groups
The Problem

Expected results

Tomáš Bartoň
Goals

- capture global trends
- absolute values (sometimes) doesn’t matter
- signals are not periodical
- discover unknown patterns
The Problem

Biological background

Anti-mitotic

Protein Synthesis Inhibitors

DNA Damaging

Calcium Modulators

Nuclear Receptor modulator
Phases of clustering process

1. data cleaning
2. data integration
3. data selection
4. data transformation
5. clustering
6. pattern evaluation
7. knowledge representation
Clustering
No “correct” clustering exists
Definition

“Those methods concerned in some way with the identification of homogeneous groups of objects”
[Arabie et al., 1996]

Definition

“A cluster is a set of entities that are alike, and entities from different clusters are not alike”
[Everitt, 1993]
- clustering can be used for understanding data
- to perform clustering you need to understand data
Determine number of clusters
$k = 3$
$k = 6$
\( k = 9 \)
$k = 14$
Determining the number of clusters in a data set is challenging [Mufti et al., 2005]
from Chinese encyklopedia Heavenly Emporium of Benevolent Knowledge. Animals are divided into [Borges, 1952]:

- those that belong to the emperor
- embalmed ones
- those that are trained
- suckling pigs
- mermaids
- fabulous ones
- those that are included in this classification
- innumerable ones
- etcetera
Clustering is ill-defined [Caruana et al., 2006]

All we care about is the “usefulness” of the clustering for achieving our final goal [Guyon et al., 2009]
Time series
Problem

- sensitive to small changes
- sum of distance does not capture shape of curve
- computationally expensive
- redundant information
Autoregressive model

- predict an output of a system based on the previous outputs

$$X_t = c + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \epsilon_t$$

- $\varphi_i$ – parameters of the AR model
- $X_t$ – amplitude of the signal
- $\epsilon_t$ – white noises
Moving average

\[ X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \phi_i \cdot \epsilon_{t-i} \]

- \( \phi_i \) – parameters of the AR model
- \( \mu \) – expectations of \( X_t \) (often assumed to equal 0)
- \( \epsilon_t \) – white noises
Autoregressive–moving-average model

putting all together:

\[ X_t = c + \epsilon_t + \sum_{i=1}^{p} \varphi_i \cdot X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \phi_i \cdot \epsilon_{t-i} \]

- ARMA(p, q) refers to the model with p autoregressive terms and q moving-average terms
- in Matlab function armax[Time-domain, data object]
Cluster 7

Cluster 8

these should be in one cluster

Cluster 9
Time series

Exponential

Polynomial
Representation of inputs

Measured values

- too many inputs
- does not represent patterns

Approximated model

\[ f(x) = p_1 x^4 + p_2 x^3 + p_3 x^2 + p_4 x + p_5 \]

- only 5 parameters describing whole curve
- represent patterns
How many parameters do we need?

2 parameters  

5 parameters
Which parameters to choose?

on previous slide input parameters were following:

- mean
- minimum
- maximum
- linear coefficient
- quadratic coefficient

for EEG clustering is [Siuly et al., 2011] using:

- minimum
- maximum
- mean
- median
- modus
- first quartile
- third quartile
- inter-quartile range
- standard deviation
Dendrogram

proportional distance between samples

input parameters

samples

-1  [-1]  [0]  [+1]  [+1]
C-Index

The C-index was reviewed in Hubert and Levin [1976]

\[ p_{c-index} = \frac{d_w - \min(d_w)}{\max(d_w) - \min(d_w)} \]

where \( d_w \) is the sum of the within cluster distances.

Gamma

\[ p_{gamma} = \frac{s(+) + s(-)}{s(+) - s(-)} \]

where \( s(+) \) represents the number of consistent comparisons involving between and within cluster distances, and \( s(-) \) represents the number of inconsistent outcomes Milligan and Cooper [1985]
An Iris dataset

![Graph 1](image1)

- Line chart with index on the y-axis and k on the x-axis, showing the performance of different indices.

![Graph 2](image2)

- Another graph with index on the y-axis and k on the x-axis, comparing different metrics:
  - Blue line: SumOfAvgPairwise
  - Red line: MinMaxCut
  - Yellow line: SSE

Legend:
- SumOfAvgPairwise
- MinMaxCut
- SSE
The strive for objectivity, repeatability, testability etc. is perfectly right attitude as long as their proper place in the “hierarchy of aims” is maintained, but becomes very harmful if these tools dominate over the purpose of scientific research. [Holynski, 2005, p. 487]
Questions?

Thanks for your attention!

tomas.barton@fit.cvut.cz
References I


