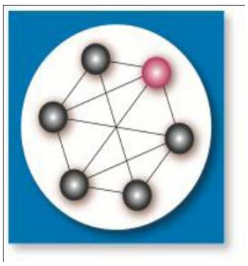


DBTech EXT
New Technology Trends and Challenges for the Database Professional – Challenges for Database Education

TIME SERIES DATA MINING

A Cluster-Based Approach to Similarity Search

Leonidas Karamitopoulos
Dept. of Applied Informatics
University of Macedonia
Thessaloniki, Greece

Helsinki, Finland, November 26th 2010

1

Time Series

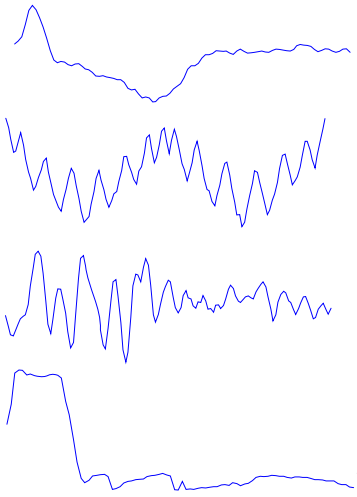
A time series is a collection of observations (variables) made sequentially in time

Time series are generated in almost every domain

- •Economy
- •Meteorology
- •Industry
- •Biology
- ...

Moreover, other types of data may be transformed into time series

- •Text
- •Images
- •Video
- ...



2

Time Series
Applications

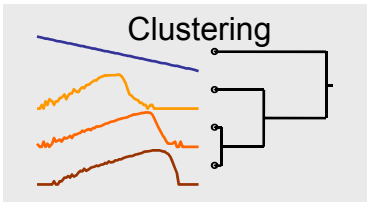
Business / Economy
- Identify stocks with similar variation through time with respect to their closing prices
- Group (cluster) customers with respect to their monthly purchase behavior

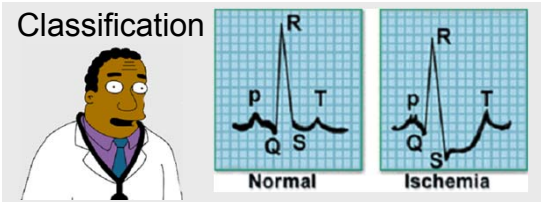
Medical / Biology
- Characterize (classify) an electrocardiograph as normal or ischemic
- Identify genes with similar functionality

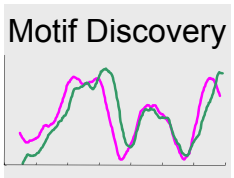
Meteorology
- Identify geographical regions with similar precipitation pattern

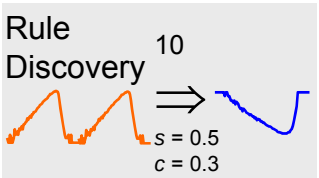
3

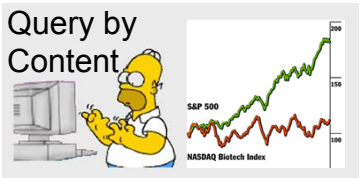
Time Series Data Mining
(c) Eamonn Keogh, eamonn@cs.ucr.edu



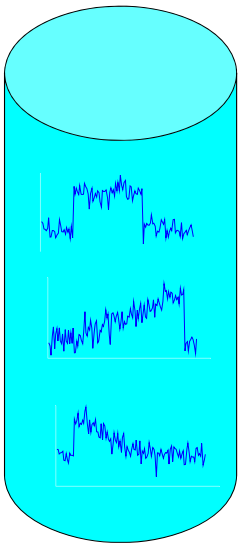






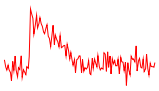


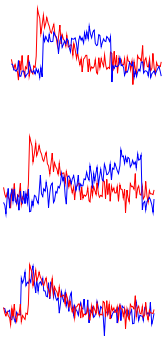
Similarity Search in Time Series
The Problem



Search the database for the k most similar time series to the query

Query



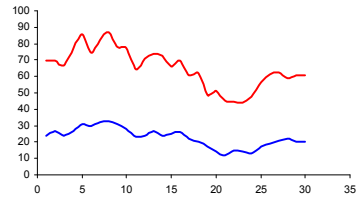


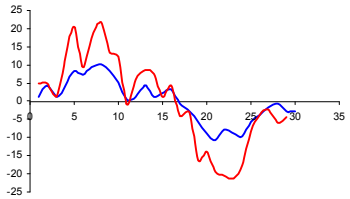
Database of time series

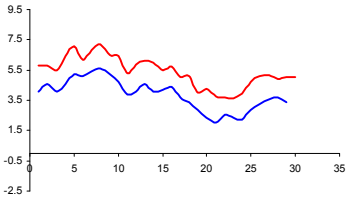
5

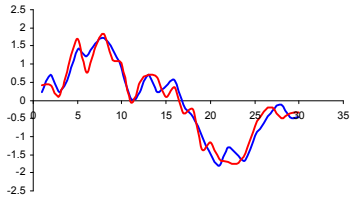
The Notion of *Similarity*

The effectiveness of similarity search depends on the appropriate definition of "*similarity*"







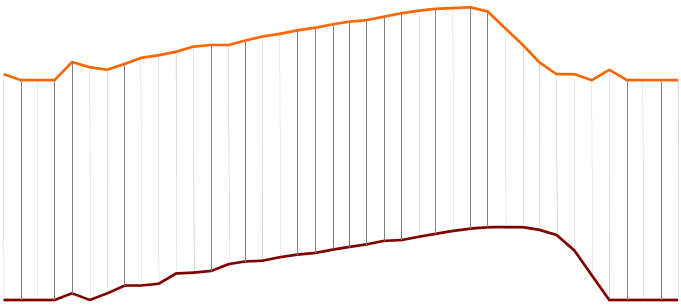


The definition of "*similarity*" is subjective and it depends on the application under consideration

6

Measuring Similarity

Euclidean Distance

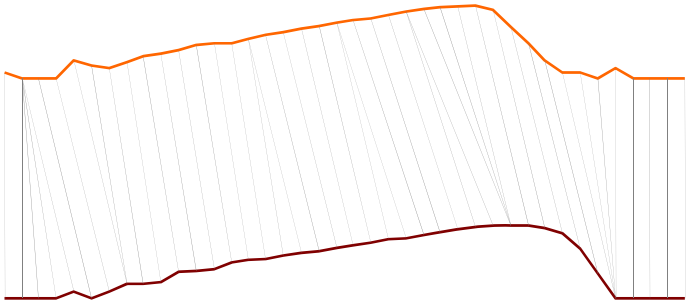


the values of the two time series are aligned in time

7

Measuring Similarity

Dynamic Time Warping (DTW)



allows local acceleration or deceleration in time

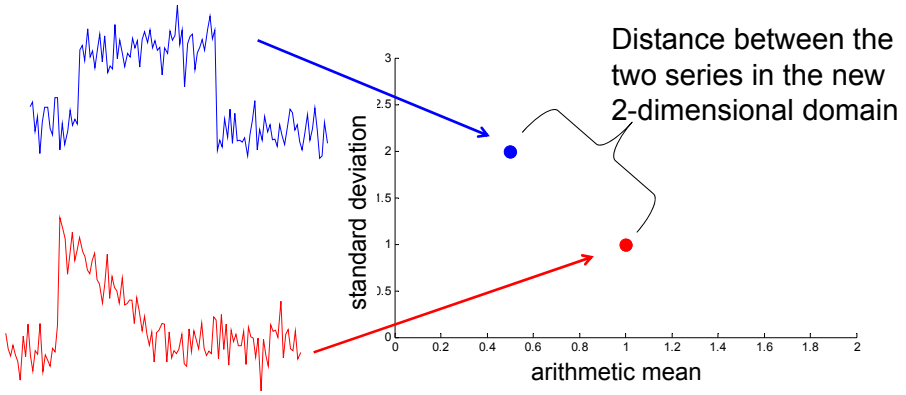
8

Similarity & Dimensionality

two time series of length (dimensions) 120

Representation of the series in two dimensions

Distance between the two series in the new 2-dimensional domain

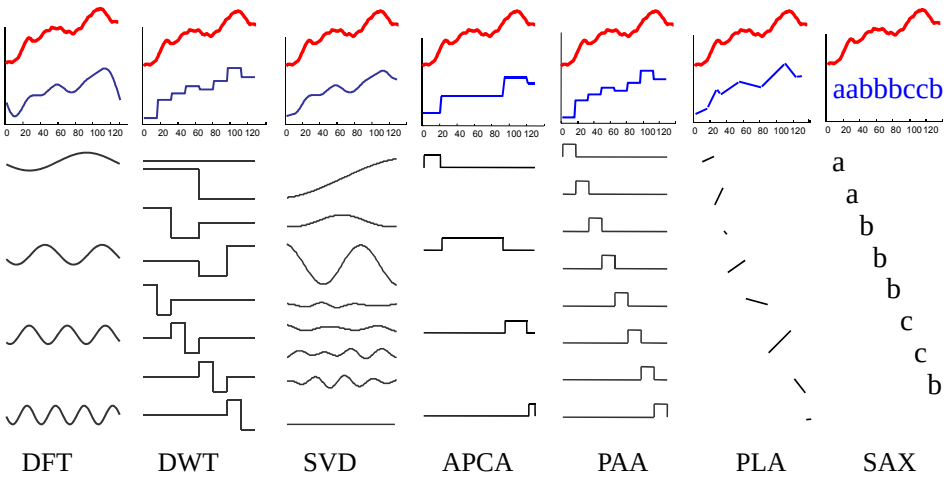


The objective is to index the database

9

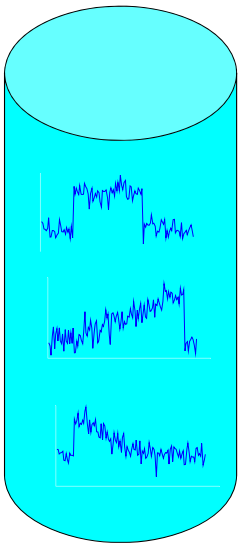
Fundamental Representations

(c) Eamonn Keogh, eamonn@cs.ucr.edu



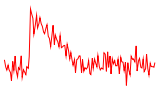
10

Fast Time Series Similarity Search
The Problem

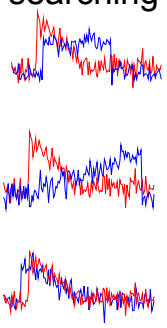


Search the database for the most similar time series to the query

query



sequential searching



Database of time series

11

Fast Time Series Similarity Search
Basic Approaches

- ✓ **Sequential Searching** – **Nearest Neighbor**
  (inefficient - prohibitive in very large databases)
- ✓ **Indexing**
  (efficiency degrades rapidly with the increase of the dimensionality – requires dimensionality reduction)
- ✓ **Clustering**
  (it depends on the quality of the derived clusters)

12

Current Research Work
A Proposed Method (CLUREP)

I. **Representation** (dimensionality reduction) of time series (eg. PAA)

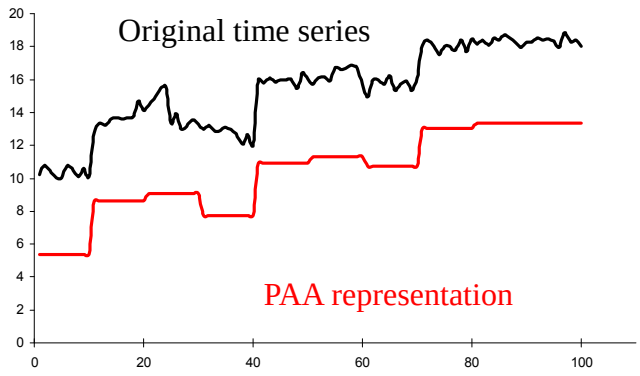
II. **Clustering** of the transformed time series by applying the k-means algorithm

III. **Sequential searching and restriction** of the search space in two ways:
(α) discarding a number of clusters
(b) restricting the search space within a cluster

*The **CLUREP** method guarantees that, under the specific representation, the nearest neighbor will be found*

13

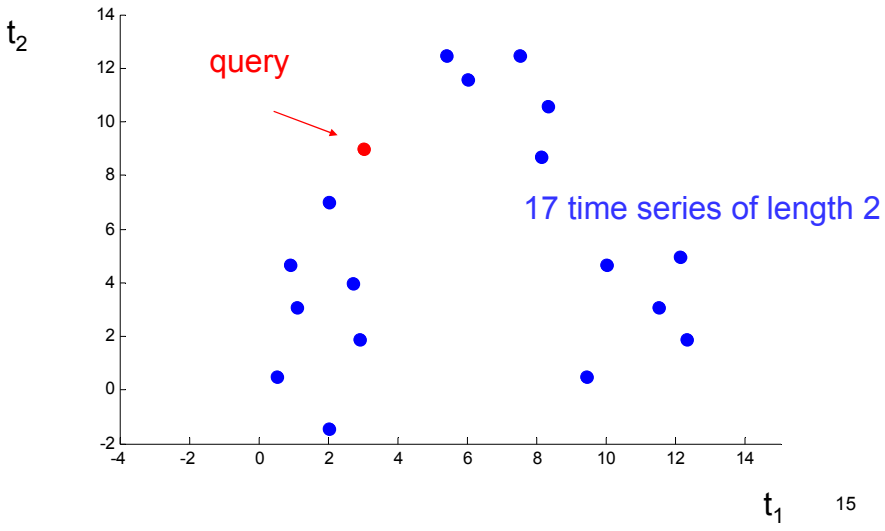
I. Representation
Piecewise Aggregate Approximation



PAA segments a time series into a number of sections and records the corresponding means

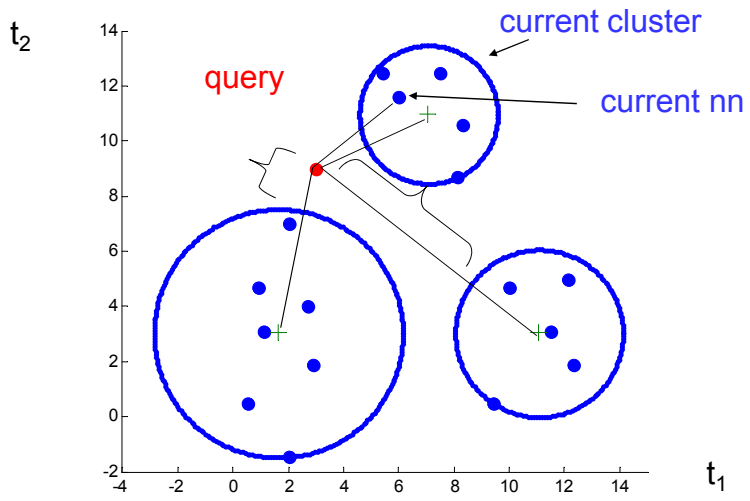
14

II. Clustering (k-means)



15

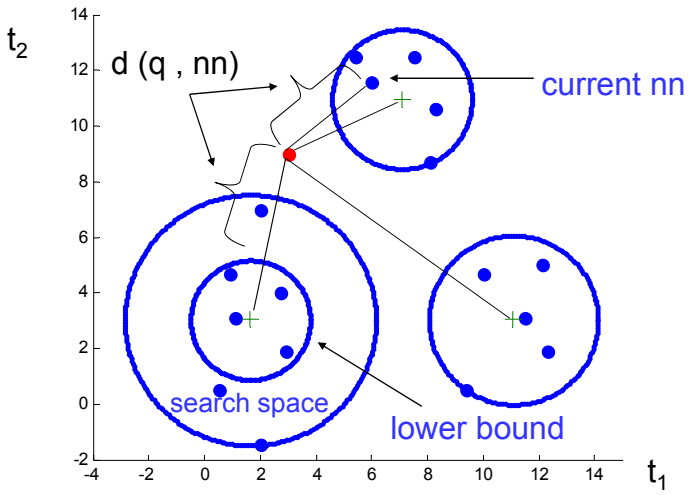
III. Similarity Searching (1)



16

determine the current (closest) cluster - find the current nearest neighbor

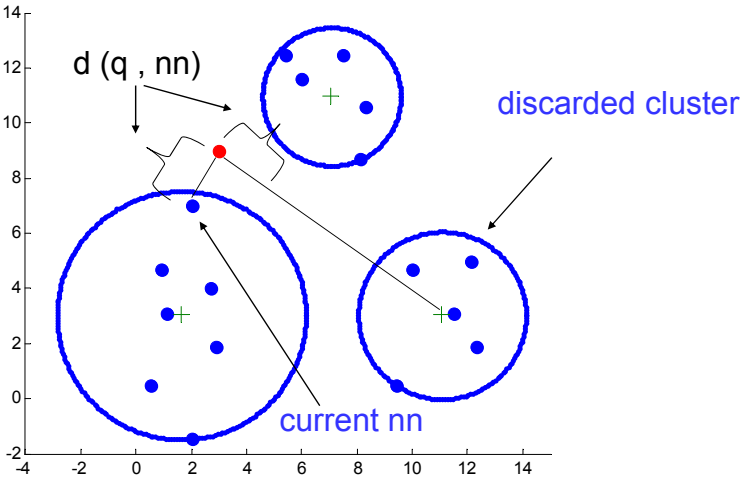
III. Similarity Searching (2)



determine the next closest cluster - check whether it should be searched - if yes, determine the search space

17

III. Similarity Searching (3)



find the current nearest neighbor - determine the next closest cluster - check of whether it should be searched - if yes, determine the search space

18

Experimental Evaluation (1)

Datasets:

- 12 synthetic and real world datasets that are commonly utilized for experimentation
- The number of classes ranges from 2 to 50, whereas the length of time series ranges from 60 to 427

Evaluation Method:

- One-nearest neighbor classification (1-NN)
- Record classification accuracy
- Record percentage of the volume of data that is searched

Validation: leave-one-out procedure

19

Experimental Evaluation (2)

Parameters:

- Regarding PAA, the number of sections is set equal to multiples of 2, ranging from 4 to 16
- Regarding k-means, the number of clusters is set equal to multiples of 2, ranging from 2 to 20

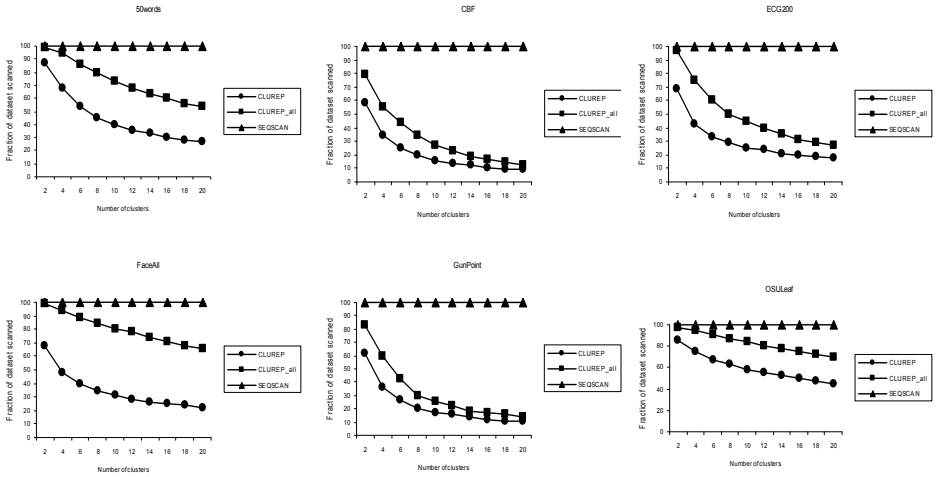
Rival Methods:

- the complete sequential search
- the method that does not restrict the search space of the visited clusters (CLUREP_all)

20

Experimental Results (1)

Fraction of data that is searched with respect to the number of clusters

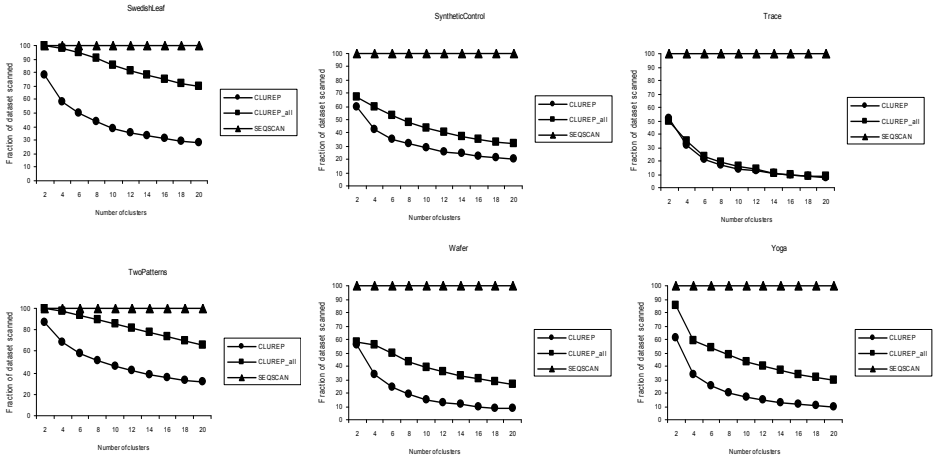


The dimensionality is set equal to that number for which the lowest error rate is achieved

21

Experimental Results (2)

Fraction of data that is searched with respect to the number of clusters



The dimensionality is set equal to that number for which the lowest error rate is achieved

22

Preliminary Findings

- ✓ The proposed method (CLUREP) is at least 5 times faster than sequential searching in the majority of the datasets upon which it was tested.
- ✓ When the number of the derived clusters increases, the effectiveness of CLUREP improves.
  The rate of improvement decreases when the number of the derived clusters is 8 to 12.
- ✓ When the dimensionality of the transformed data increases, the classification error rate decreases and the fraction of time series that need to be searched increases

23

Continuing Work

- ➢ Investigate and compare other clustering methods that may exhibit similar desirable properties to K-Means
- ➢ Evaluate the effect that the dimensionality of the time series has on cluster-based similarity search
- ➢ Experiment on very large datasets

Speeding up the similarity search process remains the main objective

24

A Notice

Some of the slides of this presentation have been copied or reproduced from

Eamonn Keogh's (eamonn@cs.ucr.edu) VLDB06 Tutorial: A Decade of Progress in Indexing and Mining Time Series Data

(**http://www.cs.ucr.edu/~eamonn/**)

25